Small Area Estimation

1 - Small area estimation problem

2 - Estimation for domains - **Direct estimators** – estimation for planned domains

3 – Coefficient of Variation and Minimum level of precision

4- Estimation for unplanned domains and/or where the sample size is not enough for the minimum level of precision

Recap

- Target parameters in the domain:

- Total of the study variable $\boxed{t = \sum_{k \in U} y_k}$
- Mean of the study variable $\boxed{\bar{y} = \sum_{k \in U} y_k \, / \, N}$
- At risk of poverty rate
- Poverty gap

$$P_0 = \frac{1}{N} \sum_{i=1}^{N} I(y_i < z).$$

$$P_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{G_i}{z}.$$

More on direct estimators for domains

There are more complex formulations of the direct estimator, other than HT estimator:

- ratio estimator
- **regression estimator**
- calibration estimator

These estimators are based on the use of auxiliary information, the HT estimator is based only on the y_i values of the sampled units

Auxiliary information

Auxiliary: providing supplementary or additional help and support to the estimation process.

- The auxiliary information about the population in the domain may include one or more known variables to which the variable of interest is **approximately** related.
- The auxiliary information typically is easy to measure, whereas the variable of interest may be expensive to measure.

Regression estimator

When the auxiliary variable *x* is **linearly** related to *y* but does not pass through the origin, a linear regression estimator would be appropriate.

This does not mean that regression estimate cannot be used when the intercept is close to zero.

The two estimates, regression and ratio may be quite close in such cases and you can choose the one you want to use.

In addition, if multiple auxiliary variables have a linear relationship with *y*, multiple regression estimates may be appropriate.

Regression estimator

To estimate the mean and total of *y*-values, denoted as μ and τ, one can use the linear relationship between *y* and known *x*-values.

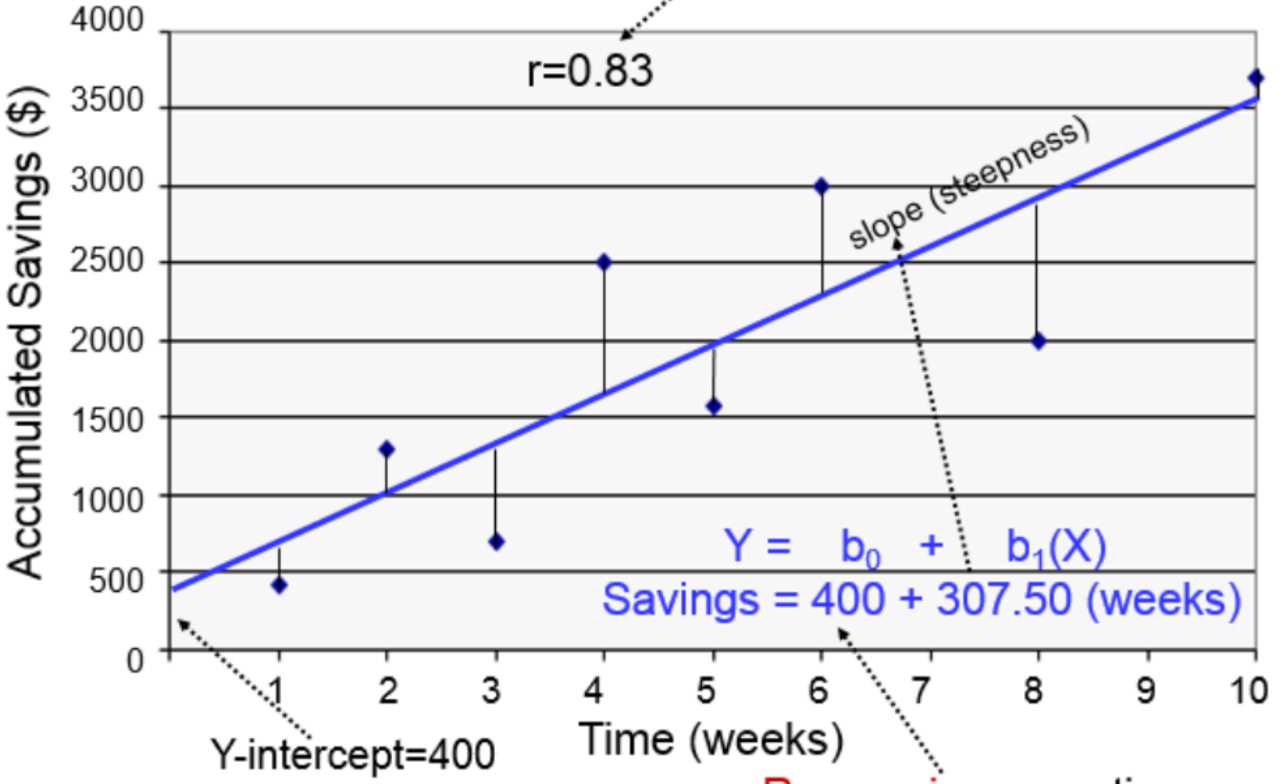
Let's start with a simple example:

$$\hat{y} = a + bx$$

This is our basic regression equation, it can be written also $Y=b_o+b_1X$, as in the following figure:

Simple linear regression

Correlation coefficient (r): a measure of how closely observation points conform to the regression line



r=0.83

$Y = b_0 + b_1(X)$

Savings = 400 + 307.50 (weeks)

Y-intercept=400

Regression: equation describing the relationship

Simple linear regression

The correlation coefficient indicates how closely these observations conform to a linear equation. The slope (b_1 or b) is the steepness of the regression line, indicating the average or expected change in Y for each unit change in X. In the illustration above the slope is 307.5, so the average savings per week is \$307.50. The intercept (b_0 or a) is the saving at the beginning of the period.

Formulas for intercept and slope

The intercept is obtained by the sample means

$$a = \bar{y} - b\bar{x}$$

The slope is the covariance (X,Y) divided by the variance of X on the sample

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Regression estimator of the mean of Y

$$\boxed{\bar{y} = \sum_{k \in U} y_k / N}$$

Recall the regression line,

$$\hat{y} = a + bx \,,$$

we calculate it in the mean:

$$\hat{\mu}_L = \underbrace{(\bar{y} - b\bar{x})}_{\text{intercept}} + \underbrace{b}_{\text{slope}}\mu_x$$

Regression estimator $\hat{\mu}_L$

$$\hat{\mu}_L = (\bar{y} - \underline{b\bar{x}) + b\mu_x} \Rightarrow \hat{\mu}_L = a + b\mu_x$$

$$\boxed{\hat{\mu}_L = \bar{y} + b(\mu_x - \bar{x}),}$$

Note that even though $\hat{\mu}_L$ is not unbiased under simple random sampling , it is roughly so (asymptotically unbiased) for large samples.

Mean Squared Error of $\hat{\mu}_L$

- It is roughly estimated by

$$\hat{V}ar(\hat{\mu}_L) = \frac{N-n}{N \times n} \cdot \frac{\sum_{i=1}^{n} (y_i - a - bx_i)^2}{n-2}$$
$$= \frac{N-n}{N \times n} \cdot MSE$$

where MSE is the MSE of the linear regression model of y on x

Confidence Interval for $\hat{\mu}_L$

Approximate 100(1-a)% confidence interval based on t student tables is

$$\hat{\mu}_L \pm t_{n-2,\alpha/2}\sqrt{\hat{V}ar(\hat{\mu}_L)}$$

Regression estimator of the total of Y

$$\boxed{t = \sum\nolimits_{k\in U} y_k}$$

It follows from $\hat{\mu}_L$ that

$$\hat{\tau}_L = N \cdot \hat{\mu}_L = N\bar{y} + b(\tau_x - N\bar{x})$$

with variance:

$$\begin{aligned}\hat{V}ar(\hat{\tau}_L) &= N^2 \hat{V}ar(\hat{\mu}_L) \\ &= \frac{N \times (N - n)}{n} \cdot MSE\end{aligned}$$

MSE of the linear regression model of y on x

Confidence Interval for $\hat{\tau}_L$:

Approximate 100(1-a)% confidence interval based on t student tables is

$$\hat{\tau}_L \pm t_{n-2,\alpha/2}\sqrt{\hat{V}ar(\hat{\tau}_L)}$$

- Examples of application of the regression estimator during the R lab