Statistical Methods, part 1

Module 2: Latent Class Analysis of Survey Error

Models for measurement errors, *continued*

Dan Hedlin
Stockholm University
November 2012

Acknowledgement: **Paul Biemer**

Stockholm University

Regression coefficients in the presence of measurement errors

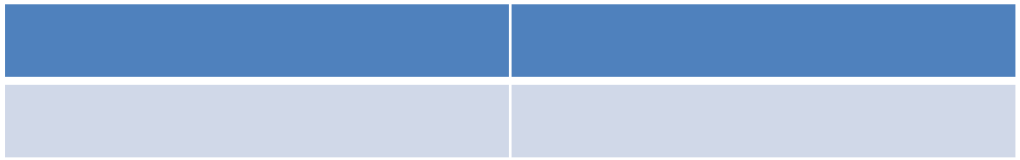
- $Var(\bar{y}) = \frac{1}{R}\frac{\sigma^2}{n}\tau$
- 'Effective sample size' is not *n*, it is *Rn*
- Recall that the estimated variance is largely ok, despite the true variance is enlarged by measurement errors
- In regression analysis measurement errors in y will enlarge residuals but not introduce bias

Bias in regression

- Measurement error in x will:
  - Attenuate slope
  - Attenuate correlation
  - Make intercept larger, if all variables are non-negative and slope positive
- Often hard to guess if measurement error affects analysis a lot or only modestly
- Sometimes hard even to guess direction, let alone how severely measurement errors influence estimates

From now on focus on dichotomous variables

- Y is either 1 or 0
- Interested in proportions
- True score is prob(y_i=1) (i.e. = E(y_i))
- m=2 yields a 2x2 table:



- Gross error and net error

Shortcut Computation

		Reinterview (y_2) 1	0	
Interview (y_1)	1	a	b	
	0	c	d	
				$n = a+b+c+d$

$$\frac{\sum_{i=1}^{n} (y_{i1} - y_{i2})^2}{n} = \frac{b+c}{n} = g \qquad \text{("Gross difference rate")}$$

$$\boxed{s_2^2 = \frac{b+c}{2n} = \frac{g}{2}}$$

Inconsistency (or "*Un*reliability") Ratio

Thus, when $\frac{(N-1)}{N} \approx 1$,

Inconsistency Ratio is

$$I = \frac{SRV}{SV + SRV}$$

Re liability Ratio

$$R = 1 - I$$

Notes on I:

1. Estimation of I or R requires two measurements or observations from the response distribution
2. $0 < I < 1$
   small I (or R is large) ==> small measurement error
   large I (or R is small) ==> large measurement error
3. U.S. Census Bureau "Rule of Thumb"
   $0.0 \leq I < 0.2$ ==> good reliability ($R \geq .8$)
   $0.2 \leq I < 0.5$ ==> moderate reliability ($.5 \leq R < .8$)
   $0.5 \leq I < 1.0$ ==> poor reliability ($0 \leq R < .5$)

U.S. Bureau of the Census (1985)

Estimation of *I* from Reinterview

Recall

$$SRV = \frac{b+c}{2n} = \frac{g}{2}$$

$$SV + SRV = p_1 q_1 \text{ or } p_2 q_2$$

$$\text{or } \frac{p_1 q_1 + p_2 q_2}{2} \text{ or } \frac{p_1 q_2 + p_2 q_1}{2}$$

Therefore, we can estimate I (or R) in a number of ways:

$$\hat{I}_1 = \frac{g}{2p_1q_1}$$

or

$$\hat{I}_2 = \frac{g}{p_1q_1 + p_2q_2}$$

or

$$\hat{I}_3 = \frac{g}{p_1q_2 + p_2q_1}$$

Best (Census Bureau 1985)

where $g = \frac{b+c}{n}$ (gross difference rate)

The estimate of I denoted by $\hat{I}_3$ is called the "index of inconsistency"

It can be shown (see Hess, Singer, and Bushery, 1999) that $\kappa = 1 - \hat{I}_3$ is identical to Cohen's kappa statistic

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

where $P_0 = \frac{a+d}{n}$ and $P_e = p_1 p_2 + q_1 q_2$

Why is this result remarkable?

NOTE:

$$E(g) = E\left[\frac{\sum_{i=1}^{n}(y_{i1} - y_{i2})^2}{n} \mid s\right]$$

$$= \text{Var}\, y_{i1} \mid s) + \text{Var}\, y_{i2} \mid s) - 2\text{Cov}(y_{i1}, y_{i2} \mid s)$$

$$\text{Thus, if } \text{Cov}(y_{i1}, y_{i2} \mid s) < 0, \text{ then } \frac{g}{2} \text{ will overestimate } SRV.$$

$$\text{Further, if } \text{Cov}(y_{i1}, y_{i2} \mid s) > 0, \frac{g}{2} \text{ will underestimate } SRV.$$

$$\text{What might lead to } \text{Cov}(y_{i1}, y_{i2} \mid s) > 0 ? \text{Cov}(y_{i1}, y_{i2} \mid s) < 0 ?$$

The Classification Probability Model

(Bross 1954, Biemer and Stokes, 1991)

Next, we consider a better specified model for measurement error in categorical data, referred to as the *classification probability error model*.

Recall that

$E_j(\bullet \mid s)$ = expectation across the "columns" of responses for a given unit on the row of the response probability matrix.

$E_s E_j(\square \mid i)$ = expectation across "rows" or all possible selections of units or samples of units from the rows.

Terminology and notation

- ϕ_i referred to as probability of a false positive
- $1-\phi_i$ referred to as "specificity"
- θ_i is probability of a false negative
- $1-\theta_i$ referred to as "sensitivity"
- $\pi = P(\mu_i = 1) =$ true population proportion

$$P_i = E_j(y_{ij} \mid s) = \Pr(y_{ij} = 1 \mid i)$$

$$\mu_i = \text{true value}$$

We now extend that notation

Again, consider

$$y_{ij} = \begin{cases} 1 \text{ if } yes \text{ or "+"} \\ 0 \text{ if } no \text{ or "–"} \end{cases}$$

”phi” = p for positive

$P(y_{ij} = 1 \mid \mu_i = 0) = \phi_i$ false positive probability

$P(y_{ij} = 0 \mid \mu_i = 1) = \theta_i$ false negative probability

Thus, we have the following for unit *i*

y_{ij}	μ_i	$P(y_{ij} \mid \mu_i)$	
0	0	$1-\phi_i$	specificity
1	0	ϕ_i	false positive probability
0	1	θ_i	false negative probability
1	1	$1-\theta_i$	sensitivity

$$E_j(y_{ij} \mid s) = P(y_{ij} = 1 \mid s)$$

$$= P(y_{ij} = 1 \mid \mu_i = 1) P(\mu_i = 1) + P(y_{ij} = 1 \mid \mu_i = 0)\, P(\mu_i = 0)$$

$$= (1-\theta_i)\mu_i + \phi_i(1-\mu_i)$$

Let $E_s(\theta_i) = \theta$

$E_s(\phi_i) = \phi$

Then

$$E(y_{ij}) = E_s E_j(y_{ij} \mid s)$$

$$= E_s[1-\theta_i)\pi + \phi_i(1-\pi)]$$

$$= (1-\theta)\pi + \phi(1-\pi)$$

Bias of *p*

From these results,

$$\text{Bias}(y_{ij}) = E(y_{ij}) - \pi = (1-\theta)\pi + \phi(1-\pi) - \pi$$

$$\boxed{= -\theta\pi + \phi(1-\pi)}$$

$$\text{Let } p = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$\text{Bias}(p) = \sum_{i=1}^{n} \frac{[E(y_i) - \pi]}{n} \qquad \boxed{= -\theta\pi + \phi(1-\pi)}$$

When is the bias 0?

Bias = 0 if and only if either:

$$\theta\pi = \phi(1-\pi)$$

or

$$\theta = \phi = 0$$

Variance of *p*

- Note that P_i from Census Model can be written as

$$P_i = (1-\theta_i) \text{ if } \mu_i = 1$$
$$= \phi_i \quad \text{if } \mu_i = 0$$

or

$$P_i = (1-\theta_i)\mu_i + \phi_i(1-\mu)$$

Variance of *p* Under the Classification Probability Model

Assumptions: Same as Census Bureau Model
Recall, under Census Bureau Model for $m = 1$

$$\mathrm{Var}(p) = (1 - f)\frac{S_1^2}{n} + \frac{S_2^2}{n}$$

where

$$S_1^2 = \sum_{i=1}^{N} \frac{(P_i - P)^2}{N - 1}$$

$$S_2^2 = \sum_{i=1}^{N} \frac{P_i Q}{N}$$

Rewriting *SV* and *SRV*

Let's find S_1^2 (or *SV*) and S_2^2 (or *SRV*) Under this Model

$$1.\ S_1^2 = \sum_{i=1}^{N} \frac{(P_i - P)^2}{N-1}$$

$$P_i = (1-\theta_i)\,\mu_i + \phi_i(1-\mu_i)$$

$$P = (1-\theta)\,\pi + \phi(1-\pi)$$

Rewriting *SV* and *SRV* (cont'd)

$$\theta = \frac{1}{N_1} \sum_{i=1}^{N_1} \theta_i, N_1 = \sum_{i=1}^{N} \mu_i$$

$$\phi = \frac{1}{N_0} \sum_{i=1}^{N} \phi_i, N_0 = N - N_1$$

$$\pi = \frac{1}{N} \sum_{i=1}^{N} \mu_i$$

Rewriting *SV* and *SRV* (cont'd)

Therefore, after some algebra

$$\left(\frac{N-1}{N}\right)SV = \pi(1-\pi)(1-\theta-\phi)^2 + \gamma_{\theta\phi}$$

where

$$\gamma_{\theta\phi} = \pi\sigma_\theta^2 + (1-\pi)\sigma_\phi^2$$

(Note: Often, $\gamma_{\theta\phi}$ is assumed to be negligible.)

Rewriting *SV* and *SRV* (cont'd)

$$2.\ S_2^2 = \frac{1}{N}\sum_{i=1}^{N} P_i Q_i = SRV$$

Where

$$P_i = (1-\theta_i)\mu_i + \phi_i(1-\mu_i)$$

After some algebra,

$$\boxed{SRV = \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) - \gamma_{\theta\phi}}$$

Remarks:

1. If f can be ignored,

$$
\begin{aligned}
nVar(p) &= SV + SRV \\
&= \pi(1-\pi)(1-\theta-\phi)^2 \\
&\quad + \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi)
\end{aligned}
$$

2. This is useful for studying how false +'s and false –'s affect variance

Remarks: (cont'd)

3. *R* (and *I*) vary as π varies.
4. When π is very small even a small false positive error rate can be important.

Eg. $\pi = .01$ $\theta = 0$

$\phi = .01$ $\gamma_{\theta\phi} = 0$

Then *I* = .50 or 50%

Summary

$$\left(\frac{N-1}{N}\right)SV = \pi(1-\pi)(1-\theta-\phi)^2 + \gamma_{\theta\phi}$$

$$SRV = \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) - \gamma_{\theta\phi}$$

$$I = \frac{\pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) - \gamma_{\theta\phi}}{PQ}$$

This expression illustrates the difficulty of interpreting *I* for categorical data.

Estimation of *SRV* and *I*

Note: As we showed before, $\frac{g}{2} = \frac{b + c}{2n}$ is unbiased

for *SRV* provided assumptions are satisfied.

Recall the key assumptions are:

- Independence of classification errors
- Equal classification probabilities across trials or repeated measurements

Estimation of Bias

Suppose reinterview is truth. This is usually assumed for reconciled reinterview

		Gold Standard	
		1	0
Interview	1	a	b
	0	c	d

Est. of bias is

$$p_1 - p_2 = \frac{a+b}{n} - \frac{a+c}{n} = \boxed{\frac{b-c}{n}}$$

net difference rate

Estimation of False Positive and False Negatives

		Gold Standard	
		1	0
Interview	1	a	b
	0	c	d

$$\hat{\phi} = \frac{b}{b+d} \quad \text{estimates } \phi, \text{ false positive}$$

$$\hat{\theta} = \frac{c}{a+c} \quad \text{estimates } \theta, \text{false negative}$$

Example

		Reconciled Reinterview: Unemp	Reconciled Reinterview: Other
Interview	Unemp	200	20
	Other	6	3000

$n = 3226$

$$Bias = \frac{20-6}{3226} = .004$$

$$\hat{\phi} = \frac{20}{3020} = .007$$

$$\hat{\theta} = \frac{6}{206} = .029$$

Problems with Reconciled Reinterview

1. Assumes agreements are correct.
2. Assumes disagreements can be reconcile accurately.
3. Biemer & Forsman; Sinclair & Gastwirth show this is not true for many surveys.

See Forsman and Schreiner (1991) for a good discussion of the issues with reinterview

Question

- For test-retest reinterview, what does a significant net difference rate indicate?

- How should the net difference rate be interpreted if the second measurement is from a “preferred” survey process?