BIOS 617 - Lecture 20

Walter Dempsey

3/28/2022

Replication methods for variance estimation

An alternative to linearization is replication or resampling: elements of the sample are dropped, a new estimator is computed using the remaining elements of the sample, and the resulting estimates resulting from repeated applications of this process are used to compute a variance estimator.

This approach is an extension of variance estimation under interpenetrated subsamples. In this setting, a sample is drawn and randomly divided into K subsamples reflecting the original sample design. The resulting K estimators $\hat{\theta}_k$ can then be viewed as an SRS from all possible samples under the design the mean $\hat{\theta} = K^{-1} \sum_{k=1}^{K} \hat{\theta}_k$ and estimated variance $v(\hat{\theta}) = K^{-1}(K-1)^{-1} \sum_{k=1}^{K} (\hat{\theta}_k - \hat{\theta})^2$

Balanced repeated replication (BRR)

- In practice actually creating interpenetrated samples can be practically onerous, especially if stratification and clustering is involved. Hence the practical methods we will discuss do not rely on actual interpenetrated sample.
- BRR is a method that assumes 2 PSUs per stratum (paired selection model).
- In practice this might not be the case, so approximations are made by collapsing/combining strata
  - Ultimate cluster sampling (ignore lower levels of clustering)
  - With-replacement approximations
  - Creating of Sampling Error Computation Units (SECUs)

Example: National Health and Nutrition Examination Survey I (1971-1974)

- Divide the US into 1900 PSUs (counties, groups of counties), combined into 65 strata:
  - 15 strata selected with certainty
  - 50 PSUs sampled from each of the non-certainty strata
  - Problem: have one PSU per stratum
  - For the certainty strata, 1213 second-stage neighborhoods were selected: these are collapsed 20 SECUs, 2 per 10 (new) strata
  - For the non-certainty strata, they are paired to create 50 SECUs in 25 (new) strata
- 1263 PSUs in 65 strata combined and collapsed to 35 strata with 2 SECUs each.

Paired selection design

- Consider estimating a total Y in a paired selection design
- Assume we have appropriate sample selection probabilities/weights so that $y_h = y_{h1} + y_{h2}$ is an unbiased estimator of the population total Y_h in the hth stratum,

$$E(y) = E\left(\sum_{h=1}^{H}(y_{h1} + y_{h2})\right) = Y$$

- Drawing one PSU per stratum α' at random to form a half-sample
- Letting the remaining PSUs α'' to form the other half-sample, yields

$$E(y') = 2E\left(\sum_{h=1}^{H} y_{h\alpha'}\right) = E(y'') = 2E\left(\sum_{h=1}^{H} y_{h\alpha''}\right) = Y$$

Alternative view

This can be viewed as case of interpenetrated subsamples with $K = 2$:

$$
\begin{aligned}
v(y) &= \frac{\sum_{k=1}^{2}(y_k - y)^2}{2 \times 1} \\
&= \frac{(y' - y)^2 + (y'' - y)^2}{2} \\
&= (y' - y)^2 = (y'' - y)^2
\end{aligned}
$$

Alternative view

- This is a very unstable estimator, as it only has a single degree of freedom based on that random split between SECUs within each stratum.
- So let's consider this estimator as a function of the underlying strata.

$$
\begin{aligned}
(y' - y)^2 &= \left(2\sum_{h=1}^{H} y_{h\alpha'} - \sum_{h=1}^{H}(y_{h\alpha'} + y_{h\alpha''})\right)^2 = \left(\sum_{h=1}^{H}(y_{h\alpha'} - y_{h\alpha''})\right)^2 \\
&= \sum_{h=1}^{H}(y_{h\alpha'} - y_{h\alpha''})^2 + 2\sum_{h=k+1}^{H}\sum_{k=1}^{H}(y_{h\alpha'} - y_{h\alpha''})(y_{k\alpha'} - y_{k\alpha''}) \\
&= \sum_{h=1}^{H}(y_{h1} - y_{h2})^2 + 2\sum_{h=k+1}^{H}\sum_{k=1}^{H}\xi_h\xi_k(y_{h1} - y_{h2})(y_{k1} - y_{k2})
\end{aligned}
$$

Alternative view

- So let's consider this estimator as a function of the underlying strata.

$$(y' - y)^2 = \sum_{h=1}^{H} (y_{h1} - y_{h2})^2 + 2 \sum_{h=k+1}^{H} \sum_{k=1}^{H} \xi_h \xi_k (y_{h1} - y_{h2})(y_{k1} - y_{k2})$$

$$= \sum_{h=1}^{H} d_h^2 + 2 \sum_{h=k+1}^{H} \sum_{k=1}^{H} \xi_h \xi_k d_h d_k$$

where $\xi_h = 1$ if $\alpha' = 1$ and $\xi_h = -1$ if $\alpha' = 2$ in stratum h, and $y_{(h1)} - y_{(h2)} = d_h$.

Alternative view

- Conditional on the sampled clustering in each stratum, ξ_h can be considered as a binary variable with probability 0.5 of taking on 1 and 0.5 of taking on -1
  - So $E(\xi_h \mid i \in s) = 0.5 - 0.5 = 0$
- Since sampling across strata are independent,

$$
\begin{aligned}
E\left[(y' - y)^2\right] &= E\left(\sum_{h=1}^{H} d_h^2\right) + 2E\left(\sum_{h=k+1}^{H}\sum_{k=1}^{H} \xi_h \xi_k d_h d_k\right) \\
&= V(y) + 2E\left[E\left[\sum_{h=k+1}^{H}\sum_{k=1}^{H} \xi_h \xi_k d_h d_k \mid i \in s\right]\right] \\
&= V(y) + 2E\left[\sum_{h=k+1}^{H}\sum_{k=1}^{H} E\left[\xi_h \xi_k \mid i \in s\right] d_h d_k\right] \\
&= V(y) + 2E\left[\sum_{h=k+1}^{H}\sum_{k=1}^{H} \underbrace{E\left[\xi_h \mid i \in s\right]}_{=0} \times \underbrace{E\left[\xi_k \mid i \in s\right]}_{=0} d_h d_k\right]
\end{aligned}
$$

More precise estimator

Thus we can obtain a more precise variance estimator by repeating the process of forming half-samples C times, and averaging over the differences between the half-sample estimator y'_k and the full sample estimator y:

$$v_c(y) = \frac{\sum_{c=1}^{C}(y'_c - y)^2}{C}$$

Half-samples without replacement:

For a specific set of draws of half-samples without replacement,

$$v_c(y) = C^{-1} \sum_{c=1}^{C} \left[\sum_{h=1}^{H} d_h^2 + 2 \sum_{h=k+1}^{H} \sum_{k=1}^{H} \xi_{ch} \xi_{ck} d_h d_k \right]$$

$$= \sum_{h=1}^{H} d_h^2 + 2 \sum_{h=k+1}^{H} \sum_{k=1}^{H} \left(\sum_{c=1}^{C} \frac{\xi_{ch} \xi_{ck}}{C} \right) d_h d_k$$

- Now $\sum_{c=1}^{C} \frac{\xi_{ch} \xi_{ck}}{C} \to 0$ as C gets large, and when $C = 2^H$ (maximum value), $\sum_{c=1}^{C} \xi_{ch} \xi_{ck} = 0$
- However, it is possible to choose samples in a balanced manner to achieve $\sum_{c=1}^{C} \xi_{ch} \xi_{ck} = 0$ for smaller values of C

Example: 3 strata and 4 replicate samples

Replicate c	ξ_{c1}	ξ_{c2}	ξ_{c3}
1	+1	+1	+1
2	+1	-1	-1
3	-1	-1	+1
4	-1	+1	-1

Can verify that cross-products cancel so that

$$(2(y_{11} + y_{21} + y_{31}) - y)^2 + \cdots (2(y_{12} + y_{21} + y_{32}) - y)^2/4$$

yields $v(y) = \sum_{h=1}^{H}(y_{h1} - y_{h2})^2$

Hadamard matrices

- A feature of this matrix is that columns are **orthogonal**: $\sum_c \xi_{ch}\xi_{ck} = 0$ for all h, k
- Matrices are called *Hadamard* matrices, and methods available to generate $C \times C$ matrices of this form for multiples of 4
  - For a 2-SECU design with H strata, use a Hadamard matrix such that $\min(C : C \geq H), \text{mod}(C, 4) = 0)$
  - Drop extra columns (remainder will still be orthogonal)
  - For example, if there are 70 strata, use a 72×72 Hadamard matrix, dropping columns 71 and 72 for the analysis.

Other option

- Of course, we could just compute $v(y)$ directly
- The value of the replication methods is that we compute the variance for a **general statistic** using this method
- Typically this is done using replication weights
- Going back to our simple example, we will generate 4 replication weights w_{ic}, $c = 1, \ldots, 4$ from the sampling weights w_i:
  - $w_{i1} = 2w_i \cdot 1[\text{SECU}_i = 1]$
  - $w_{i2} = 2w_i \cdot 1[\text{SECU}_i = 1, h_i = 1 \text{ or } \text{SECU}_i = 2, h_i = 2, 3]$
  - $w_{i3} = 2w_i \cdot 1[\text{SECU}_i = 1, h_i = 3 \text{ or } \text{SECU}_i = 2, h_i = 1, 2]$
  - $w_{i4} = 2w_i \cdot 1[\text{SECU}_i = 1, h_i = 2 \text{ or } \text{SECU}_i = 2, h_i = 1, 3]$

Other option (ctd)

We then compute a weighted estimator of our statistic $\hat{\theta}_c$ (e.g., a regression parameter) using the replication weight w_c, and estimate the variance as

$$v(\hat{\theta})_{BRR} = \frac{\sum_{c=1}^{C}(\hat{\theta}_c - \hat{\theta})^2}{C}$$

If θ is a vector, the variance-covariance can be obtained as

$$v(\hat{\theta})_{BRR} = \frac{\sum_{c=1}^{C}(\hat{\theta}_c - \hat{\theta})(\hat{\theta}_c - \hat{\theta})}{C}$$

Linear vs non-linear

Theory requires linear statistic $\bar{\theta}$ (e.g, mean, total) for exact results. But approximate results hold for non-linear statistics (e.g., regression parameters, variance components), and simulation studies generally show good behavior for non-linear statistics.

Jackknife repeated replication (JRR)

This is another replication method that is more flexible than the BRR methods in that it does not assume a 2 SECU-stratum design.

In the general setting with varying numbers of PSUs k_h per stratum,

$$v_{JRR}(\hat{\theta}) = \sum_{h=1}^{H} \frac{k_h - 1}{k_h} \sum_{i=1}^{k_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2$$

where $\hat{\theta}_{(hi)}$ is obtained dropping the ith PSU within the hth stratum, weighting up the remaining observations in the hth stratum by $k_h/(k_h - 1)$, and recomputing the estimate of θ.

JRR

Jackknife estimators of variance can be obtained for any of the general forms of sample design, by either letting $H = 1$ in the absence of stratification, or by treating each observation as a PSU in the absence of clustering.

Thus for a clustered design without stratification

$$v_{JRR}(\hat{\theta}) = \frac{k-1}{k} \sum_{i=1}^{k} \left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

where $\hat{\theta}_{(i)}$ is obtained dropping the ith PSU and weighting up the remaining observations $k/(k-1)$

JRR for stratified design

For a stratified design with independent observations

$$v_{JRR}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(\hat{\theta}_{(hi)} - \hat{\theta} \right)^2$$

where $\hat{\theta}_{(hi)}$ is obtained dropping the ith observation within the hth stratum and weighting up the remaining observations in the hth stratum by $n_h/(n_h - 1)$

JRR for SRS design

For an SRS design

$$v_{JRR}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}_{(i)} - \hat{\theta} \right)^2$$

where $\hat{\theta}_{(i)}$ is obtained dropping the ith observation and weighting up the remaining observations in the hth stratum by $n/(n-1)$

Illustration

For illustration, consider $\theta = Y$, the population total, and a stratified cluster design with k_h PSUs in the hth stratum, $h = 1, \ldots, H$. Then

$$\hat{\theta}_{(hi)} = \sum_{l=1, l \neq h}^{H} y_l + \frac{k_h}{k_h - 1} \sum_{j=1, j \neq i}^{k_h} y_{hj}$$

$$\begin{aligned}
\hat{\theta}_{(hi)} - \hat{\theta} &= \frac{k_h}{k_h - 1} \sum_{j=1, j \neq i}^{k_h} y_{hj} - y_h \\
&= \left(\sum_{j=1, j \neq i}^{k_h} y_{hj} - y_h \right) + \frac{1}{k_h - 1} \sum_{j=1, j \neq i}^{k_h} y_{hj} \\
&= \frac{1}{k_h - 1} \sum_{j=1, j \neq i}^{k_h} y_{hj} - y_{hi} = \frac{y_h}{k_h - 1} - y_{hi} \left[1 + \frac{1}{k_h - 1} \right].
\end{aligned}$$

Illustration (ctd)

$$\hat{\theta}_{(hi)} - \hat{\theta} = \frac{1}{k_h - 1} \sum_{j=1, j \neq i}^{k_h} y_{hj} - y_{hi} = \frac{y_h}{k_h - 1} - y_{hi} \left[1 + \frac{1}{k_h - 1} \right]$$

$$= \frac{k_h}{k_h - 1} \left[y_h / k_h - y_{hi} \right].$$

and thus

$$\left(\hat{\theta}_{(hi)} - \hat{\theta} \right)^2 = \left[\frac{k_h}{k_h - 1} \right]^2 \left[y_h / k_h - y_{hi} \right]^2$$

Maximizing precision

Conditioning on the sampled PSUs,

$$E\left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2 = \left[\frac{k_h}{k_h - 1}\right]^2 \frac{1}{k_h} \sum_{i=1}^{k_h} \left[y_{hi} - y_h/k_h\right]^2 = \frac{k_h}{k_h - 1} s_{yh}^2$$

where the expectation is over i uniformly chosen from $1, \ldots, k_h$. If we form C_h replicates from stratum h,

$$E\left(\frac{k_h - 1}{C_h} \sum_{c=1}^{C_h} \left(\hat{\theta}_{(hc)} - \hat{\theta}\right)^2\right) = k_h s_{yh}^2$$

We can maximize precision by using the maximum number of replicates k_h:

$$E\left(\frac{k_h - 1}{k_h} \sum_{i=1}^{k_h} \left(\hat{\theta}_{(hi)} - \hat{\theta}\right)^2\right) = k_h s_{yh}^2$$

JRR

Summing over the strata

$$E\left(\sum_{h=1}^{H}\frac{k_h-1}{k_h}\sum_{i=1}^{k_h}\left(\hat{\theta}_{(hi)}-\hat{\theta}\right)^2\right)=\sum_{h=1}^{H}k_h s_{yh}^2=v(y)$$

As with BRR, JRR requires linear statistics for this exact result to hold. Simulation studies using non-linear statistics show reasonably reliable behavior (better in larger samples).

Bootstrap

The bootstrap, like the jackknife, uses the empirical distribution function to estimate variance. In the SRS setting, the existing sample of n elements in sampled **with replacement** B times to yield a set of estimators $\{\hat{\theta}^{(b)}\}$ for $b = 1, \ldots, B$. The variance of $\hat{\theta}$ is estimated by the variance of the bootstrap estimators:

$$v_{boot}(\hat{\theta}) = (B-1)^{-1} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \hat{\theta}\right)^2$$

Bootstrap

- The sample design needs to be mirrored in the construction of the bootstrap estimator $\hat{\theta}^{(b)}$.
- When clustering is present, the resampling should be done at the clustering level, not at the element level.
- When stratification is present, the resampling should be conducted within each stratum.
- As with BRR and JRR, in unequal probability of selection sample weights are present, $\hat{\theta}^{(b)}$ needs to be computed using the (re)sample(d) weights.

Bootstrap

- Because the bootstrap is an asymptotic procedure, it requires a large sample size for the estimated variance to be relatively unbiased.
- But in multistage stratified designs, this requirement is often impossible to meet, since the number of PSUs with each stratum may be small.
- Indeed, if θ is a population total and $k_h = 2$ for all h, $v_{boot}(\hat{\theta})$ will underestimate $V(\hat{\theta})$ by a factor of 2!

Suggestions from the literature

1. Resample with replacement a sample of size $k_h - 1$ within each stratum, replacing totals with $\tilde{Y}^\star = \sum_{h=1}^{H} \frac{k_h}{k_h - 1} \sum_{i \in s^\star} w_i y_i$ for the resample $s^\star$
2. Resample with replacement a sample of size m_h within each stratum, replacing totals

$$\tilde{Y}^\star = \sum_{h=1}^{H} \left[\sqrt{\frac{m_h}{k_h - 1}} \frac{k_h}{m_h} \sum_{i \in s^\star} w_i y_i + \left(1 - \sqrt{\frac{m_h}{k_h - 1}} \right) \sum_{i \in s^\star} w_i y_i \right]$$

for the resample $s^\star$. Here m_h can be any size, although moment matching suggest $m_h = 1$ when $k_h = 2$ and $m_h \approx (k_h - 2)^2/(k_h - 1)$ for $k_h > 2$.

Example

Estimating the relationship of age on dioxin levels in the blood.

National Health and Nutrition Examination Survey (2005-2006)

- ~3100 US counties collapsed into 30 strata
- 1 PSU (county or group of counties) sampled per stratum
- Census blocks sampled within each PSU
- Blocks aggregated to 2 SECUs per stratum.
- Over/under sampling based on income, race/ethnicity, age.
- ~11,000 in full sample.

Example: NHNES

Regress TCDD level Y on age A:

$$Y_i = B_0 + B_1 A_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Sample limited to 1,250 adults who had full blood draws and mobile health check sites.

Example: R code

```
dioxin<-read.table("./data/dioxin2.dat")
tcdd<-log(dioxin[,1])
age<-dioxin[,2]
wt<-dioxin[,3]
psu<-dioxin[,4]
st<-dioxin[,5]
mypsu<-st*10+psu
designn<-svydesign(ids=mypsu,strata=st,
                   variables=tcdd~age,
                   weights=wt)
myreg<-svyglm(tcdd~age,design=designn)
```

Example: R output

```
d <- coef(summary(myreg))
knitr::kable(d)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.7228865	0.0568359	-12.71883	0
age	0.0232836	0.0014295	16.28832	0

Example: R output

```
designn2<-svydesign(ids=mypsu,strata=st,
                    variables=tcdd~age,
                    weights=rep(1,length(wt)))
myreg2<-svyglm(tcdd~age,design=designn2)
d2 <- coef(summary(myreg2))
knitr::kable(d2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.6654792	0.0436675	-15.23970	0
age	0.0226202	0.0010888	20.77536	0