Stat 257 Fall 2005: Selected Solutions to Midterm

Remark: The following numerical solutions must be adjusted if the approximation $z_{0.05} = 2$ is made.

2. (a) If we consider the estimator of $\overline{Y}$ based on the values of y alone, then we obtain

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{42\ 500}{50} = 850.$$

We also find

$$s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right) = \frac{1}{49}\cdot\left(37\ 250\ 000 - 50\cdot 850^2\right) \approx 22959$$

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

$$s^2(\overline{y}) = \frac{(1-f)}{n}s_Y^2 \approx \frac{\left(1-\frac{50}{12\ 700}\right)}{50}\cdot 22959 \approx 457.38.$$

Thus, an approximate 95% confidence interval for $\overline{Y}$ is given by

$$\overline{y} \pm z_{0.05}s(\overline{y}) \quad \text{or} \quad 850 \pm 1.96\cdot 21.39 \quad \text{or} \quad 850 \pm 42 \quad \text{or} \quad (808, 892).$$

2. (b) In order to determine the regression estimate, we begin by computing the estimated slope of the regression line, namely

$$\tilde{b} = \frac{s_{YX}}{s_X^2} = \frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{y}\,\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{4\ 351\ 000 - 50\cdot 850\cdot 100}{516\ 500 - 50\cdot 100^2} \approx 6.12.$$

This gives the regression estimate as

$$\overline{y}_L = \overline{y} + \tilde{b}(\overline{X} - \overline{x}) \approx 850 + 6.12\cdot(108-100) \approx 899.$$

We also observe that

$$s_{YX} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x}) = \frac{4\ 351\ 000 - 50\cdot 850\cdot 100}{49} \approx 2061.$$

Hence, we find that $\overline{y}_L$ has estimated variance

$$\begin{aligned} s^2(\overline{y}_L) &= \frac{(1-f)}{n}\cdot\left(s_Y^2 - \tilde{b}s_{YX}\right) \\ &\frac{\left(1-\frac{50}{12\ 700}\right)}{50}\cdot(22959 - 6.12\cdot 2061) \\ &\approx 206.03. \end{aligned}$$

Thus, an approximate 95% confidence interval is given by

$$\overline{y}_L \pm z_{0.05}s(\overline{y}_L) \quad \text{or} \quad 899 \pm 1.96\cdot 14.35 \quad \text{or} \quad 899 \pm 28 \quad \text{or} \quad (871, 927).$$

2. (c) The relative efficiency of the regression estimator to the simple random sampling estimator is simply the ratio of their variances, namely

$$\mathrm{RelEff}(\overline{y}, \overline{y}_L) = \frac{s^2(\overline{y}_L)}{s^2(\overline{y})} \approx \frac{206.03}{457.38} \approx 45\%.$$

This gives us a strong indication that the regression estimate is strongly preferrable to the simple random sampling estimate for estimating the average amount spent on books.

2. (d) If we want to be 95% confident that a simple random sample estimate of $\overline{Y}$ is within \$20 of its true value, then we need to sample n students, where

$$n \geq N\left[1 + N\left(\frac{d}{z_\alpha s_Y}\right)^2\right]^{-1}$$

and $N = 12\ 700$, $d = 20$, $z_\alpha = 1.96$, and $s_Y^2 = 22959$. Substituting these values gives a required sample size of $n \geq 216.7$. Hence, the minimum number of students we need to sample in order to be 95% confident that a simple random sample estimate of $\overline{Y}$ is within \$20 of its true value is 217.