STK4600

13.4.16 and 20.4.16

Solutions to Ex. #7 — 1

1.

$$E(\hat\sigma^2) = \frac{1}{n-1}\sum_{i\in s}\frac{1}{x_i}E(Y_i-\hat R x_i)^2 \qquad (*)$$

$$E(Y_i-\hat R x_i)^2 = Var(Y_i - \hat R x_i)$$

since $E(Y_i-\hat R x_i) = \beta x_i - E(\hat R)x_i$

$$= \beta x_i - \frac{\sum_{j\in s}E(Y_j)}{\sum_{j\in s}x_j}x_i = \beta x_i - \frac{\beta\sum_{j\in s}x_j}{\sum_{j\in s}x_j}x_i = \beta x_i - \beta x_i = 0$$

$$Var(Y_i-\hat R x_i) = Var(Y_i) + x_i^2 Var(\hat R) - 2x_i Cov(Y_i,\hat R)$$

$$= \sigma^2 v(x_i) + x_i^2\cdot\frac{\sigma^2 n\bar v_s}{(n\bar x_s)^2} - 2x_i Cov\left(Y_i, \frac{Y_i+\sum_{j\neq i, j\in s}Y_j}{n\bar x}\right)$$

$$= \sigma^2 v(x_i) + x_i^2\cdot\frac{\sigma^2 n\bar v_s}{(n\bar x_s)^2} - 2\frac{x_i}{n\bar x_s}\Big(\underbrace{Var(Y_i)}_{=\sigma^2 v(x_i)} + \underbrace{cov(Y_i,\textstyle\sum_{j\neq i}Y_j)}_{=0}\Big)$$

Fra $(*)$:

$$E(\hat\sigma^2) = \sigma^2\left[\frac{1}{n-1}\sum_{i\in s}\frac{v(x_i)}{x_i} + \frac{1}{n-1}\frac{n\bar v_s}{(n\bar x_s)^2}\sum_s x_i - \frac{2}{n-1}\frac{\sum_s v(x_i)}{n\bar x_s}\right]$$

$$= \sigma^2\left[\frac{n(v/x)_s}{n-1} + \frac{1}{n-1}\frac{\bar v_s}{\bar x_s} - \frac{2}{n-1}\frac{\bar v_s}{\bar x_s}\right]$$

$$= \sigma^2\left[(v/x)_s + \frac{1}{n-1}(v/x)_s - \frac{1}{n-1}\frac{\bar v_s}{\bar x_s}\right]$$

2. a) $Y_i^* = Y_i/x_i$. Then $E(Y_i^*) = \beta$ and $Var(Y_i^*) = \sigma^2$.

common mean model based on Y_i^*

b) $\hat t_{reg} = t_x\cdot\hat\beta_\pi$ (from p. 223)

where $\hat\beta_\pi = \dfrac{\sum_s y_i^*/\pi_i}{\underbrace{\sum_s 1/\pi_i}_{\hat N}} = \dfrac{1}{\hat N}\sum_{i\in s}\dfrac{y_i/x_i}{\pi_i}$

with $x_i=1$ for Y_i^*, $t_x = N$

and $\hat t_{reg} = \dfrac{N}{\hat N}\sum_{i\in s}\dfrac{y_i/x_i}{\pi_i}$

c) From p. 224:

For $|s| = n$ and $e_i^* = y_i^* - \hat{\beta}_\pi = \frac{y_i}{x_i} - \hat{\beta}_\pi$

$$\hat{V}(\hat{t}_{reg}) = \sum_{i \in s} \sum_{\substack{j \in s \\ j > i}} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i^*}{\pi_i} - \frac{e_j^*}{\pi_j} \right)^2$$

or $\hat{V}^*(\hat{t}_{reg}) = \left(\frac{N}{\hat{N}}\right)^2 \hat{V}(\hat{t}_{reg})$

3. a) Size of response stratum: $N_R = q_R N$

Size of response stratum h: $q_h N_h$

$$\bar{Y}_R = \frac{1}{q_R N} \sum_{U_R} y_i = \frac{1}{q_R N} \sum_{h=1}^{H} q_h N_h \bar{Y}_{rh}$$

$$= \frac{1}{q_R} \sum_{h=1}^{H} q_h \frac{N_h}{N} \bar{Y}_{rh} = \frac{1}{q_R} \sum_{h=1}^{H} q_h W_h \bar{Y}_{rh}$$

& similar for $\bar{Y}_M$: Size of $U_M = N - N_R = (1-q_R)N$

and size of nonresponse stratum h: $N_h - q_h N_h = (1-q_h)N_h$

, we get:

$$\bar{Y}_M = \frac{1}{(1-q_R)N} \sum_{U_M} y_i = \frac{1}{(1-q_R)N} \sum_{h=1}^{H} (1-q_h) N_h \bar{Y}_{Mh}$$

$$= \frac{1}{1-q_R} \cdot \sum_{h=1}^{H} (1-q_h) W_h \bar{Y}_{Mh}$$

b) $E(\bar{y}_r) - \bar{Y} = \bar{Y}_R - \bar{Y} = \bar{Y}_R - (q_R \bar{Y}_R + (1-q_R)\bar{Y}_M)$

$$= (1-q_R)(\bar{Y}_R - \bar{Y}_M)$$

$$= \frac{1-q_R}{q_R} \sum_h q_h W_h \bar{Y}_{rh} - \sum_h (1-q_h) W_h \bar{Y}_{Mh}$$

$$= \frac{1}{q_R} \sum_h (q_h - q_R) W_h \bar{Y}_{rh} + \sum_h W_h \bar{Y}_{rh} - \sum_h q_h W_h \bar{Y}_{rh} - \sum_h (1-q_h) W_h \bar{Y}_{Mh}$$

$$= \frac{1}{q_R} \sum_h (q_h - q_R) W_h \bar{Y}_{rh} + \sum_h (1-q_h) W_h (\bar{Y}_{rh} - \bar{Y}_{Mh})$$

4.

a) Estimate $= \bar{y}_s = (66\times 32 + 58\times 41 + 26\times 54)/150$

$= 5894/150 = \underline{39.3}$ hours

Standard error of $\bar{y}_s$

$$SE(\bar{y}_s) = \sqrt{\frac{s^2}{n}(1-f)}$$ (see p. 26)

$$s^2 = \frac{1}{n-1}\sum_s (y_i - \bar{y}_s)^2 \qquad f = n/N \quad , n=150,\ N=2000$$

Let s_1^2, s_2^2, s_3^2 be the sample variances in the 3 groups, and let the sample means be $\bar{y}_1, \bar{y}_2, \bar{y}_3$

Sample sizes are n_1, n_2, n_3 in response groups

Then:

$$s^2 = \frac{1}{n-1}\Big[\sum_{i\in group 1}(y_i - \bar{y}_1 + \bar{y}_1 - \bar{y}_s)^2$$

$$+ \sum_{i\in group 2}(y_i - \bar{y}_2 + \bar{y}_2 - \bar{y}_s)^2$$

$$+ \sum_{i\in group 3}(y_i - \bar{y}_3 + \bar{y}_3 - \bar{y}_s)^2\Big]$$

$$\Rightarrow s^2 = \frac{1}{n-1}\Big[(n_1-1)s_1^2 + n_1(\bar{y}_1 - \bar{y}_s)^2 + (n_2-1)s_2^2 + n_2(\bar{y}_2 - \bar{y}_s)^2$$

$$+ (n_3-1)s_3^2 + n_3(\bar{y}_3 - \bar{y}_s)^2\Big]$$

$$= \frac{1}{149}\Big[65\times 15^2 + 66(32-39.3)^2 + 57\times 19^2 + 58(41-39.3)^2$$

$$+ 25\times 25^2 + 26(54-39.3)^2\Big]$$

$$= 60130/149 = 403.56$$

$$\Rightarrow SE(\bar{y}_s) = \sqrt{\frac{403.56}{150}\left(1-\frac{150}{2000}\right)} = \sqrt{2.4886} = 1.58 \simeq \underline{1.6}$$

b) Poststrata are the GPA groups with $N_1 = 700$, $N_2 = 800$, $N_3 = 500$

Poststratified estimate of total viewing time:

$$\hat{t}_{post} = \sum_{h=1}^{3} N_h \bar{y}_h = 700 \times 32 + 800 \times 41 + 500 \times 54$$
$$= 82200$$

Estimate of mean viewing time:

$$\frac{\hat{t}_{post}}{N} = \frac{82200}{2000} = \underline{41.1}$$

c) The nonresponse rates for the 3 groups are:

0.12, 0.194 and .509

so probability of nonresponse clearly depends on GPA group. So the missing data mechanism is clearly not MCAR

So we have MAR or MNAR.

If MNAR, probability of nonresponse will also depend on viewing time, with no more information it may be enough to assume MAR

d) Other variables to use in poststratification:

Examples: gender, leisure activities, parents education, parents income.