

# Missing Covariate Data in Matched Case-Control Studies

Department of Statistics  
North Carolina State University

Paul Rathouz  
Dept. of Health Studies  
U. of Chicago  
prathouz@health.bsd.uchicago.edu

with

Glen A. Satten  
Centers for Disease Control and Prevention

Raymond J. Carroll  
Texas A& M University

**October 15, 2004**

## General Framework

### “Highly-stratified” or “Clustered” Binary Data

**Observations:**  $j = 1, \dots, n$  within **stratum**  $i$

**Strata:** (many!)

- matched set  $i$  of case-control data
- multiple subjects in cluster  $i$  (prospective)
- longitudinal observations on subject  $i$

**Response:**  $D_{ij}$  (binary disease status)

**Covariates:**  $Z_{ij}$  (vector)

## Logistic Regression Model for Stratified Data

- For observation  $j$ , stratum  $i$ , define the **odds** that  $D_{ij} = 1$ :

$$\theta(Z_{ij}) = \frac{\Pr(D_{ij} = 1|Z_{ij})}{\Pr(D_{ij} = 0|Z_{ij})}$$

- And let

$$\theta(Z_{ij}) = \exp(q_i + \beta' Z_{ij})$$

—→ **conditional logistic regression** model

- Stratum-level intercept  $q_i$  is a **nuisance**

↖  
“fixed effect”

## Conditional Likelihood for Nuisance Intercept Model

- Model:  $\theta = \exp(q_i + \beta' Z_{ij})$
- Data for stratum  $i$ :

$$\begin{aligned}\mathbf{D}_i &= (D_{i1}, \dots, D_{in})' \\ \mathbf{Z}_i &= (Z_{i1}, \dots, Z_{in})'\end{aligned}$$

- Stratum-level likelihood:  $L_i(\beta, q_i) = \Pr(\mathbf{D}_i | \mathbf{Z}_i)$   
can be written

$$L_i = \underbrace{\Pr(\mathbf{D}_i | \sum_j D_{ij}, \mathbf{Z}_i; \beta)}_{(1)} \underbrace{\Pr(\sum_j D_{ij} | \mathbf{Z}_i; \beta, q_i)}_{(2)}$$

- Important:
  - $\sum_j D_{ij}$  is a CSS for  $q_i \Rightarrow$  no  $q_i$  in (1)
- Define **conditional likelihood** for  $(\beta)$

$$L_i^c(\beta) = \Pr(\mathbf{D}_i | \sum_j D_{ij}, \mathbf{Z}_i; \beta)$$

→ **conditional logistic regression likelihood**

## What happens when some covariates may be missing?

- **Covariates:**

$X_{ij}$  ← some may be missing

$Z_{ij}$  ← always observed

- **Missing covariate indicator:**

$$R_{ij} = I(X_{ij} \text{ observed})$$

- Odds that  $D_{ij} = 1$ :

$$\theta(X_{ij}, Z_{ij}) = \exp(q_i + \beta'_z Z_{ij} + \beta'_x X_{ij})$$

- Interest on the effects of covariates given by

$$\beta = (\beta'_z, \beta'_x)'$$

- **Missing at random (MAR) assumption**

$$R_{ij} \perp\!\!\!\perp X_{ij} \mid D_{ij}, Z_{ij}$$

allows identification of  $\beta$

## Missing Data Example

### Matched case-control study of hip fracture

- 118 female hip fracture patients (cases) in Beijing, China (Huo, Lauderdale and Li)
- 2 controls per case matched on neighborhood and age (within 5 years)
- **Of interest:** “whether reproductive factors were related to risk of hip fractures in Chinese women aged 50 years and older.”
- Focus on effects ( $Z_{ij}$ 's) of:
  - **parity** (per child)
  - **breastfeeding** (average months per child)
- Important adjustors ( $X_{ij}$ 's) include:
  - **height** (surrogate for hip axis length)
  - **BMI** (a well-established risk factor)

Height and weight self-reported and hence may be missing

## Missing $X$ in CLR Model

### Three “Tricky” Features

1. Three sets of nuisance parameters to manage

- Nuisance intercept  $q_i$  in model
- Distribution of missing  $X$ :

$$\Pr(X_{ij}|Z_{ij}; \alpha)$$

- Missingness model:

$$\Pr(R_{ij}|D_{ij}, Z_{ij}; \gamma)$$

2. Loss of strata with complete record analysis

Hip Fracture Example:

- $X'_{ij}$ s **Height** and **BMI** missing for 52 (15%) subjects (52/354)
- Results in 24 (20%) matched sets dropped and 85 (24%) observations dropped (worse if only one control per case)

### 3. MAR is not “ignorable”:

- Likelihood (conditioning on  $Z$  implicit) :

$$L = \{\Pr(X|D) \Pr(R = 1|D) \Pr(D)\}^R \\ \times \{\Pr(R = 0|D) \Pr(D)\}^{1-R}$$

- Unconditional inference about  $\beta$ :

$$L = \{\Pr(D|X; \beta) \Pr(X; \alpha) \Pr(R = 1|D; \gamma)\}^R \\ \times \{\Pr(D; \beta, \alpha) \Pr(R = 0|D; \gamma)\}^{1-R}$$

$$\propto \{\Pr(D|X; \beta) \Pr(X; \alpha)\}^R \{\Pr(D; \beta, \alpha)\}^{1-R}$$



Note:

a valid likelihood but

no longer a valid pmf!

3. (cont.) MAR is not “ignorable”:

Conditional inference about  $\beta$  (given stratum):

- Temptation: Begin with

$$L = \prod_j \{\Pr(D_j|X_j; \beta) \Pr(X_j; \alpha)\}^{R_j} \{\Pr(D_j; \beta, \alpha)\}^{1-R_j}$$

and “condition on”  $\sum_j D_j$  for that stratum

- Problem:
  - $L$  not a valid probability mass function
  - conditioning does not make sense
  - $R_j$  is neither a conditioning statistic nor a random variable in  $L$

(with all these problems)  
Why use CLR anyway ?

- The odds model

$$\theta(X_{ij}, Z_{ij}) = \exp(q_i + \beta'_z Z_{ij} + \beta'_x X_{ij})$$

looks like a **random ( $q_i$ ) effects** model

- We are treating  $q_i$  as a **fixed effect**. Why?

↖  
nuisance intercept

- Retrospective data: **CLR** likelihood reflects **matched case-control sampling** design
- Prospective data (clustered or longitudinal):
  - \* no **distributional assumption** on  $q_i$
  - \* distribution  $Q_{\mathbf{Z}}$  of  $q_i$ : may **depend on**  
 $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in})$
  - \* controls for **stratum-level confounders**
  - \* each cluster acts as “own control”

## Advantage of CLR over Random Effects Models Simulated Example

- $n = 10,000$  matched pairs ( $j = 1, 2$ ) with model

$$\text{logit}\{\Pr(D_{ij} = 1|Z_{ij})\} = q_i + \beta Z_{ij}$$

where marginal  $\Pr(D_{ij} = 1) \approx 0.5$  and  $\beta = 0.5$

- **Case 1:**  $\text{corr}(q_i, Z_{ij}) = 0$   
CLR:  $\hat{\beta} = 0.52$   
RE:  $\hat{\beta} = 0.52$
- **Case 2:**  $\text{corr}(q_i, Z_{ij}) = 0.54$  ( $q_i$  a confounder)  
CLR:  $\hat{\beta} = 0.48$   
RE:  $\hat{\beta} = 0.85$

## Semiparametric Efficiency of Maximum Conditional Likelihood Estimator

- With data across strata  $i \dots$   
 $\dots$  obtain  $\hat{\beta}$  by maximizing

$$\prod_i L_i^c(\beta)$$

- Let  $q_i$  be **random** instead of fixed
- Let  $Q_{\mathbf{Z}}$  be the **non-parametric** distribution function of  $q_i$  which **may depend** on  $\mathbf{Z}$
- Semiparametric model in:  $( \underbrace{\beta}_{\uparrow}, \underbrace{Q_{\mathbf{Z}}}_{\uparrow} )$

parametric

non-  
parametric

- Then  $\hat{\beta}$  achieves Cramèr-Rao lower bound in presence of unknown  $Q_{\mathbf{Z}}$ 
  - Lindsay (1983) for fixed  $\mathbf{Z}_i = \mathbf{z}$  across  $i$
  - Extends to  $Q_{\mathbf{Z}}$  varying with  $\mathbf{Z}_i$  across  $i$
  - Key assumption:  $\sum_j D_{ij}$  is CSS for  $q_i$

## Missing $X$ in CLR Model

### Outline

- Complete record estimator
  - Bias correction by conditioning on observed missingness pattern
- Efficiency improvement
  - Elimination of ancillary information in missingness process via projection
  - Approximate projection avoids high-dimensional integral and need for exact distribution of  $X$
- Variation to problem of attrition in longitudinal analyses

## Notation

- Data for stratum  $i$ :

$$\mathbf{D}_i = (D_{i1}, \dots, D_{in})'$$

$$\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in})'$$

- For missing data, write

$$\mathbf{R}_i = (R_{i1}, \dots, R_{in})'$$

$$\mathbf{X}_i = (X_{i1}, \dots, X_{in})'$$

- Define

$$\mathbf{X}_{i,\text{obs}}, \mathbf{Z}_{i,\text{obs}}, \mathbf{D}_{i,\text{obs}}, \text{ etc.}$$

to be the observed rows of

$$\mathbf{X}_i, \mathbf{Z}_i, \mathbf{D}_i, \text{ etc.}$$

## Complete Record Estimation Exploiting a Missingness Model

- Delete records with missing  $X_j$  and model

$$(\mathbf{D}_{\text{obs}} | \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}})$$

as if this were the original data

- We **do not** need to model  $X_j$ , but ...  
... (selection) bias in  $\hat{\beta}$   
... inefficiency in  $\hat{\beta}$
- Bias correction by **conditioning** on the missingness process, modelling

$$\Pr(\mathbf{D}_{\text{obs}} | \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, \mathbf{R}) = \prod_{j=1}^n \Pr(D_j | R_j = 1, X_j, Z_j)^{R_j}$$

- Requires a **missingness model**

$$\Pr(R_j = 1 | D_j = d, X_j, Z_j) = \pi(d, Z_j; \gamma)$$

depends on **response** and **other covariates**

- Odds that  $D_j = 1$  when  $X_j$  **is observed**,

$$\theta^*(X_j, Z_j) = \frac{\Pr(D_j = 1 | R_j = 1, X_j, Z_j)}{\Pr(D_j = 0 | R_j = 1, X_j, Z_j)}$$

are just

$$\theta^*(X_j, Z_j) = \exp(q_i + \beta'_z Z_j + \beta_x^t X_j + B_j)$$

where

$$B_j = \log\{\pi(1, Z_j; \gamma) / \pi(0, Z_j; \gamma)\}$$

↑

case

↑

control

- $B_j$ 
  - does not contain  $\beta$  or  $q_i$
  - is just an **offset term**
  - depends on missingness parameter  $\gamma$

## Implications

- In the complete record likelihood

$$\Pr(\mathbf{D}_{\text{obs}} | \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, \mathbf{R}; \beta, \gamma, q_i)$$

$\sum_j R_j D_j$  is a CSS for  $q_i$

- The **complete-record conditional likelihood**

$$L_{\text{complete}}^c(\beta, \gamma) = \Pr(\mathbf{D}_{\text{obs}} | \underbrace{\sum_j R_j D_j}_{\uparrow}, \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, \mathbf{R})$$

is free of  $q_i$

# of cases among  
complete records

- Maximizing

$$\prod_i L_{i,\text{complete}}^c(\beta, \gamma)$$

will yield the **SPE** estimator for  $\beta$  among estimators **only** relying on complete records

- $L_{\text{complete}}^c(\beta, \alpha)$  written in terms of odds  $\theta^*$  is:

$$L_{\text{complete}}^c = \frac{\prod_j \theta^*(X_j, Z_j)^{R_j D_j}}{\sum_{\mathbf{d} \in \mathcal{D}^*} \prod_j \theta^*(X_j, Z_j)^{R_j d_j}}$$

where  $\mathcal{D}^* = \{\mathbf{d} : \sum_j R_j d_j = \sum_j R_j D_j\}$

$\mathcal{D}^*$  = All possible allocations of complete-data **cases** among complete data **records**

- **Notes:**

- $L_{\text{complete}}^c$  requires that missingness model  $\pi(\cdot, \cdot; \gamma)$  be **known** or **estimated**
- can use standard software for estimation via offset specification
- standard errors from standard software are **conservative** if  $\gamma$  estimated
- if  $\pi(d, Z_j; \gamma)$  only depends on **either**  $d$  or  $Z_j$ , naive complete case estimator is consistent
- Lipsitz, Parzen & Ewell, 1998

- **Example** (revisited)

## Hip Fracture Example

- Naive complete-record analysis  
(using 94 of 118 matched sets)

Coef.	Est.	SE	Z
(others)			
parity	0.066	0.132	0.50
<b>br feed (sd unit)</b>	<b>-0.469</b>	<b>0.244</b>	<b>-1.92</b>
bmi (sd unit)*	-1.142	0.237	-4.81
height (sd unit)*	-0.228	0.153	-1.49

\*possibly missing

(standard software)

- Non-missingness model ( $\pi(\cdot)$ )  
(data from all 354 observations)

	Coef.	Est.	SE	Z
(others)				
case		-0.64	0.35	-1.81
elem school		0.97	0.44	2.19
middle school		1.33	0.63	2.11
post 2nd sch		1.90	0.84	2.28
parity		-0.01	0.08	-0.11
br feed (sd unit)		0.20	0.17	1.18

(standard software – logistic regression)

- $\Pr(\text{BMI and Height missing})$  depends on some covariates (but not parity or breast feeding)
- non-missing  $\log(\text{OR; case vs. control}) = -0.64$

$B_j = \log\{\pi_j(1)/\pi_j(0)\}$  has mean  $-0.10 \pm .08$

which is not very severe

→ complete case analysis (approx) consistent

- Bias-corrected complete-record analysis

Coef.	Est.	SE	Z
(others)			
parity	0.066	0.132	0.50
<b>br feed (sd unit)</b>	<b>-0.485</b>	<b>0.245</b>	<b>-1.98</b>
bmi (sd unit)	-1.141	0.237	-4.81
height (sd unit)	-0.233	0.154	-1.52

(standard software with offset)

- Bias-corrected complete-record analysis **with** correct standard errors

Coef.	Est.	SE	Z
(others)			
parity	0.066	0.129	0.52
<b>br feed (sd unit)</b>	<b>-0.485</b>	<b>0.237</b>	<b>-2.05</b>
bmi (sd unit)	-1.141	0.234	-4.88
height (sd unit)	-0.233	0.149	-1.57

## Efficiency Improvement with $L_{\text{complete}}^c$

- Suppose all records are available
- Then  $\pi(D_j, Z_j; \gamma)$  can be **estimated** with likelihood

$$\prod_i \Pr(\mathbf{R}_i | \mathbf{D}_i, \mathbf{Z}_i; \gamma)$$

- $L_{\text{complete}}^c(\beta, \hat{\gamma})$  with **estimated**  $\hat{\gamma}$  is more efficient than  $L_{\text{complete}}^c(\beta, \gamma)$  with known  $\gamma$   
Why?
- Examine the full likelihood for stratum  $i$ :

$$p(\mathbf{X}_{\text{obs}} | \mathbf{D}, \mathbf{Z}; \beta) \underbrace{\Pr(\mathbf{R} | \mathbf{D}, \mathbf{Z}; \gamma)} \Pr(\mathbf{D} | \mathbf{Z}; \beta)$$

and note that the complete data likelihood is:

$$\Pr(\mathbf{D}_{\text{obs}} | \mathbf{X}_{\text{obs}}, \mathbf{Z}_{\text{obs}}, \mathbf{R}) = \prod_{j=1}^n \Pr(D_j | R_j = 1, X_j, Z_j)^{R_j}$$

wherein  $R_j$  is a random variable

- Heuristically,  $L_{\text{complete}}^c$  is **inefficient** because it contains **ancillary information** in

$$(\mathbf{R}|\mathbf{D}, \mathbf{Z})$$

- Estimation of  $\gamma$  removes (some) ancillary information **and** exploits information on records with missing  $X$
- **Projection** can further improve efficiency ...
- Define the score

$$U_{\text{complete}}^c = \frac{\partial \log L_{\text{complete}}^c}{\partial \beta}$$

- **Idea:** remove from  $U_{\text{complete}}^c$  the **projection\*** onto the space of functions of

$$(\mathbf{R}, \mathbf{D}, \mathbf{Z})$$

which are **unbiased** over  $\mathbf{R}$  conditional on

$$(\mathbf{D}, \mathbf{Z})$$

\* requires integration over  $\mathbf{X}$

- Define the projection of a given score  $g$ :

$$\text{Proj}(g) = E_{\mathbf{X}}(g|\mathbf{R}, \mathbf{D}, \mathbf{Z}) - E_{\mathbf{R}, \mathbf{X}}(g|\mathbf{D}, \mathbf{Z})$$

and an improved score for  $\beta$

$$U_*^c = U_{\text{complete}}^c - \text{Proj}(U_{\text{complete}}^c)$$

- **Notes:**

- $U_*^c$  is **doubly robust**

$\alpha$  incorrect **OR**  $\gamma$   
incorrect

- $\text{Proj}(U_{\text{complete}}^c)$  is (very) difficult to compute

- We employ an **approximate projection**

$$U_{\text{improved}}^c = U_{\text{complete}}^c - \text{Proj}_{\text{approx}}(U_{\text{complete}}^c)$$

exploiting a **working model** and **working integral** for

$$(X_j|D_j, Z_j; \alpha)$$

$\implies \hat{\beta}$  solving  $\sum_i U_{i,\text{improved}}^c = 0$  is consistent even if  $p$  is **wrong**

## Hip Fracture Example Efficiency Improvement

- Working model for BMI and height:
  - dichotomize BMI and height
  - 4-cell multinomial model for BMI × height
  - mean BMI and height in each category
- Bias-corrected complete-record analysis **with** efficiency improvement

Coef.	Est.	SE	Z
(others)			
parity	-0.104	0.130	-0.80
<b>br feed (sd unit)</b>	<b>-0.384</b>	<b>0.173</b>	<b>-2.22</b>
bmi (sd unit)	-1.060	0.224	-4.73
height (sd unit)	-0.210	0.167	-1.26

### Notes:

- small differences for missing covariates
- greater for non-missing covariates  
(80% improvement for br feed coefficient)

(from earlier)

- Bias-corrected complete-record analysis **with** correct standard errors

	Coef.	Est.	SE	Z
<hr/>				
(others)				
parity		0.066	0.129	0.52
<b>br feed (sd unit)</b>		<b>-0.485</b>	<b>0.237</b>	<b>-2.05</b>
bmi (sd unit)		-1.141	0.234	-4.88
height (sd unit)		-0.233	0.149	-1.57
<hr/>				

## Simulation Study

- Binary response  $D_{ij}$ , logistic regression
- $n = 4$  observations per stratum, 200 strata
- Continuous  $(X_{ij}, Z_{ij})$  with

$$\text{corr}(X_{ij}, Z_{ij}) \approx 0.5$$

$$\text{var}(X_{ij}) = \text{var}(Z_{ij}) = 1$$

- Average  $E(D_{ij}) \approx 0.3$
- Missingness probabilities depend on  $(D_{ij}, Z_{ij})$ :
  - 18% missing when  $D_{ij} = 0$
  - 45% missing when  $D_{ij} = 1$
- 1000 replicates
- Similar results for binary  $X_{ij}$
- Similar results for matched case-control study

## Simulation Results – Continuous $X_{ij}$

True values are  $\beta_z = 0.405$  and  $\beta_x = 0.262$ .

Method	$X$ -model	$R$ -model	% Bias		% Rel. MSE Eff.	
			$\beta_z$	$\beta_x$	$\beta_z$	$\beta_x$
$L^c$	$X \nmid Z$		0.3	-3.2	100	100
$L^c$	$X \parallel Z$		33.0	67.3	50	25
Naive		$\pi(Z)$	39.9	0.8	28	51
$L^c_{\text{complete}}$		$\pi(D, Z)$	0.1	2.5	59	51
$U^c_{\text{improved}}$	$X \parallel Z$	$\pi(D, Z)$	-1.4	3.3	74	46
$U^c_{\text{improved}}$	$X \nmid Z$	$\pi(D, Z)$	-1.2	2.9	77	50

 = wrong model

## Key Results

- Complete data likelihood  $L_{\text{complete}}^c$

- uses data

$$(\mathbf{D}_{i,\text{obs}} | \sum_{ij} R_{ij} D_{ij}, \mathbf{X}_{i,\text{obs}}, \mathbf{Z}_{i,\text{obs}}, \mathbf{R}_i)$$

- relies on model  $\pi$  for  $R_{ij}$

- no model for  $X_{ij}$  required

- loss of efficiency

- Efficiency improvement in  $L_{\text{complete}}^c$ :

- projection to increase efficiency of  $L_{\text{complete}}^c$

- estimating function  $U_{\text{improved}}^c$

- exploits a “working model” for  $X_{ij}$

- consistent even if this working model is wrong

- moderate efficiency gained for  $\beta_z$

- less gain for  $\beta_x$

- Better for one (or a few) pattern of missingness

- Better for missing confounder variables

## Fixed effects models for binary data with drop-outs

**Longitudinal observations:**

$$t = 1, \dots, J \text{ within subject } i$$

**Response:**  $D_{it}$  (binary disease status)

**Covariates:**  $Z_{it}$  (vector)

**Drop-out:** Subject  $i$  observed at times

$$t = 1, \dots, T_i \leq J$$

↑

drop-out time

**Response vector up to time  $t$ :**

$$\mathbf{D}_{it} = (D_{i1}, \dots, D_{it})^T$$

**Observed response vector:**

$$\mathbf{D}_{iT} = (D_{i1}, \dots, D_{iT})^T$$

## Bias-corrected complete-record model

Model of interest:

$$\theta_{it} = \theta(Z_{it}) = \frac{\Pr(D_{it} = 1|Z_{it})}{\Pr(D_{it} = 0|Z_{it})}$$

with

$$\theta_{it} = \exp(q_i + \beta^T Z_{it})$$

suppressing  $i \dots$

Drop-out hazard model:

$T$  is drop-out time

$$\lambda(t, \mathbf{d}_t; \gamma) = \Pr(T = t | T \geq t, \mathbf{D}_t = \mathbf{d}_t, \mathbf{Z})$$

Marginal drop-out probability:

$$\pi(t, \mathbf{d}_t; \gamma) = \Pr(T = t | \mathbf{D} = \mathbf{d}, \mathbf{Z})$$

↑

drop-out is “MAR”

**Condition on drop-out:**

$$L_{\text{complete}} = \Pr(\mathbf{D}_T | \mathbf{Z}, T)$$

**Now condition on # positive responses:**

$$L_{\text{complete}}^c = \Pr(\mathbf{D}_T | \sum_{t=1}^T D_t, \mathbf{Z}, T)$$

which yields

$$L_{\text{complete}}^c = \frac{\{\prod_{t=1}^T \theta_t^{D_t}\} \pi(T, \mathbf{D}_T)}{\sum_{\mathbf{d}_T \in \mathcal{D}_T} \{\prod_{t=1}^T \theta_t^{d_t}\} \pi(T, \mathbf{d}_T)}$$

where  $\mathcal{D}_T$  is the set of all possible allocations of complete-positive data responses among complete data records

## Efficiency Improvement with $L_{\text{complete}}^c$

- $L_{\text{complete}}^c$  contains drop-out time  $T$  as a random variable
- But the drop-out process

$$T \mid \mathbf{D}, \mathbf{Z}$$

is **ancillary** for parameter of interest  $\beta$

- Remove from  $U_{\text{complete}}^c$  the projection\* onto the space spanned by all scores that are functions of

$$(R_{it}, \mathbf{D}_{i,t-1}, \mathbf{Z}_i)$$

$R_{it}$  is non-drop-out indicator at  $t$

which are **unbiased** over  $R_{it}$  conditional on

$$(R_{i,t-1} = 1, \mathbf{D}_{i,t-1}, \mathbf{Z}_i)$$

\*projection requires integration over  $(D_t, D_{t+1}, \dots, D_J)^T$

- Projection requires integration over  $(D_t, D_{t+1}, \dots, D_J)^T$  given  $(D_{i1}, \dots, D_{i,t-1})^T$ :
  - ... requires model for joint distribution of  $\mathbf{D}|\mathbf{Z}$
  - ... which depends on the non-parametric distribution  $Q_{\mathbf{Z}}$  of intercepts  $q_i$
- **Approximate projection** via a **working transition model** for the vector of responses  $\mathbf{D}$
- Simulation results relative to

$$L_{\text{complete}} = \Pr(\mathbf{D}_T | \mathbf{Z}, T)$$

with correct drop-out model

- 5–10% efficiency improvement for using a **rich** drop-out model
- 15–20% improvement using approximate projection
- bias and efficiency **very robust** to working transition model for  $\mathbf{D}$
- rich drop-out model irrelevant under approximate projection

## **EXTRA SLIDES**

## **Construction of Projected Score**

## Construction of Projected Score

- Vector  $\mathbf{R}$  specifies **missingness pattern**:

$$\mathbf{r}_k = k\text{th missingness pattern, } k = 1, \dots, 2^n$$

- Pattern indicator:

$$\Delta_k = I(\mathbf{R} = \mathbf{r}_k) = I(k\text{th pattern observed})$$

- Data under  $k$ th pattern:

$$\mathbf{X}_{(k,\text{obs})} = \text{observed components of } \mathbf{X}$$

$$\mathbf{X}_{(k,\text{miss})} = \text{missing components of } \mathbf{X}$$

- Similarly

$$\mathbf{D}_{(k,\text{obs})}, \mathbf{D}_{(k,\text{miss})}, \mathbf{Z}_{(k,\text{obs})}, \mathbf{Z}_{(k,\text{miss})}$$

- Rewrite  $L_{\text{complete}}^c$  as

$$L_{\text{complete}}^c = \prod_{k=1}^{2^n} L_k^{\Delta_k}, \text{ where}$$

$$L_k = \Pr(\mathbf{D}_{(k,\text{obs})} \mid \sum_j D_j r_{kj}, \mathbf{X}_{(k,\text{obs})}, \mathbf{Z}_{(k,\text{obs})}, \mathbf{R} = \mathbf{r}_k)$$

is  $L_{\text{complete}}^c$  under the  $k$ th missingness pattern

- Similarly

$$U_{\text{complete}}^c = \sum_{k=1}^{2^n} \Delta_k U_k \text{ where } U_k = \partial \log L_k / \partial \beta$$

this sums over all possible  
missingness patterns

- Now note:

$$\Delta_k(\mathbf{R}) \quad \text{and} \quad U_k(\mathbf{D}_{(k,\text{obs})}, \mathbf{X}_{(k,\text{obs})}, \mathbf{Z}_{(k,\text{obs})})$$

no  $\mathbf{X}$                       no  $\mathbf{R}$

- Because of this, we can write

$$\text{Proj}(\Delta_k U_k) = \text{Proj}(\Delta_k) U_{*,k}$$

where

$$U_{*,k} = \mathbf{E}_{\mathbf{X}_{(k,\text{obs})}}(U_k \mid \mathbf{D}_{(k,\text{obs})}, \mathbf{Z}_{(k,\text{obs})})$$

- And

$$\text{Proj}(\Delta_k) = \Delta_k - \mathbf{E}_{\mathbf{R}}(\Delta_k | \mathbf{D}, \mathbf{Z}) = \Delta_k - \epsilon_k$$

- So that

$$\text{Proj}(U_{\text{complete}}^c) = \sum_{k=1}^{2^n} (\Delta_k - \epsilon_k) U_{*,k}$$

and

$$U_*^c = U_{\text{complete}}^c - \text{Proj}(U_{\text{complete}}^c)$$

- Important notes on  $\text{Proj}(U_{\text{complete}}^c)$ 
  - does not contain  $\mathbf{X}$
  - unbiasedness only depends on correct model  $\pi$  for  $R$
  - **Q:** Can we replace  $U_{*,k}$  by any function of  $(\mathbf{D}_{(k,\text{obs})}, \mathbf{Z}_{(k,\text{obs})})$ ? **A:** Yes!
  - Such approximate functions can be derived from a **working model** for  $X_j$

## **Modelling $X$ among the controls**

## Modelling $X_{ij}$ among the controls

### Joint Model for $D_{ij}$ and $X_{ij}$

- The model of interest can be written in terms of the **odds** that  $D_j = 1$ :

$$\theta(X_j, Z_j) = \frac{\Pr(D_j = 1 | X_j, Z_j)}{\Pr(D_j = 0 | X_j, Z_j)}$$

where

$$\theta(X_j, Z_j) = \exp(q_i + \beta'_z Z_j + \beta'_x X_j)$$

- Define the **marginal** (over  $X_j$ ) **odds** as

$$\tilde{\theta}(Z_j) = \frac{\Pr(D_j = 1 | Z_j)}{\Pr(D_j = 0 | Z_j)}$$

- Define a model for  $X_j$  via

$$p_0(X_j | Z_j; \alpha) = p(X_j | D_j = 0, Z_j)$$

new parameter

density or pmf of  $X_j$

among **controls**

## Two important facts

- The marginal (over  $X_j$ ) odds  $\tilde{\theta}(Z_j)$  are in general

$$\tilde{\theta}(Z_j) = \int \theta(x, Z_j) p_0(x|Z_j) dx$$

odds of  $D_j = 1$   
versus  $D_j = 0$

density of  $X_j$   
given  $D_j = 0$

and more specifically

$$\begin{aligned} \tilde{\theta} &= \exp(q_i + \beta'_z Z_j) \\ &\quad \times \int_x \exp(\beta'_x x) p_0(x|Z_j; \alpha) dx. \end{aligned}$$

- Density  $p(X_j|D_j = 1, Z_j)$  can be expressed generally as

$$p(X_j|D_j = 1, Z_j) = p_0(X_j|Z_j) \frac{\theta(X_j, Z_j)}{\tilde{\theta}(Z_j)}$$

odds given  $X_j$

marginal odds

and specifically as

$$p(X_j|D_j = 1, Z_j) = p_0(X_j|Z_j; \alpha) \times \left\{ \int_x \exp(\beta'_x x) p_0(x|Z_j; \alpha) dx \right\}^{-1}$$

- **Important notes:**

- role of  $q_i$  is the **same** in  $\theta$  and  $\tilde{\theta}$
- $\sum_j D_j$  is a CSS for  $q_i$  in **both**:
  - the model for  $(D_j|X_j, Z_j)$  and
  - the **marginal model** for  $(D_j|Z_j)$
- $p(X_j|D_j = 1, Z_j)$  does **not depend** on  $q_i$
- (Satten & Kupper, 1993; Satten & Carroll, 2000)

## Implications

- The **full likelihood** for stratum  $i$  is

$$p(\mathbf{X}_{\text{obs}}|\mathbf{R}, \mathbf{D}, \mathbf{Z}; \beta, \alpha)\Pr(\mathbf{R}|\mathbf{D}, \mathbf{Z}; \gamma)\Pr(\mathbf{D}|\mathbf{Z}; \beta, \alpha, q_i)$$

$\beta$  – regression parameter

$\alpha$  – parameter in  $(X_j|D_j = 0, Z_j)$

$\gamma$  – parameter in  $\Pr(R_j = 1|D_j, Z_j)$

- Important facts:

- Again,  $\sum_j D_j$  is a CSS for  $q_i$

- $\Pr(\mathbf{R}|\mathbf{D}, \mathbf{Z})$  is free of  $(\beta, \alpha)$

- MAR:  $p(\mathbf{X}_{\text{obs}}|\mathbf{R}, \mathbf{D}, \mathbf{Z}) = p(\mathbf{X}_{\text{obs}}|\mathbf{D}, \mathbf{Z})$

- By conditioning on  $\sum_j D_j$ , we obtain the joint **conditional likelihood** for  $(\beta, \alpha)$

$$L^c(\beta, \alpha) \propto p(\mathbf{X}_{\text{obs}}|\mathbf{D}, \mathbf{Z})\Pr(\mathbf{D}|\sum_j D_j, \mathbf{Z})$$

which is free of  $q_i$  and  $\gamma$

- Maximizing the conditional likelihood

$$\prod_i L_i^c(\beta, \alpha)$$

will be **SPE** for  $(\beta, \alpha)$  even when  $X_i$  may be missing

- $L^c(\beta, \alpha)$  is written:

$$L^c = \left\{ \prod_j p(X_j | D_j, Z_j)^{R_j} \right\} \left\{ \frac{\prod_j \tilde{\theta}(Z_j)^{D_j}}{\sum_{\mathbf{d} \in \mathcal{D}} \prod_j \tilde{\theta}(Z_j)^{d_j}} \right\}$$

where  $\mathcal{D} = \{\mathbf{d} : \sum_j d_j = \sum_j D_j\}$

(Satten & Kupper, 1993; Satten & Carroll, 2000)

- Pitfall of  $L^c$ :
  - **heavily** dependent on model  $p_0(X_i | Z_i; \alpha)$
  - does not reduce to standard conditional likelihood when  $X_i$  is never missing
- Simulation results ...

## **Suboptimal Estimation**

## Suboptimal Estimation

- In  $L^c$ , random variables are  $(\mathbf{D}, \mathbf{X}_{\text{obs}}, \mathbf{R})$  and  $\sum_j D_j, \mathbf{Z}$  are the only conditioning statistics
- Suggests conditioning on  $(\mathbf{X}_{\text{obs}}, \mathbf{R})$  and using likelihood  $\Pr(\mathbf{D}|\mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Z})$ :

$$\prod_j \Pr(D_j|R_j = 1, X_j, Z_j)^{R_j} \Pr(D_j|R_j = 0, Z_j)^{1-R_j}$$

- Again,  $\sum_j D_j$  is sufficient for  $q_i$ , so the **conditional likelihood**

$$L_{\text{subopt}}^c = \Pr(\mathbf{D}|\sum_j D_j, \mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Z})$$

is free of  $q_i$

- Because  $\sum_j D_j$  is **not** CSS, maximizing

$$\prod_i L_{i,\text{subopt}}^c$$

will **not be SPE** for  $(\beta, \alpha)$

- $L_{\text{subopt}}^c(\beta, \alpha, \gamma)$  is written

$$\frac{\prod_j \theta^*(X_j, Z_j)^{R_j D_j} \tilde{\theta}^*(Z_j)^{(1-R_j)D_j}}{\sum_{\mathbf{d} \in \mathcal{D}} \prod_j \theta^*(X_j, Z_j)^{R_j d_j} \tilde{\theta}^*(Z_j)^{(1-R_j)d_j}}$$

- Odds that  $D_j = 1$  when  $X_j$  **is observed**:

$$\theta^*(X_j, Z_j) = \frac{\Pr(D_j = 1 | R_j = 1, X_j, Z_j)}{\Pr(D_j = 0 | R_j = 1, X_j, Z_j)}$$

$$\theta^*(X_j, Z_j) = \theta(X_j, Z_j) \frac{\pi(1, Z_j)}{\pi(0, Z_j)}$$

and when  $X_j$  **is not observed**:

$$\tilde{\theta}^*(Z_j) = \frac{\Pr(D_j = 1 | R_j = 0, Z_j)}{\Pr(D_j = 0 | R_j = 0, Z_j)}$$

$$\tilde{\theta}^*(Z_j) = \tilde{\theta}(Z_j) \frac{1 - \pi(1, Z_j)}{1 - \pi(0, Z_j)}$$

- Missingness model for  $R_j$ :

$$\Pr(R_j = 1 | D_j = d, Z_j) = \pi(d, Z_j; \gamma)$$

- **Important notes** about  $L_{\text{subopt}}^c$ :
  - reduces to standard conditional likelihood when  $X_j$  is never missing
  - less dependent on model for  $X_j$
  - but: requires a model  $\pi$  for missingness
  - related to work by Paik & Sacco, 2000
- Implementation: we need to **pre-estimate**
  - $\hat{\alpha}$  in  $p_0(X_j|Z_j; \alpha)$  and
  - $\hat{\gamma}$  in  $\pi(D_j, Z_j; \gamma)$
 plug into  $L_{\text{subopt}}^c$  before maximizing over  $\beta$
- Simulation results ...

## **Full Simulation Results**

## Simulation Study

- Binary response  $D_{ij}$ , logistic regression
- $n = 4$  observations per stratum, 200 strata
- Continuous  $(X_{ij}, Z_{ij})$  with

$$\text{corr}(X_{ij}, Z_{ij}) \approx 0.5$$

$$\text{var}(X_{ij}) = \text{var}(Z_{ij}) = 1$$

- Average  $E(D_{ij}) \approx 0.3$
- Missingness probabilities depend on  $(D_{ij}, Z_{ij})$ :
  - 18% missing when  $D_{ij} = 0$
  - 45% missing when  $D_{ij} = 1$
- 1000 replicates
- Similar results for binary  $X_{ij}$
- Similar results for matched case-control study

## Simulation Results – Continuous $X_{ij}$

True values are  $\beta_z = 0.405$  and  $\beta_x = 0.262$ .

Method	$X$ -model	$R$ -model	% Bias		% Rel. MSE Eff.	
			$\beta_z$	$\beta_x$	$\beta_z$	$\beta_x$
$L^c$	$X \perp Z$		0.3	-3.2	100	100
$L_{\text{subopt}}^c$	$X \perp Z$	$\pi(D, Z)$	-2.1	-4.8	82	76
$L_{\text{subopt}}^c$	$X \perp Z$	$\pi(Z)$	-21.1	29.2	46	34
$L^c$	$X \parallel Z$		33.0	67.3	50	25
$L_{\text{subopt}}^c$	$X \parallel Z$	$\pi(D, Z)$	17.4	-9.4	83	69
$L_{\text{subopt}}^c$	$X \parallel Z$	$\pi(Z)$	21.1	-33.9	57	16
Naive		$\pi(Z)$	39.9	0.8	28	51
$L_{\text{complete}}^c$		$\pi(D, Z)$	0.1	2.5	59	51
$U_{\text{improved}}^c$	$X \parallel Z$	$\pi(D, Z)$	-1.4	3.3	74	46
$U_{\text{improved}}^c$	$X \perp Z$	$\pi(D, Z)$	-1.2	2.9	77	50

## Simulation Results – Binary $X_{ij}$

True values are  $\beta_z = 0.405$  and  $\beta_x = 0.693$ .

Method	$X$ -model	$R$ -model	% Rel.			
			% Bias		MSE Eff.	
			$\beta_z$	$\beta_x$	$\beta_z$	$\beta_x$
$L^c$	$X \perp Z$		1.2	-1.7	100	100
$L_{\text{subopt}}^c$	$X \perp Z$	$\pi(Y, Z)$	1.1	-0.2	95	69
$L_{\text{subopt}}^c$	$X \perp Z$	$\pi(Z)$	-2.6	19.1	88	37
$L^c$	$X \parallel Z$		20.5	12.4	66	92
$L_{\text{subopt}}^c$	$X \parallel Z$	$\pi(Y, Z)$	9.7	-1.5	91	69
$L_{\text{subopt}}^c$	$X \parallel Z$	$\pi(Z)$	10.2	-0.2	88	40
$L_{\text{complete}}^c$		$\pi(Y, Z)$	2.5	0.9	51	60
$U_{\text{improved}}^c$	$X \parallel Z$	$\pi(Y, Z)$	1.6	1.1	78	59
$U_{\text{improved}}^c$	$X \perp Z$	$\pi(Y, Z)$	1.7	0.9	79	60

## Estimation via $L_{\text{subopt}}^c$

- Recall:  $L_{\text{subopt}}^c$  conditions on

$$(\sum_j D_j, \mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Z})$$

- $L_{\text{subopt}}^c$  contains parameters

$$\alpha \text{ in } p_0(X_j|Z_j; \alpha) \text{ and}$$

$$\gamma \text{ in } \pi(d, Z_j; \gamma)$$

- Missingness parameter  $\gamma$  estimated via likelihood

$$\prod_i \Pr(\mathbf{R}_i | \mathbf{D}_i, \mathbf{Z}_i; \gamma)$$

- $X_j$ -model parameter  $\alpha$  estimated via likelihood

$$\prod_i \Pr(\mathbf{X}_{i,\text{obs}} | \mathbf{R}_i, \mathbf{D}_i, \mathbf{Z}_i; \alpha, \beta_x)$$

which results in an “extra” estimate of  $\beta_x$

this “extra” estimate is discarded

**What is  $\pi(\cdot)$  doing in  $L_{\text{subopt}}^c$ ?**

**An heuristic explanation**

- The full likelihood can be written

$$\prod_j \Pr(D_j|Z_j) \{p(X_j|D_j, Z_j) \pi(D_j, Z_j)\}^{R_j} \\ \times \{1 - \pi(D_j, Z_j)\}^{1-R_j}$$

- no  $q_i$ : factors  $\pi(D_j, Z_j)$  and  $[1 - \pi(D_j, Z_j)]$  generally drop out
- with  $q_i$ : conditioning on  $\sum_j D_j$  replaces

$$\prod_j \Pr(D_j|Z_j) \quad \text{with} \quad \Pr(\mathbf{D}|\sum_j D_j, \mathbf{Z})$$

and so  $\pi(D_j, Z_j)$  and  $[1 - \pi(D_j, Z_j)]$  still drop out of the final conditional likelihood

- This likelihood is **only** conditional on  $Z_j$ ;  
 $(D_j, R_j, X_j)$  are random variables

- Conditioning on  $\mathbf{X}_{\text{obs}}$  requires also conditioning on  $\mathbf{R}$
- The starting-point likelihood is

$$\prod_j \Pr(D_j | R_j = 1, X_j, Z_j)^{R_j} \Pr(D_j | R_j = 0, Z_j)^{1-R_j}$$

- each factor contains  $\pi(\cdot)$
- after conditioning on  $\sum_j D_j$ : the terms containing  $\pi(\cdot)$  are no longer separable

## Simulation Results – Binary $X_{ij}$

True values are  $\beta_z = 0.405$  and  $\beta_x = 0.693$ .

Method	$X$ -model	$R$ -model	% Rel.			
			% Bias		MSE Eff.	
			$\beta_z$	$\beta_x$	$\beta_z$	$\beta_x$
$L^c$	$X \amalg Z$		1.2	-1.7	100	100
$L_{\text{subopt}}^c$	$X \amalg Z$	$\pi(Y, Z)$	1.1	-0.2	95	69
$L_{\text{subopt}}^c$	$X \amalg Z$	$\pi(Z)$	-2.6	19.1	88	37
$L^c$	$X \amalg Z$		20.5	12.4	66	92
$L_{\text{subopt}}^c$	$X \amalg Z$	$\pi(Y, Z)$	9.7	-1.5	91	69
$L_{\text{subopt}}^c$	$X \amalg Z$	$\pi(Z)$	10.2	-0.2	88	40
$L_{\text{complete}}^c$		$\pi(Y, Z)$	2.5	0.9	51	60
$U_{\text{improved}}^c$	$X \amalg Z$	$\pi(Y, Z)$	1.6	1.1	78	59
$U_{\text{improved}}^c$	$X \amalg Z$	$\pi(Y, Z)$	1.7	0.9	79	60