Weighting Methods

Erin Hartman

PS239 Sp2010

April 7, 2010

Reweighting non-random samples to get back a representative sample.

What are some causes of non-random samples?

- Quota sampling
- Nonresponse
- Convenience Samples
- Oversampling Designs

What problems are caused by non-random samples?

Roadmap

Some reweighting methods:

- Post Stratification
- Raking
- Inverse Propensity Score Weighting
- Maximum Entropy Weighting

A side note on using inverse propensity score weighting to recover average treatment effects.

Post Stratification

- Post-stratification (aka cell weighting) is the simplest of ways to bring sample proportions in line with population proportions
- Stratify the sample into a number of cells (J), based on characteristics of the population deemed important (the X variables).
- Give more weight to under-represented groups and less weight to over-represented groups
- $$w_j = \frac{\text{proportion in population in cell } j}{\text{proportion in sample in cell } j}$$
- Each person in cell j receives weight w_j

Post-stratification Example

Cell	Pop Size	Pop Prop	Sample Size	Sample Prop	Cell Weight
Male 18-34	2000	0.25	30	0.15	1.67
Female 18-34	1760	0.22	40	0.20	1.10
Male 35-64	1440	0.18	50	0.25	0.72
Female 35-64	1600	0.20	60	0.30	0.67
Male 65+	400	0.05	10	0.05	1.00
Female 65+	800	0.10	10	0.05	2.00
Total:	8000	1.00	200	1.00	

Problems with Post-stratification

As numbers variables increases, number of cells increases rapidly.

You are quickly going to run out of people in your cells!

Also requires a lot of information about your population. You need the joint distributions in your population.

Raking

- Raking (aka iterative proportion weighting or rim weighting) uses only marginal distributions.
- This allows us to include more covariates.
- First used in the 1940 census to ensure that the complete census data and samples taken from it gave consistent results. Originally developed by Deming and Stephan (1940).
- Uses an iterative algorithm to rake along columns and then rows until convergence.

Raking Algorithm

There are many different algorithms, here is one (from Little and Wu 1991):

- 1. Initialize the weights by setting each equal to n_{ij}/n, which is the sample cell count over the sample size
- 2. Calculate $\hat{w}_{ij}^{(1)} = w_i * \frac{\hat{w}_{ij}^{(0)}}{\sum_i \hat{w}_i^{(0)}}$. Here you are "raking" over rows.
- 3. Calculate $\hat{w}_{ij}^{(2)} = w_j * \frac{\hat{w}_{ij}^{(1)}}{\sum_j \hat{w}_j^{(1)}}$. Here you are "raking" over columns.
- 4. Repeat steps 2 and 3 until $\sum_i \hat{w}_i = w_i$ and $\sum_j \hat{w}_j = w_j$ for each i and j, i.e. "convergence" is achieved.

Raking Example

Suppose the sum of weights in our sample are:

	Black	White	Asian	Native American	Other	Sum of Weights
Female	300	1200	60	30	30	1620
Male	150	1080	90	30	30	1380
Sum of Weights	450	2280	150	60	60	3000

But suppose that the true population counts for the marginal totals are: 1510 women and 1490 men, 600 blacks, 2120 whites, 150 asians, 100 native americans, 30 others

Raking Example

So, we adjust the rows first. We multiply each row by the $\frac{\text{true row pop}}{\text{estimated row pop}}$, or $\frac{1510}{1620}$ for the female row and $\frac{1490}{1380}$ for the male row.

	Black	White	Asian	Native American	Other	Sum of Weights
Female	279.63	1118.52	55.93	27.96	27.96	1510
Male	161.96	1166.09	97.17	32.39	32.39	1490
Sum of Weights	441.59	2284.61	153.10	60.35	60.35	3000

Now the row margins are right, but the column margins aren't. So, we rake over columns.

Raking Example

We multiply each column by the $\frac{\text{true row pop}}{\text{newly estimated row pop}}$, or $\frac{600}{441.59}$ for the first column, etc.

	Black	White	Asian	Native American	Other	Sum of Weights
Female	379.94	1037.93	54.79	46.33	13.90	1532.90
Male	220.06	1082.07	95.21	53.67	16.10	1467.10
Sum of Weights	600	2120	150	100	30	3000

And... we see the problem... so we repeat, until convergence.

Raking Example

We multiply each column by the $\frac{\text{true row pop}}{\text{newly estimated row pop}}$, or $\frac{600}{441.59}$ for the first column, etc.

	Black	White	Asian	Native American	Other	Sum of Weights
Female	375.59	1021.47	53.72	45.56	13.67	1510
Male	224.41	1098.53	96.28	54.44	16.33	1490
Sum of Weights	600	2120	150	100	30	3000

So, we see that for each cell we end up with new cell values. White males were valued at 1080 before, which was sligntly under weighted compared to the 1098.53 now. So, we weight each person in the white male cell by $1098.53/1080 \approx 1.017$ whereas other males should be given a weight of $16.33/30 \approx 0.544$

Raking Pros and Cons

- Convergence can be slow, and occasionally impossible.
- Makes an assumption about how the data is structured, namely that response probabilities depend only on the row and column and not on the specific cell.
- Should converge to post-stratification if the marginal variables are independent.

Inverse Propensity Score Weighting

- Inverse propensity score weighting allows us to use some of the advantages of raking, namely that we can condition on many covariates.
- However, requires similar data in the sample and population, so we need more information from our population.

Inverse Propensity Score Weighting Algorithm

Using the data from the sample and the population, we will estimate the propensity of being in the sample. So, instead of estimating with $T \in (0,1)$, we want to estimate the propensity for $I \in (0,1)$ where I is an indicator for being in the sample.

We then weight by the inverse of the estimated propensity score.

IPSW Example

Say we have 1000 schools in our state, but only 150 of them were chosen to be in our study.

We estimate our propensity to be in the study, and each school in the study receives a weight of $w_i = \frac{1}{p_i}$.

- Say a school in the study has a p propensity of being in the study of 0.1. What weight does this result in them having? Why?
- 10. They represent 9 other schools across the state that are similar to them but were not selected.
- Say a school in the study has a p propensity of being in the study of 0.9. What weight does this result in them having? Why?
- 1.11. Nearly all schools "like" them are likely to be in the trial, so they are doing less work.

IPSW Pros and Cons

- What do we do for propensities near 0? As $p \longrightarrow 0, w_i \longrightarrow \infty$
- How does this compare to post stratification? Raking?

Entropy

- Maximum entropy is an information theoretic concept based on Shannon's measure of entropy.

$$S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i$$

where $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ is a probability distribution s.t. $p_i \geq 0$ and $\sum_{i=1}^{n} p_i = 1$

Entropy can be thought of as a measure of the amount of information in the distribution. Maximum entropy would be obtained in a uniform distribution.

Maximum Entropy

Jaynes defined the principle of maximum entropy as:

$$\max_{\mathbf{p}} S(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln p_i$$

$$s.t. \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_r(x_i) = \sum_{i=1}^{n} p_i g_{ri} = a_r \qquad r = 1, \ldots, m \\ p_i \geq 0 \qquad i = 1, 2, \ldots, n \end{cases}$$

- Equation (1) is referred to as a natural constraint, stating that all probabilities must sum to unity.
- Equation (2), the m moment constraints, are referred to as the consistency constraints. Each a_r represents an r-th order moment, or characteristic moment, of the probability distribution.
- Equation (3) the final constraint ensures that all probabilities are non-negative. This is always met.

MaxEnt Algorithm

- Maximization problem:

$$\max_{\mathbf{p},\lambda_0,x} L_x = \sum_{i=1}^{N} p_i \ ln \ \frac{p_i}{q_i} + (\lambda_0 - 1)\Big(\sum_{i=1}^{N} p_i - 1\Big) + \sum_{r=1}^{M} \lambda_r \Big(\sum_{i=1}^{N} p_i \ g_{ri}(x_i) - a_r\Big)$$

  - $\mathbf{z} = [\lambda_1, \ldots, \lambda_M]$ and $\lambda_0 - 1$ are the Lagrange Multipliers

MaxEnt Algorithm

- First Order Conitions:

$$p_i = \frac{q_i \exp\Big(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\Big)}{\sum_{i=1}^{N} q_i \exp\Big(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\Big)}$$

$$a_r = \frac{\sum_{i=1}^{N} g_{ri}(x_i) q_i \exp\Big(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\Big)}{\sum_{i=1}^{N} q_i \exp\Big(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\Big)}$$

- Solution:

$$L_k^*(\mathbf{z}) = -ln\Big[\sum_{i=1}^{N} g_{ri}(x_i) q_i \exp\Big(-\sum_{r=1}^{M} \lambda_r g_{ri}(x_i)\Big)\Big] - \sum_{r=1}^{M} \lambda_r a_r$$

- Solution isn't closed form, so it is solved algorithmically.

Maximum Entropy

The maximum entropy is the probability distribution that meets the constraints and maximizes Shannon's entropy metric.

The solution to this problem satisfies LaPlace's "principle of indifference". It ensures that we don't use any information beyond the information that allows us to satisfy the constraints, and nothing more.

Allows us to incorporate both the joint cell distributions and marginals from other datasets. If we include just cells, should converge to cell weights, only marginals should converge to raking methods.

IPSW for Treatment Effects

IPSW can be used within one population to estimate average treatment effects.

We estimate a propensity score in the typical way, estimating propensity to treatment.

Define $\hat{r}_i = \frac{\hat{p}_i}{1-\hat{p}_i}$ as the estimated odds of the propensity of being assigned to treatment.

IPSW for Treatment Effects

$$ATT = \left(\frac{1}{n_1}\sum_{i:T_i=1} y_i\right) - \left(\frac{\sum_{i:T_i=0}\hat{r}_i y_i}{\sum_{i:T_i=0}\hat{r}_i}\right)$$

$$ATC = \left(\frac{\sum_{i:T_i=1} y_i/\hat{r}_i}{\sum_{i:T_i=1} n_1/\hat{r}_i}\right) - \left(\frac{1}{n_0}\sum_{i:T_i=0} y_i\right)$$

$$ATE = \left(\frac{1}{n}\sum_i T_i\right) ATT + \left(1 - \frac{1}{n}\sum_i T_i\right) ATC$$

IPW for Treatment Effects

$$\widehat{ATE}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i Y_i}{\hat{\pi}(\mathbf{Z_i})} - \frac{(1-T_i)Y_i}{1-\hat{\pi}(\mathbf{Z_i})}\right\}$$

Note the problems here with propensities near 0 and 1!