Impact of using grouping strategy with miss-measured exposures in logistic and Cox proportional hazard models and some improvement by Bayesian approach in logistic models

H.M. Kim, Y. Yasui, I. Burstyn

Department of Public Health Sciences

the University of Alberta

October 13, 2005

outline

- background
- group-based strategy
- objectives
- model
- results
- consequences
- Bayesian method and results
- further research

background

- there are over and/or under estimates in the slope parameter estimation in logistic and Cox proportional-hazards models in the occupational/environmental exposure-health studies

at least two issues are involved:

- all exposure measurements are not available
- the true exposures are not available, but we only have observed exposures with errors

group-based strategy

- group-based strategy is widely used in occupational health research

1. estimate the group-mean for a sample of workers from each group (department, job, task)
2. assign all workers in a group with the estimated value of exposure for that group
3. assess health outcome for each subject individually

- a single-impute "fill-in" method with group means

objective

- the objective of this study is to see the impact of using the group-based strategy in logistic and Cox proportional-hazard models

classical and Berkson error models

- classical error model

$$X = Z + u$$

where Z: true exposure, X: observed exposure, (Z, u): independent

- Berkson error model

$$Z = X + e$$

where Z: true exposure, X: observed exposure, (X, e): independent, which leads to $E[Z|X] = X$

(Berkson, 1950)

classical exposure error model

- log- transformation

$$X_{gij} = \log(\text{exposure}) = \mu_g + \gamma_{gi} + \epsilon_{gij}$$

$$X_{gij} = \mu_{gi} + \epsilon_{gij}, \quad \text{where } \mu_{gi} = \mu_g + \gamma_{gi}$$

$$\mu_{gi} \sim \mathcal{N}(\mu_g, \sigma_B^2) \text{ and } \epsilon_{gij} \sim \mathcal{N}(0, \sigma_W^2)$$

where g groups $(1, \cdots, G)$, i workers $(1, \cdots, K_g)$ and j days $(1, \cdots, N_{gi})$

where μ_{gi} and ϵ_{gij} are mutually independent .

result 1: Berkson error model

from the conditional distribution:

$$E[\mu_{gi}|\bar{X}_g] = \bar{X}_g + \left(\frac{n\sigma_B^2}{n\sigma_B^2+\sigma_W^2} - 1\right)(\bar{X}_g - \mu_g) \quad \approx \quad \bar{X}_g$$

if the number of workers (k) is large enough ($\bar{X}_g \approx \mu_g$)

- Berkson error model:

  $\mu_{gi} = \bar{X}_g + e_{gi}$

  with $E[e_{gi}|\bar{X}_g] = 0$ and $E[\mu_{gi}|\bar{X}_g] \approx \bar{X}_g$ if k is large enough.

  (note that this is not a true Berkson model)

`result 2: error variance`

if $E[\mu_{gi}|\bar{X}_g] = \bar{X}_g$

1. $cov(\bar{X}_g, \mu_{gi}) = V(\bar{X}_g)$
2. $cov(\bar{X}_g, e_{gi}) = 0$
3. $cov(\mu_{gi}, e_{gi}) = V(\mu_{gi}|\bar{X}_g) = V(e_{gi}) = \boldsymbol{\sigma}_e^2$

$$V(\mu_{gi}|\bar{X}_g) = (1 - \tfrac{1}{k})\sigma_B^2 - \tfrac{1}{nk}\sigma_W^2 \approx \sigma_B^2$$

$$V(\mu_{gi}|\bar{X}_g) = \boldsymbol{\sigma}_e^2 \approx \sigma_B^2$$

if the number of workers (k) is large enough

results 3: attenuation Equations

- logistic Model

$$\beta_1^* \approx \frac{\beta_1}{\sqrt{c^2\beta_1^2\sigma_B^2 + 1}}$$

(Burr, 1988; Reeves, 1998)

- Cox proportional-hazards model

$$\alpha^* \approx \frac{\alpha}{(1 + \frac{1}{2}\alpha^2\sigma_B^2)\sqrt{c^2\alpha^2\sigma_B^2 + 1}}$$

where $c = 0.588$: connection value between Logistic and Probit functions when $0.1 < p < 0.9$

details for logistic Model

- logistic Model

- $P(i=1|Z) = \Lambda(\beta_0 + \beta_1 Z)$, where $\Lambda(t) = \frac{1}{1+\exp(-t))}$

- $P(i=1|Z) = \Lambda(\beta_0 + \beta_1 Z) \approx \Phi[c(\beta_0 + \beta_1 Z)]$ if $0.1 < p < 0.9$,
  where $\Phi(t)$: c.d.f for the standard Normal distribution

- $E[P(i=1|Z)|X] = E[\Lambda(\beta_0 + \beta_1 Z)|X]$
  $\approx E\{\Phi[c(\beta_0 + \beta_1 Z)]|X\} \approx \Phi[c(\beta_0^* + \beta_1^* X)]$
  $\approx \Lambda(\beta_0^* + \beta_1^* X)$, where $Z|X \sim \mathcal{N}(X, \sigma_e^2)$.

$$\beta_1^* \approx \frac{\beta_1}{\sqrt{c^2 \beta_1^2 \sigma_B^2 + 1}}$$

since $\sigma_e^2 \approx \sigma_B^2$

details for Cox proportional-hazards Model

- Cox proportional hazards model

$$h(t|Z) = h_0(t)\exp(\alpha Z)$$

  a. survival function in the logistic model : $\exp(\beta_0^* + \beta_1^* X)$ (after Taylor series expansion)

  b. survival function in Cox proportional hazards model: $\lambda T \exp\left(\alpha X + \frac{1}{2}\sigma_e^2\right)$

- estimates of β_1^* and α^* are approximately equivalent when survival functions of both models with the observed exposures are approximately equivalent

- i.e. $a \approx b \Longrightarrow \exp(\beta_1^* X) \approx \exp\left((\alpha - \alpha^*)X + \alpha^* X + \frac{1}{2}\alpha^2\sigma_e^2\right)$

(Prentice, 1982; Green, 1983; Li et al., 2004)

details for Cox proportional-hazards Model

$$\Longrightarrow \quad \exp(\beta_1^* X) \approx \exp\left(\alpha^* X + \tfrac{1}{2}\alpha^2\sigma_e^2\right)$$

- apply Taylor series expansion and derivation within small error variance, σ_e^2

$$\alpha^* \approx \frac{\beta_1^*}{(1 + \frac{1}{2}\alpha^2\sigma_e^2)}$$

- $\beta_1 \approx \alpha$ and $\sigma_e^2 \approx \sigma_B^2$

$$\alpha^* \approx \frac{\alpha}{(1 + \frac{1}{2}\alpha^2\sigma_B^2)\sqrt{c^2\alpha^2\sigma_B^2 + 1}}$$

where $c = 0.588$

simulations

- proc phreg and proc logistic in SAS
  - groups: $G = 5$
  - true parameters: $\beta_0 = -4$ and $\beta_1 = 0.2, 0.4, 0.6$
  - baseline hazard: $\lambda = 0.01$
  - true group means: $\mu_g = 1.1, 2.1, \cdots, 5.1$
  - within-worker standard deviation: $\sigma_W = 0, 0.5, 1.5, 3$
  - population: $N = 1000$ workers each group and $K = 10$ days measurements for each worker
  - sample size: $k = 10, 50, 100$ workers each group and $n = 2$ days measurements for each worker
  - Number of replications: rep$=1000$ times

$$\text{bias} = \frac{1}{\text{rep}} \sum_{r=1}^{\text{rep}} (\hat{\beta}_r - \beta) \qquad \text{MSE} = \frac{1}{\text{rep}} \sum_{r=1}^{\text{rep}} (\hat{\beta}_r - \beta)^2$$

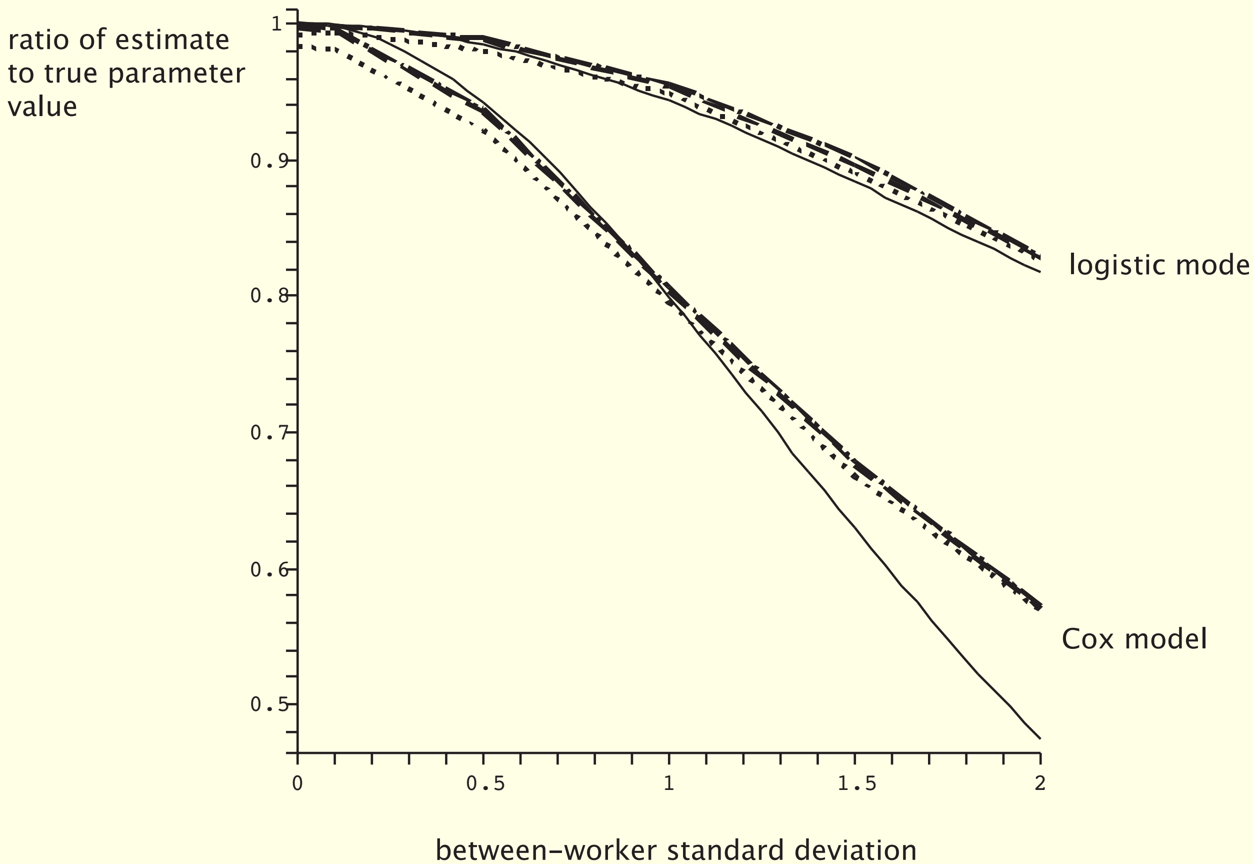

Figure 1: $\beta = 0.6$, $k = 100$, $\sigma_W = 0.5$ (dot-dash), 1.5 (dash), 3 (dot)

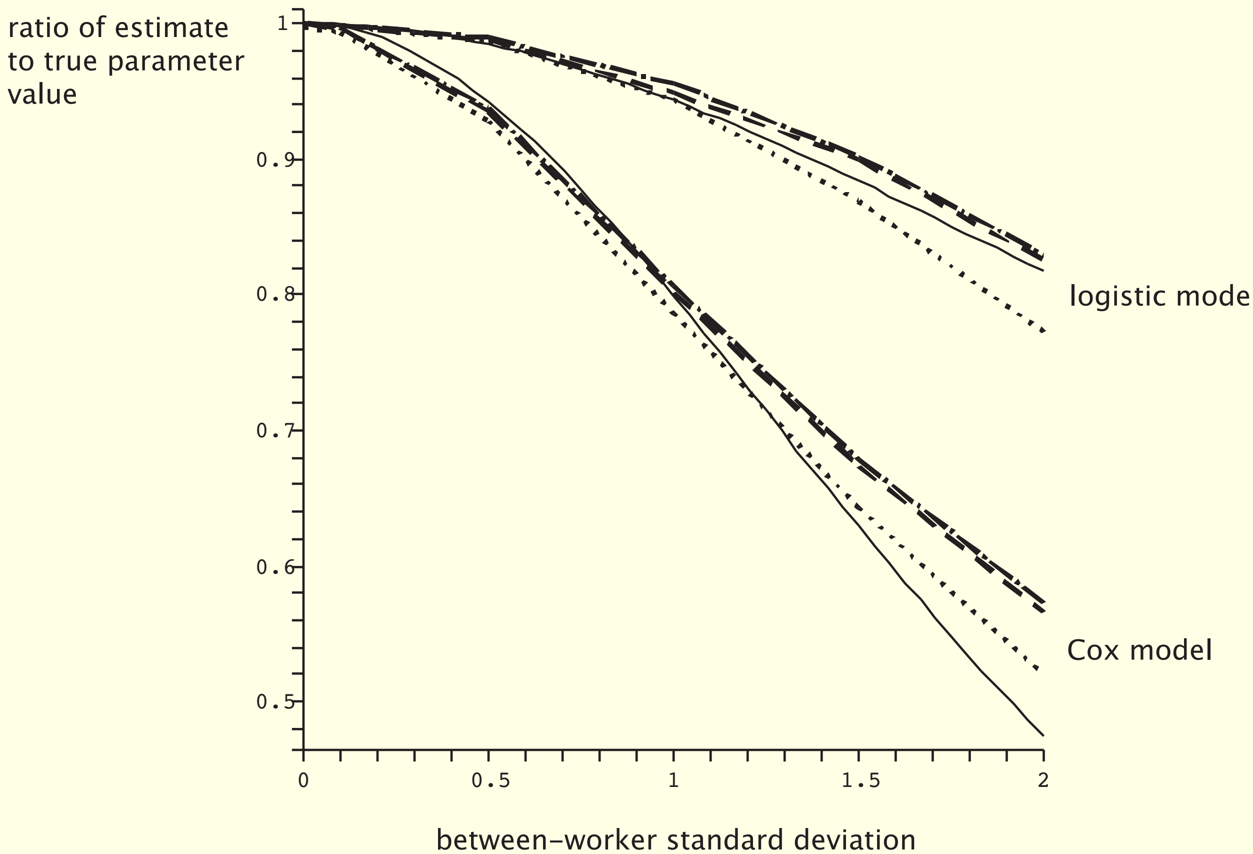

Figure 2: $\beta = 0.6$, $\sigma_W = 0.5$, $k = 100$ (dot-dash), 50 (dash), 10 (dot)

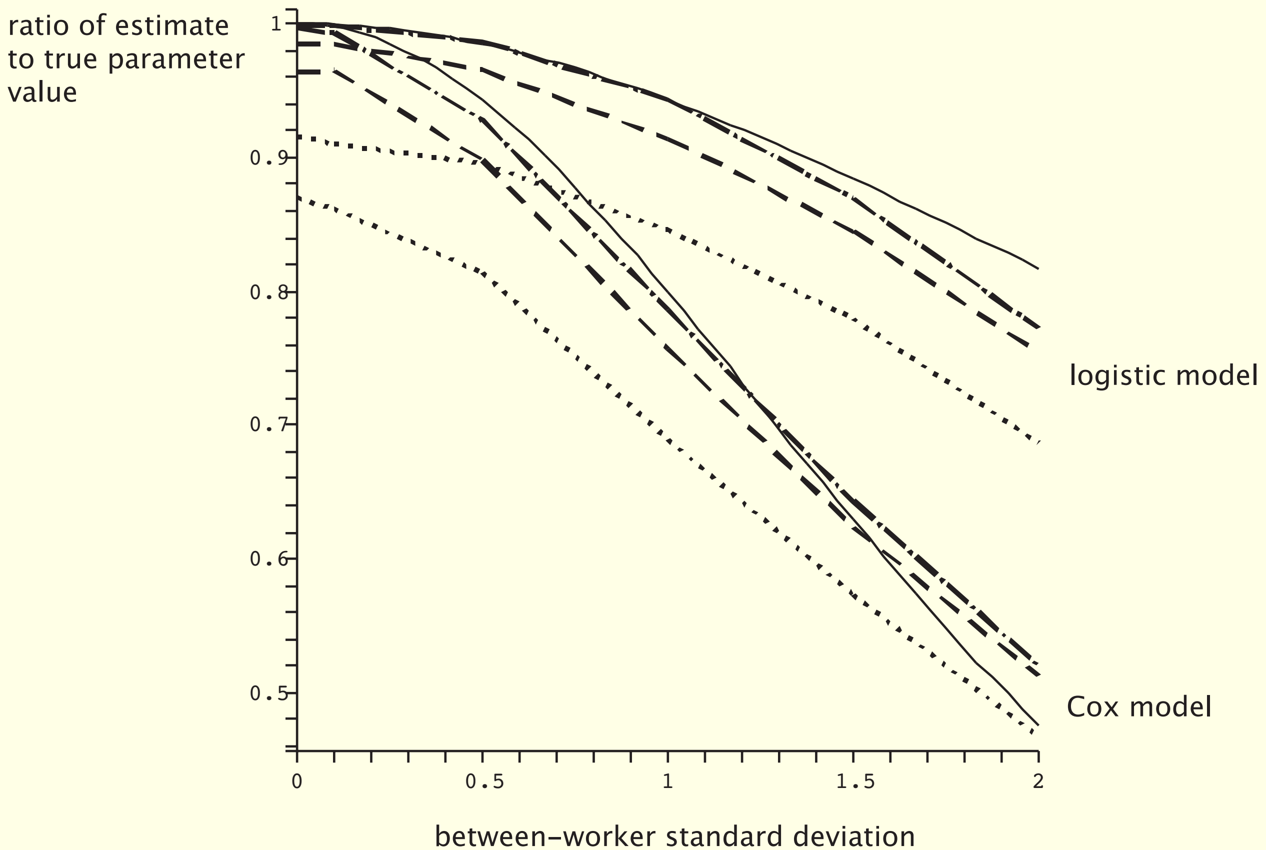

Figure 3: $\beta = 0.6$, $k = 10$, $\sigma_W = 0.5$ (dot-dash), 1.5 (dash), 3 (dot)

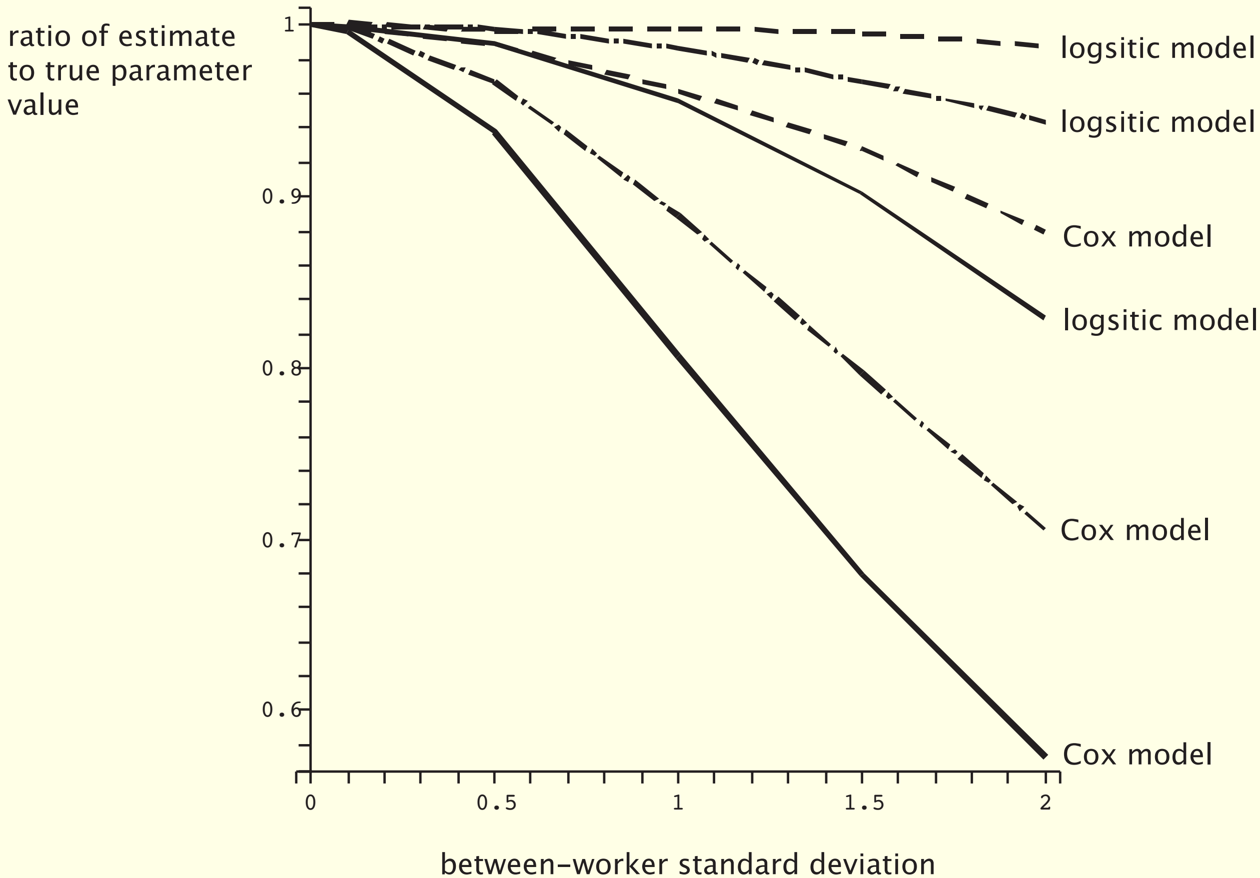

Figure 4: $k = 100$, $\beta = 0.2$ (dash), 0.4 (dot-dash), 0.6 (solid)

consequences

- with a true Berskon error structure

\- there is no or little attenuation in logistic and Cox proportional- hazards models if the sample size is moderately large (Deddens et al, 1994; Armstrong. 1990, 1998)

- grouping with the mean values (full observations)

\- unbiased estimate (Prais et al., 1953)

- a calibration method (i.e. use $E[Z|X]$ instead of X)

\- adjusting measurement error (Rosner, 1989; Spiegelman, 2001)

consequences

- with the group-based strategy

- there is attenuation with large between-worker variance
- it is severe in Cox proportional-hazard models

may be...

- leads to an approximate Berkson (it is not a true Berkson)
- don't have full observations
- a calibration method may fail to adjust when the error variance is large

Bayesian method in logistic models

- the attempt to adjust attenuation when the between-worker variance is large in logistic models: the group-based strategy in a Bayesian framework

- two-steps procedure:
  1. complete data with assigned group-means of the sample
  2. estimates the slope parameter in a Bayesian framework

Bayesian method in logistic models

- three sub-models

1. response model: logistic regression model: $[i|Z, \beta]$
2. measurement error model :
   - classical error $[X|Z, \text{parameters}]$
   - Berkson error $[Z|X, \text{parameters}]$
3. exposure model:
   - classical error: $[Z|$ parameters$]$
   - Berkson error: $[X]$, which is not needed in this framework

where Z: true and X: observed.

(Gilks et al., 1996; Gossl et.al, 2001; Gelman et al., 2004)

Bayesian method in logistic models

since the group-based strategy leads to an approximate Berkson error structure

- Bayesian framework for Berkson type error

1. response model: logistic regression model: $[i|Z, \beta]$
2. measurement error model :
   - Berkson error $[Z|X, \sigma_B]$; $\mu_{gi}|\bar{X}_g \sim \mathcal{N}(\bar{X}_g, \sigma_B^2)$

- prior for β : $f(\beta) = 1$ and σ_B is known
- initial value : $\hat{\beta}$ from GLM
- MH algorithm with random walk proposals

Bayesian method: results

$\beta = 0.6$, $\sigma_W = 1.5$, $n = 2$ repeated measurements and $K = 1000$ per group with burn=1000 and size=10000, (90% credible interval)

$\sigma_B = 1$	$k = 100$	$k = 50$
BBG	0.592 (0.53-0.66)	0.539 (0.48-0.60)
GBS	0.567 (0.52-0.62)	0.518 (0.47-0.57)
$\sigma_B = 1.5$	$k = 100$	$k = 50$
BBG	0.572 (0.51-0.64)	0.617 (0.55-0.69)
GBS	0.518 (0.47-0.56)	0.549 (0.50-0.59)
$\sigma_B = 2$	$k = 100$	$k = 50$
BBG	0.617 (0.53-0.71)	0.435 (0.38-0.49)
GBS	0.502 (0.46-0.54)	0.392 (0.36-0.43)

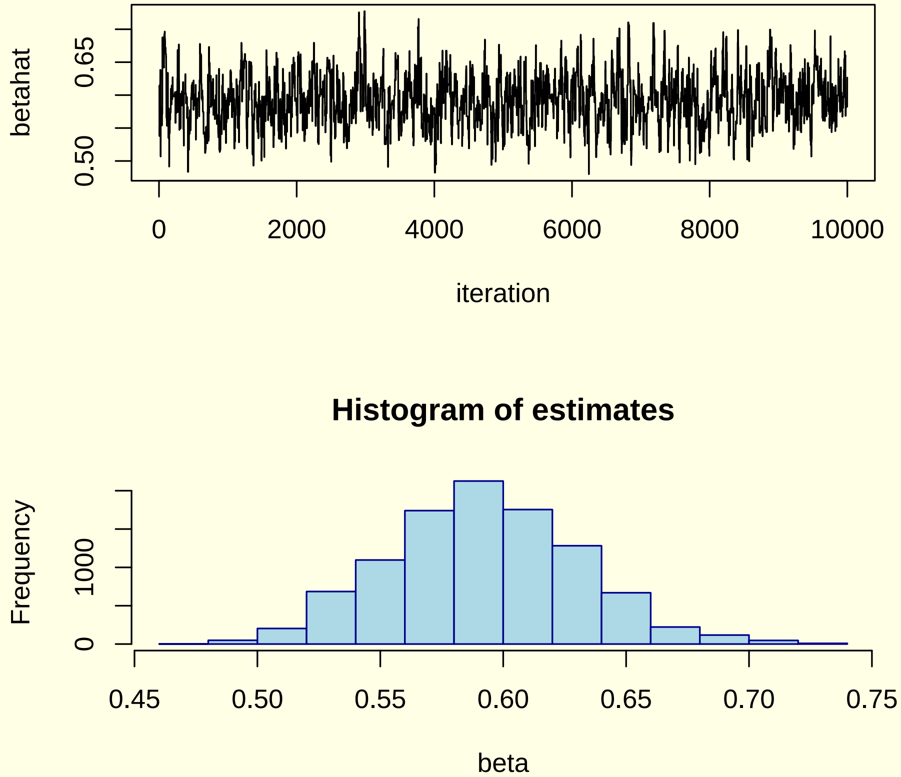

Figure 5: trajectory and histogram of estimates of MH algorithm when $\beta = 0.6$ and $\sigma_B = 1, \sigma_W = 1.5, k = 100$

further research

- unknown variance components
- different sample size in each group in the simulations
- Bayesian method for Cox proportional-hazard models
- Bayesian imputation methods

Thank you!

Questions or Comments?

"Definitely work-related."

The New Yorker, March 21, 2005, page 71