Statistica Sinica **18**(2008): Supplement, S??∼S??

DISCLOSURE RISK AND REPLICATION-BASED VARIANCE ESTIMATION

Wilson W. Lu and Randy R. Sitter

Acadia University and Simon Fraser University

Supplementary Material

Appendix: Proofs of Lemmas and SAS Code

Proof of Lemma 1.

(a) For the delete-1 jackknife, the total number of replicates is $R = \sum_{h=1}^{H} n_h = n$. Let P_{hi} denote the i-th PSU from stratum h. Assume that for a given replicate r, PSU $P_{h_r i_r}$ is deleted. Clearly, for any sample unit j:

$$\delta_{j(r)} = \begin{cases} 0, & \text{if } j \in P_{h_r i_r}; \\ 1, & \text{if } j \in P_{hi},\ h \neq h_r; \\ \frac{n_h}{n_h - 1}, & \text{if } j \in P_{h_r i},\ i \neq i_r. \end{cases}$$

Thus,

(1) if $j, l \in P_{hi}$, $\delta_{j(r)} = \delta_{l(r)}$ for all $r = 1, \ldots, R$ and $d(j,l) = 0$;

(2) if $j \in P_{hi}$, and $l \in P_{hi'}$ for $i \neq i'$, the values of $\delta_{j(r)}$ and $\delta_{l(r)}$ differ only for two values of r; the one when P_{hi} is deleted and the one when $P_{hi'}$ is deleted. The absolute difference in both of these replicates, $|\delta_{j(r)} - \delta_{l(r)}|$, is $n_h/(n_h - 1)$. Thus, $d(j,l) = \sum_{r=1}^{R} |\delta_{i(r)} - \delta_{j(r)}| = 2n_h/(n_h - 1)$;

(3) suppose $j \in P_{hi}$ and $l \in P_{h'i'}$, where $h \neq h'$ and $i \neq i'$. For all r involving PSUs not in one of strata h or h', $\delta_{j(r)} = \delta_{l(r)}$. For r deleting P_{hi}, $\delta_{j(r)} = 0$ and $\delta_{l(r)} = 1$. For r deleting $P_{h'i'}$, $\delta_{l(r)} = 0$ and $\delta_{j(r)} = 1$. For each r deleting P_{hi^*}, $i^* \neq i$, $\delta_{j(r)} = n_h/(n_h - 1)$ and $\delta_{l(r)} = 1$. For each r deleting $P_{h'i^*}$, $i' \neq i^*$, $\delta_{jl(r)} = n_{h'}/(n_{h'} - 1)$ and $\delta_{j(r)} = 1$. Thus,

$$\begin{aligned} d(j,l) &= \sum_{r=1}^{R} |\delta_{j(r)} - \delta_{l(r)}| = 0 + \sum_{r \in s_h} |\delta_{j(r)} - \delta_{l(r)}| + \sum_{r \in s_{h'}} |\delta_{j(r)} - \delta_{l(r)}| \\ &= \left[1 + (n_h - 1) \cdot \left|1 - \frac{n_h}{n_h - 1}\right|\right] + \left[1 + (n_{h'} - 1) \cdot \left|1 - \frac{n_{h'}}{n_{h'} - 1}\right|\right] = 4. \end{aligned}$$

(b) For Fay's BRR, $n_h \equiv 2$. Any cell of Table 2, $\delta_{j(r)}$, will take values ϵ or $2-\epsilon$, such that,

(1) if $j, l \in P_{hi}$ $\delta_{j(r)} = \delta_{l(r)}$ for all $r = 1, \ldots, R$ so that $d(j,l) = 0$;

(2) if $j \in P_{hi}$, and $l \in P_{hi'}$ for $i \neq i'$, the values of $\delta_{j(r)}$ and $\delta_{l(r)}$ for any replicate r will always be ϵ and $2-\epsilon$ or $2-\epsilon$ and ϵ, resulting in $|\delta_{j(r)} - \delta_{l(r)}| = 2(1-\epsilon)$. Therefore, $d(j,l) = \sum_{r=1}^{R} |\delta_{j(r)} - \delta_{l(r)}| = 2R(1-\epsilon)$;

(3) suppose $j \in P_{hi}$ and $l \in P_{h'i'}$, where $h \neq h'$ and $i \neq i'$. The R replicates for j and l are constructed by two orthogonal columns of some Hadamard matrix, respectively. Therefore, among the R replicates, the pairs of $(\delta_{j(r)}, \delta_{l(r)})$ will take each of the four possible combinations (ϵ, ϵ), $(\epsilon, 2-\epsilon)$, $(2-\epsilon, \epsilon)$ and $(2-\epsilon, 2-\epsilon)$, $R/4$ times. Thus,

$$\begin{aligned} d_(j,l) &= \sum_{r=1}^{R} |\delta_{j(r)} - \delta_{l(r)}| \\ &= \frac{R}{4}[|\epsilon - \epsilon| + |\epsilon - (2-\epsilon)| + |(2-\epsilon) - \epsilon| + |(2-\epsilon) - (2-\epsilon)|] \\ &= R(1-\epsilon). \end{aligned}$$

Proof of Lemma 2. In the bootstrap, let $n_{hi}^{*(r)}$ denote the number of times the PSU P_{hi} is resampled for replicate r. For any sample unit j from PSU P_{hi}, we have

$$\delta_{j(r)} = n_{hi}^{*(r)} n_h (n_h - 1)^{-1},$$

where $\sum_{i=1}^{n_h} n_{hi}^{*(r)} = n_h – 1$. We also know that, for any replicate r, $(n_{h1}^{*(r)}, \ldots, n_{hn_h}^{*(r)})$ follow a multinomial distribution. It thus immediately follows that $E_*[\delta_{j(r)}] = V_*[\delta_{j(r)}] = 1$ and $Cov(\delta_{j(r)}, \delta_{l(r)}) = -(n_h - 1)^{-1}$, for j, l from different PSUs within the same stratum. Thus:

(1) if $j, l \in P_{hi}$, $\delta_{j(r)} = \delta_{l(r)}$ for all $r = 1, \ldots, R$ and $d(j,l) = 0$.

(2) if $j \in P_{hi}$, $l \in P_{hi'}$ and $i \neq i'$, for any given replicate r, we have

$$\begin{aligned} E_*[(\delta_{j(r)} - \delta_{l(r)})^2] &= V_*(\delta_{j(r)} - \delta_{l(r)}) \\ &= V_*(\delta_{j(r)}) + V_*(\delta_{l(r)}) - 2 \cdot Cov_*(\delta_{j(r)}, \delta_{l(r)}) \\ &= \frac{2n_h}{n_h - 1}, \end{aligned}$$

and thus, $E_*[d(j,l)^2] = \sum_{r=1}^{R} E_*[(\delta_{j(r)} - \delta_{l(r)})^2] = 2n_h R/(n_h - 1)$.
In addition, one can show that

$$E_*(\delta_{j(r)} - \delta_{l(r)})^4$$

$$
\begin{aligned}
&= E_*(\delta_{j(r)}^4) - 4E_*(\delta_{j(r)}^3\delta_{l(r)}) + 6E_*(\delta_{j(r)}^2\delta_{l(r)}^2) - 4E_*(\delta_{j(r)}\delta_{l(r)}^3) + E_*(\delta_{l(r)}^4)\\
&= (14n_h^3 - 24n_h^2)(n_h-1)^{-3},
\end{aligned}
$$

so that

$$
\begin{aligned}
E_*[d(j,l)^4] &= E_*\left[\sum_r (\delta_{j(r)} - \delta_{l(r)})^2\right]^2\\
&= \sum_r E_*(\delta_{j(r)} - \delta_{l(r)})^4 + \sum_{r\neq r'} E_*(\delta_{j(r)} - \delta_{l(r)})^2 E_*(\delta_{j(r')} - \delta_{l(r')})^2\\
&= \frac{(14n_h^3 - 24n_h^2)R}{(n_h-1)^3} + \frac{4R(R-1)n_h^2}{(n_h-1)^2}\\
&= \frac{10n_h^2(n_h-2)R}{(n_h-1)^3} + \frac{4n_h^2R^2}{(n_h-1)^2}.
\end{aligned}
$$

Thus, it follows that

$$
V_*[d(j,l)^2] = E_*[d(j,l)^4] - \{E_*[d(j,l)^2]\}^2 = 10n_h^2 R(n_h-2)(n_h-1)^{-3}.
$$

(3) suppose $j \in P_{hi}$ and $l \in P_{h'i'}$, where $h \neq h'$ and $i \neq i'$. Notice that now $\delta_{j(r)}$ and $\delta_{l(r)}$ are independent. Hence,

$$
E_*[(\delta_{j(r)} - \delta_{l(r)})^2] = V_*(\delta_{j(r)} - \delta_{l(r)}) = V_*(\delta_{j(r)}) + V_*(\delta_{l(r)}) = 1 + 1 = 2,
$$

and thus $E_*[d(j,l)^2] = \sum_{r=1}^R E_*[(\delta_{j(r)} - \delta_{l(r)})^2] = 2R$.
Also, since

$$
\begin{aligned}
&E_*(\delta_{j(r)} - \delta_{l(r)})^4\\
&= E_*(\delta_{j(r)}^4) - 4E_*(\delta_{j(r)}^3) + 6E_*(\delta_{j(r)}^2)E_*(\delta_{l(r)}^2) - 4E_*(\delta_{l(r)}^3) + E_*(\delta_{l(r)}^4)\\
&= 24 - n_h(5n_h-6)(n_h-1)^{-2} - n_{h'}(5n_{h'}-6)(n_{h'}-1)^{-2},
\end{aligned}
$$

we have

$$
\begin{aligned}
E_*[d(j,l)^4] &= \sum_r E_*(\delta_{j(r)} - \delta_{l(r)})^4 + \sum_{r\neq r'} E_*(\delta_{j(r)} - \delta_{l(r)})^2 E_*(\delta_{j(r')} - \delta_{l(r')})^2\\
&= \left[24 - \frac{n_h(5n_h-6)}{(n_h-1)^2} - \frac{n_{h'}(5n_{h'}-6)}{(n_{h'}-1)^2}\right]\cdot R + 4R(R-1).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
V_*[d(j,l)^2] &= E_*[d(j,l)^4] - \{E_*[d(j,l)^2]\}^2\\
&= 20R + 4R^2 - \frac{n_h(5n_h-6)R}{(n_h-1)^2} - \frac{n_{h'}(5n_{h'}-6)R}{(n_{h'}-1)^2} - (2R)^2\\
&= \left[20 - \frac{n_h(5n_h-6)}{(n_h-1)^2} - \frac{n_{h'}(5n_{h'}-6)}{(n_{h'}-1)^2}\right]R.
\end{aligned}
$$

SAS Code for a Simple Clustering Analysis.

```
/*Determine the location and names of input and output files*/
FILENAME inputf "C:\...\cluster\nhanesdata.txt";
FILENAME outputf "C:\...\cluster\nhanesout.txt";
/*Input the data file with true PSU identifiers and replicate weights*/
DATA nhanes;
   INFILE inputf truncover;
   INPUT psu_id rep1-rep24;
/*Obtain cluster membership based on replicate weights*/
PROC FASTCLUS data=nhanes out=Clust0 maxc=100 noprint;
   VAR rep1-rep24;
RUN;
/*Cross-tabulate PSU and cluster indicators*/
PROC FREQ data=Clust0;
   TABLES PSU_ID*Cluster / out=summary outpct;
RUN;
/*Output result with misspecification rate of cluster membership*/
PROC SORT data=summary;
   BY Cluster psu_id;
RUN;
DATA _null_;
   SET summary;
   FILE outputf;
   IF _n_=1 THEN
      PUT @5 "Cluster_ID" @20 "PSU_ID" @35 "PCT_COL" @50 "PCT_ROW";
   PUT @5 Cluster @20 psu_id @35 PCT_COL 6.2 @50 PCT_ROW 6.2;
RUN;
```

Proof of Lemma 3. One can view the swapping of PSU identifiers as the swapping of $\boldsymbol{y}_{hik}$ values between PSUs. Let $\boldsymbol{y}'_{hi}$ be the $\boldsymbol{y}_{hi}$ for the i-th PSU in stratum h after swapping, and let $\bar{\boldsymbol{y}}'_h = \sum_{i=1}^{n_h} \boldsymbol{y}'_{hi}/n_h$. Then, if unit k_j in PSU i_j in stratum h_j is swapped with unit k_l in PSU i_l in stratum h_l, $\boldsymbol{y}'_{h_j i_j} = \boldsymbol{y}_{h_j i_j} + n_{h_j}\boldsymbol{\Delta}_{jl}$ and $\boldsymbol{y}'_{h_l i_l} = \boldsymbol{y}_{h_l i_l} - n_{h_l}\boldsymbol{\Delta}_{jl}$, so that $\bar{\boldsymbol{y}}'_{h_j} = \bar{\boldsymbol{y}}_{h_j} + \boldsymbol{\Delta}_{jl}$ and $\bar{\boldsymbol{y}}'_{h_l} = \bar{\boldsymbol{y}}_{h_l} - \boldsymbol{\Delta}_{jl}$. It is then clear that

$$\begin{aligned}\boldsymbol{\Delta}_{01} &= \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left[(\boldsymbol{y}'_{hi} - \bar{\boldsymbol{y}}'_h)(\boldsymbol{y}'_{hi} - \bar{\boldsymbol{y}}'_h)^T - (\boldsymbol{y}_{hi} - \bar{\boldsymbol{y}}_h)(\boldsymbol{y}_{hi} - \bar{\boldsymbol{y}}_h)^T \right] \\ &= \sum_{(j,l)\in\mathcal{A}_{01}} \{D_j + D_l\},\end{aligned}$$

where

$$D_j = \frac{1}{n_{h_j}(n_{h_j}-1)} \left[(\boldsymbol{y}'_{h_j i_j} - \bar{\boldsymbol{y}}'_{h_j})(\boldsymbol{y}'_{h_j i_j} - \bar{\boldsymbol{y}}'_{h_j})^T - (\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})^T \right]$$

$$= \frac{1}{n_{h_j}(n_{h_j}-1)} \Big\{ [(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j}) + (n_{h_j}-1)\boldsymbol{\Delta}_{jl}][(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j}) + (n_{h_j}-1)\boldsymbol{\Delta}_{jl}]^T$$
$$-(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})^T \Big\}$$
$$= \frac{(n_{h_j}-1)}{n_{h_j}} \boldsymbol{\Delta}_{jl}\boldsymbol{\Delta}_{jl}^T + \frac{1}{n_{h_j}} \Big[(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})\boldsymbol{\Delta}_{jl}^T + \boldsymbol{\Delta}_{jl}(\boldsymbol{y}_{h_j i_j} - \bar{\boldsymbol{y}}_{h_j})^T \Big]$$

and similarly

$$D_l = \frac{(n_{h_l}-1)}{n_{h_l}} \boldsymbol{\Delta}_{jl}\boldsymbol{\Delta}_{jl}^T - \frac{1}{n_{h_l}} \left[(\boldsymbol{y}_{h_l i_l} - \bar{\boldsymbol{y}}_{h_l})\boldsymbol{\Delta}_{jl}^T + \boldsymbol{\Delta}_{jl}(\boldsymbol{y}_{h_l i_l} - \bar{\boldsymbol{y}}_{h_l})^T \right],$$

and the result follows.

Department of Mathematics and Statistics, Acadia University, Wolfville, NS, B4P 2R6, Canada.

E-mail: wilson.lu@acadiau.ca

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC V5A-1S6, Canada.

E-mail: sitter@stat.sfu.ca

(Received June 2006; accepted May 2007)