STAT 152, Fall 2003
Summary for Chapter 3

The theme for this chapter is using known information to improve the precision of estimates of means and totals. Ratio and Regression estimation improve precision when the variability of the residuals $(y_i - \widehat{y}_i)$, where $\widehat{y}_i$ is the predicted value from the ratio or regression model, is smaller than the variability of the $(y_i - \overline{y})$.

1 Important Formulas in Ratio Estimation

For ratio estimation to apply, two quantities y_i and x_i must be measured on each sample unit. If an SRS is taken, natural estimators for ratio B, population total t_y, and population mean $\overline{y}_U$ are:

- $\widehat{B} = \frac{\overline{y}}{\overline{x}} = \frac{\widehat{t}_y}{\widehat{t}_x}$
- $\widehat{t}_{yr} = \widehat{B} t_x$
- $\widehat{\overline{y}}_r = \widehat{B} \overline{x}_U$

Bias and mean squared error of ratio estimators:

- $|Bias(\widehat{B})| = -Cov(\widehat{B}, \overline{x})/\overline{x}_U$
- $\frac{|Bias(\widehat{B})|}{[V(\widehat{B})]^{1/2}} \leq CV(\overline{x})$
- $E[\widehat{B} - B] \approx (1 - \frac{n}{N}) \frac{1}{n\overline{x}_U^2} (BS_x^2 - RS_xS_y) = \frac{1}{\overline{x}_U^2}[BV(\overline{x}) - Cov(\overline{x}, \overline{y})]$
- $E[(\widehat{B} - B)^2] \approx \frac{1}{\overline{x}_U^2} V(\overline{d})$, where $\overline{d} = \overline{y} - B\overline{x}$
- $\widehat{V}[\widehat{B}] = (1 - \frac{n}{N}) \frac{s_e^2}{n\overline{x}_U^2} = (1 - \frac{n}{N}) \frac{1}{n\overline{x}_U^2} \frac{\sum_{i \in S} (y_i - \widehat{B}x_i)^2}{n-1}$, where $e_i = y_i - \widehat{B}x_i$
- $\widehat{V}[\widehat{t}_{yr}] = \widehat{V}[t_x \widehat{B}] = N^2 (1 - \frac{n}{N}) \frac{s_e^2}{n}$
- $\widehat{V}[\widehat{\overline{y}}_r] = \widehat{V}[\overline{x}_U \widehat{B}] = (1 - \frac{n}{N}) \frac{s_e^2}{n}$

2 Regression Estimation

Ratio estimation works best if the data are well fit by a straight line through the origin. Sometimes, data appear to be evenly scattered about a straight line that does not go through the origin-that is, the data look as though the usual straight line regression model: $y = B_0 + B_1 x$

Important formulas in regression estimation:

- $\widehat{\overline{y}}_{reg} = \widehat{B}_0 + \widehat{B}_1 \overline{x}_U$, estimator of $\overline{y}_U$.
- $\widehat{B}_0$ and $\widehat{B}_1$ are the ordinary least squares regression coefficients: $\widehat{B}_1 = \frac{\sum_{i \in \mathcal{S}} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i \in \mathcal{S}} (x_i - \overline{x})^2} = \frac{r s_y}{s_x}$, $\widehat{B}_0 = \overline{y} - \widehat{B}_1 \overline{x}$.
- Bias: $E[\widehat{\overline{y}}_{reg} - \overline{y}_U] = -Cov(\widehat{B}_1, \overline{x})$
- Mean Squared Error: $\text{MSE}(\widehat{\overline{y}}_{reg}) \approx (1 - \frac{n}{N}) \frac{S_d^2}{n}$, where $S_d^2 = \sum_{i=1}^{N} \frac{(y_i - [\overline{y}_U + B_1(x_i - \overline{x}_U)])^2}{N-1}$
- Standard Error: $\text{SE}(\widehat{\overline{y}}_{reg}) = \sqrt{(1 - \frac{n}{N}) \frac{s_e^2}{n}}$, where $e_i = y_i - (\widehat{B}_0 + \widehat{B}_1 x_i)$.

3 Estimation in Domains

Suppose there are D domains. Let $\mathcal{U}_d$ be the index set of the units in the population that are in domain d and let $\mathcal{S}_d$ be the index set of the units in the sample that are in domain d, for $d = 1, 2, ..., D$. Let N_d be the number of population units in $\mathcal{U}_d$, and n_d be the number of sample units in $\mathcal{S}_d$.

Important formulas in domain estimation:

- $\overline{y}_{U_d} = \sum_{i \in \mathcal{U}_d} \frac{y_i}{N_d}$.
- $\overline{y}_d = \sum_{i \in \mathcal{S}_d} \frac{y_i}{n_d}$, a natural estimator of $\overline{y}_{U_d}$
- Standard Error of $\overline{y}_d$: $\text{SE}(\overline{y}_d) \approx \sqrt{(1 - \frac{n}{N}) \frac{s_{y_d}^2}{n_d}}$
- $\widehat{t}_{y_d} = N \overline{u}$, where $u_i = \begin{cases} y_i & \text{if } i \in \mathcal{U}_d \\ 0 & \text{if } i \notin \mathcal{U}_d \end{cases}$
- Standard Error of $\widehat{t}_{y_d}$: $\text{SE}(\widehat{t}_{y_d}) = N \sqrt{(1 - \frac{n}{N}) \frac{s_u^2}{n}}$

4 Comparison

	Population Total Estimator	Population Mean Estimator	e_i
SRS	$\widehat{t}_y$	$\overline{y}$	$y_i - \overline{y}$
Ratio	$\widehat{t}_y(\frac{t_x}{\widehat{t}_x})$	$\overline{y}(\frac{t_x}{\widehat{t}_x})$	$y_i - \widehat{B}x_i$
Regression	$N[\overline{y} + \widehat{B}_1(\overline{x}_U - \overline{x})]$	$\overline{y} + \widehat{B}_1(\overline{x}_U - \overline{x})]$	$y_i - \widehat{B}_0 - \widehat{B}_1 x_i$

The population total estimator variance is $N^2(1 - \frac{n}{N})\frac{s_e^2}{n}$.

The population mean estimator variance is $(1 - \frac{n}{N})\frac{s_e^2}{n}$.