

Identifiability Assumptions for Missing Covariate Data in Failure Time Regression Models

Northeastern Illinois Chapter
American Statistical Association
Northbrook, IL

Paul Rathouz
Department of Health Studies
University of Chicago
prathouz@uchicago.edu

October 19, 2006

Overview

- Problem set-up: missing X in failure-time models
- MAR for missing X
- Is MAR really what we want?
- A critical look at how MAR might arise
- Alternative identifiability assumptions: CIMAR and FIMAR
- Model checking under CIMAR and FIMAR
- Lymphoma data example

Failure Time Regression Models with Missing X Notation

- **Failure Time:** T
- **Censoring Time:** C
- **Observed Response:** $Y = \min(T, C)$ and $D = I(T \leq C)$
- **Covariates:**
 - $X \leftarrow$ some may be missing
 - $Z \leftarrow$ always observed
- **Missing covariate indicator:** $R = I(X \text{ observed})$

Failure Time Regression Models with Missing X Model with MAR

- Interest on the effects of covariates (X, Z) on T ,
i.e., on the **conditional distribution, hazard or survivor function**

$$f(T|X, Z; \beta) \text{ or } h(T|X, Z; \beta) \text{ or } S(T|X, Z; \beta)$$

parameterized, say, by $\beta = (\beta'_z, \beta'_x)'$ (β_z including baseline hazard)

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

$$R \perp\!\!\!\perp X | Y, D, Z$$

allows identification of β

- MAR: The **value** realized by X is independent of the **fact** that X is missing (or not), given the observed data (Y, D, Z)

MAR in Failure Time Regression Models Existing Approaches and Additional Assumptions

- Two broad approaches:
 - Model distribution $(X|Z)$ and pursue full likelihood estimation of β :
 - joint estimation of (β, α) in

$$f(T|X, Z; \beta) \quad \text{and} \quad \Pr(X|Z; \alpha)$$

- Model missingness probability $\Pr(R = 1|Y, D, X; \gamma)$ and pursue IPW estimation (and variants / SPE extensions):
 - * estimation of γ , then β
 - * no need to model X -distribution
- See Rathouz (2006, *Biostatistics*) for references

- All methods assume **independent censoring**:

$$T \perp\!\!\!\perp C | X, Z$$

(I will, too)

- And **MAR**:

$$R \perp\!\!\!\perp X | (Y, D, Z)$$

- In addition, most methods modeling X assume “ **X -ignorable censoring**”:

$$X \perp\!\!\!\perp C | Z$$

- Take a critical look at MAR ...

MAR in Failure Time Regression Models

- Missingness (of X) is related to the **observed response**, which is **not** T or C or (T, C) , but **rather** is (Y, D)
- Thus, the **MAR assumption** in failure time models is

$$R \perp\!\!\!\perp X | (Y, D, Z)$$

- MAR may hold when missingness is **by design**, e.g., case-cohort or nested-case-control designs
- But is MAR reasonable when missingness is **by chance** ... ?

MAR in Failure Time Regression Models? Missingness by Chance

- “Missingness (of X) is related to severity of disease as well as some components of Z , so we posit that R depends on (T, Z) ”

e.g.:

- invasive/expensive test X only ordered for severely-ill patients
- test X unsafe in patients with poor prognosis

or

- “Missingness (of X) is related to censoring process as well as some components of Z , so we posit that R depends on (C, Z) ”

e.g.: rolling enrollment over several years, with fixed calendar time for censoring and secular trends in usage of X

- In these situations, we might posit something like

$$R \perp\!\!\!\perp X | (T, Z) \quad \text{or} \quad R \perp\!\!\!\perp X | (C, Z)$$

- **Critical question:**

“(How) Can assumptions such as these yield MAR?”

- More specifically: “Under MAR and possibly X -ignorable censoring, what **additional assumptions** about the relationship of R to (T, C, Z) and possibly X will yield MAR?”

... consider 3 scenarios ...

Generating MAR

Scenario 1: Missingness primarily related to failure

- Suppose:

$$R \perp\!\!\!\perp (C, X) | (T, Z)$$

i.e, **censoring-ignorable missing at random (CIMAR)**

- CIMAR will yield MAR if in addition we assume: $T \perp\!\!\!\perp X | Z$,
→ restrictive (maybe ok for testing?)
- Alternatively, we could assume: $R \perp\!\!\!\perp (T, C, X) | Z$,
i.e., that X is MCAR → highly-restrictive
- Caveat: Pathological exceptions exist (see paper)

Generating MAR

Scenario 2: Missingness primarily related to censoring

- Suppose:

$$R \perp\!\!\!\perp (T, X) \mid (C, Z)$$

i.e, **failure-ignorable missing at random (FIMAR)**

- FIMAR will yield MAR if in addition we assume: $C \perp\!\!\!\perp X \mid Z$,
i.e., X -ignorable censoring
- Q: Plausible that C depend on R , but not on value of X itself?
- Again, MCAR is a (highly-restrictive) option and pathological exceptions exist

Generating MAR

Scenario 3: Missingness jointly related to failure and censoring

- Can R depend **jointly** on (T, C, X, Z) such that MAR holds?
- Yields odd results: The allowable dependencies of R on (T, C, X, Z) depend on nature of dependence of C on X given Z
e.g.: if $C \perp\!\!\!\perp X|Z$, then MAR holds **only** under MCAR
- One possibility might hold in some settings:
 - CIMAR holds
 - No censoring before some $\tau > 0$: $C \geq \tau$ w.p. 1
 - $R \perp\!\!\!\perp T|T > \tau, Z$

Summary of Ways to Generate MAR under Missingness by Chance

- When missingness is thought to be related to **censoring**, MAR is practically equivalent to FIMAR along with X -ignorable censoring
 - X -ignorable censoring may be unappealing
 - FIMAR \Rightarrow naive CR (likelihood) analysis is consistent (even without X -ignorable censoring)
 - X -ignorable censoring checkable
 - none of the existing methods acknowledge or exploit these facts
- When missingness is thought to be related to **failure**, MAR is difficult to obtain, unless censoring is delayed to a certain time τ for all subjects
- What about **alternative identifiability assumptions** for missing X ?

Alternative Identifiability Assumptions I

CIMAR

- Assume CIMAR:

$$R \perp\!\!\!\perp (C, X) \mid (T, Z)$$

(censoring-ignorable missing at random)

- Q: (Under what conditions) Is the FT distribution $f(T|X, Z)$ identifiable?
A: Consider complete record (CR) estimation
- Steps:
 1. Obtain the CR failure time distribution $f(T|R = 1, X, Z)$
 2. Ask if it is estimable

CIMAR Identifiability

- **Complete-record failure time distribution:**

$$\begin{aligned} f(T|R = 1, X, Z) &= \frac{f(T|X, Z)\Pr(R = 1|T, X, Z)}{\Pr(R = 1|X, Z)} \\ &= \frac{f(T|X, Z; \beta)\Pr(R = 1|T, Z; \gamma)}{\int_t \Pr(t|X, Z; \beta)\Pr(R = 1|t, Z; \gamma) dt} \\ &= f(T|R = 1, X, Z; \beta, \gamma) \end{aligned}$$

where:

- failure time distribution parameterized by β (e.g., $\beta = (\beta'_z, \beta_z)'$)
 - missingness model parameterized by γ
(e.g., logistic regression of R on $(Z, T, Z \times T)$)
- **Implication:** If one is willing to **model** the missingness process R , then the distribution of X is **ignorable** in the complete-record analysis

CIMAR Identifiability: Two Key Questions

- Q: Does independent censoring still hold, given $R = 1, X, Z$?

A: Yes (see paper).

- Q: Is it possible to estimate γ in $\Pr(R = 1|T, Z; \gamma)$?

A: Yes, maximize likelihood from subjects with **observed failure**:

$$L_\gamma = \prod_{i=1}^n \Pr(R_i|T_i, T_i \leq C_i, Z_i)^{I(T_i \leq C_i)} = \prod_{i=1}^n \Pr(R_i|T_i, Z_i; \gamma)^{I(T_i \leq C_i)},$$

yielding $\hat{\gamma}$

- **Conclusion:** Failure time parameter β estimable via pseudo-ML with

$$L_{CR} = \prod_{i=1}^n f(Y_i|R_i = 1, X_i, Z_i; \beta, \hat{\gamma})^{R_i D_i} S(Y_i|R_i = 1, X_i, Z_i; \beta, \hat{\gamma})^{R_i(1-D_i)},$$

with no need to model the censoring distribution

Alternative Identifiability Assumptions

FIMAR

- Suppose assume:

$$R \perp\!\!\!\perp (T, X) \mid (C, Z)$$

i.e, **failure-ignorable missing at random** (FIMAR)

- Argument is symmetric to CIMAR, including independent censoring given $R = 1$
- However, can show in this case that

$$f(T \mid R = 1, X, Z) = f(T \mid X, Z; \beta)$$

- **Conclusion:** The naive complete-record estimator for β is consistent under FIMAR!

Checking Model Assumptions under CIMAR and FIMAR

- As with MAR, impossible to **confirm** CIMAR or FIMAR with data
- However, under CIMAR or FIMAR:
 - certain “independencies” should hold
 - can be checked to detect conflicts with CIMAR (FIMAR)
- First, under either CIMAR or FIMAR:

$$C \perp\!\!\!\perp X|Z \implies C \perp\!\!\!\perp X|R = 1, Z$$

so that X -independent censoring can be checked by examining only the complete records (those with $R = 1$)

- Second, under CIMAR **and** X -independent censoring,

$$C \perp\!\!\!\perp T|R, Z \quad \text{and} \quad C \perp\!\!\!\perp R|Z$$

(with symmetric result for FIMAR)

Model Checking (cont.)

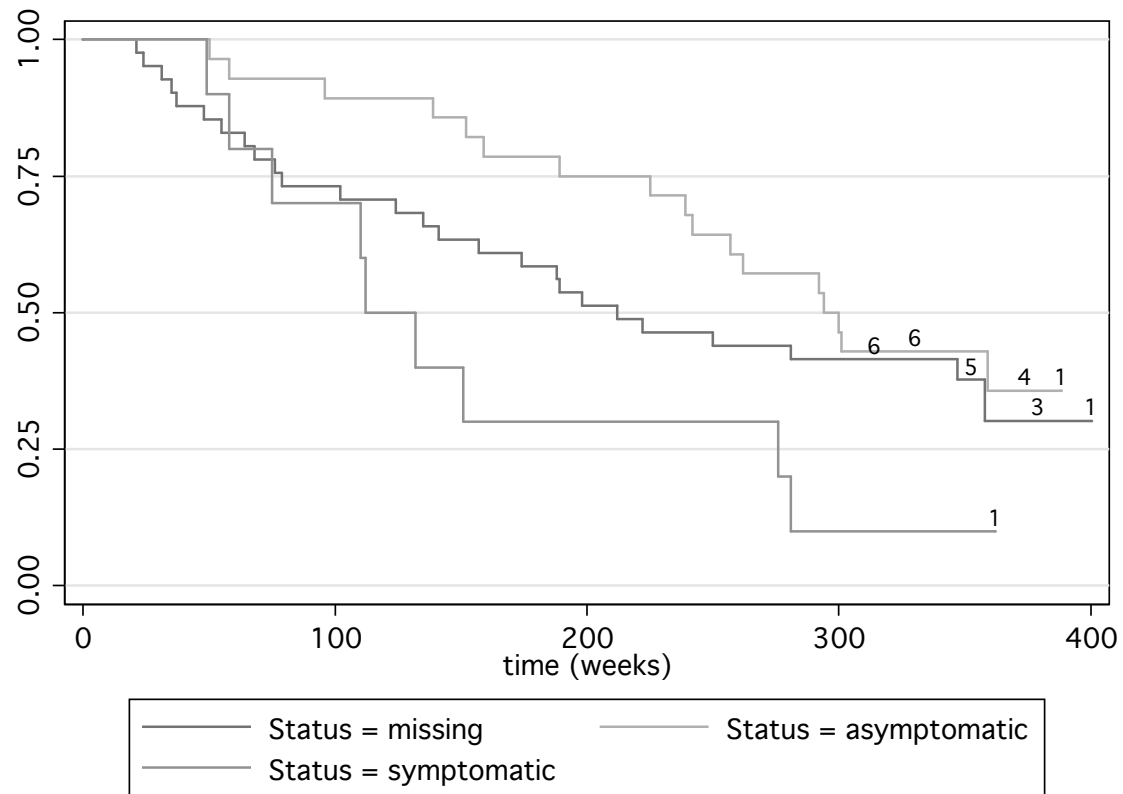
- **Implication:** can test CIMAR (FIMAR) by modeling $C(T)$ as a function of (R, Z) with $T(C)$ as “censoring” time
- Suggests **analysis plan** for checking data consistency with CIMAR (FIMAR):
 1. posit CIMAR (FIMAR) as a working assumption
 2. in complete records, evaluate whether C depends on X given Z , treating T as “censoring” time
 3. if not, evaluate whether $C(T)$ is independent of R given Z which should hold under CIMAR (FIMAR)

Worked Example

Survival of Lymphoma Patients

- 79 male patients with non-Hodgkins lymphoma
- 38 have non-missing X ; 26 (68%) fail
41 have missing X ; 26 (63%) fail
- All censoring beyond median survival time
- Research question: Is survival associated w presence of symptoms (X) at start of treatment?
- “with uncensored data, a mechanism where missingness depended on the underlying true survival time . . . would be ignorable . . .”^{*}
→ suggests CIMAR as a working assumption

^{*}Schluchter & Jackson (1989; JASA)



Lymphoma Patients

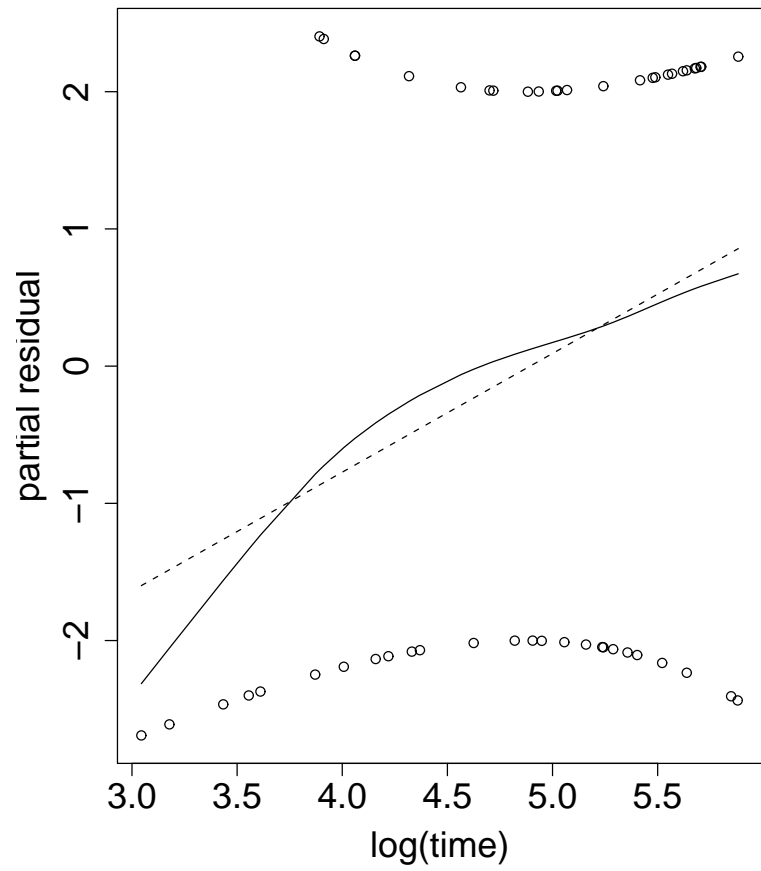
Check CIMAR Working Assumption

- Is $C \perp\!\!\!\perp X$?
 - 12 censoring events in group with $R = 1$
 - 11 in asymptomatic patients
 - expected number under independence: 11.05
- No relationship of C to R
- Conclude: CIMAR is reasonable

Lymphoma Patients

Parametric Failure Time Model

- **Step 1:** Logistic model for R given $\log(T)$ among (non-censored) subjects with $T \leq C$ ($Z = 1.97$ for effect of $\log(T)$)
 - partial residual plot suggested non-linear effect of $\log(T)$
 - added $\{\log(T)\}^2$ (not significant, but captures of R to T relationship better)
- **Step 2:** Failure time model among CR subjects: Exponential with constant hazard ratio comparing symptomatic to asymptomatic



Lymphoma Patients Results

- Naive (MCAR) complete-record analysis:
 - 2.26 and 5.60 deaths / 1000 person-weeks in asymptomatic and symptomatic groups
 - risk ratio: 2.48
- Corrected complete-record analysis under CIMAR:
 - 2.41 and 7.95 deaths / 1000 person-weeks in two groups
 - risk ratio: 3.30
- Not including $\{\log(T)\}^2$ in model for R :
 - 3.27 and 8.09 deaths / 1000 person-weeks in two groups
 - irreconcilable with overall rate of 2.09 deaths / 1000 person-weeks, ignoring symptom status!

Summary / Conclusions

Missing X in Failure Time Regression Models

- MAR assumption made in much of existing work difficult to justify when missingness is related to **either** failure **or** to censoring
- Two new identifiability assumptions: CIMAR and FIMAR
- In a practical sense, FIMAR is weaker than MAR
- Model of interest is identifiable under CIMAR and FIMAR
- Naive complete-record analysis valid under FIMAR
- Corrected CR analysis available under CIMAR
- Suggested model-checking procedures for CIMAR / FIMAR

Further Work

- How badly do MAR procedures do when CIMAR or FIMAR hold?
- Vice-versa?
- Estimators under CIMAR or FIMAR for the proportional hazards model
- Method that exploit models for $(X|Z)$
- Semi-parametric efficiency theory

EXTRA SLIDES

General Regression Models

Preliminary Notation and Set-up

- **Response:** Y

- **Covariates:**

X ← some may be missing

Z ← always observed

- **Missing covariate indicator:** $R = I(X \text{ observed})$
- Interest on the effects of covariates (X, Z) on Y ,
i.e., on the **conditional distribution**

$$f(Y|X, Z; \beta)$$

perhaps parameterized by $\beta = (\beta'_z, \beta'_x)'$

Regression Models with Missing X

What is MAR?

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

$$R \perp\!\!\!\perp X | Y, Z$$

allows identification of β

- MAR: The **value** realized by X is independent of the **fact** that X is missing or not, given the observed data (Y, Z)

Why (how) does MAR work?

Full Likelihood Analysis

- **Likelihood function** (conditioning on Z implicit):

$$\begin{aligned}L_F &= \{\Pr(X|R=1, Y)\Pr(R=1|Y)\Pr(Y)\}^R \{\Pr(R=0|Y)\Pr(Y)\}^{1-R} \\ &= \{\Pr(X|Y)\Pr(R=1|Y)\Pr(Y)\}^R \{\Pr(R=0|Y)\Pr(Y)\}^{1-R} \\ &= \{\Pr(R=1|Y; \gamma)\Pr(Y|X; \beta)\Pr(X; \alpha)\}^R \\ &\quad \times \{\Pr(R=0|Y; \gamma)\Pr(Y; \beta, \alpha)\}^{1-R} \\ &\propto \{\Pr(Y|X; \beta)\Pr(X; \alpha)\}^R \{\Pr(Y; \beta, \alpha)\}^{1-R}\end{aligned}$$

↑ for purposes of inferences on β

- **Implications:**

- If one is willing to **model** the distribution of X , then the missingness model is **ignorable**
- β can be estimated via joint estimation of (α, β) using L_F

Why (how) does MAR work? Complete Record Likelihood Analysis

- **Complete record likelihood function**, conditional on X (and Z):

$$\begin{aligned} L_C &= \{\Pr(Y|X, R = 1)\}^R = \left\{ \frac{\Pr(Y|X)\Pr(R = 1|Y, X)}{\Pr(R = 1|X)} \right\}^R \\ &= \left\{ \frac{\Pr(Y|X; \beta)\Pr(R = 1|Y; \gamma)}{\int_y \Pr(y|X; \beta)\Pr(R = 1|y; \gamma) dy} \right\}^R \end{aligned}$$

- **Implication:** If one is willing to **model** the missingness process R , then the distribution of X is **ignorable**
- However, this ignorability does not extend over to non-likelihood-based approaches.

Complete Record Estimation under MAR

- **Estimate:** γ via

$$L_\gamma = \Pr(R|Y; \gamma) \longrightarrow \hat{\gamma}$$

- **Estimate:** β via

$$L_C(\beta, \hat{\gamma}) \longrightarrow \hat{\beta}$$

- **Compute:** Standard errors for $\hat{\beta}$ accounting for variability in estimation of γ