STK4600

Solutions final exam Spring 2015

1a) Sampling unit = household

Variable of interest: y = cost of food per week

Auxiliary information: x = no of persons in hh

b) $\bar{y}_S = \frac{1}{n}\sum_{i\in s} y_i = \frac{1}{4}\times 7080 = \underline{1770}$

Estimated variance of $\bar{y}_S$:

$$\hat{V}(\bar{y}_S) = \frac{s^2}{4}\left(1-\frac{n}{N}\right)$$

where $s^2 = \frac{1}{3}\sum_{i\in s}(y_i - \bar{y}_S)^2 = \underline{243600}$

$$\Rightarrow \hat{V}(\bar{y}_S) = \frac{243600}{4}\cdot\frac{24096}{25000} = \underline{60890.256}$$

$$SE(\bar{y}_S) = \sqrt{\hat{V}(\bar{y}_S)} = \underline{246.76}$$

c) Ratio-estimator for the total:

$$\hat{t}_R = X_0 \cdot \frac{\sum_s y_i}{\sum_s x_i}$$

Estimator for μ: $\hat{\mu}_R = \frac{\hat{t}_R}{N} = \bar{X}\cdot \underbrace{\frac{\sum_s y_i}{\sum_s x_i}}_{\hat{R}}$

$$\bar{X} = \frac{72500}{25000} = 2.9$$

$$\hat{R} = 7080/13 = 544.6154$$

$$\Rightarrow \hat{\mu}_R = 2.9\times 544.6154 = \underline{1579.4}$$

$$\hat{V}(\hat{\mu}_R) = \left(\frac{\bar{X}}{\bar{x}_S}\right)^2 \cdot \frac{1-f}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{4}(y_i - \hat{R}x_i)^2$$

$$\sum_{i=1}^{4}(y_i - \hat{R}x_i)^2 = 206711.7586$$

$$\hat{V}(\hat{\mu}_R) = \left(\frac{2.9}{3.25}\right)^2 \cdot \frac{.9984}{4} \times \frac{1}{3} \times 206711.7586$$

$$= 13713.355$$

and $\underline{SE(\hat{\mu}_R) = 117.10}$

[Accepts also: $\hat{V}(\hat{\mu}_R) = \frac{1-f}{n} \cdot \frac{1}{n-1} \Sigma (y_i - \hat{R} x_i)^2$

$= 17223.2$

and $SE(\hat{\mu}_R) = 131.2$]

d) Prefer $\hat{\mu}_R$ based on smaller SE.

(or scatter plot)

2a) With MCAR, the response sample is a random sample

and $\hat{p} = \frac{60+280+80}{500} = \frac{420}{500} = \underline{0.84}$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \cdot \frac{N-n}{N}} \approx$$

Here $N = 8000000$ and $n = 500$.

$$\Rightarrow SE(\hat{p}) = \sqrt{\frac{.84 \times .16}{499} \times .999938} = 10^{-2}\sqrt{2.6932} = .0164$$

95% CI for p: $\hat{p} \pm 1.96 \cdot SE(\hat{p})$

$= .84 \pm .032 = \underline{(0.808, 872)}$

True $p = .78$

$\Rightarrow$ Estimate & CI have large bias

The nonresponse is not MCAR.

b) $\hat{p}_{post} = \hat{t}_{post} / N$

where $\hat{t}_{post} = \sum_{h=1}^{3} N_h \cdot \hat{p}_h$

$\hat{p}_1 = 60/90 = 2/3 = 0.667$

$\hat{p}_2 = 280/305 = 0.918$

$\hat{p}_3 = 80/105 = 0.762$

$\Rightarrow \hat{t}_{post} = 6482500$ and $\hat{p}_{post} = 0.810$

c) Standard error could be on ~~p_{1h}, p_{2h}, p_{3h}~~ sample mean for respondents in poststrata:

$$SE(\hat{p}_{post}) = \sqrt{\hat{V}(\hat{t}_{post})} / N^2 = \sqrt{\hat{V}(\hat{p}_{post})}$$

where $\hat{V}(\hat{t}_{post}) = \sum_{h=1}^{3} N_h^2 \cdot \left(1 - \frac{n_{rh}}{N_h}\right) \frac{s_h^2}{n_{rh}}$

$n_{r1} = 90, \; n_{r2} = 305, \; n_{r3} = 105$

$s_h^2 = \frac{n_{rh}}{n_{rh}-1} \hat{p}_h (1 - \hat{p}_h)$

$$\hat{V}(\hat{p}_{post}) = \sum_{h=1}^{3} W_h^2 \left(1 - \frac{n_{rh}}{N_h}\right) \cdot \frac{\hat{p}_h(1-\hat{p}_h)}{n_{rh}-1} \qquad W_h = \frac{N_h}{N}$$

$$= \left(\frac{2.5}{8}\right)^2 \left(1 - \frac{90}{N_1}\right) \cdot \frac{\frac{2}{3} \cdot \frac{1}{3}}{89} + \left(\frac{4}{8}\right)^2 \left(1 - \frac{305}{N_2}\right) \cdot \frac{0.918 \times 0.082}{304}$$

$$+ \left(\frac{1.5}{8}\right)^2 \left(1 - \frac{105}{N_3}\right) \cdot \frac{0.762 \times 0.238}{104}$$

$$= 2.438 \times 10^{-4} + 0.619 \times 10^{-4} + 0.613 \times 10^{-4} = 3.67 \times 10^{-4}$$

$\Rightarrow SE(\hat{p}_{post}) = 1.916 \times 10^{-2} = \underline{0.0192}$

95% CI : $0.810 \pm 1.96 \times 0.0192 = 0.810 \pm 0038$

$= \underline{(0.772, 0.848)}$

d) The poststratified estimator is unbiased if response sample is representative of nonresponse sample within each poststratum. It is not the case here. Need to assume that the prob. of response depends on the value of y (=1 if votes, 0 if not). Or: find other poststrata variables

3a Model 2, because:
- intercept with y-axis not at 0
- variance do not seem to increase with x

3b $\hat{t}_R = t_x \cdot \frac{\bar{y}_s}{\bar{x}_s} \Rightarrow \hat{\mu}_R = \frac{\hat{t}_R}{N} = \bar{x} \cdot \frac{\bar{y}_s}{\bar{x}_s}$

$\hat{r} = \frac{\bar{y}_s}{\bar{x}_s} = 223.60$, $\bar{x} = 2.7935$

$\Rightarrow \underline{\hat{\mu}_R = 624.6}$

Model-based SE $= \sqrt{\hat{V}(\hat{\mu}_R - \frac{T}{N})}$

$\hat{V}(\hat{\mu}_R - \frac{T}{N}) = \hat{\sigma}^2 \cdot \frac{1-f}{n} \cdot \frac{\bar{x}_r \cdot \bar{x}}{\bar{x}_s}$ $\quad \bar{x}_r = \frac{1}{N-n} \sum_{i \notin s} x_i$

$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i \in s} \frac{1}{x_i} (y_i - \hat{r} x_i)^2 = 2571.6$

$f = n/N$, $\bar{x}_r = (N \cdot \bar{x} - n \cdot \bar{x}_s)/(N-n) = 2.7940$

$\Rightarrow \hat{V}(\hat{\mu}_R - \frac{T}{N}) = 380.1237$

$\Rightarrow \underline{SE = 19.5}$

95% CI $= 624.6 \pm 1.96 \times 19.5 = 624.6 \pm 38.2$

$= \underline{(586.4, 662.8)}$

3c

$$\hat{\mu}_{reg} = \hat{T}_{reg}/N = \frac{1}{N} \cdot N\left(\bar{y}_S + \hat{\beta}_2(\bar{x} - \bar{x}_S)\right)$$

$$= \bar{y}_S + \hat{\beta}_2(\bar{x} - \bar{x}_S)$$

$\bar{y}_S = 588.4 \rightarrow$

$$\hat{\mu}_{reg} = 588.4 + 162.5 \cdot (2.7935 - 2.6315)$$

$$= 588.4 + 26.3 = \underline{614.7}$$

$\hat{\sigma}^2 = 4209.3$

$$\hat{V}(T_{reg}/N - T/N) = \frac{\hat{\sigma}^2}{n} \cdot \left[\left(1 - \frac{n}{N}\right) + \frac{n(\bar{x} - \bar{x}_S)^2}{\sum_S (x_i - \bar{x}_S)^2}\right]$$

$$= \frac{4209.3}{20} \cdot \left[\frac{6174}{6194} + \frac{20 \cdot (0.162)^2}{19 \times 0.628}\right]$$

$$= \frac{4209.3}{20} \times 1.04076 = 219.044$$

$\Rightarrow$ $SE = \sqrt{219.044} = \underline{14.8}$

95% CI: $614.7 \pm 1.96 \times 14.8 = 614.7 \pm 29.0$

$= \underline{(585.7, 643.7)}$

3d $\bar{y}_S = 588.4$

$$\hat{V}(\bar{y}_S - \frac{T}{N}) = \frac{1 - \frac{1}{5}}{n} \cdot \hat{\sigma}^2 \quad ; \quad \hat{\sigma}^2 = 20572.5$$

$$= 1025.3$$

$\Rightarrow$ $SE = \underline{32.0}$

Including x is important, since clearly x and y are positively correlated.

It corrects estimate and reduces SE by close to 50%.

e) $\bar{y}_S$ is farthest away because sample mean of x is too low compared to $\bar{x}$.

Even though model 2 is the most appropriate model, as also seen by smaller SE than ratio-estimator, $\hat{\mu}_{LR}$ is by chance closest to 632. Would exped $\hat{\mu}_{reg}$ to be closest on average.

(Any sensible explanation gets credit)

```
schooldata=data.frame(y[s],x[s])
> schooldata
   y.s. x.s.
1   595 2.54
2   354 1.34
3   431 2.06
4   695 2.09
5   608 3.11
6   555 2.21
7   484 2.09
8   414 2.04
9   528 2.71
10  549 2.17
11  527 2.03
12  806 3.63
13  594 2.65
14  890 4.49
15  627 2.92
16  692 2.97
17  372 1.79
18  562 2.33
19  708 3.46
20  777 4.00
> regbeta=sum(x[s]*y[s])-20*mean(y[s])*mean(x[s])
> regbeta=regbeta/var(x[s])
> regbeta=regbeta/19
> regbeta
[1] 162.4926
> ratio=sum(y[s])/sum(x[s])
> ratio
[1] 223.5987
> alfa=mean(y[s])-regbeta*mean(x[s])
> alfa
[1] 160.8007
> regest=mean(y[s])+regbeta*(meanx-mean(x[s]))
> regest
[1] 614.7223
> meanx
[1] 2.793491
> var(x[s])
[1] 0.6281187
> mean(x[s])
[1] 2.6315
> 1-20/6194
[1] 0.9967711
> 6174/6194
[1] 0.9967711
> sigmasq=sum((y[s]-mean(y[s])-regbeta*(x[s]-mean(x[s])))^2)
> sigmasq
[1] 75766.62
> sigmasq=sigmasq/18
> sigmasq
[1] 4209.257
> sigm=sqrt(sigmasq)
> sigm
```

```
> var(y[s])
[1] 20572.46
> 20572/20
[1] 1028.6
> sum(y[s])
[1] 11768
> mean(y[s])*20
[1] 11768
> ratio
[1] 223.5987
> sigmratio=sum((y[s]-ratio*x[s])^2/x[s])
> sigmr=sigmratio/19
> sigmr
[1] 2571.591
> varestr=sigmr*(6194*meanx-sum(x[s]))*meanx/(6194*mean(x[s]))
> ser=sqrt(varestr)
> serr=sqrt(varestr/20)
> serr
[1] 19.4971
> sig=var(y[s])-regbeta^2*var(x[s])
> sig
[1] 3987.717
> sig=19*sig/18
> sig
[1] 4209.257
> d=20*(meanx-mean(x[s]))^2/(19*var(x[s]))
> d
[1] 0.04397591
> a=d+1-20/6194
> var=a*sign
> sqrt(var)
[1] 14.79995
> sum((x[s]-mean(x[s]))^2)
[1] 11.93425
> 0.512/11.93
[1] 0.04291702
> var(x[s])
[1] 0.6281187
> lm(formula=y[s]~x[s],data=schools)
Error in is.data.frame(data) : object 'schools' not found
> lm(formula=y[s]~x[s],data=schooldata)

Call:
lm(formula = y[s] ~ x[s], data = schooldata)
Coefficients:
(Intercept)        x[s]
      160.8       162.5

> lm(formula=y[s]/sqrt(x[s])~sqrt(x[s])-1)
Call:
lm(formula = y[s]/sqrt(x[s]) ~ sqrt(x[s]) - 1)

Coefficients:
sqrt(x[s])
     223.6
```