Stat 257: Solutions to Assignment #3

(3.1) Let X denote the assessed yields, and let Y denote the actual yields. Our goal is to estimate Y_T. In addition to the data clearly given in the problem, note that we also know the following: $N = 280$, $n = 25$, and $X_T = 439.5$.

simple random sample estimator

If we consider the estimator based solely on the values of the actual yields, then we obtain

$$y_T = N\overline{y} = \frac{N}{n}\sum_{i=1}^{n} y_i = 280 \cdot \frac{39.8}{25} = 280 \cdot 1.592 = 445.76$$

with estimated variance

$$\begin{aligned} s^2(y_T) = N^2 s^2(\overline{y}) = N^2 \frac{(1-f)}{n(n-1)}\sum_{i=1}^{n}(y_i - \overline{y})^2 &= N^2 \frac{(1-f)}{n(n-1)}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) \\ &= 280^2 \cdot \frac{\left(1-\frac{25}{280}\right)}{25\cdot 24} \cdot \left(69.08 - 25 \cdot 1.592^2\right) \\ &= 680.4896. \end{aligned}$$

ratio estimator

The method of ratio estimation provides us with the estimate

$$y_{TR} = rX_T = \frac{y_T}{x_T} \cdot X_T = \frac{39.8}{41.4} \cdot 439.5 \approx 422.51$$

which has estimated variance

$$\begin{aligned} s^2(y_{TR}) = N^2 \cdot \frac{(1-f)}{n} \cdot \sum_{i=1}^{n}\frac{(y_i - rx_i)^2}{n-1} &= N^2 \cdot \frac{(1-f)}{n(n-1)} \cdot \left(\sum_{i=1}^{n} y_i^2 - 2r\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} x_i^2\right) \\ &\approx \frac{280^2 \cdot \left(1 - \frac{25}{280}\right)}{25 \cdot 24} \cdot \left(69.08 - 2 \cdot \frac{39.8}{41.4} \cdot 70.64 + \left(\frac{39.8}{41.4}\right)^2 \cdot 73.47\right) \\ &\approx 138.1581. \end{aligned}$$

regression estimator

In order to determine the regression estimate, we begin by computing the estimated slope of the regression line, namely

$$\tilde{b} = \frac{s_{YX}}{s_X^2} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{y}\,\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{70.64 - 25 \cdot \frac{39.8}{25} \cdot \frac{41.4}{25}}{73.47 - 25 \cdot \left(\frac{41.4}{25}\right)^2} \approx 0.963.$$

This gives the regression estimate as

$$\begin{aligned} y_{LT} = N \cdot \overline{y}_L &= N\left(\overline{y} + \tilde{b}(\overline{X} - \overline{x})\right) \\ &\approx 280 \cdot \left(1.592 + 0.963 \cdot \left(\frac{439.5}{280} - \frac{41.4}{25}\right)\right) \\ &\approx 422.47. \end{aligned}$$

We find

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \frac{69.08 - 25\cdot\left(\frac{39.8}{25}\right)^2}{24} \approx 0.2383$$

so that y_{TL} has estimated standard variance

$$\begin{aligned} s^2(y_{TL}) &= N^2 \cdot \frac{(1-f)}{n}\cdot\left(s_Y^2 - \tilde{b}s_{YX}\right) \\ &\approx 280^2\cdot\frac{\left(1-\frac{25}{280}\right)}{25}\cdot(0.2383 - 0.963\cdot 0.1971) \\ &\approx 138.1559. \end{aligned}$$

Hence, approximate 95% confidence intervals for Y_T are given by

- $445.76 \pm 2\sqrt{680.4896}$ or 445.8 ± 52.2 (simple random sampling estimation),
- $422.51 \pm 2\sqrt{138.1581}$ or 422.5 ± 23.5 (ratio estimation),
- $422.47 \pm 2\sqrt{138.1559}$ or 422.5 ± 23.5 (regression estimation).

Note that the estimated standard errors are simply the square roots of the estimated variances, namely

- $s(y_T) \approx \sqrt{680.4896} \approx 26.09$,
- $s(y_{TR}) \approx \sqrt{138.1581} \approx 11.75$,
- $s(y_{TL}) \approx \sqrt{138.1559} \approx 11.75$.

The estimated relative efficiencies are the ratios of the estimated variances. That is,

$$\text{RelEff}(y_{TR}, y_T) = \frac{s^2(y_{TR})}{s^2(y_T)} \approx \frac{138.1581}{680.4896} \approx 20.3\%$$

and

$$\text{RelEff}(y_{TL}, y_T) = \frac{s^2(y_{TL})}{s^2(y_T)} = \frac{138.1559}{680.4896} \approx 20.3\%.$$

(3.3) It appears that the most appropriate method for estimating $\overline{Y}$ is regression estimation. This is arguably the best choice because we are observing bivariate data (X =height, and Y) and we have complete knowledge about X. Furthermore, it appears that there is a rough linear relationship between X and Y which does not pass through the origin. From the data presented, we observe that $N = 560$, $n = 10$, $\overline{X} = 173.2$, and we calculate that

$$\sum_{i=1}^{n} y_i = 34.1, \quad \sum_{i=1}^{n} y_i^2 = 117.67, \quad \sum_{i=1}^{n} x_i = 1707, \quad \sum_{i=1}^{n} x_i^2 = 292069, \quad \sum_{i=1}^{n} y_i x_i = 5813.4.$$

In order to determine the regression estimate, we begin by computing the estimated slope of the regression line, namely

$$\tilde{b} = \frac{s_{YX}}{s_X^2} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{5813.4 - 10\cdot\frac{1707}{10}\cdot\frac{34.1}{10}}{292069 - 10\cdot\left(\frac{1707}{10}\right)^2} \approx -0.0109.$$

This gives the regression estimate as

$$\begin{aligned}\overline{y}_L &= \overline{y} + \tilde{b}(\overline{X} - \overline{x}) \\ &\approx 3.41 - 0.0109 \cdot (173.2 - 170.7) \\ &\approx 3.38.\end{aligned}$$

We find

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = \frac{117.67 - 10 \cdot \left(\frac{34.1}{10}\right)^2}{9} \approx 0.1543$$

so that $\overline{y}_L$ has estimated standard variance

$$\begin{aligned}s^2(\overline{y}_L) &= \frac{(1-f)}{n} \cdot \left(s_Y^2 - \tilde{b} s_{YX}\right) \\ &\approx \frac{\left(1 - \frac{10}{560}\right)}{10} \cdot [0.1543 - (-0.0109) \cdot (-0.83)] \\ &\approx 0.0143.\end{aligned}$$

This gives an estimated standard error of $s(\overline{y}_L) \approx 0.119$ so that an approximate 95% confidence interval for $\overline{Y}$ is

$$3.38 \pm 2 \cdot 0.119 \quad \text{or} \quad (3.14, 3.61).$$