Seven Model Motivated Rules of Thumb or Equations

Fritz Scheuren

Scheuren@aol.com

March 17, 2005

Notes on Nonresponse Presentation

In the slides that follow much has been left unwritten because it was intended to be said. Those spoken words are not included here, although not to be too cryptic, I have added a few notes to hint at what I was trying to say.

Notes on Nonresponse Presentation

First, the (obviously vast) topic of nonresponse has been confined to unit nonresponse. No mention is made of item nonresponse, for example. Arguably many points carry over nonetheless.

Notes on Nonresponse Presentation

Second, the topic has been treated from an historical perspective, following the approach I have been taking in my History Corner series in The American Statistician (TAS). In this connection the November 2004 TAS issue may be of particular value, as it brings out some of my points in more detail.

Notes on Nonresponse Presentation

Third, the formulas provided are more LOGOS that prescriptions to be used in real life applications. They were part of the arguments that the original framers of these ideas used but are oversimplified. For example, most of these ideas would be imbedded in pre or post strata to give them more plausibility and content.

Notes on Nonresponse Presentation

Fourth, in some cases the formula has been associated in the slides with their author. But this has not always been done so each of authors of these formulas or rules has been given mention below.

Notes on Nonresponse Presentation

Fifth, the bias/variance tradeoff slide (No.1) was taken from the introductory chapters of the HHM (Hansen, Hurwitz and Madow) and Cochran sampling texts. Their result has been repeated here to introduce how our professional (over) emphasis on sampling was sold. Those texts carved out the sampling role from the much larger data collection issues - all too successfully it might be added.

Notes on Nonresponse Presentation

Sixth, the response (and implicitly nonresponse) bias factors formula (No.2), original with me, was an attempt to succinctly indicate how wrongly our practice continues to overemphasize sampling, even 50+ years later. We need to add significantly to our records the extensive paradata that is routinely assembled in our surveys and bring it forward to the client and to future survey developers.

Notes on Nonresponse Presentation

Seventh, the missingness mixture slides (Nos. 3 to 5) are my interpretation of Rubin's seminal work on missingness (Biometrika 1976), where the emphasis is on the ubiquity with which all forms of missing unit nonresponse existing simultaneously in nearly all practical settings. Our failure to act on this knowledge is a weakness in our practice.

Notes on Nonresponse Presentation

Eighth, the Hansen-Hurwitz subsampling idea for nonresponse is discussed next (No.6). Here the attempt is to illustrate how we could reinterpret their work using Rubin's idea of the multiple forms of missingness and, thereby, reduce the variance penalty if we have sufficient confidence in our missingness model. The interview mode change made as part of the original idea is sometimes lost and I regret that I did not say more about it last week.

Notes on Nonresponse Presentation

Ninth, the Sekar-Deming capture/recapture application to nonresponse (No. 7) concludes the talk, except for some discussion of implications. Since these ideas are familiar in their dual systems incarnation, there is less to say here that elsewhere.

Notes on Nonresponse Presentation

Tenth, the concluding discussion slides were just my take on our history Many good suggestions were made for better terminology by those present. Several objected to Rubin's use of the phrase "Missing at Random," abbreviated as MAR. Why not say, instead, they argued, that if we have the right variables present (say in the frame), then we can condition on them and thus eliminate some of the bias.

Seven Rules or Equations

Seven Rules or Equations

1. Bias versus Sampling Error

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors
3. Mixtures of Missingness

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors
3. Mixtures of Missingness
4. Data Collection Missingness

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors
3. Mixtures of Missingness
4. Data Collection Missingness
5. Data Analysis Missingness

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors
3. Mixtures of Missingness
4. Data Collection Missingness
5. Data Analysis Missingness
6. Subsamples for Nonresponse

Seven Rules or Equations

1. Bias versus Sampling Error
2. Response Bias Factors
3. Mixtures of Missingness
4. Data Collection Missingness
5. Data Analysis Missingness
6. Subsamples for Nonresponse
7. "Bias" or Hard-Core Non-Response Rates

1. Bias versus Sampling Error

1. Bias versus Sampling Error

$$\begin{aligned} \mathrm{MSE}[f(Y)] &= \sqrt{\frac{\sigma^2}{n} + B^2} \\ &= \sqrt{\frac{\sigma^2}{n}(1 + R^2)} \end{aligned}$$

(e.g., as shown in the Hansen, Hurwitz and Madow text.)

2. Survey Bias Factors

$$\int \int \int f(Y, M, R, C, D) dM \, dR \, dC = g(Y, D)$$

Implicitly, this formal integration symbolizes that final survey data files usually treat process fixes as complete. Leaving out the paradata concerning these flaws continues in a way the historical theme in slide No. 1.

3. Mixtures of Missingness

In general

$$m = m_{\mathrm{MCAR}} + m_{\mathrm{MAR}} + m_{\mathrm{NMAR}}$$

can become specialized at various survey stages

4. Data Collection Missingness

4. Data Collection Missingness

$$m = m_{\mathrm{MCAR}} + m_{\mathrm{MAR}} + m_{\mathrm{NMAR}}$$

Usually becomes $m = m_{\mathrm{NMAR}}$!

Implicitly, given what is done

4. Data Collection Missingness

Some Alternatives

4. Data Collection Missingness

Some Alternatives

- Subsampling Solutions, $m = m_{\text{NMAR}}$

4. Data Collection Missingness

Some Alternatives

- Subsampling Solutions, $m = m_{\text{NMAR}}$
- Previous Experience, See Example

4. Data Collection Missingness

Some Alternatives

- Subsampling Solutions, $m = m_{\text{NMAR}}$
- Previous Experience, See Example
- Adaptive Approach, Decide on Fly

4. Data Collection Missingness

Some Alternatives

- Subsampling Solutions, $m = m_{\text{NMAR}}$
- Previous Experience, See Example
- Adaptive Approach, Decide on Fly
- Enrich the Frame, Make More $m = m_{\text{MAR}}$

4. Data Collection Missingness

Some Alternatives

- Subsampling Solutions, $m = m_{\text{NMAR}}$
- Previous Experience, See Example
- Adaptive Approach, Decide on Fly
- Enrich the Frame, Make More $m = m_{\text{MAR}}$
- Use a Composite

5. Data Analysis Missingness

5. Data Analysis Missingness

$$m = m_{\mathrm{MCAR}} + m_{\mathrm{MAR}} + m_{\mathrm{NMAR}}$$

Usually becomes $m = m_{\mathrm{MAR}}$!

Again, given what is done

5. Data Analysis Missingness

Some Alternatives

5. Data Analysis Missingness

Some Alternatives

- Noncontact Nonresponse $m = m_{\mathrm{NMAR}}$?

5. Data Analysis Missingness

Some Alternatives

- Noncontact Nonresponse $m = m_{\mathrm{NMAR}}$?
- Refused Nonresponse $m = m_{\mathrm{MAR}}$!

5. Data Analysis Missingness

Some Alternatives

- Noncontact Nonresponse $m = m_{\mathrm{NMAR}}$?
- Refused Nonresponse $m = m_{\mathrm{MAR}}$!
- Adaptively Balanced Adjustment

5. Data Analysis Missingness

Some Alternatives

- Noncontact Nonresponse $m = m_{\mathrm{NMAR}}$?
- Refused Nonresponse $m = m_{\mathrm{MAR}}$!
- Adaptively Balanced Adjustment
- Sensitivity Analysis

5. Data Analysis Missingness

Some Alternatives

- Noncontact Nonresponse $m = m_{\mathrm{NMAR}}$?
- Refused Nonresponse $m = m_{\mathrm{MAR}}$!
- Adaptively Balanced Adjustment
- Sensitivity Analysis
- Weight Trimming

6. Hansen-Hurwitz Formula

...But Reinterpreted

6. Hansen-Hurwitz Formula

...But Reinterpreted

$$\bar{y} = p_R \bar{y}_R + p_M \bar{y}_M$$

where

- $p_R =$ "responding" fraction of the original sample
- $p_M =$ remaining "non-responding" fraction of the original sample

6. Hansen-Hurwitz Formula

...But Reinterpreted (cont'd)

6. Hansen-Hurwitz Formula

...But Reinterpreted (cont'd)

$$\hat{v}(\hat{\bar{y}}) \approx p_R \frac{s_R^2}{n} + p_M \frac{s_M^2}{vn} + \frac{1}{n-1}\left[p_R(\bar{y}_R - \hat{\bar{y}})^2 + p_M(\bar{y}_M - \hat{\bar{y}})^2\right]$$

where $v =$ sub-sampling rate and other terms, e.g., s_R^2, are just the standard notation

6. Subsamples for Nonresponse

Mixture Example

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts
- Treat Most (80%?) as $m_N = m_{\text{NMAR}}$

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts
- Treat Most (80%?) as $m_N = m_{\text{NMAR}}$
- Lightly Subsample m_R Refusals

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts
- Treat Most (80%?) as $m_N = m_{\text{NMAR}}$
- Lightly Subsample m_R Refusals
- If there are good covariates present

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts
- Treat Most (80%?) as $m_N = m_{\text{NMAR}}$
- Lightly Subsample m_R Refusals
- If there are good covariates present
- Treat "Bulk" as $m_R = m_{\text{MAR}} + m_{\text{MCAR}}$

6. Subsamples for Nonresponse

Mixture Example

- Mainly Subsample m_N Noncontacts
- Treat Most (80%?) as $m_N = m_{\text{NMAR}}$
- Lightly Subsample m_R Refusals
- If there are good covariates present
- Treat "Bulk" as $m_R = m_{\text{MAR}} + m_{\text{MCAR}}$
- "Bulk" can sometimes be estimated?

7. Sekar-Deming "Bias" Rates

7. Sekar-Deming "Bias" Rates

- Employ two samples on similar subjects

7. Sekar-Deming "Bias" Rates

- Employ two samples on similar subjects
- Include cases from first survey in second

7. Sekar-Deming "Bias" Rates

- Employ two samples on similar subjects
- Include cases from first survey in second
- Unlike Hansen-Hurwitz, obtain both Respondents and Nonrespondents

7. Sekar-Deming "Bias" Rates

- Employ two samples on similar subjects
- Include cases from first survey in second
- Unlike Hansen-Hurwitz, obtain both Respondents and Nonrespondents
- Create a 2×2 Table where the cells are
  - $rr =$ responded both times
  - $rn =$ responded first time only
  - $nr =$ responded second time only
  - $nn =$ never responded

7. Capture/Recapture Table

$$\left[\begin{array}{c|c} rr & rn \\ \hline nr & nn \end{array}\right]$$

7. Potential Bias Rate

7. Potential Bias Rate

$$dd = \frac{(rn)(nr)}{rr}$$

7. Potential Bias Rate

$$dd = \frac{(rn)(nr)}{rr}$$

Then

$$bb = \frac{nn - dd}{m}$$

can be considered the potential bias rate

7. Application Experiences

7. Application Experiences

- UI Nonprofit Nonresponse Rate = 27 %
- UI Nonprofit Potential Bias Rate = 11 %

7. Application Experiences

- UI Nonprofit Nonresponse Rate = 27 %
- UI Nonprofit Potential Bias Rate = 11 %
- NSAF Nonresponse Rate = 38 %
- NSAF Potential Bias Rate = 15 %

7. Application Experiences

- UI Nonprofit Nonresponse Rate = 27 %
- UI Nonprofit Potential Bias Rate = 11 %
- NSAF Nonresponse Rate = 38 %
- NSAF Potential Bias Rate = 15 %
- NORC Sales Tax Nonresponse Rate = 79 %
- NORC Sales Tax Potential Bias Rate = 28 %

Potential Next Steps

Potential Next Steps

- These Rules of Thumb obviously will not work everywhere but have been good starting points in many practical settings

Potential Next Steps

- These Rules of Thumb obviously will not work everywhere but have been good starting points in many practical settings
- Some reallocation of resources in survey execution and adjustment seems worthy of consideration in such settings

Potential Next Steps

- These Rules of Thumb obviously will not work everywhere but have been good starting points in many practical settings
- Some reallocation of resources in survey execution and adjustment seems worthy of consideration in such settings
- Estimating the missingness mixture fractions is key here. One way to do this has been proposed but more are needed

Provide to End Users and Future Survey Designers

Provide to End Users and Future Survey Designers

- Metadata/Paradata should include all relevant circumstances of the interview

Provide to End Users and Future Survey Designers

- Metadata/Paradata should include all relevant circumstances of the interview
- For example number of Calls, who was pesent, Language used, Length of interview, Data on case from Frame, etc

Provide to End Users and Future Survey Designers

- Metadata/Paradata should include all relevant circumstances of the interview
- For example number of Calls, who was pesent, Language used, Length of interview, Data on case from Frame, etc
- Nonrespondent cases should also go forward s a final deliverable, with details like point of Breakoff, Frame characteristics, number and conditions around Attempts, etc

Kish-Hess Data Files

Kish-Hess Data Files

- Employ Kish-Hess idea of consciously reusing nonrespondents but expand

Kish-Hess Data Files

- Employ Kish-Hess idea of consciously reusing nonrespondents but expand
- Keep safe identifiers of both respondents and nonrespondents for possible later use

Kish-Hess Data Files

- Employ Kish-Hess idea of consciously reusing nonrespondents but expand
- Keep safe identifiers of both respondents and nonrespondents for possible later use
- For bias reduction/analyses, as originally proposed, or for potential bias rate calculations, as described here

Kish-Hess Data Files

- Employ Kish-Hess idea of consciously reusing nonrespondents but expand
- Keep safe identifiers of both respondents and nonrespondents for possible later use
- For bias reduction/analyses, as originally proposed, or for potential bias rate calculations, as described here
- Could greatly lower long-term costs, while improving quality and enhancing credibility

Reanalysis Opportunities

Reanalysis Opportunities

- Consider public archiving of the augmented files

Reanalysis Opportunities

- Consider public archiving of the augmented files
- Revisit the augmentation regularly, as our understanding of what are the key paradata items can grow or change

Reanalysis Opportunities

- Consider public archiving of the augmented files
- Revisit the augmentation regularly, as our understanding of what are the key paradata items can grow or change
- Examine whether partnerships across organizations would have merit in measuring potential bias rates

Unintended Consequences and Complications

Unintended Consequences and Complications

- Look at statements made to potential respondents, should a sample of them be purposely returned

Unintended Consequences and Complications

- Look at statements made to potential respondents, should a sample of them be purposely returned
- Examine confidentiality issues that might arise even if former response status of a case was to be disclosed

Unintended Consequences and Complications

- Look at statements made to potential respondents, should a sample of them be purposely returned
- Examine confidentiality issues that might arise even if former response status of a case was to be disclosed
- Run cognitive tests of these ideas, for potential respondents and clients

THANK YOU...