Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 4: Ratio and Regression Estimation

Ratio Estimation

Two quantities y_i and x_i are measured on each sample unit

- y_i: response variable, x_i: auxiliary variable, or subsidiary variable
- Let $t_y = \sum_{i=1}^{N} y_i$ and $t_x = \sum_{i=1}^{N} x_i$ and their ratio be

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U}$$

Example 4.1: Suppose the population consists of agricultural fields of different sizes. Let
y_i = bushels of grain harvested in field i
x_i = acreage of field i
then B = average yield in bushels per acre
$\bar{y}_U$ = average yield in bushels per field
t_y = total yield in bushels

If an SRS is taken, natural estimators for B, t_y, and $\bar{y}_U$ are:

$$\hat{B} = \frac{\bar{y}}{\bar{x}}, \quad \hat{t}_{yr} = \hat{B} t_x \quad \hat{\bar{y}}_r = \hat{B} \bar{x}_U$$

- Ratio estimation take advantage of the correlation of x and y in the population; the higher the correlation, the better they work. Define the population correlation coefficient of x and y be

$$R = \frac{\sum_{i=1}^{N} (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1) S_x S_y}$$

—- S_x is the population standard deviation of the x_i's
—- S_y is the population standard deviation of the y_i's
—- R is simply the pearson correlation coefficient of x and y for the N units

Why use ratio estimation?

1. Want to estimate a ratio
   Example: interested in the percentage of pages in Good Housekeeping magazines that contain at least one advertisement

- Take an SRS of 10 issues
- let x_i be the total number of pages in issue i
- let y_i be the total number of pages in issue i that contain at least one advertisement
- $\hat{B} = \dfrac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}$

2. Want to estimate a population total, but population size N is unknown

- $\hat{t}_y = N\bar{y}$, but N is unknown
- $N = \dfrac{t_x}{\bar{x}_U}$
- $\hat{t}_y = \dfrac{t_x}{\bar{x}_U}\bar{y}$
- $\hat{t}_{yr} = \dfrac{t_x}{\bar{x}}\bar{y} = \dfrac{\bar{y}}{\bar{x}}t_x = \hat{B}t_x$

Example: Apple Juice

For a juice company, the price they are paid for apples in large shipments is based on the amount of apple juice from the load.

- Need to determine the amount of apple juice in the whole load prior to extraction.
- We can sample n apples and find $y_1, \cdots y_n$, the amount of apple juice in those apples.
- $N\bar{y}$ is hard to get in this case because N is hard to count. But total weight of apple is easy to get.
  ——use the relationship between weight of the load and weight of the apple juice one obtains
  ——let x be the weight of each apple in the sample, $\bar{x}$ is the average weight of each apple in the sample.
  ——number of apples is estimated by $t_x/\bar{x}$
  The total weight t_x is easy to get for the entire shipment. We can thus estimate the total apple juice by:

$$\hat{t}_{yr} = \frac{\bar{y}}{\bar{x}} t_x$$

Example: Want to estimate the total # of fish in a haul that are longer than 12 cm

- Take an SRS, estimate the proportion and multiply by the total # of fish N. But N is unknown
- Take an SRS, consider the fact that having a length of more than 12 cm (y) is related to weight (x), introduce an auxiliary variable x_i: weight of fish
- y_i: fishes longer than 12cm,
  x_i: weight of fish,
  t_x: total weight of haul,
  t_y =?
- $\hat{t}_{yr} = \hat{B} t_x = \frac{\bar{y}}{\bar{x}} t_x$

3. Ratio estimation is often used to increase the precision of estimated means and totals
   Let y_i be the # of persons in commune i
   and x_i be the # of registered births in commune i
   want to estimate # of persons in France

- Randomly select 30 communes
- Estimator 1 $= N\bar{y} = \frac{t_x}{\bar{x}_U}\bar{y}$
  # of communes in France $*$ average number of persons in the 30 communes
- Estimator 2$= \hat{B}t_x = \frac{\bar{y}}{\bar{x}}t_x = \frac{t_x}{\bar{x}}\bar{y}$

- When $\bar{y}$ and $\bar{x}$ are positively correlated, the sampling distribution of $\frac{\bar{y}}{\bar{x}}$ exhibits less variability compared to the sampling distribution of $\frac{\bar{y}}{\bar{x}_U}$.
- Consequently, the ratio estimator (Estimator 2) has a smaller Mean Squared Error (MSE), i.e.,
  MSE(Estimator 2) $<$ MSE(Estimator 1).

4. Adjust estimators from the sample so that they reflect demographic totals.
   Example: An SRS of 400 students is taken at a university with a total of 4,000 students. The sample contains:
   - 240 women and 160 men
   - 84 women and 40 men in the sample plan to pursue careers in teaching

   We aim to estimate the total number of students who plan to become teachers.

- Estimator 1: Using only the information from the SRS,

$$N\bar{y} = 4000 \times \frac{124}{400} = 1240$$

- Estimator 2: Incorporating demographic information (college has 2,700 women and 1,300 men), a better estimate is:

$$\frac{84}{240} \times 2700 + \frac{40}{160} \times 1300 = 1270$$

- Highlights:
  - Ratio estimation is applied within each gender.
  - In the sample, 60% are women, but women comprise 67.5% of the population. The estimator is adjusted accordingly to better reflect the demographic proportions.

5. Adjust for nonresponse
   Example: a sample of businesses,
   y_i: amount spent on health insurance by business i
   x_i: number of employees in business i, x_i known
   want to estimate total insurance expenditures

- Estimator 1: $N\bar{y}$
  —companies with few employees are less likely to respond to the survey
  —y_i is proportional to x_i
  —Estimator 1 overestimate the total insurance expenditures t_y
- Estimate 2: $t_x \frac{\bar{y}}{\bar{x}}$
  — $\frac{t_x}{\bar{x}} < N$, since companies with many employees are more likely to respond to the survey
  —Thus a ratio estimate of total heath care insurance expenditures may help to compensate for the nonresponse of companies with few employees

Example 4.2

- Dataset: SRS of $n = 300$ counties selected from a total of $N = 3078$ counties (U.S. Census of Agriculture, file `agsrs.dat`).
- Context:
  - Total acreage for 1987 is known.
  - 1992 acreage data is only available for the sampled 300 counties.
  - Goal: Estimate the population total $\hat{t}_y$ and mean $\hat{\bar{y}}$ for 1992.
- Estimate 1: Using only the SRS data from 1992,

$$\hat{t}_{y,\text{srs}} = N\bar{y} = 3078 \cdot \bar{y} = 916,927,110$$

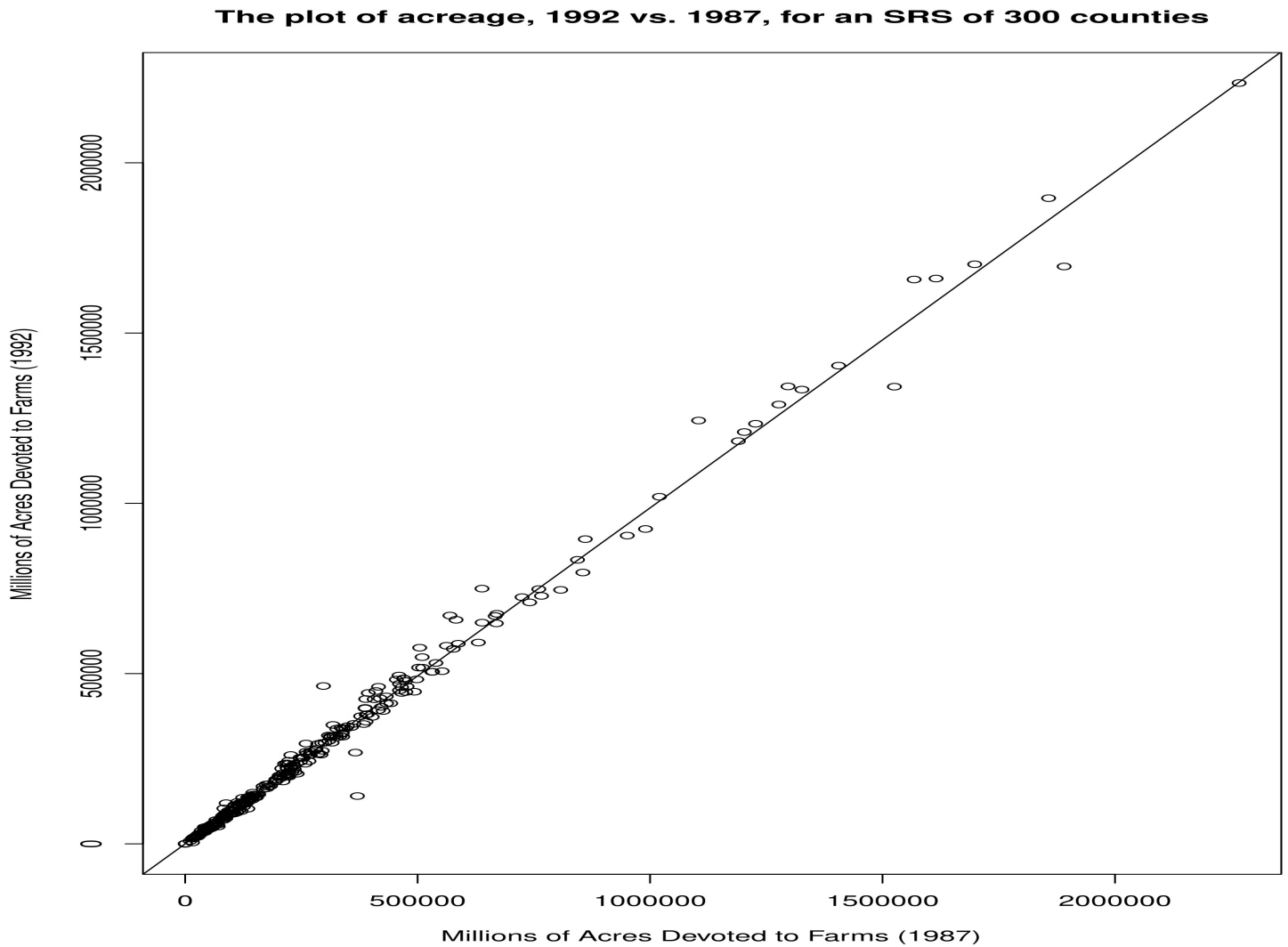

Figure 1: A scatter plot of 1992 acreage (y-axis) against 1987 acreage (x-axis) for a simple random sample (SRS) of 300 counties.

As shown in Figure 1, the line of best fit passes through the origin with a slope of $\hat{b} = 0.9866$, demonstrating a strong positive correlation between 1992 and 1987 acreages. We will now apply ratio estimation using 1987 acreage as an auxiliary variable.

- Estimate 2: ratio estimation
  —y_i = total acreage of farms in county i in 1992
  —x_i = total acreage of farms in county i in 1987
  —For 1987, $t_x = 964,470,625$,
  $\bar{x}_U = 964,470,625/3078 = 313343.3$
  —

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{297897.0467}{301953.7233} = .986565$$

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U = 309,133.6$$

$$\hat{t}_{yr} = \hat{B}t_x = .986565 \times 964470625 = 951,513,191$$

Comments:

- When the same quantity is measured at different times, the response of interest at an earlier time often serves as an excellent auxiliary variable.
- The sample mean $\bar{x}$ is slightly smaller than the true population mean $\bar{x}_U$, indicating that the SRS of size 300 slightly underestimates the true population mean of the x's.
- Since x and y are positively correlated, it is reasonable to expect that $\bar{y}$ may also underestimate the true population mean $\bar{y}_U$.
- Ratio estimation improves the precision of $\bar{y}_U$ by adjusting $\bar{y}$ with the factor $\bar{x}_U/\bar{x}$.

Properties of Ratio Estimators

- For an SRS, $\bar{y}$ is unbiased:
  — If we calculate $\bar{y}_S$ for every possible sample, the average of all the sample means will equal the population mean $\bar{y}_U$.
- For the ratio estimator, $\hat{\bar{y}}_r = \frac{\bar{y}}{\bar{x}}\bar{x}_U$:
  — The bias of the ratio estimator arises because $\bar{y}$ is adjusted by the factor $\bar{x}_U/\bar{x}$.
  — If we compute $\hat{\bar{y}}_r$ for all possible samples, the average of these estimates will generally be close to $\bar{y}_U$, but not exactly equal to it.
- For large samples, the sampling distributions of both $\bar{y}$ and $\hat{\bar{y}}_r$ are approximately normal.
- Ratio estimators are biased but typically have smaller variance compared to $\bar{y}$.

Bias of $\hat{B}$

$$
\begin{aligned}
\text{Bias}[\hat{B}] &= E[\hat{B}] - B = E\left[\frac{\bar{y}}{\bar{x}} - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= E\left[\frac{\bar{y}}{\bar{x}_U} \times \frac{\bar{x}_U}{\bar{x}} - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= E\left[\frac{\bar{y}}{\bar{x}_U} \times \left(1 - \frac{\bar{x} - \bar{x}_U}{\bar{x}}\right) - \frac{\bar{y}_U}{\bar{x}_U}\right] \\
&= -E\left[\frac{\bar{y}(\bar{x} - \bar{x}_U)}{\bar{x}_U \bar{x}}\right] \\
&\vdots \\
&\approx \frac{1}{\bar{x}_U^2}[BV(\bar{x}) - \text{Cov}(\bar{x}, \bar{y})] \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}_U^2}(BS_x^2 - RS_xS_y)
\end{aligned}
$$

where R is the correlation between x and y.

$$
\begin{aligned}
\text{Bias}[\hat{\bar{y}}_r] &= E[\hat{\bar{y}}_r - \bar{y}_U] \\
&= E[\hat{B}\bar{x}_U - B\bar{x}_U] \\
&= \bar{x}_U E[\hat{B} - B] \\
&\approx \frac{1}{\bar{x}_U}\left[BV(\bar{x}) - \text{Cov}(\bar{x}, \bar{y})\right] \\
&= \left(1 - \frac{n}{N}\right)\frac{1}{n\bar{x}_U}\left(BS_x^2 - RS_xS_y\right)
\end{aligned}
$$

Bias of $\hat{\bar{y}}_r$ is small if:

- The sample size n is large.
- The sampling fraction n/N is large.
- The standard deviation S_x is small.
- The correlation R is close to 1.

  Note: If all x values are identical ($S_x = 0$), the ratio estimator becomes equivalent to the SRS estimator $\bar{y}$, and the bias is zero.

MSE of $\hat{B}$

$$
\begin{aligned}
E[(\hat{B}-B)^2] &= E\left[\left(\frac{\bar{y}}{\bar{x}}-B\frac{\bar{x}}{\bar{x}}\right)^2\right] \\
&= E\left[\left(\frac{\bar{y}-B\bar{x}}{\bar{x}}\right)^2\right] \\
&= E\left[\left(\frac{\bar{y}-B\bar{x}}{\bar{x}_U}\right)\left(1-\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right)\right]^2 \\
&= E\left[\left(\frac{\bar{y}-B\bar{x}}{\bar{x}_U}\right)^2+\left(\frac{\bar{y}-B\bar{x}}{\bar{x}_U}\right)^2\times\left(-2\frac{\bar{x}-\bar{x}_U}{\bar{x}}+\left(\frac{\bar{x}-\bar{x}_U}{\bar{x}}\right)^2\right)\right] \\
&\approx E\left[\left(\frac{\bar{y}-B\bar{x}}{\bar{x}_U}\right)^2\right] \\
&= \frac{1}{\bar{x}_U^2}E[(\bar{y}-B\bar{x})^2]
\end{aligned}
$$

where the approximation is from the fact that the second and third term is negligible relative to the first term.

Let

- $d_i = y_i - Bx_i$
- $e_i = \hat{d}_i = y_i - \hat{B}x_i$
- $\bar{d} = \bar{y} - B\bar{x}$

$$\begin{aligned} MSE(\hat{B}) &\approx \frac{1}{\bar{x}_U^2} E[(\bar{y} - B\bar{x})^2] \\ &= \frac{1}{\bar{x}_U^2} E\left[\frac{1}{n}\sum_{i\in S}(y_i - Bx_i)\right]^2 \\ &= \frac{1}{\bar{x}_U^2} V(\bar{d}) \\ &= \frac{1}{\bar{x}_U^2}\left(1 - \frac{n}{N}\right)\frac{S_d^2}{n} \end{aligned}$$

So

$$\widehat{MSE}(\hat{B}) \approx \frac{1}{\bar{x}^2}\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}$$

Variance of $\hat{B}$

In large sample, bias of $\hat{B}$ is typically small relative to $V(\hat{B})$, $MSE(\hat{B}) \approx V(\hat{B})$

$$V(\hat{B}) \approx \frac{1}{\bar{x}_U^2}\left(1-\frac{n}{N}\right)\frac{S_d^2}{n}$$

$$\hat{V}(\hat{B}) \approx \frac{1}{\bar{x}^2}\left(1-\frac{n}{N}\right)\frac{s_e^2}{n}$$

Variance of $\hat{\bar{y}}_r$

$$\begin{aligned} \hat{V}(\hat{\bar{y}}_r) &= \bar{x}_U^2 \hat{V}(\hat{B}) \\ &\approx \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n} \end{aligned}$$

Variance of $\hat{V}(\hat{\bar{y}}_r)$ is small if

- the sample size n is large
- the sampling fraction n/N is large
- the deviations $y_i - Bx_i$ are small
- the correlation R is close to 1

Compare ratio estimator to SRS estimator

$$V(\hat{\bar{y}}_r) \approx \left(1 - \frac{n}{N}\right) \frac{S_d^2}{n}$$

$$V(\bar{y}_{srs}) = \left(1 - \frac{n}{N}\right) \frac{S_y^2}{n}$$

If $S_d^2 < S_y^2$ then $V(\hat{\bar{y}}_r) < V(\bar{y}_{srs})$, ratio estimation is more efficient

$$
\begin{aligned}
(N-1)S_d^2 &= \sum_{i=1}^{N}(y_i - Bx_i)^2 \\
&= \sum_{i=1}^{N}\left((y_i - \bar{y}_U) + (\bar{y}_U - Bx_i)\right)^2 \\
&= \sum_{i=1}^{N}(y_i - \bar{y}_U)^2 + \sum_{i=1}^{N}(B\bar{x}_U - Bx_i)^2 \\
&\quad +2\sum_{i=1}^{N}(y_i - \bar{y}_U)(B\bar{x}_U - Bx_i) \\
&= (N-1)S_y^2 + (N-1)B^2S_x^2 \\
&\quad -2(N-1)BRS_xS_y
\end{aligned}
$$

Where R is the population correlation coefficient, defined as:

$$
R = \frac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_xS_y}
$$

$$\begin{aligned} V(\hat{\bar{y}}_r) &\approx \left(1-\frac{n}{N}\right)\frac{1}{n}\frac{\sum_{i=1}^{N}(y_i - Bx_i)^2}{N-1} \\ &= \left(1-\frac{n}{N}\right)\frac{1}{n}\left(S_y^2 + B^2 S_x^2 - 2BRS_xS_y\right) \end{aligned}$$

The variance for the simple random sample (SRS) estimator is given by:

$$V(\bar{y}_{srs}) = \left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$$

Ratio estimation is more efficient when:

$$S_y^2 + B^2 S_x^2 - 2BRS_xS_y < S_y^2$$

This simplifies to:

$$B^2 S_x^2 < 2BRS_xS_y, \quad BS_x < 2RS_y$$

Or equivalently:

$$\frac{\bar{y}_U}{\bar{x}_U}S_x < 2RS_y, \quad \frac{S_x}{\bar{x}_U} < 2R\frac{S_y}{\bar{y}_U}$$

Thus, for ratio estimation to be more efficient:

$$R > \frac{\frac{S_x}{\bar{x}_U}}{2\frac{S_y}{\bar{y}_U}} = \frac{\mathrm{CV}(\bar{x})}{2\mathrm{CV}(\bar{y})}$$

- $\mathrm{CV}(\bar{y}) = \frac{Sd(\bar{y})}{\bar{y}_U}$
- $\mathrm{CV}(\bar{x}) = \frac{Sd(\bar{x})}{\bar{x}_U}$
- The absolute values of $\mathrm{CV}(\bar{x})$ and $\mathrm{CV}(\bar{y})$ often do not make a significant difference.

As a result, **Ratio estimation is more efficient than an SRS when**:

$$R > \frac{1}{2}$$

Examples 4.2 and 4.3 (using R handout 4)

- Ratio estimator
  —y_i = total acreage of farms in county i in 1992
  —x_i = total acreage of farms in county i in 1987
  —For 1987, $t_x = 963,464,412$
  —

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{297897.0467}{301953.7233} = .986565, \text{ and } R = 0.995806$$

$$\hat{t}_{yr} = \hat{B} t_x = .986565 \times 963,464,412 = 950,520,496$$

$$\text{SE}(\hat{t}_{yr}) = 5,540,376$$

- SRS estimator
  $\hat{t}_y = N\bar{y} = 3078 * 297897.0467 = 916,927,110$
  $\text{SE}(\hat{t}_y) = 58,169,381$, this is almost 10 times as large as the SE from ratio estimation ($\text{SE}(\hat{t}_{yr}) = 5,540,376$)

- Coefficient of Variation (CV) comparison
  Recall, Coefficient of Variation (CV)
  when $\bar{y}_U \neq 0$,

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})}, \quad \widehat{CV}(\bar{y}) = \frac{\mathrm{SE}(\bar{y})}{\bar{y}}$$

  —-Measure of relative variability
  —-Does not depend on the unit of measurement
  —-$CV(\hat{t}) = CV(\bar{y})$
  —-If the CV of $\bar{y}$ is small, that is, if $\bar{y}_U$ is estimated with high relative precision, the bias is small relative to the square root of the variance.
  —- A small $CV(\bar{y})$ also means that $\bar{y}$ is stable from sample to sample.

Table 1: Comparisons of ratio estimator and SRS estimator

	Ratio estimation	SRS estimation
SE of $\hat{t}$	5,540,376	58,169,381
Estimated CV	$\frac{5,540,376}{950,520,496} = 0.0058$	$\frac{58,169,381}{916,927,110} = 0.0634$

- Incorporating the 1987 data through the ratio estimator has significantly increased precision.
- If all quantities to be estimated are highly correlated with the 1987 acreage, using ratio estimators instead of $N\bar{y}$ could substantially reduce the sample size while maintaining high precision.

Regression Estimation in Simple Random Sampling

Review: Regression Analysis

A nutritionist aims to explore the relationship between age and muscle mass in women. It is hypothesized that muscle mass decreases with age. To investigate this, she randomly selects 15 women from each 10-year age group, starting from age 40 and ending at age 79, for a total of 60 women.

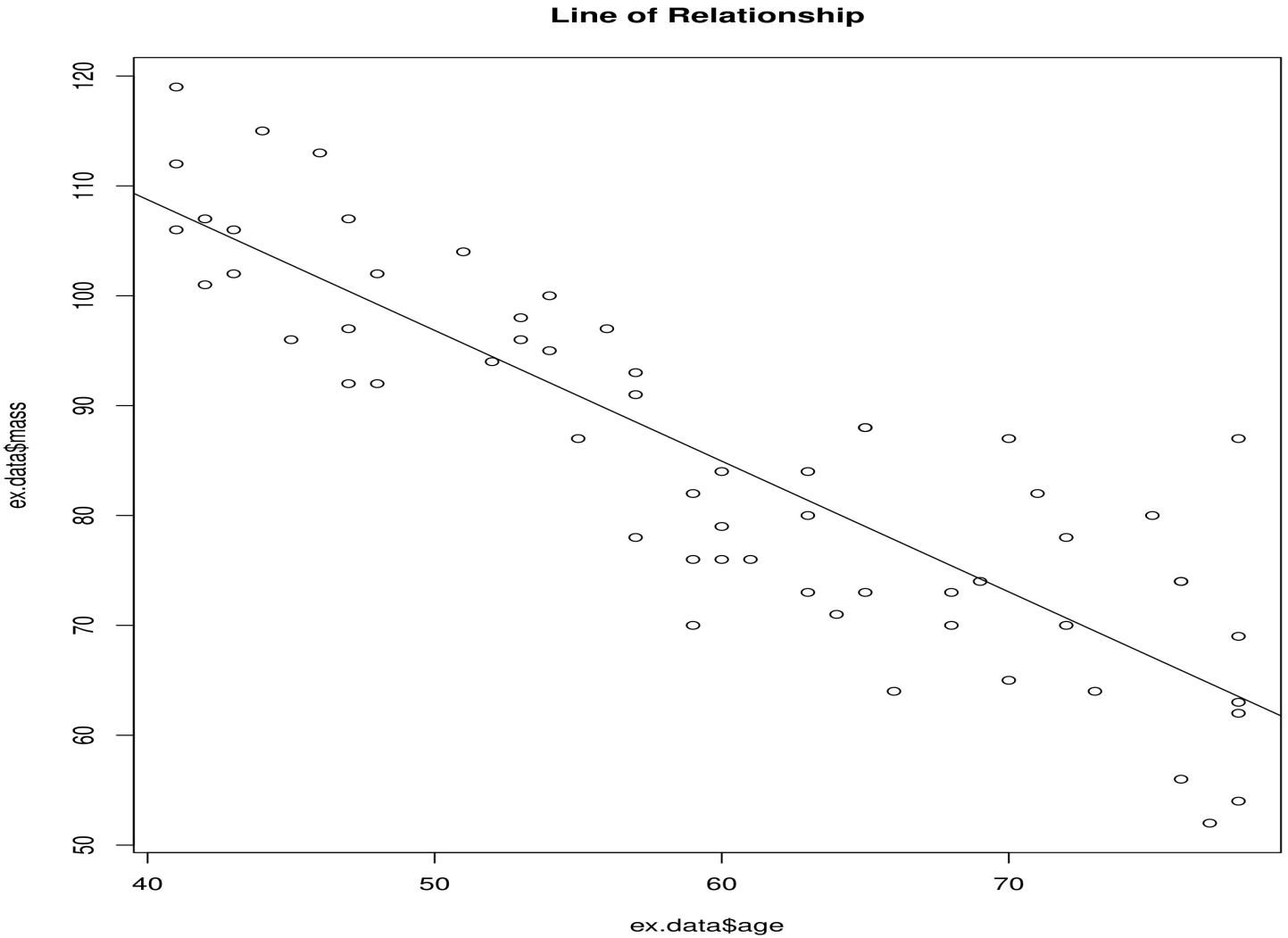

Figure 2: Age vs. Muscle Mass in Women with fitted regression line

Normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i: response of the ith trial
- X_i: a known constant, the level of the predictor variable in the ith trial
- β_0 and β_1: parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$
- $E(Y_i) = \beta_0 + \beta_1 X_i$

Least square estimators:

- Consider the deviation of Y_i from its expected value

$$[Y_i - (\beta_0 + \beta_1 X_i)]$$

- Least Square Measures:

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- Objective: to find estimators b_0 and b_1 for β_0 and β_1 respectively, for which Q is minimum
-

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}$$

- Regression line: $\widehat{E(Y)} = b_0 + b_1 X$

Regression in Simple Random Sampling (SRS)

Want to estimate population mean and population total

Assumptions:

- The relationship between $E(y)$ and x is a straight line

$$E(y) = B_0 + B_1 x$$

- The population mean of x's, $\bar{x}_U$ is known

Population quantities	Estimators
$B_1 = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\sum_{i=1}^{N}(x_i - \bar{x}_U)^2}$	$\hat{B}_1 = \dfrac{\sum_{i\in S}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in S}(x_i - \bar{x})^2}$
$B_0 = \bar{y}_U - B_1\bar{x}_U$	$\hat{B}_0 = \bar{y} - \hat{B}_1\bar{x}$

- B_1 and B_0 are the least squares regression slope and intercept calculated from all the data in the population respectively
- The regression estimator of $\bar{y}_U$ is

$$\begin{aligned} \hat{\bar{y}}_{\text{reg}} &= \hat{B}_0 + \hat{B}_1\bar{x}_U \\ &= \bar{y} - \hat{B}_1\bar{x} + \hat{B}_1\bar{x}_U \\ &= \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) \end{aligned}$$

Properties of the Estimators

Notation

- $d_i = y_i - (B_0 + B_1 x_i)$
- $e_i = y_i - (\hat{B}_0 + \hat{B}_1 x_i)$ called residuals
- $R = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x}_U)(y_i - \bar{y}_U)}{(N-1)S_x S_y}$, population correlation coefficient of x and y

Bias of $\hat{\bar{y}}_{\text{reg}}$

$$\text{bias}(\hat{\bar{y}}_{\text{reg}}) = -\text{cov}(\hat{B}_1, \bar{x})$$

Proof:

$$\begin{aligned}
\text{Bias}(\hat{\bar{y}}_{\text{reg}}) &= E[\hat{\bar{y}}_{\text{reg}} - \bar{y}_U] \\
&= E[\hat{B}_0 + \hat{B}_1 \bar{x}_U - \bar{y}_U] \\
&= E[\bar{y} - \hat{B}_1 \bar{x} + \hat{B}_1 \bar{x}_U - \bar{y}_U] \\
&= E[\bar{y} - \bar{y}_U] - E[\hat{B}_1(\bar{x} - \bar{x}_U)] \\
&= -\text{cov}(\hat{B}_1, \bar{x})
\end{aligned}$$

$\hat{\bar{y}}_{\text{reg}}$ is biased for $\bar{y}_U$

- If the regression line goes through all of the points (x_i, y_i) in the population, then the bias is zero since $\hat{B}_1 = B_1$ for every sample, so $\text{cov}(\hat{B}_1, \bar{x}) = 0$

MSE of $\hat{\bar{y}}_{\text{reg}}$: $MSE(\hat{\bar{y}}_{\text{reg}}) = \left(1 - \frac{n}{N}\right)\frac{S_d^2}{n}$

$$\begin{aligned} d_i &= y_i - (B_0 + B_1 x_i) \\ &= y_i - (B_0 + B_1 x_i - B_1 \bar{x}_U + B_1 \bar{x}_U) \\ &= y_i - (B_0 + B_1 \bar{x}_U + B_1 (x_i - \bar{x}_U)) \\ &= y_i - [\bar{y}_U + B_1 (x_i - \bar{x}_U)] \end{aligned}$$

$$\begin{aligned} MSE(\hat{\bar{y}}_{\text{reg}}) &= E(\hat{\bar{y}}_{\text{reg}} - \bar{y}_U)^2 \\ &= E[\bar{y} + \hat{B}_1(\bar{x}_U - \bar{x}) - \bar{y}_U]^2 \\ &= E\{\bar{y} - [\bar{y}_U + \hat{B}_1(\bar{x} - \bar{x}_U)]\}^2 \\ &\approx \text{Var}(\bar{d}) \\ &= \left(1 - \frac{n}{N}\right)\frac{S_d^2}{n} \end{aligned}$$

Another expression of MSE of $\hat{\bar{y}}_{\mathbf{reg}}$

Notice

$$B_1 = R \cdot \frac{S_y}{S_x}$$

and

$$\begin{aligned} S_d^2 &= \sum_{i=1}^{N} \frac{(y_i - \bar{y}_U - B_1[x_i - \bar{x}_U])^2}{N-1} \\ &= S_y^2(1 - R^2) \end{aligned}$$

$$MSE(\hat{\bar{y}}_{\text{reg}}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} S_y^2 (1 - R^2)$$

$MSE(\hat{\bar{y}}_{\text{reg}})$ is small if

- n is large, n/N is large
- The correlation R is close to either -1 or $+1$

Variance of $\hat{\bar{y}}_{\mathbf{reg}}$

For large SRSs

- Bias is often negligible in large samples
- The MSE for regression estimation is approximately equal to the variance

Estimator for total: $\hat{t}_{\text{yreg}}$

$$\begin{aligned}
\hat{t}_{\text{yreg}} &= \sum_{i\in S} y_i + \sum_{i\notin S} y_i \\
&= \sum_{i\in S} y_i + \sum_{i\notin S} (\hat{B}_0 + \hat{B}_1 x_i) \\
&= \sum_{i\in S} y_i + (N-n)\hat{B}_0 + \hat{B}_1 \left(t_x - \sum_{i\in S} x_i \right)
\end{aligned}$$

If $n << N$

$$\begin{aligned}
\hat{t}_{\text{yreg}} &\approx N\hat{B}_0 + \hat{B}_1 t_x \\
&= N\hat{B}_0 + \hat{B}_1 N\bar{x}_U \\
&= N(\hat{B}_0 + \hat{B}_1 \bar{x}_U) \\
&= N\hat{\bar{y}}_{\text{reg}}
\end{aligned}$$

Confidence Intervals:

$$SE(\hat{\bar{y}}_{\text{reg}}) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$

$$SE(\hat{t}_{\text{yreg}}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$

A $100(1-\alpha)\%$ CI for $\bar{y}_U$ is

$$\hat{\bar{y}}_{\text{reg}} \pm t_{n-2}(\alpha/2)\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$

A $100(1-\alpha)\%$ approximate CI for t is

$$\hat{t}_{\text{yreg}} \pm t_{n-2,\alpha/2}N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}}$$

Example 4.7: Estimating the Number of Dead Trees

To estimate the number of dead trees in a specific area, the following sampling procedure is implemented:

1. Divide the area into 100 square plots.
2. Conduct photo counts:
   - Count the number of dead trees in each plot using aerial photographs.
   - Photo counts are efficient but may include misclassifications or missed detections.
3. Select a simple random sample (SRS) of 25 plots for field verification:
   - Perform on-the-ground counts of dead trees in these 25 plots to assess and correct for any inaccuracies in the photo counts.
4. Calculate the population mean:
   - The mean number of dead trees per plot, based on photo counts, is 11.3.

This methodology combines aerial photographic analysis with field verification to enhance the accuracy of the estimated mean number of dead trees per plot.

```
print.data.frame(deadtrees)
   photo field
1     10    15
2     12    14
3      7     9
4     13    14
5     13     8
6      6     5
7     17    18
8     16    15
9     15    13
10    10    15
11    14    11
12    12    15
13    10    12
14     5     8
15    12    13
.........
25    10     8
```

```
dtree<- svydesign(id = ~1, weight=rep(4,25),
 fpc=rep(100,25), data = deadtrees)
> dtree
Independent Sampling design
svydesign(id = ~1, weight = rep(4, 25), fpc = rep(100, 25),
data = deadtrees)
> myfit1 <- svyglm(field~photo, design=dtree)
> summary(myfit1) # displays regression coefficients
Call:
svyglm(formula = field ~ photo, design = dtree)
Survey design:
svydesign(id = ~1, weight = rep(4, 25), fpc = rep(100, 25),
data = deadtrees)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0593     1.3930   3.632   0.0014 **
photo         0.6133     0.1259   4.870 6.44e-05 ***
```

From R output,

$$\hat{B}_0 = 5.0593, \hat{B}_1 = 0.6133$$

Fitted regression line

$$\hat{y} = 5.0593 + 0.6133x$$

```
newdata <- data.frame(photo=11.3)
> predict(myfit1, newdata)
    link    SE
1 11.989 0.418
> confint(predict(myfit1, newdata),df=23)
     2.5 %   97.5 %
1 11.12455 12.85404
```

The regression estimate of the mean is

$$\hat{y}_{reg} = 5.0593 + 0.6133 * 11.3 = 11.99$$

For these data, $\bar{x} = 10.6, \bar{y} = 11.56, s_y^2 = 9.09$, and the sample correlation between x and y is $r = 0.62420$.

$$SE(\hat{\bar{y}}_{reg}) = \sqrt{\left(1 - \frac{25}{100}\right)\frac{1}{25} * 9.09 * (1 - 0.62420^2)} = 0.408$$

```
newdata2 <- data.frame(photo=1130)
> predict(myfit1, newdata2, total=100)
    link     SE
1 1198.9 41.802
> confint(predict(myfit1, newdata2,total=100),df=23)
     2.5 %   97.5 %
1 1112.455 1285.404
```

The estimate of the total number of dead trees

$$\hat{t}_{yreg} = 100 * 11.99 = 1199$$

The estimated standard error of $\hat{t}_{yreg}$ is 41.80, with CI of $[1112.455, 1285.404]$.

- Because of the relatively small sample size, we used the t distribution percentile (with $n - 2 = 23$ degrees of freedom) of 2.07 in the CI rather than the normal distribution percentile of 1.96.

SRS and Ratio estimation using weights

SRS

- $w_i = N/n$
- $\bar{y} = \dfrac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}$
- $\hat{t}_y = \sum_{i \in S} w_i y_i = N\bar{y}$
- $\sum_{i \in S} w_i = \sum_{i \in S} \dfrac{N}{n} = n \cdot \dfrac{N}{n} = N$

Ratio Estimation

$$
\begin{aligned}
\hat{t}_{yr} &= \frac{\bar{y}}{\bar{x}} t_x = \frac{\hat{t}_y}{\hat{t}_x} t_x \\
&= \sum_{i \in S} w_i y_i \frac{t_x}{\hat{t}_x} = \sum_{i \in S} \left(w_i \cdot \frac{t_x}{\hat{t}_x} \right) y_i \\
&\overset{g_i = t_x / \hat{t}_x}{=} \sum_{i \in S} w_i g_i y_i \\
&\overset{w_i^* = w_i g_i}{=} \sum_{i \in S} w_i^* y_i
\end{aligned}
$$

- w_i^* depend upon values from the sample
- The weight adjustments g_i calibrate the estimates on the x variable. Since $\sum_{i \in S} w_i g_i x_i = t_x$, the adjusted weights force the estimated total for the x variable to equal the known population total t_x. The factors g_i are called the calibration factors.

Example 4.6: Census of Agriculture Data (Examples 4.2 and 4.3 Continued)

- For each observation:

$$g_i = \frac{t_x}{\hat{t}_x} = \frac{964{,}470{,}625}{929{,}413{,}560} \approx 1.0377$$

- Since $\hat{t}_x < t_x$, each observation's sampling weight is increased by a small amount.
- The sampling weight for the Simple Random Sample (SRS) design is:

$$w_i = \frac{3078}{300} \approx 10.26$$

- The ratio-adjusted weight for each observation is:

$$w_i^* = w_i \times g_i = 10.26 \times 1.0377 \approx 10.65$$

- $$\sum_{i \in S} w_i g_i x_i = \sum_{i \in S} 10.64700262 x_i = 964,470,625 = t_x$$

- $$\sum_{i \in S} w_i g_i y_i = \sum_{i \in S} 10.64700262 y_i = 951,513,191 = \hat{t}_{yr}$$

- The adjusted weights, however, no longer sum to $N = 3078$

$$\sum_{i \in S} w_i g_i = (300)(10.64700262) = 3194$$

- The ratio estimator is calibrated to the population total t_x of the x variable, but is no longer calibrated to the population size N.

Regression Estimation

$$
\begin{aligned}
\hat{t}_{\text{yreg}} &= N[\hat{B}_0 + \hat{B}_1 \bar{x}_U] \\
&= N\hat{B}_0 + N\hat{B}_1 \bar{x}_U \\
&= \hat{B}_0 N + \hat{B}_1 t_x \\
&= (\bar{y} - \hat{B}_1 \bar{x})N + \hat{B}_1 t_x \\
&= \hat{t}_y - \hat{B}_1 \hat{t}_x + \hat{B}_1 t_x \\
&= \hat{t}_y + \hat{B}_1 (t_x - \hat{t}_x) \\
&= \sum_{i \in S} w_i y_i + \frac{\sum_{i \in S} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in S} (x_i - \bar{x})^2} \cdot (t_x - \hat{t}_x)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i\in S} w_i y_i + \frac{\sum_{i\in S} y_i(x_i-\bar{x})}{\sum_i (x_i-\bar{x})^2}\cdot(t_x-\hat{t}_x)\\
&= \sum_{i\in S} w_i y_i + \frac{\sum_{i\in S} w_i y_i(x_i-\bar{x})}{\sum_i w_i(x_i-\bar{x})^2}\cdot(t_x-\hat{t}_x)\\
&= \sum_{i\in S} w_i \left(1+\frac{(x_i-\bar{x})(t_x-\hat{t}_x)}{\sum_i w_i(x_i-\bar{x})^2}\right) y_i\\
&= \sum_{i\in S} w_i g_i y_i
\end{aligned}
$$

where $g_i = \left(1+\frac{(x_i-\bar{x})(t_x-\hat{t}_x)}{\sum_i w_i(x_i-\bar{x})^2}\right)$ called g-weight

If $y_i = x_i$

$$\begin{aligned}
\hat{t}_{\text{yreg}} &= \sum_{i \in S} w_i \left(1 + \frac{(x_i - \bar{x})(t_x - \hat{t}_x)}{\sum_i w_i (x_i - \bar{x})^2} \right) x_i \\
&= \sum_{i \in S} w_i x_i + \frac{\sum_{i \in S} w_i (x_i - \bar{x})^2}{\sum_{i \in S} w_i (x_i - \bar{x})^2} \cdot (t_x - \hat{t}_x) \\
&= \hat{t}_x + (t_x - \hat{t}_x) \\
&= t_x
\end{aligned}$$

Comparison of Estimation Methods

- Both **ratio** and **regression** estimation utilize an auxiliary variable highly correlated with the variable of interest.
- The ratio and regression estimators discussed are special cases of a generalized regression estimator.
- Ratio estimation is particularly useful in cluster sampling.
- For a Simple Random Sample (SRS) of size n, the estimators are summarized in the following table.

	Estimator for Mean $\bar{y}_U$	Estimator for Total t_y	Residual e_i
SRS	$\bar{y}$	$N\bar{y}$	$y_i - \bar{y}$
Ratio	$\hat{B}\bar{x}_U$	$\hat{B}t_x$	$y_i - \hat{B}x_i$
Regression	$\hat{B}_0 + \hat{B}_1\bar{x}_U$ $= \bar{y} + \hat{B}_1(\bar{x}_U - \bar{x})$	$N(\hat{B}_0 + \hat{B}_1\bar{x}_U)$	$y_i - \hat{B}_0 - \hat{B}_1 x_i$

Table 2: Estimators and Residuals for SRS of Size n

Estimated variance for $\hat{\bar{y}}_U$: $\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}$

Estimated variance for $\hat{t}$: $N^2\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}$

Estimation in Domains

- **Domain:** A subpopulation of interest.
- Objective: Obtain separate estimates for subpopulations.
  - Example: Estimating the average income for women in a Simple Random Sample (SRS).
- Considerations:
  - The number of women in the sample is a random variable.
  - Membership in a domain (e.g., women) is unknown for the entire population until sampled.
  - Therefore, the number of individuals in each domain within an SRS is a random variable, with its value unknown at the survey design stage.
- Estimating domain means is a special case of ratio estimation.

Estimating Mean for a Specific Domain

Suppose there are D domains in the population:

- U_d: Index set of units in the population for domain d, where $d = 1, 2, \ldots, D$.
- S_d: Index set of units in the sample for domain d, where $d = 1, 2, \ldots, D$.
- N_d: Number of population units in U_d.
- n_d: Number of sample units in S_d.

To estimate the mean salary for a specific domain (e.g., women):

$$\bar{y}_{U_d} = \frac{\sum_{i \in U_d} y_i}{N_d} = \frac{\text{Total salary for all women in population}}{\text{Number of women in population}}$$

A natural estimator for $\bar{y}_{U_d}$ is the sample mean:

$$\bar{y}_d = \frac{\sum_{i \in S_d} y_i}{n_d} = \frac{\text{Total salary for women in sample}}{\text{Number of women in sample}}$$

Note: n_d is a random variable.

Let:

$$y_i : \text{Income for person } i, \quad x_i = \begin{cases} 1 & \text{if women} \\ 0 & \text{otherwise} \end{cases}$$

$$u_i = x_i y_i = \begin{cases} y_i & \text{if women} \\ 0 & \text{otherwise} \end{cases}$$

- $t_x = \sum_{i=1}^{N} x_i = N_d$: Total number of women in the population, $\bar{x}_U = N_d/N$.
- $t_u = \sum_{i=1}^{N} u_i$: Total income for women in the population.
- $\bar{y}_{U_d} = t_u/t_x = B$: Average income for women in the population.
- $\bar{y}_d = \bar{u}/\bar{x} = \hat{B}$: Average income for women in the sample d, where:

$$\bar{u} = \frac{\sum_{i \in S} x_i y_i}{n}, \quad \bar{x} = \frac{\sum_{i \in S} x_i}{n} = \frac{n_d}{n}.$$

$$
\begin{aligned}
\hat{V}(\hat{B}) &= \frac{1}{\bar{x}^2}\left(1-\frac{n}{N}\right)\frac{s_e^2}{n} \\
&= \frac{1}{\bar{x}^2}\left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in S}(u_i-\hat{B}x_i)^2 \\
&= \frac{1}{\bar{x}^2}\left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in S}(y_ix_i-\hat{B}x_i)^2 \\
&= \frac{1}{\bar{x}^2}\left(1-\frac{n}{N}\right)\frac{1}{n}\cdot\frac{1}{n-1}\sum_{i\in S_d}(y_i-\bar{y}_d)^2 \\
&\overset{\bar{x}=n_d/n}{=} \left(1-\frac{n}{N}\right)\frac{n}{n_d^2}\cdot\frac{n_d-1}{n-1}s_{yd}^2 \\
&\approx \left(1-\frac{n}{N}\right)\frac{s_{yd}^2}{n_d}
\end{aligned}
$$

$$
SE(\bar{y}_d) \approx \sqrt{\left(1-\frac{n}{N}\right)\frac{s_{yd}^2}{n_d}}
$$

General Case of Domain Estimation

Estimates for different subpopulations:

- The mean for a subpopulation is expressed as a ratio.
- The sample size of the domain is a random variable.
- The estimate for the mean, $\hat{B}$, is given by:

$$\hat{B} = \frac{\text{Sum of } y_i\text{'s in the domain}}{\text{Total number of observations in the domain}} = \bar{y}_d$$

Estimating Totals in Domains

- If N_d (total population size of the domain) is known:

$$\hat{t}_{yd} = N_d \bar{y}_d$$

- If N_d is unknown:

$$\hat{N}_d = N \cdot \frac{n_d}{n}$$

Then, the total estimate is:

$$\hat{t}_{yd} = N \cdot \frac{n_d}{n} \cdot \frac{\sum_{i \in S} u_i}{n_d} = N\bar{u}$$

- The standard error of $\hat{t}_{yd}$ is:

$$SE(\hat{t}_{yd}) = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_u^2}{n}}$$

Example 4.8: Estimating Acres Devoted to Farming
In an SRS of size 300 from the Census of Agriculture (see Examples 2.6, 4.2, and 4.3):

- 129 counties have at least 600 farms.
- 171 counties have fewer than 600 farms.
- Objective: Estimate the average and total number of acres devoted to farming in each domain.

Summary Statistics for the Two Domains:

Domain, d	n_d	$\bar{y}_d$ (Average Acres)	s_d (SE)
1. At least 600 farms	129	316,565.65	258,249.74
2. Fewer than 600 farms	171	283,813.71	397,643.92

```
> # direct calculation
> agsrsnew1<-agsrsnew[which(agsrsnew$farmcat=='large'),]
> nrow(agsrsnew1)
[1] 129
> mean(agsrsnew1$acres92)  #\bar y_1
[1] 316565.7
> sqrt(var(agsrsnew1$acres92)) #s_{1}
[1] 258249.7
> sum(agsrsnew1$acres92)/300 # u_1
[1] 136123.2
> 3078*sum(agsrsnew1$acres92)/300  # \hat t_u
[1] 418987302
> agsrsnew$is_large = ifelse(agsrsnew$farmcat == "large",
agsrsnew$acres92,0)
> sqrt(var(agsrsnew$is_large))  #s_u=230641.2 page 141
[1] 230641.2
```

```
> n<-300
> agsrsnew$farmcat<-rep("large",n)
> agsrsnew$farmcat[agsrsnew$farms92 < 600] <- "small"
> head(agsrsnew)
> print.data.frame(head(agsrsnew[,c(1,2,6,16)]))
             county state farms92 farmcat
1     COFFEE COUNTY    AL     760   large
2    COLBERT COUNTY    AL     488   small
3      LAMAR COUNTY    AL     299   small
4    MARENGO COUNTY    AL     434   small
5     MARION COUNTY    AL     566   small
6 TUSCALOOSA COUNTY    AL     436   small
```

```
> sampwt <- rep(3078/n,n)
> dsrsnew <- svydesign(id = ~1, weights=~sampwt,
fpc=rep(3078,300), data=agsrsnew)
> dsrsnew
Independent Sampling design
svydesign(id = ~1, weights = ~sampwt,
fpc = rep(3078, 300), data = agsrsnew)
> dsub1<-subset(dsrsnew,farmcat=='large')
> smean1<-svymean(~acres92,design=dsub1)
> smean1
          mean    SE
acres92 316566 21553
> df1<-sum(agsrsnew$farmcat=='large')-1 #domain df
> df1
[1] 128
> confint(smean1, level=.95,df=df1)
            2.5 %   97.5 %
acres92 273918.9 359212.4
```

Suppose we don't know N_d: how many counties in the population are in each domain

```
> stotal1<-svytotal(~acres92,design=dsub1)
> stotal1
            total       SE
acres92 418987302 38938277
> confint(stotal1, level=.95,df=df1)
            2.5 %    97.5 %
acres92 341941269 496033335
```

Please refer to detailed calculation on page 141

Thus the standard error for the estimated mean in domain 1 is:

$$\text{SE}\ (\bar{y}_1) = \sqrt{\left(1 - \frac{300}{3078}\right)\left(\frac{300}{299}\right)\left(\frac{128}{129}\right)}\frac{258{,}249.74}{\sqrt{129}} = 21{,}553.$$

An approximate 95% CI for the mean farm acreage for counties in domain 1, using the t critical value with 128 df, is $316{,}565.65 \pm 1.979\,(21{,}553)$, or [273,919, 359,212]. A similar calculation for domain 2 yields $\text{SE}\,(\bar{y}_2) = 28{,}852.24$ and an approximate 95% CI of [226,859, 340,769].

Suppose that we do not know how many counties in the population are in each domain. To estimate the total in domain 1, define

$$x_i = \begin{cases} 1, & \text{if county } i \text{ is in domain 1} \\ 0, & \text{otherwise} \end{cases}$$

and $u_i = y_i x_i$. Then

$$\hat{t}_{y1} = \hat{t}_u = \sum_{i \in \mathcal{S}} \frac{3078}{300} u_i = 418{,}987{,}302. \tag{4.25}$$

The standard error is

$$\text{SE}(\hat{t}_{y1}) = N\sqrt{1 - \frac{n}{N}}\frac{s_u}{\sqrt{n}} = 3078\sqrt{\left(1 - \frac{300}{3078}\right)}\frac{230{,}641.22}{\sqrt{300}} = 38{,}938{,}277$$

and a 95% CI for the population total in domain 1, using a t critical value with 128 df, is $418{,}987{,}302 \pm 1.979(38{,}938{,}277) = [341{,}941{,}269,\ 496{,}033{,}335]$. Similarly, a 95% CI for the population total in domain 2 is

$$497{,}939{,}808 \pm 1.974(3078)\sqrt{\left(1 - \frac{300}{3078}\right)}\frac{331{,}225.43}{\sqrt{300}} = [387{,}553{,}731,\ 608{,}325{,}884]\,. \blacksquare$$

Figure 3: Example 4.8, more calculations

Source: page 141 of Sampling: Design and Analysis, 3rd edition, by Sharon L. Lohr

Poststratification

Poststratification: Stratification performed after the selection of the sample.

- *Stratification* is an element of survey design that divides the population into distinct subgroups before sampling.
- *Poststratification*, on the other hand, is an analytical technique applied after data collection to adjust for differences between the sample and the population structure.

Example of Poststratification Consider a public opinion survey aiming to stratify respondents by gender:

- If the survey is conducted via random digit dialing of telephone numbers, it's not possible to assign potential respondents to male or female strata until they have been contacted and identified.
- Poststratification can then be used to weight the responses according to known population demographics once the data has been collected.

Example: Estimating Monthly Food Expenditure

To estimate the average amount spent on food in a month, one desirable stratification variable might be household size, since larger households are likely to have higher food expenditures than smaller ones. Given this, we can use the distribution of household sizes from U.S. Census data for the region as follows:

# of persons in household	percentage of household
1	25.75
2	31.17
3	17.50
4	15.58
5	10.00

Table 3: Distribution of Household Sizes in the Region

Poststratification:

- Collect a Simple Random Sample (SRS) and record both the amount spent on food and the household size for each household in your sample.
- When the sample size n is sufficiently large, it is likely to resemble a stratified sample with proportional allocation. That is:
  - Approximately 26% of the sample would be one-person households,
  - Around 31% would be two-person households,
  - And so on, following the known distribution of household sizes in the population.
- Treat different household-size groups as distinct domains. Using this approach, ratio estimation can be applied to estimate the average amount spent on groceries within each domain.

By employing poststratification, we can improve the precision of our estimates by accounting for the structure of the population after data collection.

Poststratified estimator Let $n_1, n_2, \cdots, n_H$ be the numbers of units sampled in the various household-size groups (domains), n_h is random
and $\bar{y}_1, \cdots, \bar{y}_H$ be the sample means for the groups
Let $x_{ih} = 1$ if observation i is in poststratum h and 0 otherwise
Let $u_{ih} = y_i x_{ih}$

$$t_{xh} = \sum_{i=1}^{N} x_{ih} = N_h$$

$$t_{uh} = \sum_{i=1}^{N} u_{ih} = \text{population total of variable } y \text{ in poststratum } h$$

For each poststratum h, estimate the total in the poststratum by

$$\hat{t}_{uh} = \sum_{i \in S} \frac{N}{n} \cdot u_{ih} \text{ and } \hat{N}_h = \sum_{i \in S} \frac{N}{n} \cdot x_{ih}$$

$$\hat{t}_{uhr} = \frac{t_{xh}}{\hat{t}_{xh}} \cdot \hat{t}_{uh} = \frac{N_h}{\hat{N}_h} \cdot \hat{t}_{uh} = N_h \cdot \bar{y}_h$$

Poststratified estimator of the population total is

$$\hat{t}_{ypost} = \sum_{h=1}^{H} \hat{t}_{uhr} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \cdot \hat{t}_{uh} = \sum_{h=1}^{H} N_h \bar{y}_h$$

ratio estimation is used within each poststratum to estimate the population total in that poststratum.

The poststratified estimator of $\bar{y}_U$

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \bar{y}_h$$

where N_h/N known, $n_h \geq 30$ and n large

Approximately proportional allocation

$$\hat{V}(\bar{y}_{\text{post}}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \cdot \frac{s_h^2}{n},$$

when the expected sample sizes in each poststratum are large.

Poststratification using weights

Poststratification modifies the weights so that they sum to N_h in poststratum h. Let w_i represent the sampling weight for unit i. For an SRS, for example, $w_i = N/n$ for each unit in the sample. Recall that

$$\hat{t}_{ypost} = \sum_{h=1}^{H} \hat{t}_{uhr} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} \cdot \hat{t}_{uh} = \sum_{h=1}^{H} \frac{N_h}{\hat{N}_h} w_i x_{ih} y_{ih}$$

Define

$$w_i^* = (N_h/\hat{N}_h) w_i$$

when unit i is in poststratum h. Then estimate the population total t_y by

$$\hat{t}_{ypost} = \sum_{i \in S} w_i^* y_i$$

and the population mean by

$$\bar{y}_{post} = \frac{\sum_{i \in S} w_i^* y_i}{\sum_{i \in S} w_i^*}.$$

For an SRS, the above reduces to

$$\bar{y}_{\text{post}} = \sum_{h=1}^{H} \frac{N_h}{N} \cdot \bar{y}_h, \quad \sum_{i \in S} w_i^* = N$$

Example 4.9. Example 3.2 displayed estimates for a stratified random sample from the Census of Agriculture population. The stratified sample, taken with proportional allocation, produced estimates with smaller variances than the SRS in Example 2.6.

But what if you took an SRS and only later realized that you should have taken a stratified sample? Or if you did not have region membership available for the counties in the sampling frame? Let's poststratify the SRS from Example 2.6 and find out. The quantities needed for the calculation are given in Table 4.4.

TABLE 4.4
Weight adjustments for poststratification in Example 4.9. The last two columns, $\bar{y}_h$ and s_h, give the poststratum mean and standard deviation, respectively.

Region	N_h	n_h	$\hat{N}_h$	w_i	w_i^*	$\bar{y}_h$	s_h
Northeast	220	24	246.24	10.26	9.1667	71970.83	65000.06
North Central	1054	107	1097.82	10.26	9.8505	350292.01	294715.13
South	1382	130	1333.80	10.26	10.6308	206246.35	277433.61
West	422	39	400.14	10.26	10.8205	598680.59	516157.67
Total	3078	300	3078.00				

The poststratification-adjusted weights w_i^* differ from the original sampling weights $w_i = 3078/300 = 10.26$. The poststratified weight for every county in the Northeast, where the SRS contained more units than would have been drawn in a stratified sample with proportional allocation, is

$$w_i^* = \frac{N_{\text{Northeast}}}{\hat{N}_{\text{Northeast}}} w_i = \frac{220}{246.24}(10.26) = 9.1667.$$

The poststratified weight for Northeast counties is smaller than 10.26 to account for the fact that the randomly selected sample contained, by chance, more counties in the poststratum than its share of the population. The poststratified weights in the West poststratum are larger than 10.26 to correct for the sample having more Western counties than it would under proportional allocation.

The weight adjustments force the estimated counts from each poststratum to equal the true poststratum count, N_h. Thus, $\sum_{i\in\mathcal{S}} w_i^* x_{i1} = (24)(9.1667) = 220$, $\sum_{i\in\mathcal{S}} w_i^* x_{i2} = 1054$, $\sum_{i\in\mathcal{S}} w_i^* x_{i3} = 1382$, and $\sum_{i\in\mathcal{S}} w_i^* x_{i4} = 422$.

The poststratified estimate of the population mean is

$$\bar{y}_{\text{post}} = \frac{\sum_{i\in\mathcal{S}} w_i^* y_i}{\sum_{i\in\mathcal{S}} w_i^*} = \frac{15833583 + 369207778.5 + 285032455.7 + 252643209}{3078} = 299{,}778.$$

From (4.27), the standard error of $\bar{y}_{\text{post}}$ is

$$\text{SE}\left(\bar{y}_{\text{post}}\right) = \sqrt{\left(1 - \frac{300}{3078}\right) \sum_{h=1}^{H} \frac{N_h}{3078} \frac{s_h^2}{300}} = 17{,}443.$$

By contrast, the standard error of the sample mean, $\bar{y}$, from Example 2.7, is 18,898. The poststratification reduces the standard error because the weighted average of the within-poststratum variances, $\sum_{h=1}^{H}(N_h/N)s_h^2$, is smaller than s^2. ■

Difference between stratification and poststratification. In both stratification and post-

Ratio Estimation with Stratified Samples

Combined ratio estimator

- First, estimate t_x and t_y using all the data
- Next ratio estimation is applied

$$\hat{t}_{yrc} = \hat{B} t_x, \text{ where } \hat{B} = \frac{\hat{t}_{y,str}}{\hat{t}_{x,str}}$$

$$\hat{t}_{y,str} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}$$

and

$$\hat{t}_{x,str} = \sum_{h=1}^{H} N_h \bar{x}_h = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} x_{hj}$$

with $w_{hj} = N_h / n_h$

$$MSE(\hat{t}_{yrc}) \approx V(\hat{t}_{y,str} - B\hat{t}_{x,str}) = V\left[\sum_{h=1}^{H}\sum_{j\in S_h} w_{hj}(y_{hj} - Bx_{hj})\right]$$

$$\begin{aligned}
\hat{V}(\hat{t}_{yrc}) &= \left(\frac{t_{x,str}}{\hat{t}_{x,str}}\right)^2 \hat{V}\left(\sum_{h=1}^{H}\sum_{j\in S_h} w_{hj}e_{hj}\right) \\
&= \left(\frac{t_{x,str}}{\hat{t}_{x,str}}\right)^2 \hat{V}(\hat{t}_{e,str}) \\
&= \left(\frac{t_{x,str}}{\hat{t}_{x,str}}\right)^2 [\hat{V}(\hat{t}_{y,str}) + \hat{B}^2\hat{V}(\hat{t}_{x,str}) - 2\hat{B}\widehat{Cov}(\hat{t}_{y,str}, \hat{t}_{x,str})]
\end{aligned}$$

where $e_{hj} = y_{hj} - \hat{B}x_{hj}$.

Separate ratio estimator

- Ratio estimation is applied separately in each stratum
- Next the strata estimators are combined to estimate the population total

$$\hat{t}_{yrs} = \sum_{h=1}^{H} \hat{t}_{yhr} = \sum_{h=1}^{H} t_{xh} \cdot \frac{\hat{t}_{yh}}{\hat{t}_{xh}} = \sum_{h=1}^{H} t_{xh} \cdot \hat{B}_h$$

with

$$\hat{V}(\hat{t}_{yrs}) = \sum_{h=1}^{H} \hat{V}(\hat{t}_{yhr})$$

Comments on Estimators

- A separate ratio estimator can improve efficiency if the ratio $\hat{B}_h = \frac{\hat{t}_{yh}}{\hat{t}_{xh}}$ varies significantly between strata.
- A separate ratio estimator should not be used when strata sample sizes are small, as each ratio is biased, and the bias may propagate across strata.
- Poststratification is a special case of the separate ratio estimator.
- The combined estimator has less bias when sample sizes in some strata are small.
- However, when ratios vary greatly between strata, the combined estimator does not leverage the efficiency gain from stratification that the separate ratio estimator does.

Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.