2005 JSM Presentation

Bayesian Models for Adjusting Response Bias in Survey Data: An Example in Estimating Rape and Domestic Violence from the NCVS

Qingzhao Yu

Elizabeth A. Stasny

Statistics Department

The Ohio State University

Aug. 10, 2005

MOTIVATIONS

- Survey data can be biased as the consequence of some known factors. If we take these factors into account, the results of the analysis of survey data will be more reliable.
- Lots of surveys collect data every year, for example, panel surveys. Using the data collected previously, we can take advantages of more information.

2005 JSM Presentation

The National Crime Victimization Survey (NCVS)

- Administered by the Census Bureau for the Bureau of Justice Statistics
- Supplements Police Data-Learn about unreported crimes
- The survey categorizes crimes as “personal” or “property” crime

Research Plan

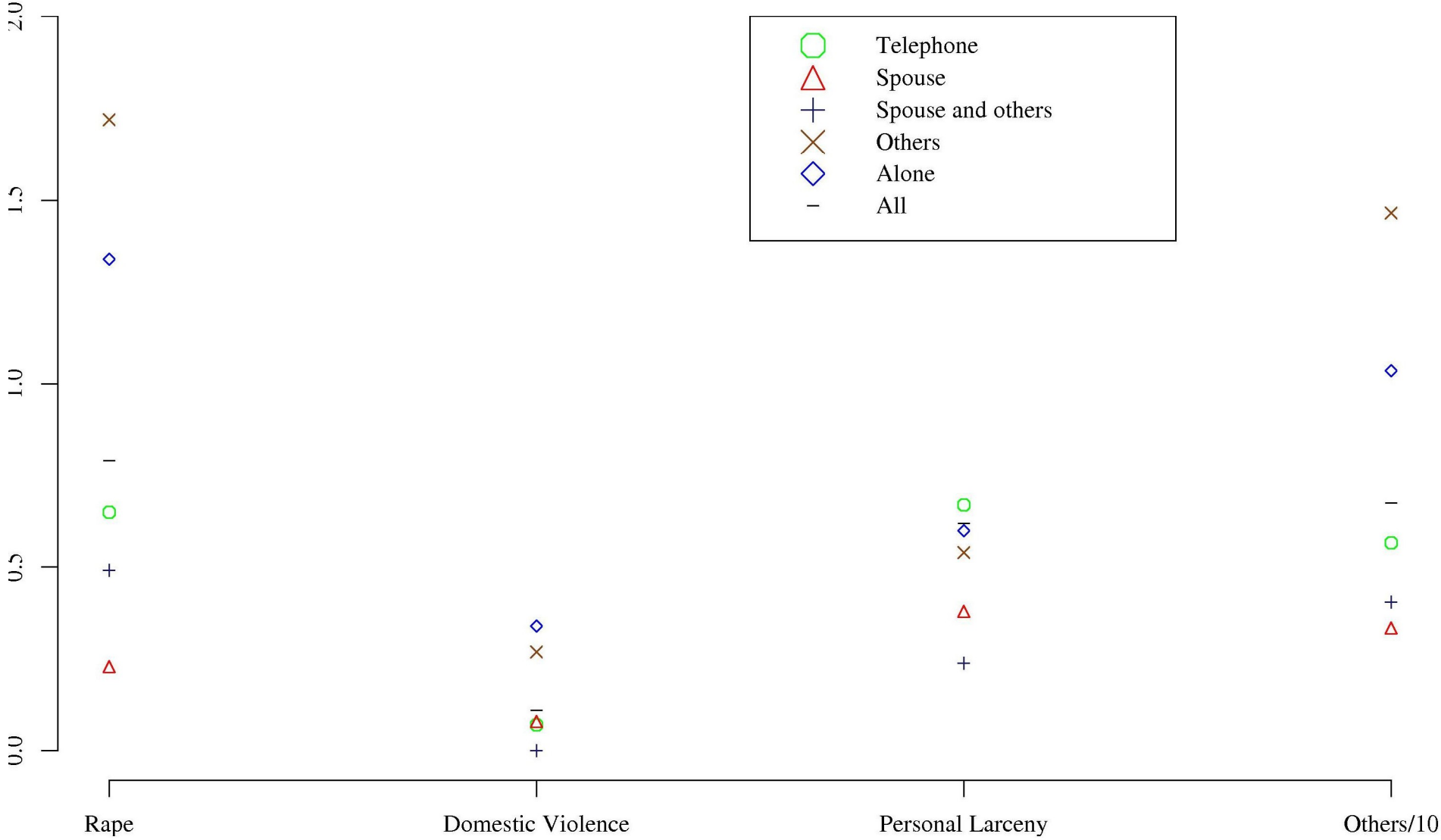
- Use 1998 to 2002 NCVS data to estimate the rape rate, domestic violence rate and other crime rates.
- Build Bayesian model, use data from 1993 to 1997 as prior information.
- Account for the Response Biases.

Survey Data

TABLE 1. Frequencies and Rates of Crimes Reported by Settings of the Interviews: NCVS 1998-2002

Type of Interview			Numbers of Incidents Reported by Type of Personal Crimes (Rates per 1000 Interviews)				
	Who was Present During Interview	Number of Interviews	Rape	Domestic Violence	Other Assault	Personal Larceny	No Crime Reported
Telephone	Unknown	292984	189 (0.65)	21 (0.07)	1660 (5.67)	192 (0.67)	290922 (992.96)
Personal	Spouse	13115	3 (0.23)	1 (0.08)	44 (3.35)	5 (0.38)	13062 (995.96)
	Spouse and Other	8178	4 (0.49)	0 (0)	33 (4.04)	2 (0.24)	8139 (995.23)
	Other	29616	51 (1.72)	8 (0.27)	434 (14.65)	16 (0.54)	29107 (982.81)
	Alone	41680	56 (1.34)	14 (0.34)	432 (10.36)	25 (0.60)	41153 (987.35)
All Personal		92589	114 (1.23)	23 (0.25)	943 (10.18)	48 (0.52)	91461 (987.82)
All Interviews		385573	303 (0.79)	44 (0.11)	2603 (6.75)	240 (0.62)	382383 (991.73)

Figue 1: Rates of Various Crimes by Type of Interview



Note: For the other assaults, we use the crime rate per 100 person to make the scale fit.

THE OHIO STATE UNIVERSITY

2005 JSM Presentation

A MODEL FOR RESPONSE BIAS ADJUSTMENT

Observed Data

	Personal Interview		Telephone Interview
	Spouse is present	Spouse is not Present	
Rape	7	107	189
Domestic violence	1	22	21
Other Crimes	84	907	14534
No Crime	21201	70260	290922

Assumptions

- The only reasons that a crime was not reported are that a spouse was present, or that the interview was conducted over the telephone.
- Spouse's presence has influence only on the reporting of rape and domestic violence.
- The presence of a spouse dominates the use of a telephone interview in determining whether or not a woman reports such an incident.

Notations

- ω_{ij} = probability of crime status i and interview status j
  - i = 1 if rape, 2 if domestic violence, 3 if other crime, 4 if no crime
  - j = 1 if spouse is present, 2 if spouse is not present
  - summation of ω_{ij} is 1
- π = probability of a telephone interview
- $1 - \tau$ = probability of crimes not reported because of telephone interview
- $1 - \rho$ = probability of rape not reported because spouse is present
- $1 - \delta$ = probability of domestic violence not reported because spouse is present

Probabilities Underlying Unobserved Complete Data

Personal Interviews

		Spouse is Present	Spouse is not Present
Rape	Reported	$(1-\pi)\rho\omega_{11}$	$(1-\pi)\ \omega_{21}$
	Not Reported - Spouse Present	$(1-\pi)(1-\rho)\omega_{11}$	-
Domestic Violence	Reported	$(1-\pi)\delta\omega_{12}$	$(1-\pi)\omega_{22}$
	Not Reported - Spouse Present	$(1-\pi)(1-\delta)\omega_{12}$	-
Other Crime	Reported	$(1-\pi)\ \omega_{13}$	$(1-\pi)\ \omega_{23}$
No Crime	Reported	$(1-\pi)\ \omega_{14}$	$(1-\pi)\ \omega_{24}$

Probabilities Underlying Unobserved Complete Data (Cont.)

Telephone Interviews

		Spouse is Present	Spouse is not Present
Rape	Reported	$\pi\tau\rho\omega_{11}$	$\pi\tau\omega_{21}$
	Not Reported - Spouse Present	$\pi(1-\rho)\ \omega_{11}$	-
	Not Reported - Phone Interview	$\pi\rho(1-\tau)\omega_{11}$	$\pi(1-\tau)\omega_{21}$
Domestic Violence	Reported	$\pi\tau\delta\omega_{12}$	$\pi\tau\omega_{22}$
	Not Reported - Spouse Present	$\pi(1-\delta)\omega_{12}$	-
	Not Reported - Phone Interview	$\pi\delta(1-\tau)\omega_{12}$	$\pi(1-\tau)\omega_{22}$
Other Crime	Reported	$\pi\tau\omega_{13}$	$\pi\tau\omega_{23}$
	Not Reported - Phone Interview	$\pi(1-\tau)\omega_{13}$	$\pi(1-\tau)\omega_{23}$
No Crime	Reported	$\pi\omega_{14}$	$\pi\omega_{24}$

Probabilities Underlying the Observed Data

Personal Interviews

	Spouse is Present	Spouse is not Present
Rape Reported	$(1-\pi)\rho\omega_{11}$	$(1-\pi)\omega_{21}$
Domestic Violence Reported	$(1-\pi)\delta\omega_{12}$	$(1-\pi)\omega_{22}$
Other Crime Reported	$(1-\pi)\omega_{13}$	$(1-\pi)\omega_{23}$
No Crime Reported	$(1-\pi)(1-\rho)\omega_{11}+(1-\pi)(1-\delta)\omega_{12}+(1-\pi)\omega_{14}$	$(1-\pi)\omega_{24}$

Telephone Interviews

Rape Reported	$\pi\tau\rho\omega_{11}+\pi\tau\omega_{21}$
Domestic Violence Reported	$\pi\tau\delta\omega_{12}+\pi\tau\omega_{22}$
Other Crime Reported	$\pi\tau\omega_{13}+\pi\tau\omega_{23}$
No Crime Reported	$\pi(1-\rho)\ \omega_{11}+\pi\rho(1-\tau)\omega_{11}+\pi(1-\tau)\omega_{21}+\pi(1-\delta)\omega_{12}+\pi\delta(1-\tau)\omega_{12}+\pi(1-\tau)\omega_{22}+\pi(1-\tau)\omega_{13}+\pi(1-\tau)\omega_{23}+\pi\omega_{14}+\pi\omega_{24}$

2005 JSM Presentation

BAYESIAN INFERENCE FOR THE BIAS-ADJUSTING MODEL

Purpose

- Goal: To estimate the parameters π, τ, ρ, δ, and ω based on observations y in the model described above.
- Adopt a Bayesian view point and treat the parameters τ, ρ, δ, and ω as random variables.
- Use prior information provided by the NCVS data from 1993 to 1997.
- We wouldn't estimate π in this way since it is a fixed value and cannot be influenced by the prior survey.

Prior Distribution

- All parameters to be estimated are probabilities. Choose Dirichlet distributions as prior distributions.
- X=(x1,…,xt): multinomial (N, P=(p1,…,pt))

  Prior Distribution for P: Dirichlet(β_1,...,β_t)

  Posterior distribution for P: Dirichlet(B+X)

  where B+X=(β1+x1,…, βt+xt).
- Use the squared error loss, Bayesian estimator for P: E(P|K,λ,X)= (X/N)N/(N+K)+λK/(N+K). Where K= β_1+…+β_t and λ_i=β_i/K.

Prior Parameters(1)

- $E(P_i|K,\lambda)=\lambda_i$.
- Use the estimators for the parameters from 1993 to 1997 NCVS data as λ.
- Get the estimators using the model described early and the method described by Stasny and Coker,1997.

Prior Parameters(2)

- K denotes how much the estimator depends on the prior information.
- We intend to use the pseudo-Bayesian method for K.
- Detailed explanation and assessment of this method can be found in Bishop, Fienberg, and Holland, 1975

Prior Parameters (3)

- phat: Bayesian estimator for P
- Risk function:

$R(phat,p)=(N/N+K)^2(1-||P||^2)+(K/N+K)^2N||P-\lambda||^2$

- $K=(1-||P||)^2/||p-\lambda||^2$ minimizes R(phat,p).
- Use the MLE for P in the function, we get the estimated optimal value for K.
- Use the optimal K in the prior distribution, we get the Bayesian estimators for P.

An EB Algorithm for the Pseudo-Bayesian Model

- EB algorithm is a variant of the EM algorithm (see, for example, Dempster, Laird, and Rubin, 1977).
- We change the M-step into B-step, where we use the prior information to get the posterior distribution of the parameters and get the Bayesian estimators for the parameters.
- It can be proved that under general conditions, the EB algorithm will generate estimators converge to the Bayesian estimators.

Likelihood

$$\pi^{y2+++}(1-\pi)^{y1+++}$$
$$\times \rho^{y1111+y2111}(1-\rho)^{y1121+y2121}$$
$$\times \delta^{y1211+y2211+y2231}(1-\delta)^{y1221+y2221}$$
$$\times \tau^{y2111+y2112+y2211+y2212+y2311+y2312}$$
$$\times (1-\tau)^{y2131+y2132+y2231+y2232+y2331+y2332}$$
$$\times \omega_{11}^{y1111+y1121+y2111+y2121+y2131}$$
$$\times \omega_{12}^{y1211+y1221+y2211+y2221+y2231}$$
$$\times \omega_{13}^{y1311+y2311+y2331} \times \omega_{14}^{y1411+y2411}$$
$$\times \omega_{21}^{y1112+y2112+y2132} \times \omega_{22}^{y1212+y2212+y2232}$$
$$\times \omega_{23}^{y1312+y2312+y2332} \times \omega_{24}^{y1412+y2412}$$
$$\equiv (1-\pi)^{y1+++} \times \pi^{y2+++} \times \rho^{a1} \times (1-\rho)^{a2} \times \delta^{b1} \times (1-\delta)^{b2} \times \tau^{c1}$$
$$\times (1-\tau)^{c2} \times \{\prod_{i=1}^{4}\prod_{j=1}^{2} \omega_{ij}^{y+i+j}\} \quad (3)$$

Estimates(1)

The closed form MLEs for these parameters are as follows:

$\hat{\pi}$ = $y_{2+++} / (y_{1+++} + y_{2+++})$
$\hat{\rho}$ = $a_1 / (a_1 + a_2)$
$\hat{\delta}$ = $b_1 / (b_1 + b_2)$
$\hat{\tau}$ = $c_1 / (c_1 + c_2)$
$\hat{\omega}_{ij}$ = y_{+i+j} / y_{++++}

where a_1, a_2, b_1, b_2, c_1, and c_2 are as defined by equation (3). Then use the formula (1) and (2), we get the pseudo-Bayesian estimators τ^*, ρ^*, δ^*, and ω^*. For example,

$$\hat{K}_\rho = 2a_1a_2 /(a_1^2 + a_2^2 - 2*(a_1 + a_2)(a_1\lambda + a_2(1-\lambda)) + (a_1 + a_2)^2(\lambda^2 + (1-\lambda)^2))$$
$$\rho^* = (a_1 + a_2)/(a_1 + a_2 + \hat{K}_\rho) \times a_1/(a_1 + a_2) + \hat{K}_\rho/(N + \hat{K}_\rho)\lambda_\rho.$$

Where λ_ρ is the estimated value of ρ using the 93 to 97 NCVS data.

Estimates(2)

- The E- and B-steps of the EB-algorithm are repeated until parameter estimates converging to the desired degree of accuracy.
- In our case when the sum of the relative differences of all estimated probabilities between two iterations is less than 0.0001.
- Convergence occurred in about 80 iterations for all our applications.

Result(1)

Estimates for Crimes

$\hat{\omega}_{ij}$

	Spouse is present	Spouse is not present
Rape	0.000343	0.001154
Domestic Violence	0.000050	0.000169
Other Crime	0.002531	0.008506
No Crime	0.226360	0.760887

$\hat{\pi} = 0.76$ $\rho^* = 0.07$ $\delta^* = 0.05$ $\tau^* = 0.56$

Result(2)

Estimates for Crimes (Not Using the Prior information)

$\hat{\omega}_{ij}$

	Spouse is present	Spouse is not present
Rape	0.000337	0.001128
Domestic Violence	0.000048	0.000161
Other Crime	0.002499	0.008368
No Crime	0.227119	0.760340

$\hat{\pi} = 0.76 \quad \rho^* = 0.09 \quad \delta^* = 0.15 \quad \tau^* = 0.58$

Conclusions

- We have shown that estimated rates of rape and domestic violence among women are increased under a model that allows for "gag" factors in reporting such crimes based on the type of interview and who is present for the interview.
- Type of interviews and who is present during interview may have different influence on different women.

Future Research

- To account for characteristics of women in the model for reporting rapes and domestic violence.
- To account for the potential correlation in responses from the same woman over time.
- Similar models may be useful in other survey sampling settings where some known factors may result in response bias.

References

- Berger, J.O., (1985). *Statistical Decision Theory and Bayesian Analysis*, New York: Springer.
- Bishop, Y. M. M., Fienberg S. E., and Holland, P. W., (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.
- Dempster, A. P., Laird, N. M., and Rubin, D. B., (1977). "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, **39**, 1-38.
- Little, R.J. and Rubin, D.B., (2002). *Statistical Analysis with Missing Data*, New York: Wiley.
- Stasny, E. A. and Coker, A. L. (1997), *Adjusting the National Crime Victimization Survey's Estimates of Rape and Domestic Violence for "Gag" Factors in Reporting,* Technical Report #592, Department of Statistics, Ohio State University.