Multiple imputation in Cox regression when there are time-varying effects of exposures

Ruth Keogh

Department of Medical Statistics
London School of Hygiene and Tropical Medicine

Bristol, 5 February 2016

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Multiple imputation in general

- Aim: To fit an analysis model $Y \sim X, Z$
- Missing data in explanatory variables is a very common problem in epidemiology
- Basic approach: Complete case analysis

Multiple imputation (MI)

For a partially missing exposure X, fully observed covariates Z

1. Draw values of X from $X|Z, Y$
2. Obtain several imputed data sets
3. Fit the analysis model in each imputed data set and combine parameter estimates using Rubin's Rules

Multiple imputation in general

- Aim: To fit an analysis model $Y \sim X, Z$
- Missing data in explanatory variables is a very common problem in epidemiology
- Basic approach: Complete case analysis

Multiple imputation (MI)

For a partially missing exposure X, fully observed covariates Z

1. Draw values of X from $X|Z, Y$
2. Obtain several imputed data sets
3. Fit the analysis model in each imputed data set and combine parameter estimates using Rubin's Rules

Multiple imputation in general

- Aim: To fit an analysis model $Y \sim X, Z$
- Missing data in explanatory variables is a very common problem in epidemiology
- Basic approach: Complete case analysis

Multiple imputation (MI)

For a partially missing exposure X, fully observed covariates Z

1. Draw values of X from $X|Z, Y$
2. Obtain several imputed data sets
3. Fit the analysis model in each imputed data set and combine parameter estimates using Rubin's Rules

Multiple imputation in general

Main challenge

What is the distribution of $X|Z, Y$?

Example: Linear regression

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

- If $Y|X, Z \sim \mathrm{Normal}$ and $X|Z \sim \mathrm{Normal}$ then $X|Z, Y \sim \mathrm{Normal}$
- Imputation model: $X = \alpha_0 + \alpha_1 Z + \alpha_2 Y + \delta$

Steps

1. Obtain estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_\delta$, and their variances/covariances
2. Draw values $\hat{\alpha}_0^{(m)}, \hat{\alpha}_1^{(m)}, \hat{\alpha}_2^{(m)}, \hat{\sigma}_\delta^{2(m)}$ from their estimated distn
3. The mth imputation of X is

$$X^{(m)} = \hat{\alpha}_0^{(m)} + \hat{\alpha}_1^{(m)} Z + \hat{\alpha}_2^{(m)} Y + \delta^*$$

Multiple imputation in general

Main challenge

What is the distribution of $X|Z, Y$?

Example: Linear regression

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

- If $Y|X, Z \sim \text{Normal}$ and $X|Z \sim \text{Normal}$ then $X|Z, Y \sim \text{Normal}$
- Imputation model: $X = \alpha_0 + \alpha_1 Z + \alpha_2 Y + \delta$

Steps

1. Obtain estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_\delta$, and their variances/covariances
2. Draw values $\hat{\alpha}_0^{(m)}, \hat{\alpha}_1^{(m)}, \hat{\alpha}_2^{(m)}, \hat{\sigma}_\delta^{2(m)}$ from their estimated distn
3. The mth imputation of X is

$$X^{(m)} = \hat{\alpha}_0^{(m)} + \hat{\alpha}_1^{(m)} Z + \hat{\alpha}_2^{(m)} Y + \delta^*$$

Multiple imputation in general

Main challenge

What is the distribution of $X|Z,Y$?

Example: Linear regression

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

- If $Y|X,Z \sim \mathrm{Normal}$ and $X|Z \sim \mathrm{Normal}$ then $X|Z,Y \sim \mathrm{Normal}$
- Imputation model: $X = \alpha_0 + \alpha_1 Z + \alpha_2 Y + \delta$

Steps

1. Obtain estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_\delta$, and their variances/covariances
2. Draw values $\hat{\alpha}_0^{(m)}, \hat{\alpha}_1^{(m)}, \hat{\alpha}_2^{(m)}, \hat{\sigma}_\delta^{2(m)}$ from their estimated distn
3. The mth imputation of X is

$$X^{(m)} = \hat{\alpha}_0^{(m)} + \hat{\alpha}_1^{(m)} Z + \hat{\alpha}_2^{(m)} Y + \delta^*$$

Multiple imputation in general

Main challenge

What is the distribution of $X|Z, Y$?

Example: Linear regression

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

- If $Y|X, Z \sim \mathrm{Normal}$ and $X|Z \sim \mathrm{Normal}$ then $X|Z, Y \sim \mathrm{Normal}$
- Imputation model: $X = \alpha_0 + \alpha_1 Z + \alpha_2 Y + \delta$

Steps

1. Obtain estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_\delta$, and their variances/covariances
2. Draw values $\hat{\alpha}_0^{(m)}, \hat{\alpha}_1^{(m)}, \hat{\alpha}_2^{(m)}, \hat{\sigma}_\delta^{2(m)}$ from their estimated distn
3. The mth imputation of X is

$$X^{(m)} = \hat{\alpha}_0^{(m)} + \hat{\alpha}_1^{(m)} Z + \hat{\alpha}_2^{(m)} Y + \delta^*$$

Multiple imputation in general

Main challenge

What is the distribution of $X|Z, Y$?

Example: Linear regression

$$Y = \beta_0 + \beta_X X + \beta_Z Z + \varepsilon$$

- If $Y|X, Z \sim \mathrm{Normal}$ and $X|Z \sim \mathrm{Normal}$ then $X|Z, Y \sim \mathrm{Normal}$
- Imputation model: $X = \alpha_0 + \alpha_1 Z + \alpha_2 Y + \delta$

Steps

1. Obtain estimates $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2_\delta$, and their variances/covariances
2. Draw values $\hat{\alpha}_0^{(m)}, \hat{\alpha}_1^{(m)}, \hat{\alpha}_2^{(m)}, \hat{\sigma}_\delta^{2(m)}$ from their estimated distn
3. The *m*th imputation of X is

$$X^{(m)} = \hat{\alpha}_0^{(m)} + \hat{\alpha}_1^{(m)} Z + \hat{\alpha}_2^{(m)} Y + \delta^*$$

Multiple imputation in Cox Regression

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

- T: Event or censoring time
- D: Event indicator

Distribution of interest for the imputation:

$$X|Z, \underline{\text{outcome}}$$

Event/censoring time T, event indicator D

What is this distribution $X|Z,T,D$??

Multiple imputation in Cox Regression

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

- T: Event or censoring time
- D: Event indicator

Distribution of interest for the imputation:

$$X|Z, \underline{\text{outcome}}$$

Event/censoring time T, event indicator D

What is this distribution $X|Z,T,D$??

Multiple imputation in Cox Regression

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

- T: Event or censoring time
- D: Event indicator

Distribution of interest for the imputation:

$$X|Z, \underline{\text{outcome}}$$

Event/censoring time T, event indicator D

What is this distribution $X|Z, T, D$??

Multiple imputation in Cox Regression

Previously suggested imputation models:

$$X \sim Z + D + T, \qquad X \sim Z + D + \log T$$

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1982–1998
Published online 19 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3618

Imputing missing covariate values for the Cox model

Ian R. White[1,*,†] and Patrick Royston[2]

[1]*MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K.*
[2]*MRC Clinical Trials Unit, Cancer Group, London, U.K.*

White and Royston imputation model

$$X \sim Z + D + H_0(T)$$

Cumulative baseline hazard

Multiple imputation in Cox Regression

Previously suggested imputation models:

$$X \sim Z + D + T, \qquad X \sim Z + D + \log T$$

STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1982–1998
Published online 19 May 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3618

Imputing missing covariate values for the Cox model

Ian R. White[1,*,†] and Patrick Royston[2]

[1] *MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, U.K.*
[2] *MRC Clinical Trials Unit, Cancer Group, London, U.K.*

White and Royston imputation model

$$X \sim Z + D + H_0(T)$$

Cumulative baseline hazard

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Modelling time-varying effects in Cox regression

Standard Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

- Sometimes we want to study how the effect of the exposure changes over time

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T) X + \beta_Z Z}$$

- This also enables a test of the proportional hazards assumption

Modelling time-varying effects in Cox regression

Standard Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

- Sometimes we want to study how the effect of the exposure changes over time

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T) X + \beta_Z Z}$$

- This also enables a test of the proportional hazards assumption

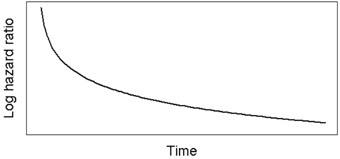
Modelling time-varying effects in Cox regression

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

- Smooth pre-specified form:

$$\beta_X(T) = \beta_X + \beta_{XT} log(T)$$



- Step function:

$$\beta_X(T) = \begin{cases} \beta_{X1} & 0 < T \leq u_1 \\ \beta_{X2} & u_1 < T \leq u_2 \\ \beta_{X3} & u_2 < T \leq u_3 \\ \beta_{X4} & T > u_3 \end{cases}$$

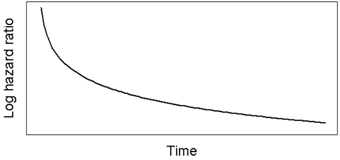
Modelling time-varying effects in Cox regression

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

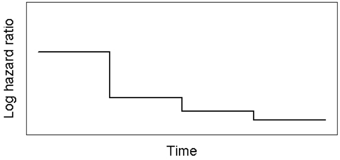
- Smooth pre-specified form:

$$\beta_X(T) = \beta_X + \beta_{XT}log(T)$$



- Step function:

$$\beta_X(T) = \begin{cases} \beta_{X1} & 0 < T \leq u_1 \\ \beta_{X2} & u_1 < T \leq u_2 \\ \beta_{X3} & u_2 < T \leq u_3 \\ \beta_{X4} & T > u_3 \end{cases}$$



Modelling time-varying effects in Cox regression

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

- Restricted cubic spline

$$\beta_X(T) = \beta_{X0} + \beta_{X1}T + \beta_{X2}\left\{(T-u_1)_+^3 - \left(\frac{(T-u_2)_+^3(u_3-u_1)}{(u_3-u_2)}\right) + \left(\frac{(T-u_3)_+^3(u_2-u_1)}{(u_3-u_2)}\right)\right\}$$

STATISTICS IN MEDICINE, VOL. 13, 1045–1062 (1994)

ASSESSING TIME-BY-COVARIATE INTERACTIONS IN PROPORTIONAL HAZARDS REGRESSION MODELS USING CUBIC SPLINE FUNCTIONS

KENNETH R. HESS

Department of Patient Studies, Box 214, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, Texas 77030-4095, U.S.A.

Modelling time-varying effects in Cox regression

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

- Restricted cubic spline

$$\beta_X(T) = \beta_{X0} + \beta_{X1}T + \beta_{X2}\left\{(T-u_1)_+^3 - \left(\frac{(T-u_2)_+^3(u_3-u_1)}{(u_3-u_2)}\right) + \left(\frac{(T-u_3)_+^3(u_2-u_1)}{(u_3-u_2)}\right)\right\}$$

STATISTICS IN MEDICINE, VOL. 13, 1045–1062 (1994)

ASSESSING TIME-BY-COVARIATE INTERACTIONS IN PROPORTIONAL HAZARDS REGRESSION MODELS USING CUBIC SPLINE FUNCTIONS

KENNETH R. HESS

Department of Patient Studies, Box 214, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd., Houston, Texas 77030-4095, U.S.A.

How do we handle missing data in this situations?

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

What is the distribution of $X|T,D,Z$?

$$p(X|T,D,Z)$$

Aims

1. Derive an (approximate) imputation model
   - By extending the work of White & Royston
2. Assess the performance of the imputation model using simulations

Motivation

- Investigation of the long-term efficacy of the BCG vaccine for TB
- Time-varying effect investigated for vaccination status:
  - 0-5 yrs
  - 5-10 yrs
  - 10-15 yrs
  - 15+ yrs post-vaccination
- Missing data on vaccination status
- Also missing data on adjustment variables

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Derivation of imputation model

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

$$p(X|T,D,Z) = p(X|Z)p(T,D|X,Z)/p(T,D|Z)$$

We will specify

$\propto h(T|X,Z)^D \times Pr(\text{survive to time } T|X,Z)$

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

Derivation of imputation model

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

$$p(X|T,D,Z) = p(X|Z)p(T,D|X,Z)/p(T,D|Z)$$

We will specify

$\propto h(T|X,Z)^D \times Pr(\text{survive to time } T|X,Z)$

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

Derivation of imputation model

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

$$p(X|T,D,Z) = p(X|Z)p(T,D|X,Z)/p(T,D|Z)$$

We will specify

$\propto h(T|X,Z)^D \times Pr(\text{survive to time } T|X,Z)$

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

Derivation of imputation model

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

$$p(X|T,D,Z) = p(X|Z)p(T,D|X,Z)/p(T,D|Z)$$

We will specify

$\propto h(T|X,Z)^D \times Pr(\text{survive to time } T|X,Z)$

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

Derivation of imputation model

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

$$p(X|T,D,Z) = p(X|Z)p(T,D|X,Z)/p(T,D|Z)$$

We will specify

$\propto h(T|X,Z)^D \times Pr(\text{survive to time } T|X,Z)$

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

Derivation of imputation model

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X+\beta_Z Z} \mathrm{d}u + const$$

How do we apply this when...?

1. X is binary

$$\text{logit } \Pr(X=1|Z) = \zeta_0 + \zeta_1 Z$$

2. X is Normally distributed given Z

$$X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$

$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta_X'(\bar{u})e^{\beta_X(\bar{u})}$

Imputation model: Z categorical

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

$H_0(T) = \int_0^T h_0(u)du$

$H_0^{(1)}(T) = \int_0^T u h_0(u)du$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

$$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$$

$$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta_X'(\bar{u})e^{\beta_X(\bar{u})}$$

Imputation model: Z categorical

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

$$H_0(T) = \int_0^T h_0(u)du$$

$$H_0^{(1)}(T) = \int_0^T u h_0(u)du$$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z$$
$$- \int_0^T h_0(u) e^{\beta_X(u)X + \beta_Z Z} \mathrm{d}u + const$$

$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$

$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u - \bar{u})\beta_X'(\bar{u})e^{\beta_X(\bar{u})}$

Imputation model: Z categorical

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T)$$
$$+ \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

$H_0(T) = \int_0^T h_0(u)du$

$H_0^{(1)}(T) = \int_0^T u h_0(u)du$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z$$
$$- \int_0^T h_0(u) e^{\beta_X(u)X+\beta_Z Z} \mathrm{d}u + const$$

$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$

$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta'_X(\bar{u})e^{\beta_X(\bar{u})}$

Imputation model: Z categorical

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T)$$
$$+\alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

$H_0(T) = \int_0^T h_0(u)du$

$H_0^{(1)}(T) = \int_0^T u h_0(u)du$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z$$
$$- \int_0^T h_0(u) e^{\beta_X(u)X + \beta_Z Z} \mathrm{d}u + const$$

$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$

$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta'_X(\bar{u})e^{\beta_X(\bar{u})}$

Imputation model: Z categorical

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T)$$
$$+\alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

$H_0(T) = \int_0^T h_0(u)du$

$H_0^{(1)}(T) = \int_0^T u h_0(u)du$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X + \beta_Z Z} \mathrm{d}u + const$$

$$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$$

$$e^{\beta_Z Z} \approx e^{\beta_Z \bar{Z}} + (Z-\bar{Z})\beta_Z e^{\beta_Z \bar{Z}}$$

$$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta'_X(\bar{u}) e^{\beta_X(\bar{u})}$$

Imputation model: Z continuous

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X+\beta_Z Z} \mathrm{d}u + const$$

$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$

$e^{\beta_Z Z} \approx e^{\beta_Z \bar{Z}} + (Z-\bar{Z})\beta_Z e^{\beta_Z \bar{Z}}$

$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u-\bar{u})\beta'_X(\bar{u}) e^{\beta_X(\bar{u})}$

Imputation model: Z continuous

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

Binary X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X + \beta_Z Z} \mathrm{d}u + const$$

$$\text{logit } p(X=1|Z) = \zeta_0 + \zeta_1 Z$$

$$e^{\beta_Z Z} \approx e^{\beta_Z \bar{Z}} + (Z - \bar{Z})\beta_Z e^{\beta_Z \bar{Z}}$$

$$e^{\beta_X(u)} \approx e^{\beta_X(\bar{u})} + (u - \bar{u})\beta_X'(\bar{u}) e^{\beta_X(\bar{u})}$$

Imputation model: Z continuous

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

Normally distributed X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X+\beta_Z Z} \mathrm{d}u + const$$

$X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$

linear or quadratic approximation for $e^{\beta_X(u)X+\beta_Z Z}$

Imputation model: a linear regression

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T) + \varepsilon$$

- Approximation assumes that $\beta_X(u)$, β_Z or $H_0(T)$ is small

Normally distributed X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X+\beta_Z Z} \mathrm{d}u + const$$

$X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$

linear or quadratic approximation for $e^{\beta_X(u)X+\beta_Z Z}$

Imputation model: a linear regression

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T) + \varepsilon$$

- Approximation assumes that $\beta_X(u)$, β_Z or $H_0(T)$ is small

Normally distributed X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u) e^{\beta_X(u)X + \beta_Z Z} \mathrm{d}u + const$$

$X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$

linear or quadratic approximation for $e^{\beta_X(u)X+\beta_Z Z}$

Imputation model: a linear regression

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T) + \varepsilon$$

- Approximation assumes that $\beta_X(u)$, β_Z or $H_0(T)$ is small

Normally distributed X

General result

$$\log p(X|T,D,Z) = \log p(X|Z) + D\beta_X(T)X + D\beta_Z Z - \int_0^T h_0(u)e^{\beta_X(u)X+\beta_Z Z}\mathrm{d}u + const$$

$X|Z \sim N(\zeta_0 + \zeta_1 Z, \sigma^2)$

linear or quadratic approximation for $e^{\beta_X(u)X+\beta_Z Z}$

Imputation model: a linear regression

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T) + \varepsilon$$

- Approximation assumes that $\beta_X(u)$, β_Z or $H_0(T)$ is small

Summary

Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T) + \alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

- Breslow's estimate

$$\widehat{H}_0(T) = \sum_{t \le T} \frac{1}{\sum_{R(t)} e^{\hat{\beta}_X(t)X + \hat{\beta}_Z Z}}$$

- The Nelson-Aalen estimate

$$\widehat{H}(T) = \sum_{t \le T} \frac{\text{number of events at } t}{\text{number at risk at } t}$$

- A Nelson-Aalen-type estimate

$$\widehat{H}^{(1)}(T) = \sum_{t \le T} \frac{t \times \text{number of events at } t}{\text{number at risk at } t}$$

Summary

Imputation model

$$\text{logit } p(X = 1|T, D, Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T)$$
$$+\alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

- Breslow's estimate

$$\widehat{H}_0(T) = \sum_{t \leq T} \frac{1}{\sum_{R(t)} e^{\hat{\beta}_X(t)X + \hat{\beta}_Z Z}}$$

- The Nelson-Aalen estimate

$$\widehat{H}(T) = \sum_{t \leq T} \frac{\text{number of events at } t}{\text{number at risk at } t}$$

- A Nelson-Aalen-type estimate

$$\widehat{H}^{(1)}(T) = \sum_{t \leq T} \frac{t \times \text{number of events at } t}{\text{number at risk at } t}$$

Summary

Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T)$$
$$+\alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

- Breslow's estimate

$$\widehat{H}_0(T) = \sum_{t \leq T} \frac{1}{\sum_{R(t)} e^{\hat{\beta}_X(t)X + \hat{\beta}_Z Z}}$$

- The Nelson-Aalen estimate

$$\widehat{H}(T) = \sum_{t \leq T} \frac{\text{number of events at } t}{\text{number at risk at } t}$$

- A Nelson-Aalen-type estimate

$$\widehat{H}^{(1)}(T) = \sum_{t \leq T} \frac{t \times \text{number of events at } t}{\text{number at risk at } t}$$

Summary

Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 H_0(T)$$
$$+\alpha_5 Z H_0(T) + \alpha_4 H_0^{(1)}(T) + \alpha_6 Z H_0^{(1)}(T)$$

- Breslow's estimate

$$\widehat{H}_0(T) = \sum_{t\leq T} \frac{1}{\sum_{R(t)} e^{\hat{\beta}_X(t)X + \hat{\beta}_Z Z}}$$

- The Nelson-Aalen estimate

$$\widehat{H}(T) = \sum_{t\leq T} \frac{\text{number of events at } t}{\text{number at risk at } t}$$

- A Nelson-Aalen-type estimate

$$\widehat{H}^{(1)}(T) = \sum_{t\leq T} \frac{t \times \text{number of events at } t}{\text{number at risk at } t}$$

Summary

Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 \widehat{H}(T)$$
$$+\alpha_5 Z\widehat{H}(T) + \alpha_4 \widehat{H}^{(1)}(T) + \alpha_6 Z\widehat{H}^{(1)}(T)$$

- Breslow's estimate

$$\widehat{H}_0(T) = \sum_{t \le T} \frac{1}{\sum_{R(t)} e^{\hat{\beta}_X(t)X + \hat{\beta}_Z Z}}$$

- The Nelson-Aalen estimate

$$\widehat{H}(T) = \sum_{t \le T} \frac{\text{number of events at } t}{\text{number at risk at } t}$$

- A Nelson-Aalen-type estimate

$$\widehat{H}^{(1)}(T) = \sum_{t \le T} \frac{t \times \text{number of events at } t}{\text{number at risk at } t}$$

Summary

Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 \widehat{H}(T) + \alpha_5 Z\widehat{H}(T) + \alpha_4 \widehat{H}^{(1)}(T) + \alpha_6 Z\widehat{H}^{(1)}(T)$$

`mice` in R, `mi impute` in Stata

In simulations we investigate...

- What happens if we ignore the time-varying effect in the imputation (White & Royston method)?
- When are the $\widehat{H}^{(1)}$ terms needed?
- When are the interactions terms needed?
- Does the approximation required for the linear regression situation perform well?

Summary

Imputation model

$$\text{logit}\ p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_2 D\beta_X(T) + \alpha_3 \widehat{H}(T)$$
$$+\alpha_5 Z\widehat{H}(T) + \alpha_4 \widehat{H}^{(1)}(T) + \alpha_6 Z\widehat{H}^{(1)}(T)$$

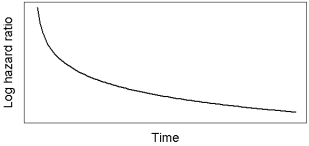
`mice` in R, `mi impute` in Stata

In simulations we investigate...

- What happens if we ignore the time-varying effect in the imputation (White & Royston method)?
- When are the $\widehat{H}^{(1)}$ terms needed?
- When are the interactions terms needed?
- Does the approximation required for the linear regression situation perform well?

Specific example: log time interaction analysis

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_{XT}\log(T)X + \beta_Z Z}$$



Imputation model

$$\text{logit } p(X=1|T,D,Z) \approx \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T)$$
$$+\alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z\widehat{H}(T) + \alpha_6 Z\widehat{H}^{(1)}(T)$$

Extension to missingness in several variables

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

- Often we will have missing data in Z as well as X
- This can be handled using multiple imputation by chained equations (MICE), aka fully conditional specification (FCS)
- We specify models for
  - $X|Z,T,D$
  - $Z|X,T,D$

Outline

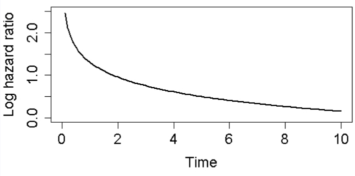
1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Simulation study

- Cohort of 5000 people followed for 10 years
- Binary or normally distributed exposure X
- Normally distributed covariate Z: $corr(X,Z) = 0.5$

Hazard model

$$h(T|X,Z) = \lambda \exp\{\beta_X X + \beta_{XT} X(\log T - \log 5) + \beta_Z Z\}$$



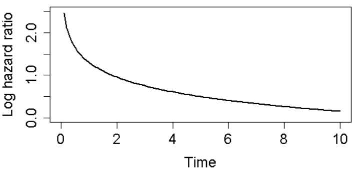
- 10% have the event
- Missing data in 20% of X and 20% of Z (MCAR)

Simulation study

- Cohort of 5000 people followed for 10 years
- Binary or normally distributed exposure X
- Normally distributed covariate Z: $corr(X,Z) = 0.5$

Hazard model

$$h(T|X,Z) = \lambda \exp\{\beta_X X + \beta_{XT} X(\log T - \log 5) + \beta_Z Z\}$$



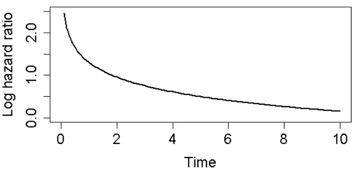
- 10% have the event
- Missing data in 20% of X and 20% of Z (MCAR)

Simulation study

- Cohort of 5000 people followed for 10 years
- Binary or normally distributed exposure X
- Normally distributed covariate Z: $corr(X,Z) = 0.5$

Hazard model

$$h(T|X,Z) = \lambda \exp\{\beta_X X + \beta_{XT} X(\log T - \log 5) + \beta_Z Z\}$$



- 10% have the event
- Missing data in 20% of X and 20% of Z (MCAR)

Simulation study

Modelling the time-varying effect

1. Log-time analysis: $\beta_X(T) = \beta_X + \beta_{XT}\{\log T - \log 5\}$
2. Step function analysis: using 4 time periods

Analyses performed

1. Complete-data analysis
2. Complete-case analysis
3. MI non-time-varying approach: White & Royston method
4. MI time-varying approach

Imputation model

$$X = \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T)$$
$$+ \alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z \widehat{H}(T) + \alpha_6 Z \widehat{H}^{(1)}(T) + \varepsilon$$

Simulation study

Modelling the time-varying effect

1. Log-time analysis: $\beta_X(T) = \beta_X + \beta_{XT}\{\log T - \log 5\}$
2. Step function analysis: using 4 time periods

Analyses performed

1. Complete-data analysis
2. Complete-case analysis
3. MI non-time-varying approach: White & Royston method
4. MI time-varying approach

Imputation model

$$X = \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T)$$
$$+ \alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z \widehat{H}(T) + \alpha_6 Z \widehat{H}^{(1)}(T) + \varepsilon$$

Simulation study

Modelling the time-varying effect

1. Log-time analysis: $\beta_X(T) = \beta_X + \beta_{XT}\{\log T - \log 5\}$
2. Step function analysis: using 4 time periods

Analyses performed

1. Complete-data analysis
2. Complete-case analysis
3. MI non-time-varying approach: White & Royston method
4. MI time-varying approach

Imputation model

$$X = \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T)$$
$$+ \alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z \widehat{H}(T) + \alpha_6 Z \widehat{H}^{(1)}(T) + \varepsilon$$

Simulation study

Modelling the time-varying effect

1. Log-time analysis: $\beta_X(T) = \beta_X + \beta_{XT}\{\log T - \log 5\}$
2. Step function analysis: using 4 time periods

Analyses performed

1. Complete-data analysis
2. Complete-case analysis
3. MI non-time-varying approach: White & Royston method
4. MI time-varying approach

Imputation model

$$X = \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T)$$
$$+ \alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z \widehat{H}(T) + \alpha_6 Z \widehat{H}^{(1)}(T) + \varepsilon$$

Simulation study

Modelling the time-varying effect

1. Log-time analysis: $\beta_X(T) = \beta_X + \beta_{XT}\{\log T - \log 5\}$
2. Step function analysis: using 4 time periods

Analyses performed

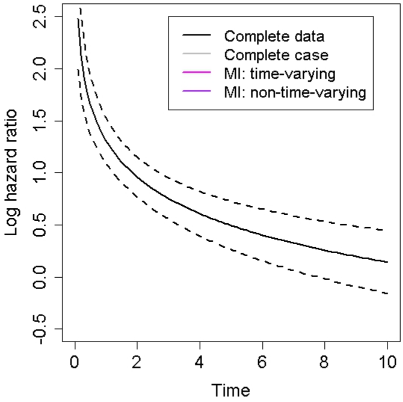
1. Complete-data analysis
2. Complete-case analysis
3. MI non-time-varying approach: White & Royston method
4. MI time-varying approach

Imputation model

$$X = \alpha_0 + \alpha_1 Z + \alpha_{21} D + \alpha_{22} D \log(T) + \alpha_3 \widehat{H}(T) + \alpha_4 \widehat{H}^{(1)}(T) + \alpha_5 Z \widehat{H}(T) + \alpha_6 Z \widehat{H}^{(1)}(T) + \varepsilon$$

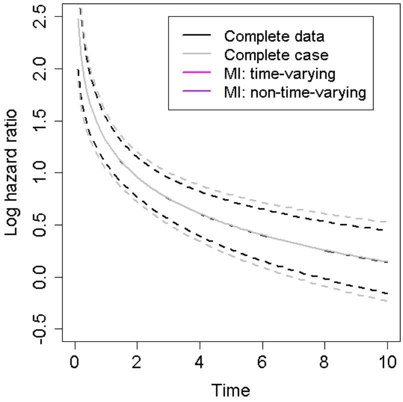
Results: Log-time analysis, Normal X

	$\beta_X = 0.5, \beta_{XT} = -0.5$ $\beta_Z = 0.5$		
	Est	bias	eff
Complete data			
β_X	0.501	0.001	100
β_{XT}	-0.502	-0.002	100
β_Z	0.498	-0.002	100
Complete case			
β_X	0.493	-0.007	63
β_{XT}	-0.505	-0.005	67
β_Z	0.502	0.002	63
Time-varying MI			
β_X	0.497	-0.003	74
β_{XT}	-0.506	-0.006	87
β_Z	0.500	-0.000	76
Non-time-varying MI			
β_X	0.492	-0.008	75
β_{XT}	-0.425	0.075	123
β_Z	0.500	0.000	77



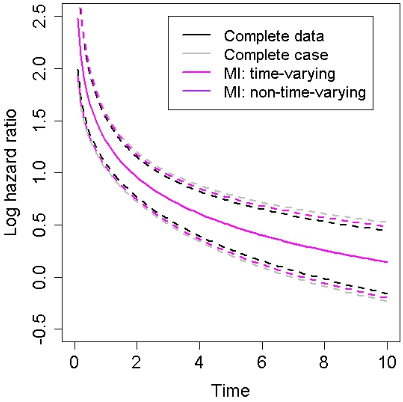
Results: Log-time analysis, Normal X

	$\beta_X = 0.5, \beta_{XT} = -0.5$ $\beta_Z = 0.5$		
	Est	bias	eff
Complete data			
β_X	0.501	0.001	100
β_{XT}	-0.502	-0.002	100
β_Z	0.498	-0.002	100
Complete case			
β_X	0.493	-0.007	63
β_{XT}	-0.505	-0.005	67
β_Z	0.502	0.002	63
Time-varying MI			
β_X	0.497	-0.003	74
β_{XT}	-0.506	-0.006	87
β_Z	0.500	-0.000	76
Non-time-varying MI			
β_X	0.492	-0.008	75
β_{XT}	-0.425	0.075	123
β_Z	0.500	0.000	77



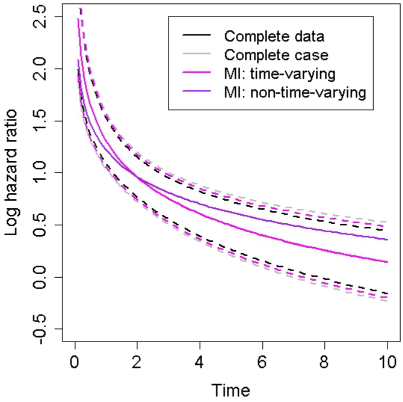
Results: Log-time analysis, Normal X

	$\beta_X = 0.5, \beta_{XT} = -0.5$		
	$\beta_Z = 0.5$		
	Est	bias	eff
Complete data			
β_X	0.501	0.001	100
β_{XT}	-0.502	-0.002	100
β_Z	0.498	-0.002	100
Complete case			
β_X	0.493	-0.007	63
β_{XT}	-0.505	-0.005	67
β_Z	0.502	0.002	63
Time-varying MI			
β_X	0.497	-0.003	74
β_{XT}	-0.506	-0.006	87
β_Z	0.500	-0.000	76
Non-time-varying MI			
β_X	0.492	-0.008	75
β_{XT}	-0.425	0.075	123
β_Z	0.500	0.000	77



Results: Log-time analysis, Normal X

	$\beta_X = 0.5, \beta_{XT} = -0.5$		
	$\beta_Z = 0.5$		
	Est	bias	eff
Complete data			
β_X	0.501	0.001	100
β_{XT}	-0.502	-0.002	100
β_Z	0.498	-0.002	100
Complete case			
β_X	0.493	-0.007	63
β_{XT}	-0.505	-0.005	67
β_Z	0.502	0.002	63
Time-varying MI			
β_X	0.497	-0.003	74
β_{XT}	-0.506	-0.006	87
β_Z	0.500	-0.000	76
Non-time-varying MI			
β_X	0.492	-0.008	75
β_{XT}	-0.425	0.075	123
β_Z	0.500	0.000	77



Results: Log-time analysis, Normal X

	$\beta_X = 0.5, \beta_{XT} = -0.5$		
	$\beta_Z = 0.5$		
	Est	bias	eff
Time-varying MI			
β_X	0.497	-0.003	74
β_{XT}	-0.506	-0.006	87
β_Z	0.500	-0.000	76
Time-varying MI: + $\widehat{H}^{(1)}(T)$"			
β_X	0.497	-0.003	75
β_{XT}	-0.508	-0.008	86
β_Z	0.499	-0.001	75
Time-varying MI: + interactions			
β_X	0.497	-0.003	74
β_{XT}	-0.507	-0.007	86
β_Z	0.500	0.000	75

Results: Log-time analysis, Normal X

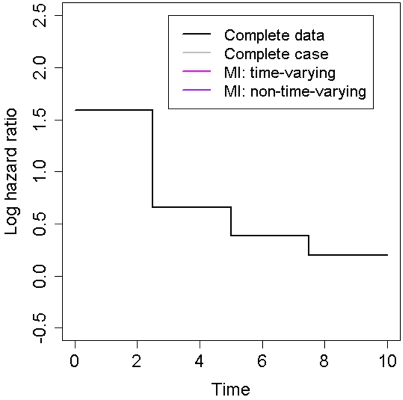
	$\beta_X = 0.5, \beta_{XT} = -0.5$ $\beta_Z = 0.5$			$\beta_X = 1.5, \beta_{XT} = -0.5$ $\beta_Z = 0.5$		
	Est	bias	eff	Est	bias	eff
Time-varying MI						
β_X	0.497	-0.003	74	1.485	-0.015	75
β_{XT}	-0.506	-0.006	87	-0.506	-0.006	89
β_Z	0.500	-0.000	76	0.498	-0.002	78
Time-varying MI: + interactions						
β_X	0.497	-0.003	74	1.488	-0.012	75
β_{XT}	-0.507	-0.007	86	-0.501	-0.001	88
β_Z	0.500	0.000	75	0.500	0.000	78

Results: Log-time analysis, Normal X

	10% have event			30% have event		
	Est	bias	eff	Est	bias	eff
Time-varying MI						
β_X	0.497	-0.003	74	0.494	-0.006	78
β_{XT}	-0.506	-0.006	87	-0.500	-0.000	87
β_Z	0.500	-0.000	76	0.502	0.002	75
Time-varying MI: + interactions						
β_X	0.497	-0.003	74	0.496	-0.004	79
β_{XT}	-0.507	-0.007	86	-0.498	0.002	87
β_Z	0.500	0.000	75	0.503	0.003	76

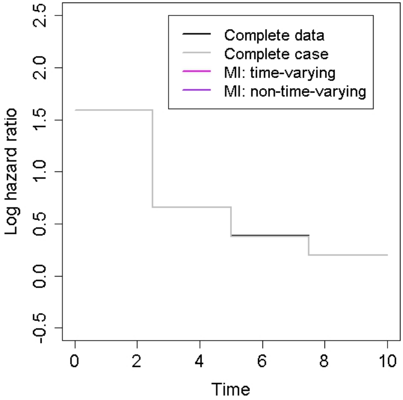
Results: step function analysis

Parameter	Est	% Bias	cov	eff
Complete case				
β_{X1}	1.116	-0.005	62	
β_{X2}	0.650	-0.005	64	
β_{X3}	0.376	-0.009	65	
β_{X4}	0.214	-0.010	61	
β_{Z}	0.502	0.002	63	
MI: time-varying method				
β_{X1}	1.121	-0.001	85	
β_{X2}	0.648	-0.007	80	
β_{X3}	0.379	-0.006	79	
β_{X4}	0.218	-0.006	78	
β_{Z}	0.500	-0.000	76	
MI: non-time-varying method				
β_{X1}	1.019	-0.103	110	
β_{X2}	0.619	-0.036	102	
β_{X3}	0.394	0.009	98	
β_{X4}	0.256	0.033	95	
β_{Z}	0.500	0.000	76	



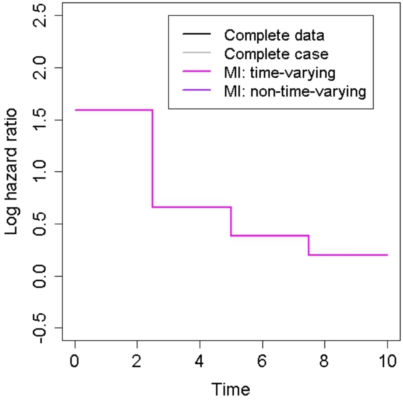
Results: step function analysis

Parameter	Est	% Bias	cov	eff
Complete case				
β_{X1}	1.116	-0.005	62	
β_{X2}	0.650	-0.005	64	
β_{X3}	0.376	-0.009	65	
β_{X4}	0.214	-0.010	61	
β_{Z}	0.502	0.002	63	
MI: time-varying method				
β_{X1}	1.121	-0.001	85	
β_{X2}	0.648	-0.007	80	
β_{X3}	0.379	-0.006	79	
β_{X4}	0.218	-0.006	78	
β_{Z}	0.500	-0.000	76	
MI: non-time-varying method				
β_{X1}	1.019	-0.103	110	
β_{X2}	0.619	-0.036	102	
β_{X3}	0.394	0.009	98	
β_{X4}	0.256	0.033	95	
β_{Z}	0.500	0.000	76	



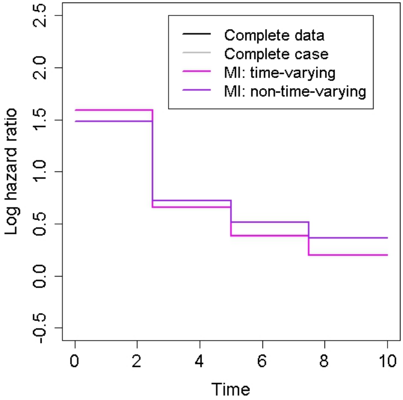
Results: step function analysis

Parameter	Est	% Bias	cov	eff
Complete case				
β_{X1}	1.116	-0.005	62	
β_{X2}	0.650	-0.005	64	
β_{X3}	0.376	-0.009	65	
β_{X4}	0.214	-0.010	61	
β_{Z}	0.502	0.002	63	
MI: time-varying method				
β_{X1}	1.121	-0.001	85	
β_{X2}	0.648	-0.007	80	
β_{X3}	0.379	-0.006	79	
β_{X4}	0.218	-0.006	78	
β_{Z}	0.500	-0.000	76	
MI: non-time-varying method				
β_{X1}	1.019	-0.103	110	
β_{X2}	0.619	-0.036	102	
β_{X3}	0.394	0.009	98	
β_{X4}	0.256	0.033	95	
β_{Z}	0.500	0.000	76	



Results: step function analysis

Parameter	Est	% Bias	cov	eff
Complete case				
β_{X1}	1.116	-0.005	62	
β_{X2}	0.650	-0.005	64	
β_{X3}	0.376	-0.009	65	
β_{X4}	0.214	-0.010	61	
β_{Z}	0.502	0.002	63	
MI: time-varying method				
β_{X1}	1.121	-0.001	85	
β_{X2}	0.648	-0.007	80	
β_{X3}	0.379	-0.006	79	
β_{X4}	0.218	-0.006	78	
β_{Z}	0.500	-0.000	76	
MI: non-time-varying method				
β_{X1}	1.019	-0.103	110	
β_{X2}	0.619	-0.036	102	
β_{X3}	0.394	0.009	98	
β_{X4}	0.256	0.033	95	
β_{Z}	0.500	0.000	76	



Testing the proportional hazards assumption

Hazard model

Data generated using the hazard model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_{XT}(\log T - \log 5)X + \beta_Z Z}$$

with $\beta_{XT} = 0$

Percentage of simulations in which the null hypothesis $\beta_{XT} = 0$ was rejected:

Complete data	5.0%
Complete case	5.3%
MI: time-varying method	5.3%
MI: non-timevarying method	2.2%

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Illustration: Rotterdam breast Cancer Study

- 2982 individuals with primary breast cancer from the Rotterdam tumour bank
- Individuals followed-up for death/disease recurrence (51%)
- **Sauerbrei et al (2007), Royston & Sauerbrei (2008)**: time-varying effects of two variables
  - tumour size: $log(T)$
  - number of progesterone receptors ($log(pgr + 1)$): $log(T)$
- I generated missing data for 20% of individuals in both variables

Illustration: Rotterdam breast Cancer Study

- 2982 individuals with primary breast cancer from the Rotterdam tumour bank
- Individuals followed-up for death/disease recurrence (51%)
- **Sauerbrei et al (2007), Royston & Sauerbrei (2008)**: time-varying effects of two variables
  - tumour size: $log(T)$
  - number of progesterone receptors ($log(pgr + 1)$): $log(T)$
- I generated missing data for 20% of individuals in both variables

Illustration: Rotterdam breast Cancer Study

- 2982 individuals with primary breast cancer from the Rotterdam tumour bank
- Individuals followed-up for death/disease recurrence (51%)
- **Sauerbrei et al (2007), Royston & Sauerbrei (2008)**: time-varying effects of two variables
  - tumour size: $log(T)$
  - number of progesterone receptors ($log(pgr + 1)$): $log(T)$
- I generated missing data for 20% of individuals in both variables

Illustration: Rotterdam Breast Cancer Study

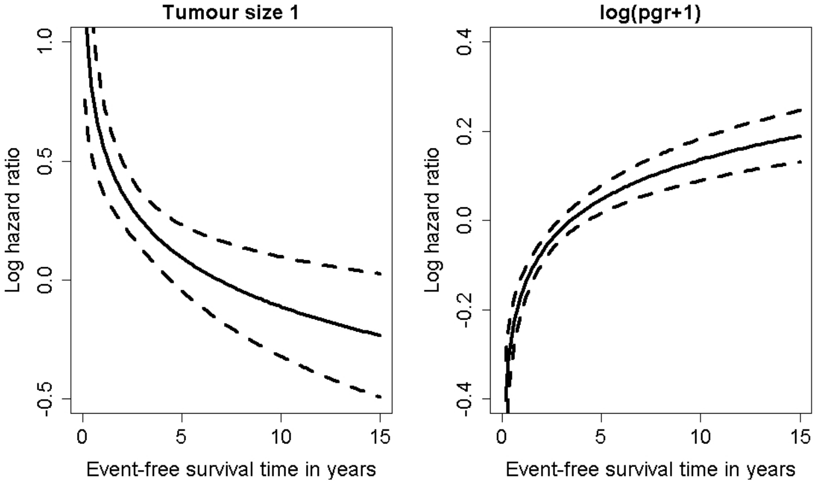


Illustration: Rotterdam Breast Cancer Study

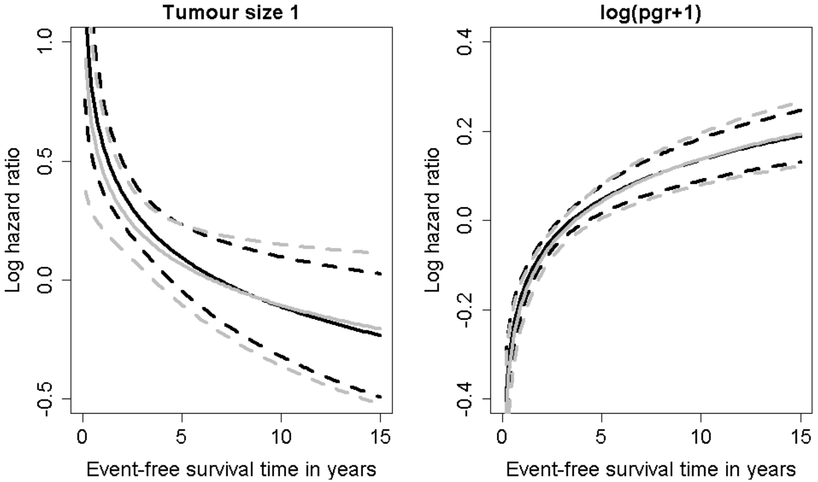


Illustration: Rotterdam Breast Cancer Study

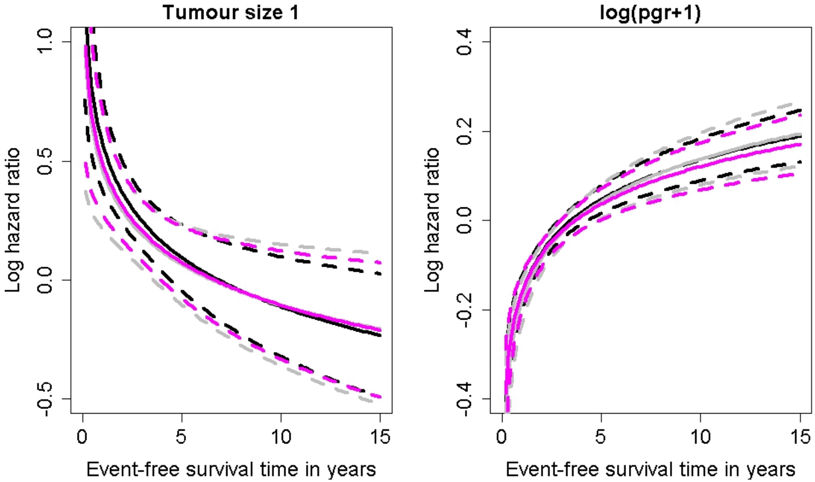
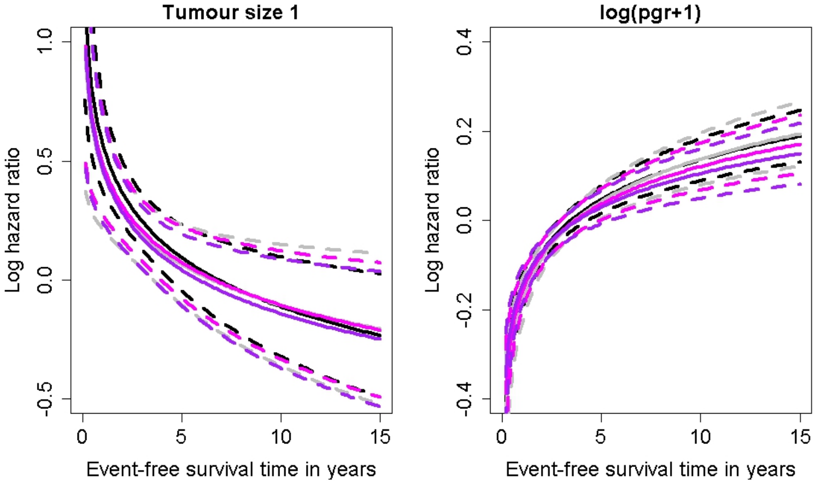


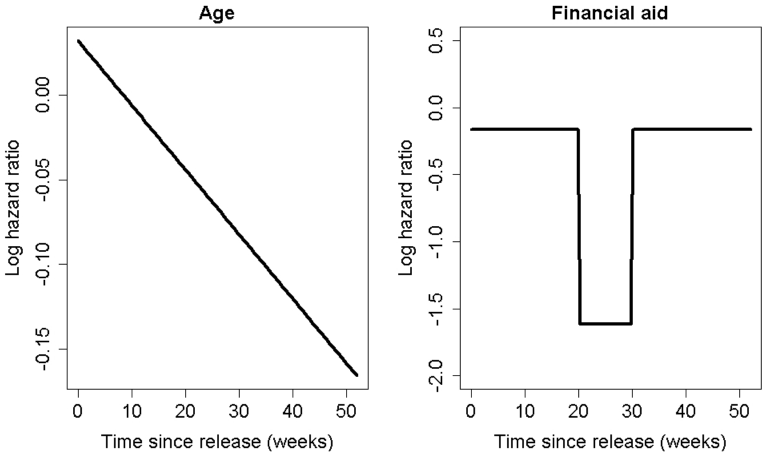
Illustration: Rotterdam Breast Cancer Study



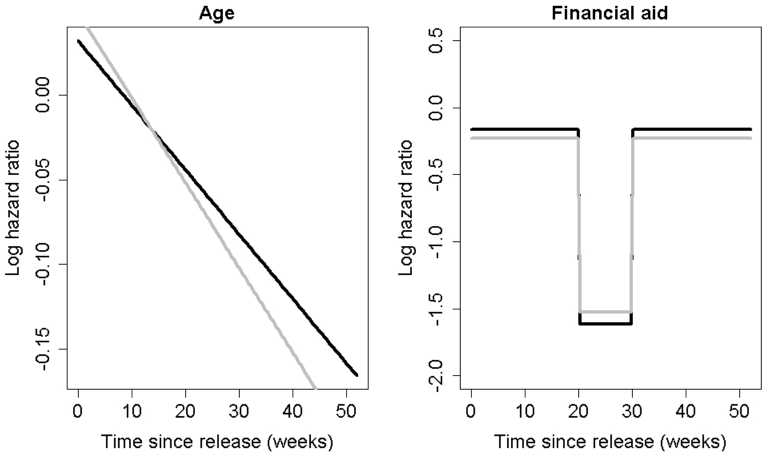
Another example: Arrest after release from prison

- 432 inmates released from state prison followed up for 1 year (Allison et al (2010))
- Factors associated with re-arrest:
  - Age: time-varying effect (linear with time since release)
  - Financial aid: step function, with a step 20-30 weeks after release
  - Prior arrests: no time-varying effect
- 20% missingness introduced in age and financial aid

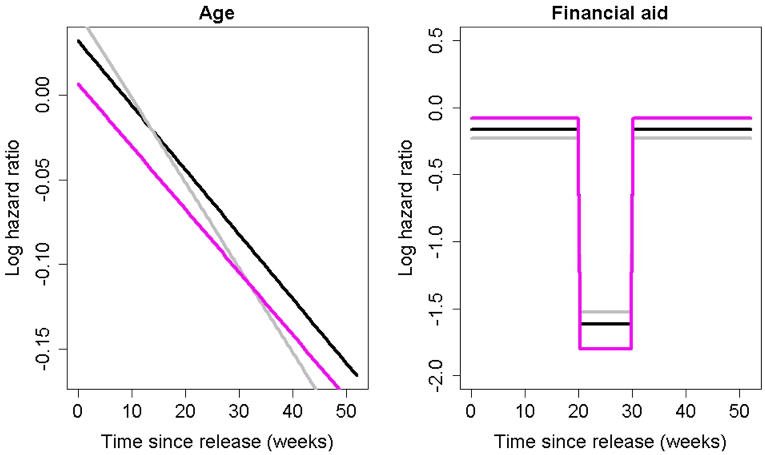
Another example: Arrest after release from prison



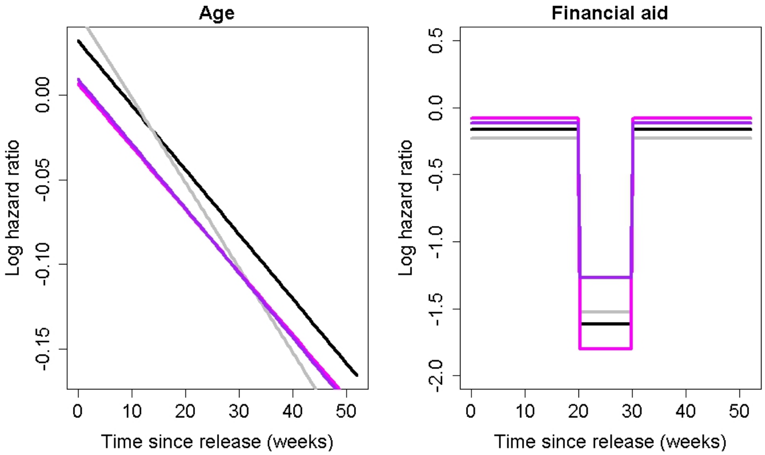
Another example: Arrest after release from prison



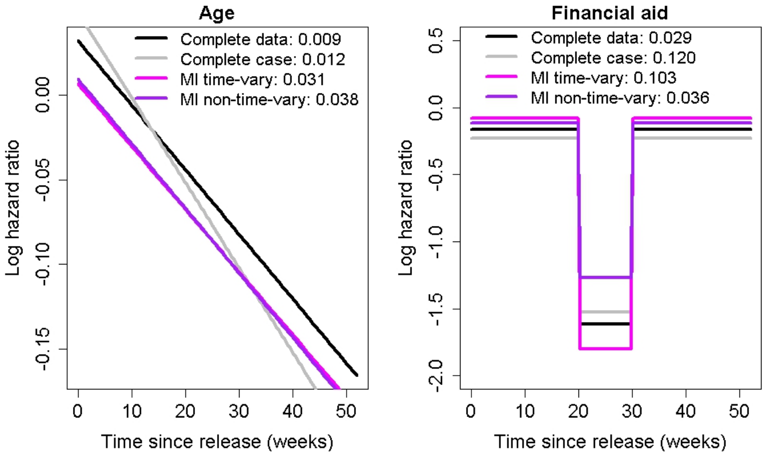
Another example: Arrest after release from prison



Another example: Arrest after release from prison



Another example: Arrest after release from prison



Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Another approach for imputation under the Cox model

- We have focused on an *approximate* imputation model for $p(X|T, D, Z)$
- This does not extend to allowing non-linear terms (e.g. X^2) or interaction terms

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–26

DOI: 10.1177/0962280214521348
smm.sagepub.com
SAGE

Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model

**Jonathan W Bartlett,[1] Shaun R Seaman,[2]
Ian R White[2] and James R Carpenter[1,3] for the Alzheimer's Disease Neuroimaging Initiative***

Another approach for imputation under the Cox model

- We have focused on an *approximate* imputation model for $p(X|T, D, Z)$
- This does not extend to allowing non-linear terms (e.g. X^2) or interaction terms

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–26

DOI: 10.1177/0962280214521348
smm.sagepub.com
SAGE

Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model

Jonathan W Bartlett,[1] Shaun R Seaman,[2] Ian R White[2] and James R Carpenter[1,3] for the Alzheimer's Disease Neuroimaging Initiative*

The Bartlett et al. approach

Aim

- For variable X with missing data and fully-observed variable Z
- To impute missing values of X by drawing from the true distribution $p(X|T,D,Z)$

The basic idea...

- Draw *potential* values of X from a proposal distribution $p(X|Z)$
- Use a rejection rule to decide whether or not to accept the potential imputed values of X as imputed values from the desired distribution $p(X|T,D,Z)$

The method does not currently accommodate time-varying effects of exposures

The Bartlett et al. approach

Aim

- For variable X with missing data and fully-observed variable Z
- To impute missing values of X by drawing from the true distribution $p(X|T,D,Z)$

The basic idea...

- Draw *potential* values of X from a proposal distribution $p(X|Z)$
- Use a rejection rule to decide whether or not to accept the potential imputed values of X as imputed values from the desired distribution $p(X|T,D,Z)$

The method does not currently accommodate time-varying effects of exposures

The Bartlett et al. approach

Aim

- For variable X with missing data and fully-observed variable Z
- To impute missing values of X by drawing from the true distribution $p(X|T,D,Z)$

The basic idea...

- Draw *potential* values of X from a proposal distribution $p(X|Z)$
- Use a rejection rule to decide whether or not to accept the potential imputed values of X as imputed values from the desired distribution $p(X|T,D,Z)$

The method does not currently accommodate time-varying effects of exposures

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

1. Obtain initial estimates for β_X, β_Z and their covariance
2. Draw values $\beta_X^{(m)}, \beta_Z^{(m)}$ from their estimated distribution
3. Fit the proposal distribution $p(X|Z)$ and draw parameter values from their estimated joint distribution
4. Draw a value X^* from the proposal distribution
5. Draw a value $U \sim \mathrm{Uniform}(0,1)$. Accept the value X^* if...

$$\begin{cases} U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)} X^* + \beta_Z^{(m)} Z}\} & \text{if } D = 0 \\ U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)} X^* + \beta_Z^{(m)} Z - H_0^{(m)}(T)e^{\beta_X^{(m)} X^* + \beta_Z^{(m)} Z}\} & \text{if } D = 1 \end{cases}$$

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

1. Obtain initial estimates for β_X, β_Z and their covariance
2. Draw values $\beta_X^{(m)}, \beta_Z^{(m)}$ from their estimated distribution
3. Fit the proposal distribution $p(X|Z)$ and draw parameter values from their estimated joint distribution
4. Draw a value X^* from the proposal distribution
5. Draw a value $U \sim \mathrm{Uniform}(0,1)$. Accept the value X^* if...

$$\begin{cases} U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 0 \\ U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 1 \end{cases}$$

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

1. Obtain initial estimates for β_X, β_Z and their covariance
2. Draw values $\beta_X^{(m)}, \beta_Z^{(m)}$ from their estimated distribution
3. Fit the proposal distribution $p(X|Z)$ and draw parameter values from their estimated joint distribution
4. Draw a value X^* from the proposal distribution
5. Draw a value $U \sim \mathrm{Uniform}(0,1)$. Accept the value X^* if...

$$\begin{cases} U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 0 \\ U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 1 \end{cases}$$

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

1. Obtain initial estimates for β_X, β_Z and their covariance
2. Draw values $\beta_X^{(m)}, \beta_Z^{(m)}$ from their estimated distribution
3. Fit the proposal distribution $p(X|Z)$ and draw parameter values from their estimated joint distribution
4. Draw a value X^* from the proposal distribution
5. Draw a value $U \sim \mathrm{Uniform}(0,1)$. Accept the value X^* if...

$$\begin{cases} U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D=0 \\ U \leq H_0^{(m)}(T)\exp\{1+\beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D=1 \end{cases}$$

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

1. Obtain initial estimates for β_X, β_Z and their covariance
2. Draw values $\beta_X^{(m)}, \beta_Z^{(m)}$ from their estimated distribution
3. Fit the proposal distribution $p(X|Z)$ and draw parameter values from their estimated joint distribution
4. Draw a value X^* from the proposal distribution
5. Draw a value $U \sim \mathrm{Uniform}(0,1)$. Accept the value X^* if...

$$\begin{cases} U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 0 \\ U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} & \text{if } D = 1 \end{cases}$$

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

$$\begin{aligned} &U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} && \text{if } D = 0 \\ &U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} && \text{if } D = 1 \end{aligned}$$

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X + \beta_Z Z}$$

$$\begin{aligned} U \leq &\quad \exp\{-\textstyle\int_0^T h_0(u)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(u)X^*}\mathrm{d}u\} \\ U \leq &\quad h_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(T)X^* - \textstyle\int_0^T h_0(u)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(u)X^*}\mathrm{d}u\} \end{aligned}$$

- The standard approach can be applied in R and Stata using Jonathan Bartlett's package `smcfcs`

Extending the Bartlett et al. approach

Cox proportional hazards model

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_Z Z}$$

$$U \leq \exp\{-H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} \quad \text{if } D = 0$$
$$U \leq H_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z - H_0^{(m)}(T)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z}\} \quad \text{if } D = 1$$

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X + \beta_Z Z}$$

$$U \leq \exp\{-\int_0^T h_0(u)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(u)X^*}\mathrm{d}u\}$$
$$U \leq h_0^{(m)}(T)\exp\{1 + \beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(T)X^* - \int_0^T h_0(u)e^{\beta_X^{(m)}X^* + \beta_Z^{(m)}Z + \beta_{XT}^{(m)}(u)X^*}\mathrm{d}u\}$$

- The standard approach can be applied in R and Stata using Jonathan Bartlett's package `smcfcs`

Extending the Bartlett et al. approach: Some results

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X X + \beta_{XT}\log(T)X + \beta_Z Z}$$

- I simulated data for binary X and continuous Z
- Missing data on X were generated for 20% of individuals

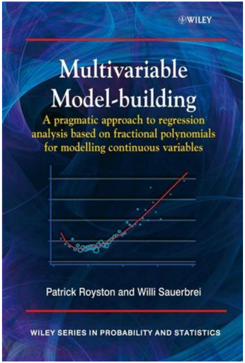
	β_X	β_{XT}	β_Z
Complete data	0.47 (0.32)	-0.53 (0.18)	0.51 (0.30)
Complete case	0.46 (0.36)	-0.55 (0.22)	0.44 (0.34)
MI: non-time-varying	0.57 (0.34)	-0.42 (0.15)	0.51 (0.31)
MI: Approx method	0.46 (0.36)	-0.54 (0.21)	0.51 (0.31)
MI: Extended Bartlett	0.47 (0.31)	-0.53 (0.18)	0.51 (0.30)

Outline

1. Handling missing data on explanatory variables in Cox regression
2. Modelling time-varying effects in Cox regression
3. Derive an imputation model which handles time-varying effects
4. Simulation study
5. An application
6. An alternative approach
7. Further work

Allowing a flexible functional form

- Everything so far requires us to specify the functional form for the time-varying effects $\beta_X(T)$
- An alternative is to somehow select a 'best' functional form
- Sauerbrei et al. (2007), Royston & Sauerbrei (2008): Using fractional polynomials to model time-varying effects



Allowing a flexible functional form

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

Using a fractional polynomial of degree 1

$$\beta_X(T) = \beta_{X0} + \beta_{X1}\, T^p.$$

The best power p is selected from set $\{-2,-1,-0.5,0,0.5,1,2,3\}$

Aim

- Incorporate MI within this approach
- By allowing accommodating a flexible functional form for $\beta_X(T)$ in the imputation model
- By selecting the best fitting FP using the imputed data sets

Allowing a flexible functional form

Extended Cox model with time-varying effects

$$h(T|X,Z) = h_0(T)e^{\beta_X(T)X+\beta_Z Z}$$

Using a fractional polynomial of degree 1

$$\beta_X(T) = \beta_{X0} + \beta_{X1}T^p.$$

The best power p is selected from set $\{-2,-1,-0.5,0,0.5,1,2,3\}$

Aim

- Incorporate MI within this approach
- By allowing accommodating a flexible functional form for $\beta_X(T)$ in the imputation model
- By selecting the best fitting FP using the imputed data sets

Allowing a flexible functional form

Research Article

Statistics in Medicine

Received 13 November 2014, Accepted 19 May 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6553

Combining fractional polynomial model building with multiple imputation

Tim P. Morris,[a,b*†] Ian R. White,[c] James R. Carpenter,[a,b] Simon J. Stanworth[d] and Patrick Royston[a]

Summary comments

- We should incorporate time-varying effects into the imputation model to get unbiased estimates of time-varying effects
- ... and correct tests for proportional hazards
- The approximate approach can be easily applied in standard software and works well in many circumstances
- The extended Bartlett et al. approach has advantages in some situations
  - ...it also allows for nonlinear terms e.g. X^2
- We aim to show how these methods can be used in conjunction with model selection and fractional polynomials

Summary comments

- We should incorporate time-varying effects into the imputation model to get unbiased estimates of time-varying effects
- ... and correct tests for proportional hazards
- The approximate approach can be easily applied in standard software and works well in many circumstances
- The extended Bartlett et al. approach has advantages in some situations
  - ...it also allows for nonlinear terms e.g. X^2
- We aim to show how these methods can be used in conjunction with model selection and fractional polynomials

Summary comments

- We should incorporate time-varying effects into the imputation model to get unbiased estimates of time-varying effects
- ... and correct tests for proportional hazards
- The approximate approach can be easily applied in standard software and works well in many circumstances
- The extended Bartlett et al. approach has advantages in some situations
  - ...it also allows for nonlinear terms e.g. X^2
- We aim to show how these methods can be used in conjunction with model selection and fractional polynomials

Summary comments

- We should incorporate time-varying effects into the imputation model to get unbiased estimates of time-varying effects
- ... and correct tests for proportional hazards
- The approximate approach can be easily applied in standard software and works well in many circumstances
- The extended Bartlett et al. approach has advantages in some situations
  - ...it also allows for nonlinear terms e.g. X^2
- We aim to show how these methods can be used in conjunction with model selection and fractional polynomials

Thanks

For comments on this work

Ian White, Mike Kenward, Tim Morris, Jonathan Bartlett

Funding

MRC Methodology Fellowship