

6 Ratio, Regression, and Difference Estimation

6.1 Introduction

Previously we studied *simple random sampling* and *stratified random sampling*.

We estimated the population mean (μ_y), population total (τ_y), and population proportion (p_y), based on just ONE variable Y .

Now, we will consider the relationship between two variables, X and Y .

We might still be interested in just one parameter, say μ_y , but data from both variables X and Y might help us estimate μ_y better than data based merely on variable Y .

6.2 Surveys That Require the Use of Ratio Estimators

(Easy) Example: Suppose we want to estimate the total number of women at some university who plan to pursue careers in teaching.

- ⊙ Suppose there are 10,000 women at this university.
- ⊙ Take a sample of size $n = 1,000$ PEOPLE (i.e., including men and women).

- ⊙ Out of the 1000 people sampled, 600 are women, of whom 90 plan to pursue careers in teaching.
- ⊙ How should we estimate τ_y , the total number of women from this university who plan to pursue careers in teaching?

- ⊙ Recall that a proportion is a special case of a mean using zeros and ones.

- ⊙ Let X_i be 1 if the i th person sampled is a woman, and 0 otherwise.
- ⊙ Let Y_i be 1 if the i th person sampled is a woman who plans a career in teaching, and 0 otherwise.
- ⊙ In terms of X_i and Y_i , what is the sample proportion of women who want to teach?

- ⊙ What is τ_x ?

□

General formula for estimating the total (p. 182):

$$\hat{\tau}_y = \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \right) \tau_x = \left(\frac{n\bar{Y}}{n\bar{X}} \right) \tau_x = \left(\frac{\bar{Y}}{\bar{X}} \right) \tau_x$$

In the above example involving women and teaching,

$$\hat{\tau}_y = \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \right) \tau_x = (90/600) \times 10000 = 1500$$

Example: Predict total church donations (or revenue from a business or taxes) for the current calendar year.

- ⊙ Are church donations uniform throughout a given calendar year?

- ⊙ Suppose that \$220,000 has been collected thus far during the current calendar year, and that today is April 2. Therefore, 13 weeks (or 91 days or 3 months or 1/4 of the calendar year) has transpired.
- ⊙ Do these 13 weeks represent a simple random sample of church donations for the entire calendar year?

- ⊙ Should we estimate the *total* donations for the current calendar year to be $\$220,000 \times 4 = \$880,000$?
- ⊙ Suppose that the donations from the previous calendar year totaled $\$1,000,000$.
- ⊙ Should we estimate the *total* donations for the current year to be $\$1,000,000$ also?

- ⊙ How should we estimate the *total* donations for the current calendar year?

Suppose the church donations from January 1 to April 2 of LAST year total $\$200,000$.

Notation: Let Y_i be the church donation for the i th week THIS year, for $i = 1, \dots, 13$ (or 52).

Let X_i be the church donation for the i th week LAST year, for $i = 1, \dots, 52$.

What are τ_x and τ_y ?

□

Really COOL Example: France wanted a population census in 1802, and Pierre-Simon Laplace intended to estimate the number of people living there.

- ⊙ Laplace recognized the extreme difficulty in taking an accurate census.
- ⊙ However, the registers of births were kept with great accuracy.

- ⊙ Let Y be the population total for a particular *commune* in France.
- ⊙ Let N be the number of communes in France.
- ⊙ *Goal:* Estimate $\tau_y = \sum_{i=1}^N y_i$, the population total of France in 1802.
- ⊙ Let X be the number of births in the past year for a particular *commune* in France.

- ⊙ $\tau_x = \sum_{i=1}^N x_i$, the total number of registered births, from the previous year was KNOWN, and was approximately 1,000,000.

For some reason, Laplace used $\tau_x = 1,000,000$, “which is nearly correct” according to Laplace.

- ⊙ Laplace took a sample of $n = 30$ communes (where the mayor was both “intelligent” and “zealous”) throughout the country.

Did these 30 communes represent a simple random sample?

- ⊙ For each commune, Laplace determined Y (the population total of the commune) and X (the number of registered births in the commune) for the previous year.

Actually, Laplace determined the number of births in the past three years in the commune, and divided this number by 3.

- ⊙ Laplace determined that $n\bar{Y} = 2,037,615$.

- ⊙ Laplace also determined that $n\bar{X} = 71,866$.

- ⊙ What does the ratio $2,037,615/71,866 = 28.353$ represent?

- ⊙ Now, estimate τ_y , the population total of France in 1802.

- ⊙ Find a second way to estimate the population total of France in 1802, if there were, say, 450 communes in France.

- ⊙ Laplace reasoned that the ratio estimator would be more precise, due to the positive correlation between X and Y . □

In general, a ratio estimator works well when there is a strong linear association (correlation) through the origin between X and Y ; i.e., the ratio Y/X is approximately the same for all (X, Y) .

6.3 Ratio Estimation Using Simple Random Sampling

Notation: $R = \tau_y/\tau_x$

Estimate the ratio R using a ratio estimator.

We estimate R by

$$(6.1) \quad r = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \frac{n^{-1} \sum_{i=1}^n Y_i}{n^{-1} \sum_{i=1}^n X_i} = \frac{\bar{Y}}{\bar{X}}$$

Sample ratios are NOT unbiased for their population values.

For many types of situations when the sample size is large, the standard deviation of the estimator is much larger than the absolute value of the bias.

Hence, in many situations, for large n , the bias is negligible.

The estimated variance of r is

$$(6.2) \quad \hat{V}(r) = \left(\frac{N-n}{nN} \right) \mu_x^{-2} s_r^2,$$

where

$$(6.3) \quad s_r^2 = \frac{\sum_{i=1}^n (Y_i - rX_i)^2}{n-1}$$

Exercise 6.10, p. 221: Members of a teachers' association are concerned about the salary increases given to high school teachers in a particular school system. A simple random sample of $n = 15$ teachers is selected from an alphabetical listing of all $N = 750$ high school teachers in the system. All 15 teachers are interviewed to determine their salaries for this year and the previous year.

(a) Plot the data, and comment.

```
> salaries = scan2( "EXER6_10.DAT", T, T )
```

```
> x = salaries[ , 2] # past year's salary
```

```
> y = salaries[ , 3] # present year's salary
```

(b) Determine the correlation coefficient between x and y , and comment.

(c) Estimate R , the rate of change in average salary for the 750 high school teachers in the community school system, and interpret your result.

```
> # Use equation (6.1), p. 184.
```

(d) Estimate the variance of r .

```
> # Use equations (6.2) and (6.3), p. 184.
```

(e) Estimate B , the bound on the error of estimation.

(f) Construct a 95% confidence interval on R .

We are 95% confident that the rate of change in average salary is between 1.034 and 1.042. \square

Estimate the population total τ_y using a ratio estimator.

Note that in section 6.2, we considered the estimator

$$(6.4) \quad \hat{\tau}_y = \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \right) \tau_x = (\bar{Y}/\bar{X})\tau_x = r\tau_x$$

The estimated variance of $\hat{\tau}_y$ ($= r\tau_x$) is

$$(6.5) \quad \hat{V}(r\tau_x) = \tau_x^2 \hat{V}(r) = \tau_x^2 \left(\frac{N-n}{nN} \right) \mu_x^{-2} s_r^2 = N^2 \left(\frac{N-n}{nN} \right) s_r^2$$

Exercise 6.1, p. 218: A forester is interested in estimating the total volume of trees in a timber sale (perhaps from George W. Bush's profitable timber company - "Need some wood?"). The forester records the volume for each tree in a simple random sample. In addition, he measures the basal area for each tree marked for sale. He then uses a ratio estimator of total volume.

The forester decides to take a simple random sample of $n = 12$ from the $N = 250$ trees marked for sale. Let x denote basal area and y the cubic-foot volume for a tree. The total basal area, τ_x , for all 250 trees is 75 square feet.

(a) Plot the data, and comment.

```
> trees = scan2( "EXER6_1.DAT", T )
```

```
> x = trees[ , 2] # area
```

```
> y = trees[ , 3] # volume
```

(b) Determine the correlation coefficient between x and y , and comment.

(c) Estimate R , the ratio of the total volume to the total basal area of the 250 trees marked for sale.

```
> # Use equation (6.1), p. 184.
```

If the trees were about the same height and shaped like cylinders, then this height would be about 21.19 feet.

(d) Estimate τ_y , the total volume of 250 trees marked for sale.

```
> # Use equation (6.4) on p. 188.
```

(e) Estimate the variance of $\hat{\tau}_y$.

```
> # Use equation (6.5) on p. 188.
```

(f) Estimate B , the bound on the error of estimation.

(g) Construct a 95% confidence interval on τ_y .

We are 95% confident that the total volume of the trees is between 1403.235 cubic feet and 1775.87 cubic feet. \square

Exercise 6.2, p. 219: Consider the forestry data in Exercise 6.1.

(a) Compute an estimate of τ_y using $N\bar{Y}$.

(b) Why is the estimate $N\bar{Y}$, which does not use any basal-area data, much larger than the ratio estimate?

> x = trees[, 2] # area

\square

Estimate the population mean μ_y using a ratio estimator.

What is the relationship between μ_y and τ_y ?

What is the relationship between estimators $\hat{\mu}_y$ and $\hat{\tau}_y$?

What is the relationship between the estimated variances of estimators $\hat{\mu}_y$ and $\hat{\tau}_y$?

$$(6.6) \quad \hat{\mu}_y = \left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} \right) \mu_x = (\bar{Y}/\bar{X})\mu_x = r\mu_x$$

The estimated variance of $\hat{\mu}_y$ ($= r\mu_x$) is

$$(6.7) \quad \hat{V}(r\mu_x) = \mu_x^2 \hat{V}(r) = \mu_x^2 \left(\frac{N-n}{nN} \right) \mu_x^{-2} s_r^2 = \left(\frac{N-n}{nN} \right) s_r^2$$

where

$$s_r^2 = \frac{\sum_{i=1}^n (Y_i - rX_i)^2}{n-1}$$

Exercise 6.6, p. 220: An investigator has a colony of $N = 763$ rats that have been subjected to a *standard* drug. The average length of time to thread a maze correctly under influence of the *standard* drug was found to be $\mu_x = 17.2$ seconds. The investigator now would like to subject a random sample of 11 rats to the *new* drug.

(a) Plot the data, and comment.

```
> drugs = scan2( "EXER6_6.DAT", T )
> x = drugs[ , 2] # time, due to standard drug
> y = drugs[ , 3] # time, due to new drug
```

(b) What is the most unusual feature of the scatter plot for these data?

(c) Determine the correlation coefficient between x and y , and comment.

(d) Estimate R , the ratio of the mean time to thread the maze under the influence of the *new* drug to the mean time to thread the maze under the influence of the *standard* drug.

```
> # Use equation (6.1), p. 184.
```

(e) Estimate μ_y , the average time required to thread the maze while under the influence of the *new* drug.

> # Use equation (6.6) on p. 190.

(f) Estimate the variance of $\hat{\mu}_y$.

> # Use equation (6.7) on p. 190.

(g) Estimate B , the bound on the error of estimation.

(h) Construct a 95% confidence interval on μ_y .

We are 95% confident that the average length of time to thread a maze correctly under influence of the *new* drug is between 17.32 seconds and 17.86 seconds. \square

6.4 Selecting the Sample Size

When using a ratio estimator and a fixed value of B (the bound on the error of estimation), solve for n , the required sample size.

This required sample size is

$$(6.19) \quad n = \frac{N\sigma^2}{ND + \sigma^2}, \text{ where}$$

$$D = \begin{cases} B^2\mu_x^2/4, & \text{when estimating } R = \mu_y/\mu_x, \\ B^2/4, & \text{when estimating } \mu_y, \\ B^2/(4N^2), & \text{when estimating } \tau_y. \end{cases}$$

σ^2 measures the variability in y for a fixed x value.

Note the typos in the numerator for (6.19), (6.21) and (6.23).

How should we estimate σ^2 ?

Revisit Exercise 6.6, p. 220: An investigator has a colony of $N = 763$ rats that have been subjected to a *standard* drug. The average length of time to thread a maze correctly under influence of the *standard* drug was found to be $\mu_x = 17.2$ seconds. What sample size n is needed to estimate μ_y (the average length of time to thread a maze correctly under influence of the *new* drug) with a bound on the error of estimation equal to 0.1 seconds?

□

6.5 Ratio Estimation in Stratified Random Sampling

Example: Suppose we want to compare the drinking (i.e., alcohol consumption) rates of spring semester vs. fall semester for college freshmen (who completed both fall and spring semesters) at some university while classes were in session. A simple random sample of freshmen was taken.

At the end of a freshman's *fall* semester, the freshman was asked: "On average how many drinks did you consume per week while classes were in session this semester?" The same question was asked at the end of the freshman's *spring* semester.

Let X and Y be the responses for the *fall* and *spring* semesters, respectively. Are ratio estimators reasonable for this example?

$$R = \mu_y / \mu_x$$

We want to estimate R .

Stratify according to *gender*, since alcohol consumption is likely to differ between *females* (stratum F) and *males* (stratum M).

Suppose the university consists of 60% *females* and 40% *males*.

Take a stratified sample of 100 students, where stratification may depend on cost.

Take a simple random sample of 56 *females* and a simple random sample of 44 *males*.

In the sample of *females*, the average weekly consumption of alcohol was $\bar{X}_F = 3.3$ drinks in the *fall* and $\bar{Y}_F = 3.2$ drinks in the *spring*.

In the sample of *males*, the average weekly consumption of alcohol was $\bar{X}_M = 4.9$ drinks in the *fall* and $\bar{Y}_M = 4.2$ drinks in the *spring*.

What are two reasonable estimators of R ?

⊙ Take a weighted average of the two *separate* ratio estimators.

- ⊙ Take the ratio of the two *combined* (weighted, stratified) estimators.

Which estimator is preferred?

Estimate the population mean μ_y using a *separate ratio estimator*.

Recall the (non-stratified) ratio estimator of the population mean μ_y .

$$(6.6) \quad \hat{\mu}_y = (\bar{Y}/\bar{X})\mu_x = r\mu_x$$

The estimated variance of $\hat{\mu}_y$ ($= r\mu_x$) is

$$(6.7) \quad \hat{V}(r\mu_x) = \mu_x^2 \hat{V}(r) = \left(\frac{N-n}{nN} \right) s_r^2$$

where

$$s_r^2 = \frac{\sum_{i=1}^n (Y_i - rX_i)^2}{n-1}$$

For two strata A and B , the **SEPARATE ratio estimator** of μ_y is

$$\begin{aligned} \hat{\mu}_{yRS} &= \left(\frac{N_A}{N} \right) r_A \mu_{xA} + \left(\frac{N_B}{N} \right) r_B \mu_{xB} \\ &= \left(\frac{N_A}{N} \right) \left(\frac{\bar{Y}_A}{\bar{X}_A} \right) \mu_{xA} + \left(\frac{N_B}{N} \right) \left(\frac{\bar{Y}_B}{\bar{X}_B} \right) \mu_{xB}, \end{aligned}$$

and the estimated variance is

$$\begin{aligned} \hat{V}(\hat{\mu}_{yRS}) &= \left(\frac{N_A}{N} \right)^2 \left(\frac{N_A - n_A}{N_A n_A} \right) \frac{\sum_{i=1}^{n_A} (Y_{iA} - r_A X_{iA})^2}{n_A - 1} \\ &\quad + \left(\frac{N_B}{N} \right)^2 \left(\frac{N_B - n_B}{N_B n_B} \right) \frac{\sum_{i=1}^{n_B} (Y_{iB} - r_B X_{iB})^2}{n_B - 1} \end{aligned}$$

Estimate the population mean μ_y using a combined ratio estimator.

For two strata A and B , the **COMBINED ratio estimator** of μ_y is

$$\hat{\mu}_{yRC} = \frac{\bar{Y}_{st}}{\bar{X}_{st}} (\mu_x),$$

and the estimated variance is

$$\hat{V}(\hat{\mu}_{yRC}) = \left(\frac{N_A}{N}\right)^2 \left(\frac{N_A - n_A}{N_A n_A}\right) s_{rA}^2 + \left(\frac{N_B}{N}\right)^2 \left(\frac{N_B - n_B}{N_B n_B}\right) s_{rB}^2,$$

where s_{rA}^2 is the sample variance of $(Y_i - r_C X_i)$ for stratum A , and s_{rB}^2 is the sample variance of $(Y_i - r_C X_i)$ for stratum B .

Since $\tau = N\mu$, then $\hat{\tau} = N \hat{\mu}$, and $\hat{V}(\hat{\tau}) = N^2 \hat{V}(\hat{\mu})$.

Exercise 6.25, p. 225: A certain manufacturing firm produces a product that is packaged under two brand names, for marketing purposes. These two brands serve as strata for estimating potential sales volume for the next quarter. A simple random sample of customers of each brand is contacted and asked to provide a potential sales figure y (in number of units) for the coming quarter. Last year's true sales figure, for the same quarter, is available for each of the sampled customers and is denoted by x . The sample for brand I was taken from a list of 120 customers for whom the total sales in the same quarter of last year was 24,500 units. The brand II sample came from 180 customers with a total quarterly sales last year of 21,200 units.

(a) Find a ratio estimate of the total potential sales for the next quarter.

> # Should we use *separate* ratio estimation or *combined* ratio estimation?

> sales = scan2("EXER6.25.DAT", T)

```
> brand = sales[, 1] # Brand number
> x = sales[, 2] # Last year's sales
> y = sales[, 3] # This year's sales
```

(b) Estimate the variance of your estimator from part (a).

$$\hat{V}(\hat{\tau}_{yRC}) = N_A^2 \left(\frac{N_A - n_A}{N_A n_A} \right) s_{rA}^2 + N_B^2 \left(\frac{N_B - n_B}{N_B n_B} \right) s_{rB}^2,$$

where s_{rA}^2 is the sample variance of $(Y_{iA} - r_C X_{iA})$ for stratum A , and s_{rB}^2 is the sample variance of $(Y_{iB} - r_C X_{iB})$ for stratum B .

Equation is on page 202, modulo a factor of N^2 .

(c) Compute a bound on the error of estimation.

(d) Compute a 95% confidence interval on τ_y , the total potential sales for the next quarter.

We are 95% confident that the total potential sales for the next quarter are between 46,717.06 and 49,702.62 units. \square

For estimating a **ratio** under stratified sampling:

Note, for the non-stratified case, $r = \bar{Y}/\bar{X}$ and $\hat{\mu}_y = r\mu_x$.

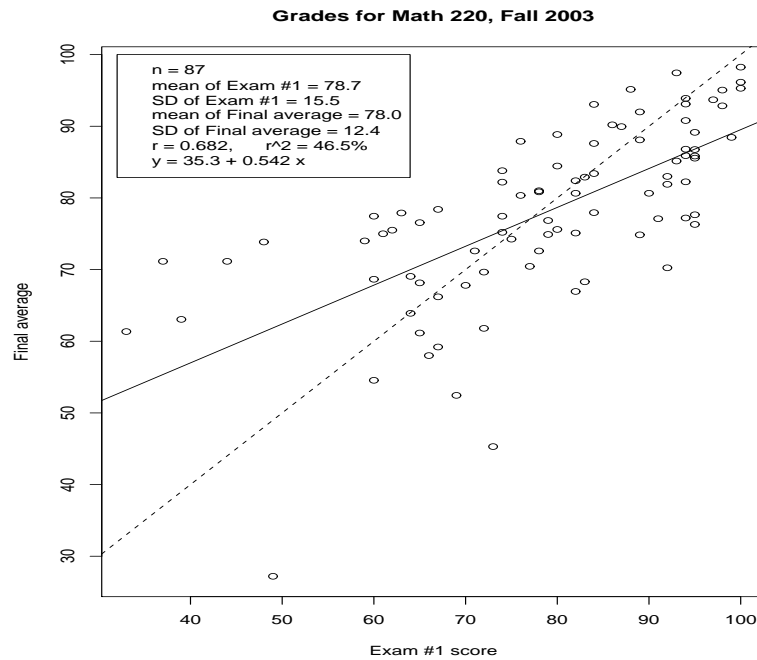
Estimate the ratio by $\bar{Y}_{st}/\bar{X}_{st}$, noting the equation on p. 202 for $\hat{\mu}_{yRC}$.

6.6 Regression Estimation

When does *ratio* estimation (from section 6.3) work well?

Regression estimation works well when a linear relationship between x and y exists but not necessarily through the origin.

The purpose of regression is to *explain* and *predict*.



Recall from Math 220 (or 318): The fitted regression line always goes through what point?

The equation of the fitted regression line:

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

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

and $a = (\text{estimated intercept}) = \bar{y} - b\bar{x}$, since $\bar{y} = a + b\bar{x}$ must be a valid solution to the regression equation.

Substituting for a into $\hat{y} = a + bx$

results in $\hat{y} = \bar{y} + b(x - \bar{x})$.

Suppose the population in X has mean μ_x , then

$$(6.24) \quad \hat{\mu}_{yL} = \bar{y} + b(\mu_x - \bar{x}).$$

The **estimated variance** of $\hat{\mu}_{yL}$ is:

$$(6.25) \quad \hat{V}(\hat{\mu}_{yL}) = \left(\frac{N-n}{Nn}\right) \left(\frac{1}{n-2}\right) \left[\sum_{i=1}^n (y_i - \bar{y})^2 - b^2 \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

$$(6.26) \quad = \left(\frac{N-n}{Nn}\right) \text{MSE},$$

where MSE is the mean square error from the standard simple linear regression of y on x .

Where is the *finite population correction* (fpc) in the above equation?

Exercise 6.33, p. 227: The Florida Game and Freshwater Commission is interested in estimating weights of alligators from much more easily observed lengths. Data on the lengths (in inches) and weights (in pounds) of 25 alligators are given in the accompanying table.

In equation (6.25) for the estimated variance, how should $(N-n)/(Nn)$ be approximated?

(a) First, scan in the data.

```
> alligators = scan2( "EXER6_33.DAT", T )
```

(b) Next, plot the data.

```
> x = alligators[ , 2] # lengths of alligators in inches
```

```
> y = alligators[ , 3] # weight of alligators in pounds
```

(c) Estimate the average weight of a population of alligators for which the average length is 100 inches.

```
> # First find the slope.
```

```
> lm( y ~ x )
```

```
> mu.x = 100
```

```
> Apply equation (6.24).
```

(d) Estimate the variance of the estimator $\hat{\mu}_{yL}$.

```
> # Use equation (6.25).
```

```
> # Alternative method: Obtain the MSE and apply equation (6.26), and note that  
the finite population correction is approximated by one.
```

```
> summary( lm( y ~ x ) )
```

(e) Compute B , the bound on the error of estimation.

(f) Construct a 95% confidence interval on μ_{yL} .

We are 95% confident that the average weight of alligators is between 175.3668 pounds and 218.576 pounds, for which the average length is 100 inches.

□

6.7 Difference Estimation

Difference estimation is similar to **regression** estimation, except **difference** estimation uses slope equal to one.

Interpretation: The average *difference* between X and Y is constant, and does not depend on the size of X (or Y).

Difference estimation is often used in *auditing*.

Example 6.10, p. 208: A population contains 180 inventory items with a (total) stated book value of \$13,320. Let x_i denote the *book* value and y_i denote the *audit* value.

```
> items = scan2( "EXPL6_10.DAT", T )
```

```
> y = items[ , 2] # audit value
```

```
> x = items[ , 3] # book value
```

(a) Plot the data, and determine if the *difference* estimator is appropriate for this data set.

(b) Define the difference to be $(y - x)$, and compute the sample mean difference.

(c) How should we estimate the mean *audit* value?

In general,

$$(6.27) \quad \hat{\mu}_{yD} = \mu_x + \bar{d}, \text{ where } \bar{d} = \bar{y} - \bar{x}.$$

(d) Determine the estimated variance of $\hat{\mu}_{yD}$.

Thus, what is the variance of $\mu_x + \bar{d}$?

Note: \bar{d} is a sample mean, based on a simple random sample of variable d_i (where d_i is based on the paired observations of x_i and y_i).

What is the estimated variance of a sample mean?

Therefore, the estimated variance of $\hat{\mu}_{yD}$ is:

$$(6.28) \quad \hat{V}(\hat{\mu}_{yD}) = \left(\frac{N-n}{Nn} \right) \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1},$$

where $d_i = y_i - x_i$.

(e) Determine the bound on the error of estimation.

(f) Construct a 95% confidence interval on μ_y .

Interpretation: We are 95% confident that the population mean *audit* value is between \$72.86 and \$75.94. \square

Suppose we are estimating τ_y (as in exercise 6.23c) rather than μ_y .

How should equations (6.27) and (6.28) be modified?

6.9 Summary

Read this section on p. 217. It is only 1/2 page long, and it solves exercise 6.19 on p. 223!

A *regression* estimator is valid when a strong linear relationship between the two variables exists.

A *ratio* estimator tends to work well when a strong linear relationship between the two variables through the origin exists.

A *difference* estimator is similar to a *regression* estimator, but the *difference* estimator uses slope equal to one.