Weighting Issues for the Hospital Survey

Elizabeth Merwin and Steven Stern

July 2003

1 Weighting without Other Merged Data

1.1 Weighting with No Inappropriate Observations

Let \( p_i = p(z_i) \) be the probability that hospital \( i \) was picked from our mailing list where \( z_i \) is a variable that determines over- and undersampling. Normalize the probabilities so that \( \sum p_i = 1 \). Let \( y_i \) be the realization of the random variable \( Y(z_i) \) for hospital \( i \). \( Y(z_i) \) is explicitly written as a function of \( z_i \) to recognize that its conditional distribution depends on \( z \). Let \( P \) be the population of observations with size \( n_P \) in the mailing list, and let \( S \subset P \) be the set of observations chosen from \( P \). Consider estimating

\[
EY = \int \int y f(y | z) dy g(z) dz.
\]

Assuming that our mailing list was a representative sample of hospitals, an unbiased estimate of \( EY \) is

\[
\sum_{i \in S} w_i y_i = \sum_{i \in P} w_i y_i 1(i \in S)
\]

where

\[
w_i = p_i^{-1} / \sum_{i \in S} p_i^{-1}
\]

**Proof.** Note that

\[
E \sum_{i \in P} w_i Y_i 1(i \in S) = \sum_{i \in P} p_i^{-1} EY_i p_i = \sum_{i \in P} EY_i \frac{1}{\sum_{i \in S} p_i^{-1}}.
\]

Note that the denominator has an expected value of

\[
E \sum_{i \in S} p_i^{-1} = \sum_{i \in P} p_i^{-1} p_i = n_P
\]
and no variance (because of the way we sampled). So

\[ E \sum_{i \in P} w_i Y_i (i \in S) = \frac{1}{n_P} \sum_{i \in P} E Y_i = E Y. \]

Note that a consistent and unbiased estimate of \( n_P E Y \) is \( n_P \sum_{i \in S} w_i y_i \).

Assume there is a fuller set of variables \( x_i \) observed for all \( i \in P^* \) with \( P \subseteq P^* \) that provides significant information about the distribution of \( Y_i \). Then we can do better. First we can estimate the relationship

\[ E Y_i = g(x_i) \]

using the observations in our sample (assuming no selection bias). Let \( \hat{g}(\cdot) \) be an unbiased estimator such that

\[ \sum_{i \in S} [y_i - \hat{g}(x_i)] = 0; \]

e.g., OLS with a constant. Then we can estimate

\[ \hat{\mu} = \frac{1}{n_{P^*}} \sum_{i \in P^*} \hat{g}(x_i). \]

Ignoring oversampling and weighting, note that

\[
\text{Var} \hat{\mu} = \text{Var} \left[ \frac{1}{n_{P^*}} \sum_{i \in P^*} \hat{g}(x_i) \right] \\
= \left( \frac{1}{n_{P^*}} \right)^2 \left[ \text{Var} \left( \frac{1}{n_S} \sum_{i \in S} \hat{g}(x_i) \right) \\
+ (n_{P^*} - n_S)^2 \text{Var} \left( \frac{1}{n_{P^*} - n_S} \sum_{i \in P^* \setminus S} \hat{g}(x_i) \right) \right] \\
= \left( \frac{1}{n_{P^*}} \right)^2 \left[ \left( n_S^2 + (n_{P^*} - n_S) n_S \right) \text{Var} \left( \frac{1}{n_S} \sum_{i \in S} y_i \right) \right] \\
= \left( \frac{1}{n_{P^*}} \right)^2 \left[ n_S n_{P^*} \text{Var} \left( \frac{1}{n_S} \sum_{i \in S} y_i \right) \right] \\
= \frac{n_S}{n_{P^*}} \text{Var} \left( \frac{1}{n_S} \sum_{i \in S} y_i \right) < \text{Var} \left( \frac{1}{n_S} \sum_{i \in S} y_i \right).
\]

Note that a consistent and unbiased estimate of \( n_P E Y \) is \( n_P \hat{\mu} \).
1.2 Weighting with Inappropriate Observations

Let \( q_i \) be an indicator if hospital \( i \) is appropriate. It would be inappropriate if it performs no surgeries. Let

\[
q_i P = \sum_{i \in P} q_i.
\]

We still pick hospitals with probability \( p_i = p(z_i) \) which is independent of \( q_i \). Again consider estimating

\[
EY = \int \int y f(y \mid z) d y g(z) \, dz.
\]

Still assuming that our mailing list was a representative sample of hospitals, an unbiased estimate of \( EY \) is

\[
\sum_{i \in S} w_i y_i = \sum_{i \in P} w_i y_i 1(i \in S)
\]

where

\[
w_i = \frac{q_i p_i^{-1}}{\sum_{i \in S} q_i p_i^{-1}}.
\]

A consistent and unbiased estimate of \( q_i EY \) is \( n_p = \sum_{i \in S} w_i y_i \). Similarly, if we use another data set to construct \( \tilde{\mu} \), then a consistent and unbiased estimate of \( n_p \tilde{\mu} \) is \( n_p \tilde{\mu} \).

1.3 Variance of Estimates with Weighting

In general, our estimator can be written as \( \sum_{i \in S} w_i y_i \) for some set of weights \( \{w_i, i \in S\} \). For purposes of computing variances of estimates, we treat the weights as nonrandom. Thus, assuming the observations are independent, the variance is

\[
Var \left[ \sum_{i \in S} w_i y_i \right] = \sum_{i \in S} w_i^2 \text{Var}(y_i).
\]

Note that, since \( \sum_{i \in S} w_i = 1 \) and \( w_i > 0 \) \( \forall i \in S \), \( \sum_{i \in S} w_i^2 \rightarrow 0 \) as \( \#S \rightarrow \infty \) as long as no finite set of observations dominates the sample.

In the case of vacancies, a reasonable assumption is that \( y_i \sim \text{Poisson}(\lambda_i) \) where \( \log \lambda_i = X_i \beta \). A reasonable choice for \( X_i \) is a constant, some measure of size, and a dummy for whether the hospital uses CRNAs. The results of such a regression are reported in Table 1.
Table 1
Poisson Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>wo/ State Dummies</th>
<th>w/ State Dummies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.594**</td>
<td>-2.761**</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
<td>(0.277)</td>
</tr>
<tr>
<td>Beds</td>
<td>0.252**</td>
<td>0.223**</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>Hires CRNAs</td>
<td>1.851**</td>
<td>1.800**</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
<td>(0.139)</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-1141.1</td>
<td>-1015.3</td>
</tr>
</tbody>
</table>

Notes:
1. Numbers in parentheses are standard errors.
2. Double-starred items are significant at the 5% level.

2 Computing Aggregate Vacancies with Merged AHA (and ARF) Data

An alternative with the potential to provide much more precise estimates is to merge our hospital data with AHA and ARF data.\(^1\) Assume that

\[ y_i = X_i \beta + u_i \]

where \( y_i \) is the number of vacancies in hospital \( i \) and \( X_i \) is a vector of hospital \( i \) characteristics including variables observable in either the AHA or ARF. Let \( S \) be the set of hospitals from our sample that can be merged with AHA and ARF data. Then the OLS estimate of \( \beta \) using the elements in \( S \) is

\[ \hat{\beta} = \left( \sum_{i \in S} X_i'X_i \right)^{-1} \left( \sum_{i \in S} X_iy_i \right) . \]

\( \hat{\beta} \) is consistent and unbiased as long as sampling into \( S \) is exogenous and efficient as long as \( u_i \sim iid (0, \sigma^2) \).\(^2\)

Let \( P \) be the population of hospitals in AHA, and let \( P_k \) be an arbitrary subset (indexed by \( k \)) of \( P \). Since each hospital in \( P \) has observable \( X_i \), we can construct a predicted value of \( y_i \):

\[ \hat{y}_i = X_i\hat{\beta} . \]

\(^1\)This needs to be done by matching names of hospitals.
\(^2\)Correcting for heteroskedasticity is a straightforward application of weighted least squares where the weights are inverse variances.
An unbiased estimate of
\[ y(P_k) = \sum_{i \in P_k} y_i \]
is
\[ \hat{y}(P_k) = \sum_{i \in P_k} \hat{y}_i. \]

For example, we define \( P_k \) to be the set of hospitals in state \( k \); then \( \hat{y}(P_k) \) is the predicted number of vacancies in state \( k \).

The variance of our estimate based on this regression is smaller than the estimate without regression results because we are only adding information in the regression framework. Only if all of the \( X_i \) are noninformative about \( y_i \) does the weighting procedure result in an estimate with the same variance.\(^3\)

\(^3\)The weighting procedure is equivalent to using the regression methodology where \( X_i \) includes only a constant.