

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 :

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

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

$$\Rightarrow \hat{V}(\bar{y}_s) = \frac{243600}{4} \cdot \frac{24996}{25000} = \underline{60890,256}$$

$$SE(\bar{y}_s) = \sqrt{\hat{V}(\bar{y}_s)} = \underline{246,76}$$

c) Ratio-estimator for the total:

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

$$\text{Estimator for } \mu: \hat{\mu}_R = \frac{\hat{t}_R}{N} = \bar{x} \cdot \frac{\sum y_i}{\sum x_i}$$

$$\bar{x} = \frac{72500}{25000} = 2,9$$

$$\hat{R} = 7080/13 = 544,6154$$

$$\Rightarrow \hat{\mu}_R = 2,9 \times 544,6154 = \underline{1579,4}$$

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

$$\sum_{i=1}^4 (y_i - \hat{R}x_i)^2 = 206711,7586$$

$$\hat{V}(\hat{\mu}_R) = \left(\frac{2.9}{3.25}\right)^2 \cdot \frac{99984}{4} \times \frac{1}{3} \times 206711.7586$$

$$= 13713.355$$

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

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

$$= 17223.2$$

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

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

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.

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

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

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

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

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

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

c) Standard error could be on ~~the~~ ~~the~~ ~~the~~ sample sizes for respondents in poststrata:

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

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

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

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

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

$$W_h = \frac{N_h}{N}$$

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

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

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

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

$$95\% \text{ CI} : 0.810 \pm 1.96 \times 0.0192 = 0.810 \pm 0.038$$

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

$$\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$$

$$\sum_{i=1}^n \hat{y}_i^2 = \sum_{i=1}^n y_i^2$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

Standard error

variance

$$\hat{y}_i = \sqrt{\frac{1}{n} \sum_{j=1}^n y_j^2}$$

$$\frac{\partial \hat{y}_i}{\partial y_j} = \frac{1}{2} \left(\frac{1}{n} \sum_{k=1}^n y_k^2 \right)^{-1/2} \cdot 2 y_j = \frac{y_j}{\hat{y}_i}$$

$$0.18 = 0.18, 0.18 = 0.18, 0.18 = 0.18$$

$$\frac{\partial \hat{y}_i}{\partial y_j} = \frac{y_j}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial y_j} = \frac{y_j}{\hat{y}_i} = \frac{0.18}{0.18} = 1$$

$$\frac{\partial \hat{y}_i}{\partial y_j} = \frac{y_j}{\hat{y}_i} = \frac{0.18}{0.18} = 1$$

$$\frac{\partial \hat{y}_i}{\partial y_j} = \frac{y_j}{\hat{y}_i} = \frac{0.18}{0.18} = 1$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

$$0.18 \cdot 0 = 0.18 \cdot 0 = 0.18$$

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 voting, 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

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

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

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

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

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

$$\lambda^2 = \frac{1}{n-1} \sum_{i \in S} \frac{1}{x_i} (y_i - \hat{\tau} x_i)^2 = 2571.6$$

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

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

$$\Rightarrow \underline{\text{SE}} = 19.5$$

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

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

of the poststratified estimator is unbiased
 if response sample is representative of
 population sample within each stratum
 it is not the case here. Need to assume that the
 bias of unadjusted weights on the order of
 N^{-1} if relative of total CV. total error
 is small.

3a. Model 2: we can write

- intercept with fixed part of μ
- variance of μ is σ^2

$$\hat{\mu} = \frac{\sum_{h=1}^H \mu_h}{H} = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h$$

Model: $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{H}$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

$$\hat{\mu} = \frac{1}{H} \sum_{h=1}^H \mu_h = \frac{1}{H} \sum_{h=1}^H \mu_h$$

3c

$$\begin{aligned}\hat{\mu}_{\text{reg}} &= \hat{T}_{\text{reg}}/N = \frac{1}{N} \cdot N (\bar{Y}_s + \hat{\beta}_2 (\bar{X} - \bar{X}_s)) \\ &= \bar{Y}_s + \hat{\beta}_2 (\bar{X} - \bar{X}_s)\end{aligned}$$

$$\bar{Y}_s = 588.4 \rightarrow$$

$$\begin{aligned}\hat{\mu}_{\text{reg}} &= 588.4 + 162.5 \cdot (7.7935 - 2.6315) \\ &= 588.4 + 26.3 = \underline{614.7}\end{aligned}$$

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

$$\hat{V}(\hat{T}_{\text{reg}}/N - T/N) = \frac{\hat{\sigma}^2}{n} \cdot \left[\left(1 - \frac{n}{N}\right) + \frac{n(\bar{X} - \bar{X}_s)^2}{\sum_s (X_s - \bar{X}_s)^2} \right]$$

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

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

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

$$\begin{aligned}95\% \text{ CI: } 614.7 \pm 1.96 \times 14.8 &= 614.7 \pm 29.0 \\ &= \underline{(585.7, 643.7)}\end{aligned}$$

3d $\bar{Y}_s = 588.4$

$$\begin{aligned}\hat{V}(\bar{Y}_s - \bar{Y}) &= \frac{1-f}{n} \cdot \hat{\sigma}^2 \quad ; \quad \hat{\sigma}^2 = 20572.5 \\ &= 1025.3\end{aligned}$$

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

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

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

$$\hat{\mu} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$= \bar{y} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

$$\hat{\mu} = 288.4$$

$$\hat{\mu} = 288.4 + \frac{102.2}{n-1} = 288.4 + \frac{102.2}{10} = 298.6$$

$$F.M.I. = 298.6$$

$$\hat{\sigma}^2 = 102.2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{102.2}{10} = 10.22$$

$$\hat{\sigma} = \sqrt{10.22} = 3.197$$

$$F.M.I. = 298.6$$

$$\hat{\sigma} = 3.197$$

$$F.M.I. = 298.6$$

$$F.M.I. = 298.6$$

$$\hat{\mu} = 288.4$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = \frac{102.2}{10} = 10.22$$

$$= 10.22$$

$$\hat{\sigma} = 3.197$$

Interpretation: The F.M.I. is 298.6, which is higher than the mean value of 288.4, indicating a right-skewed distribution.

The standard deviation is 3.197, which is relatively low compared to the mean.

If we consider the variance, it is 10.22, which is also relatively low.

F.M.I.

e) \bar{y}_1 is furthest 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, $\hat{\mu}_p$ is by chance closest to 632. Would expect $\hat{\mu}_{reg}$ to be closest on average.

(Any sensible explanation gets credit)

(c) If $\hat{\beta}$ is the least squares estimate of β , then

$\hat{\beta}$ is the value of β that minimizes the sum of squares

of $\hat{\beta}$.

Even though model 5 is the best

in terms of R^2 , we also have

that the variance of $\hat{\beta}$ is larger than that of $\hat{\beta}$ in

model 4. In fact, the variance of $\hat{\beta}$ is

larger than that of $\hat{\beta}$ in model 4.

(This result is a consequence of the fact that

```

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
```