260677676 MATH525 Assignment2_Part2

Dan Yunheum Seol

2/27/2019

```r
library(formatR)
```

#4.18 Use covariances derived in Appendix A to show the result in (4.8)

Covariances derived: Let Z_i be our sampling indicator random variables for ith unit in the population.

$$V(Z_i) = Cov(Z_i, Z_i) = \frac{n(N-n)}{N^2}$$

if $i \neq j$

$$Cov(Z_i, Z_j) = -\frac{n(N-n)}{N^2(N-1)}$$

Then we should show that

$$E[(\overline{y} - B\overline{x})^2] = V(\frac{1}{n}\sum_{i \in S}(y_i - Bx_i)^2) = \left(1 - \frac{n}{N}\right)\frac{S_y^2 - 2BRS_xS_y + B^2S_x^2}{n}$$

Recall that

$$R = \frac{\sum_{i=1}^{N}(y_i - \overline{y}_u)(x_i - \overline{x}_u)}{(N-1)S_xS_y}$$

$$\frac{1}{n^2}V(\sum_{i \in S}(y_i - Bx_i)^2) = \frac{1}{n^2}V(\sum_{i=1}^{N} Z_i(y_i - Bx_i)^2) =$$

$$\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{N} Cov((y_i - Bx_i)Z_i, (y_j - Bx_j)Z_j)$$

$$\frac{1}{n^2}\{\sum_{i=1}^{N}(y_i - Bx_i)^2 V(Z_i) + \sum_{i=1}^{N}\sum_{i \neq j}(y_i - Bx_i)(y_j - Bx_j)Cov(Z_i, Z_j)\} =$$

$$\frac{1}{n^2}\{\frac{n(N-n)}{N^2}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{i=1}^{N}\sum_{i \neq j}(y_i - Bx_i)(y_j - Bx_j)\} =$$

$$\frac{1}{n^2}\{\frac{n(N-n)N}{N^2(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{i=1}^{N}\sum_{i \neq j}(y_i - Bx_i)(y_j - Bx_j)\} =$$

$$\frac{1}{n^2}\{\frac{n(N-n)}{N(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \sum_{i=1}^{N}\sum_{j=1}^{N}(y_i - Bx_i)(y_j - Bx_j)\} =$$

$$\frac{1}{n^2}\{\frac{n(N-n)}{N(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{n(N-n)}{N^2(N-1)}\left(\sum_{i=1}^{N}(y_i - Bx_i)\right)^2\} =$$

$$\frac{1}{n}\{\frac{(N-n)}{N(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{(N-n)}{N^2(N-1)}\left(\sum_{i=1}^{N}(y_i - Bx_i)\right)^2\} =$$

$$\frac{1}{n}\left(1 - \frac{n}{N}\right)\{\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{1}{N(N-1)}\left(\sum_{i=1}^{N}(y_i - Bx_i)\right)^2\}$$

Now, remark that $\sum_{i=1}^{N}(y_i - Bx_i)$

$$\sum_{i=1}^{N}(y_i - Bx_i) = \sum_{i=1}^{N}(y_i) - \sum_{i=1}^{N}(Bx_i) = N\overline{y}_U - B\overline{x}_U = N\overline{y}_U - N\frac{\overline{y}_U}{\overline{x}_U} = 0\overline{x}_U$$

$$\frac{1}{n}\left(1 - \frac{n}{N}\right)\{\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - \frac{1}{N(N-1)}\left(\sum_{i=1}^{N}(y_i - Bx_i)\right)^2\} =$$

$$\frac{1}{n}\left(1 - \frac{n}{N}\right)\{\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 - 0\} = \frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2$$

Also remark

$$\sum_{i=1}^{N}(y_i - Bx_i)^2 = \sum_{i=1}^{N}(y_i - \overline{y}_U + \overline{y}_U - Bx_i)^2 =$$

$$\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 + \sum_{i=1}^{N}(B\overline{x}_U - Bx_i)^2 + 2\sum_{i=1}^{N}(y_i - \overline{y}_U)(B\overline{x}_U - Bx_i) =$$

$$\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 + B^2\sum_{i=1}^{N}(\overline{x}_U - Bx_i)^2 + +2B\sum_{i=1}^{N}(y_i - \overline{y}_U)(\overline{x}_U - x_i) =$$

$$= B^2S_x^2(N-1) + S_y^2(N-1) - 2(N-1)BRS_xS_y$$

So

$$\frac{1}{n}\left(1 - \frac{n}{N}\right)\frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - Bx_i)^2 =$$

$$\left(1 - \frac{n}{N}\right)\frac{\{B^2S_x^2(N-1) + S_y^2(N-1) - 2(N-1)BRS_xS_y\}}{n}$$

Proving our claim.

#4.22

Use (4.5)

$$\overline{y}_r - \overline{y}_U = \frac{\overline{x}_U(\overline{y} - B\overline{x})}{\overline{x}} = (\overline{y} - B\overline{x})\frac{(\overline{x} - \overline{x}_U)}{\overline{x}}$$

and (A.10)

$$Cov(\overline{x}, \overline{y}) = \left(1 - \frac{n}{N}\right)\frac{RS_xS_y}{n}$$

To show

$$Bias(\overline{y}_r) = E[\overline{y}_r - \overline{y}_U] \approx \frac{1}{\overline{x}_U}[BV(\overline{x}) - Cov(\overline{x}, \overline{y})] = \left(1 - \frac{n}{N}\right)\frac{1}{n\overline{x}_U}(BS_x^2 - RS_xS_y)$$

We start with

$$E[\overline{y}_r - \overline{y}_U] = E[(\overline{y} - B\overline{x})(\frac{\overline{x}_U}{\overline{x}})] = E[\overline{y} - B\overline{x}] - E[(\overline{y} - B\overline{x})\frac{\overline{x} - \overline{x}_U}{\overline{x}}] \approx$$

$$\overline{y}_U - \overline{y}_U - E[(\overline{y} - \overline{x}B)(\frac{\overline{x}}{\overline{x}_U})] =$$

$$-\frac{1}{\overline{x}_U}E[\overline{yx}] + \frac{1}{\overline{x}_U}BE[\overline{x}^2] =$$

$$-\frac{1}{\overline{x}_U}\left(Cov(\overline{x},\overline{y})-E[\overline{x}]E[\overline{y}]\right)+\frac{1}{\overline{x}_U}\{B(V[\overline{x}]+(E[\overline{x}])^2)\}=$$

$$\frac{1}{\overline{x}_U}\{BV[\overline{x}]-Cov(\overline{x},\overline{y})\}+\frac{1}{\overline{x}_U}\{\overline{x}_U^2-\overline{x}_U\overline{y}_U\}=$$

$$\frac{1}{\overline{x}_U}\{BV[\overline{x}]-Cov(\overline{x},\overline{y})\}+\overline{x}_U-\overline{y}_U$$

Since the last term is a constant, we can claim

$$Bias(\overline{y}_r)\approx\frac{1}{\overline{x}_U}[BV(\overline{x})-Cov(\overline{x},\overline{y})]=\left(1-\frac{n}{N}\right)\frac{1}{n\overline{x}_U}\left(BS_x^2-RS_xS_y\right)$$

, which is a sufficient answer.

#4.24 Suppose there are two domains, defined by the indicator variable

$$x_i=\begin{cases}1 & i\in D_1\\ 0 & i\notin D_2\end{cases}$$

Letting $u_i:=x_iy_i$, the population values of the two domain means are

$$\overline{y}_{U1}=\frac{\sum_{i=1}^N x_iy_i}{\sum_{i=1}^N x_i}=\frac{t_u}{t_x}=\frac{\overline{u}_U}{\overline{x}_U}$$

and

$$\overline{y}_{U2}=\frac{\sum_{i=1}^N(1-x_i)y_i}{\sum_{i=1}^N(1-x_i)}=\frac{t_y-t_u}{N-t_x}=\frac{\overline{y}_U-\overline{u}_U}{1-\overline{x}_U}$$

If an SRS of size is taken from the population of size N, the population domain menas may be estimated by

$$\overline{y}_1=\frac{\widehat{t}_u}{\widehat{t}_x}=\frac{\overline{u}}{\overline{x}}$$

$$\overline{y}_2=\frac{\widehat{t}_y-\widehat{t}_u}{N-\widehat{t}_x}=\frac{\overline{y}-\overline{u}}{1-\overline{x}}$$

#(a) Use an argument similar to that in the discussion following (4.5) to show that

$$Cov(\overline{y}_1,\overline{y}_2)\approx\frac{1}{\overline{x}_U(1-\overline{x}_U)}Cov[\left(\overline{u}-\frac{t_u}{t_x}\overline{x}\right),\{\overline{y}-\overline{u}-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\}]$$

We have

$$Cov(\overline{y}_1,\overline{y}_2)\approx E[(\overline{y}_1-\overline{y}_{1U})(\overline{y}_2-\overline{y}_{2U})]=$$

Remark that

$$\overline{y}_1-\overline{y}_{1U}=\frac{1}{\overline{x}}(\overline{u}-\overline{x}\overline{y}_{1U})=\frac{1}{\overline{x}}(\overline{u}-\frac{t_u}{t_x}\overline{x})\approx\frac{1}{\overline{x}_U}(\overline{u}-\frac{t_u}{t_x}\overline{x})$$

$$\overline{y}_2-\overline{y}_{2U}=(\frac{\overline{y}-\overline{u}}{1-\overline{x}}-\frac{t_y-t_u}{N-t_x})=\frac{1}{(1-\overline{x})}\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right)\approx\frac{1}{(1-\overline{x}_U)}\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right)$$

So we obtain

$$Cov(\overline{y}_1,\overline{y}_2)\approx\frac{1}{\overline{x}_U(1-\overline{x}_U)}E[\left(\overline{u}-\frac{t_u}{t_x}\overline{x}\right)\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right)]$$

Now, remark that

$$E[\frac{1}{\overline{x}_U}(\overline{u}-\frac{t_u}{t_x}\overline{x})]=\frac{1}{\overline{x}_U}E[(\overline{u}-\frac{\overline{u}_U}{\overline{x}_U}\overline{x})]=\frac{1}{\overline{x}_U}[\overline{u}_U-\frac{\overline{u}_U}{\overline{x}_U}\overline{x}_U]=0$$

It follows that

$$\frac{1}{\overline{x}_U(1-\overline{x}_U)}E[\left(\overline{u}-\frac{t_u}{t_x}\overline{x}\right)\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right)]=\frac{1}{\overline{x}_U(1-\overline{x}_U)}Cov(\left(\overline{u}-\frac{t_u}{t_x}\overline{x}\right),\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right))$$

which is the result we wanted.

#(b) Show that

$$\frac{1}{\overline{x}_U(1-\overline{x}_U)}Cov(\left(\overline{u}-\frac{t_u}{t_x}\overline{x}\right),\left((\overline{y}-\overline{u})-\frac{t_y-t_u}{N-t_x}(1-\overline{x})\right))=0$$

Using A.10, which is

$$Cov(\overline{x},\overline{y})=(1-\frac{n}{N})(\frac{RS_xS_y}{n})=(1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{N}(y_i-\overline{y}_U)(x_i-\overline{x}_U)$$

$$\sum_{i=1}^{N}\left\{\left(u_i-\frac{t_u}{t_x}x_i\right)-\left(\overline{u}_U-\frac{t_u}{t_x}\overline{x}_U\right)\right\}\left\{\left((y_i-u_i)-\frac{t_y-t_u}{N-t_x}(1-x_i)\right)-\left((\overline{y}_U-\overline{u}_U)-\frac{t_y-t_u}{N-t_x}(1-\overline{x}_U)\right)\right\}=$$

$$\sum_{i=1}^{N}\left\{\left(u_i-\overline{u}_U\right)-\frac{t_u}{t_x}\left(x_i-\overline{x}_U\right)\right\}\left\{\left((y_i-\overline{y}_U)-(u_i-\overline{u}_U)\right)+\frac{t_y-t_u}{N-t_x}\left(x_i-\overline{x}_U\right)\right\}=$$

We show that the "RS_xS_y" part is zero. \For any constant k,

$$k\sum_{i=1}^{N}\left\{\left(u_i-\overline{u}_U\right)-\frac{t_u}{t_x}\left(x_i-\overline{x}_U\right)\right\}=\overline{y}_U\{(N\overline{u}_U-N\overline{u}_U)-\frac{t_u}{t_x}(N\overline{x}_U-N\overline{x}_U)\}=0$$

and likewise

$$k\sum_{i=1}^{N}\left\{\left((y_i-\overline{y}_U)-(u_i-\overline{u}_U)\right)+\frac{t_y-t_u}{N-t_x}\left(x_i-\overline{x}_U\right)\right\}=0$$

So we have

$$\sum_{i=1}^{N}\left\{\left(u_i-\frac{t_u}{t_x}x_i\right)\left((y_i-u_i)-\frac{t_y-t_u}{N-t_x}x_i\right)\right\}=$$

$$\sum_{i=1}^{N}\left\{u_i(y_i-u_i)-\frac{t_u}{t_x}x_i(y_i-u_i)-u_i\frac{t_y-t_u}{N-t_x}x_i+\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}x_i^2\right\}$$

Now remark that since x_i either takes value 0 or value 1, $x_i^2=x_i$, $x_iu_i=u_i$, and $u_iy_i=u_i^2$. Thus we would have \ $\sum_{i=1}^{N}u_i(y_i-u_i)=\sum_{i=1}^{N}(u_i^2-u_i^2)=0$ \ $\sum_{i=1}^{N}x_i*(y_i-u_i)=\sum_{i=1}^{N}(u_i-u_i)=0$ \ $\sum_{i=1}^{N}u_ix_i=\sum_{i=1}^{N}u_i=t_u$ \ $\sum_{i=1}^{N}x_i^2=\sum_{i=1}^{N}x_i=t_x$. \ It follows that

$$\sum_{i=1}^{N}\left\{u_i(y_i-u_i)-\frac{t_u}{t_x}x_i(y_i-u_i)-u_i\frac{t_y-t_u}{N-t_x}x_i+\frac{t_u}{t_x}\frac{t_y-t_u}{N-t_x}x_i^2\right\}=0-0-t_u\frac{t_y-t_u}{N-t_x}+\frac{t_u}{t_x}t_x\frac{t_y-t_u}{N-t_x}=0$$

proving our claim.

#4.44(a)

```
## -- Attaching packages ------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------------------- t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: grid

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':
##
##     expand

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

##
## Attaching package: 'srvyr'

## The following object is masked from 'package:stats':
##
##     filter
```

```r
ipums <- read_csv('ipums.csv')
```

```
## Parsed with column specification:
## cols(
##   Stratum = col_double(),
##   Psu = col_double(),
##   Inctot = col_double(),
##   Age = col_double(),
##   Sex = col_double(),
##   Race = col_double(),
##   Hispanic = col_double(),
##   Marstat = col_double(),
##   Ownershg = col_double(),
##   Yrsusa = col_double(),
##   School = col_double(),
##   Educrec = col_double(),
##   Labforce = col_double(),
##   Occ = col_double(),
##   Classwk = col_double(),
##   VetStat = col_double()
## )
```

```r
head(ipums)
```

```
## # A tibble: 6 x 16
##   Stratum   Psu Inctot   Age   Sex  Race Hispanic Marstat Ownershg Yrsusa
##     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
## 1       1     1   4105    18     1     2        0       5        0      0
## 2       1     1   7795    20     1     1        0       5        2      0
## 3       1     1  16985    24     1     1        0       1        1      0
## 4       1     1   7045    21     1     1        0       1        2      0
## 5       1     1   2955    23     1     1        0       5        2      0
## 6       1     1      0    17     1     1        0       5        1      0
## # ... with 6 more variables: School <dbl>, Educrec <dbl>, Labforce <dbl>,
## #   Occ <dbl>, Classwk <dbl>, VetStat <dbl>
```

```r
dim(ipums)
```

```
## [1] 53461    16
```

```r
ipums_complete = ipums %>% filter(!is.na(Inctot))
#Verify that none of the data is missing
dim(ipums_complete)
```

```
## [1] 53461    16
```

```r
ipums_totals = ipums_complete %>% summarise(sum_Ages = sum(Age), N_Inctot = sum(Inctot),
    B = N_Inctot/sum_Ages)
ipums_totals
```

```
# A tibble: 1 x 3
  sum_Ages  N_Inctot     B
     <dbl>     <dbl> <dbl>
1  2200842 491533095  223.
```

Since we finished setting up our B, let's obtain a simple random sample of 500

```r
set.seed(329)
srs_500=ipums_complete %>% slice(sample(1:nrow(ipums_complete),size=500,replace=F)) %>%   mutate(fpc = n
dim(srs_500)
```

```
## [1] 500  17
```

```r
srs_500 %>%
  summarise(SampleMean=mean(Inctot),
                            SampleVar = var(Inctot),
                            SampleSD = sd(Inctot))  %>%
  gather(stat,val) %>%
  kable(.,format="latex",digits=0) %>%
  kable_styling(.)
```

stat	val
SampleMean	9664
SampleVar	120894835
SampleSD	10995

```r
srs_design = svydesign(ids=~1,data=srs_500,fpc=~fpc)
svytotal(~Inctot, srs_design)
```

```
##            total       SE
## Inctot 516636946 26164684
```

```r
r = svyratio(~Inctot, ~Age, srs_design)
r
```

```
Ratio estimator: svyratio.survey.design2(~Inctot, ~Age, srs_design)
Ratios=
            Age
Inctot 230.2003
SEs=
            Age
Inctot 11.95726
```

```r
confint(r)
```

```
              2.5 %   97.5 %
Inctot/Age 206.7645 253.6361
```

```r
predict_r = predict(r, total = ipums_totals %>% pull(sum_Ages))
predict_r
```

```
$total
             Age
Inctot 506634562

$se
            Age
Inctot 26316047
```

```r
svytotal(~Inctot, srs_design)
```

```
           total       SE
Inctot 516636946 26164684
```

```r
predict_r$total + c(qnorm(0.025), qnorm(0.975)) * predict_r$se
```

```
[1] 455056059 558213066
```

```r
confint(svytotal(~Inctot, srs_design))
```

```
           2.5 %    97.5 %
Inctot 465355108 567918785
```

```r
ipums_totals  ## True total
```

```
# A tibble: 1 x 3
  sum_Ages  N_Inctot     B
     <dbl>     <dbl> <dbl>
1  2200842 491533095  223.
```

The standard error does not decrease but rather increases, this can be due to low correlation between age and total income.

```r
R <- srs_500 %>% summarise(cor=cor(Age,Inctot))
R
```

```
## # A tibble: 1 x 1
##     cor
##   <dbl>
## 1 0.132
```

Another criterion to check is to see whether

$$Cor(X,Y) = R \geq \frac{CV(X)}{2CV(Y)}$$

In case where this does not hold, we have no guarantee that the standard error will decrease.

```
CVx = ipums_complete %>% summarise(CV_Age = (sd(Age)/sqrt(500))/mean(Age))
CVx
```

```
# A tibble: 1 x 1
  CV_Age
   <dbl>
1 0.0205
```

```
CVy = ipums_complete %>% summarise(CV_Inctot = (sd(Inctot)/sqrt(500))/mean(Inctot))
CVy
```

```
# A tibble: 1 x 1
  CV_Inctot
      <dbl>
1    0.0526
```

```
#If true we will have a higher variance
(R < CVx[1]/(2*CVy[1]))
```

```
##       cor
## [1,] TRUE
```

```
ggplot(srs_500, aes(x = Age, y = Inctot)) + geom_point() + geom_abline(intercept = 0,
    slope = r[[1]], color = "red")
```