Statistical Inference Under Latent Class Models, With Application To Cancer Survivorship Study

X. Joan Hu

Department of Statistics and Actuarial Science
Simon Fraser University, British Columbia, Canada

Presentation at BIRS Workshop, Banff
August 16, 2016
Joint work with Huijing Wang and John Spinelli

Outline

1. Introduction: Large Administrative Health Data

Readily Available

- Canadian Provincial Medical Insurance Databases

- Canada Health Care System: universally accessible, government-sponsored

  – Thomas C. (Tommy) Douglas was voted "*the greatest Canadian of all time*"

1. Introduction: Large Administrative Health Data

Readily Available

- Canadian Provincial Medical Insurance Databases

- Canada Health Care System: universally accessible, government-sponsored

  – Thomas C. (Tommy) Douglas was voted "*the greatest Canadian of all time*"

- Canadian Disease/Patient Registries: e.g. BC Cancer Registry

1. Introduction: to Address Public Health Issues with Such Data

McBride et al (2010) on Cancer Survivorship

- The cancer survivor population has been increasing rapidly due to improvements in cancer treatments.
- These survivors are often at risk of subsequent and ongoing problems that are mainly treatment related.
- The evaluation/development of strategies for long-term management requires risk assessment, particularly for those diagnosed at a young age, e.g. at age 0 to 19.

1. Introduction: to Address Public Health Issues with Such Data

McBride et al (2010) on Cancer Survivorship

- The cancer survivor population has been increasing rapidly due to improvements in cancer treatments.
- These survivors are often at risk of subsequent and ongoing problems that are mainly treatment related.
- The evaluation/development of strategies for long-term management requires risk assessment, particularly for those diagnosed at a young age, e.g. at age 0 to 19.

To address the survivorship issues, the Childhood, Adolescent, Young Adult Cancer Survivorship (CAYACS) research program uses population-based data (Registry+MSP):
e.g. physician claims of the survivors.

1. Introduction: to Address Public Health Issues with Such Data

McBride et al (2010) on Cancer Survivorship

- The cancer survivor population has been increasing rapidly due to improvements in cancer treatments.
- These survivors are often at risk of subsequent and ongoing problems that are mainly treatment related.
- The evaluation/development of strategies for long-term management requires risk assessment, particularly for those diagnosed at a young age, e.g. at age 0 to 19.

BC Cancer Agency
CARE + RESEARCH

To address the survivorship issues, the Childhood, Adolescent, Young Adult Cancer Survivorship (CAYACS) research program uses population-based data (Registry+MSP):
e.g. **physician claims** of the survivors.

1. Introduction: CAYACS Physician Claims Study

CAYACS Data Extraction

- **CAYACS survivor cohort**: diagnosed 1981-1999, under the age of 20, in BC and having survived $\geq$ 5 yrs
  - information from Cancer Registry (a total of 1962)
  - physician claims from MSP (Medical Services Plan), starting from 5 yrs after diagnosis till 2006

1. Introduction: CAYACS Physician Claims Study

CAYACS Data Extraction

- **CAYACS survivor cohort**: diagnosed 1981-1999, under the age of 20, in BC and having survived $\geq$ 5 yrs
  - information from Cancer Registry (a total of 1962)
  - physician claims from MSP (Medical Services Plan), starting from 5 yrs after diagnosis till 2006
- **CAYACS general population sample**: selected from BC general population to match in sex and birth year, 10 times the size of survivor cohort.
  - physician claims from MSP

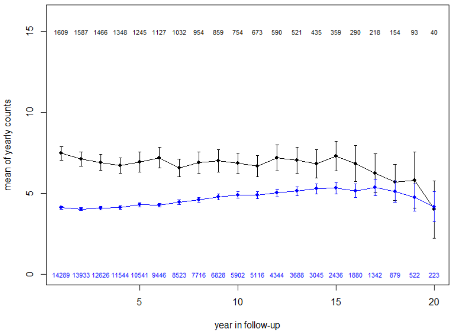
- **Objectives of the CAYACS's physician claims project:**
  - to evaluate the cohort's physician visit frequency and medical cost
  - to identify factors of risk to later effects
  - to compare it with the general population

- **Objectives of the CAYACS's physician claims project:**
  - to evaluate the cohort's physician visit frequency and medical cost
  - to identify factors of risk to later effects
  - to compare it with the general population

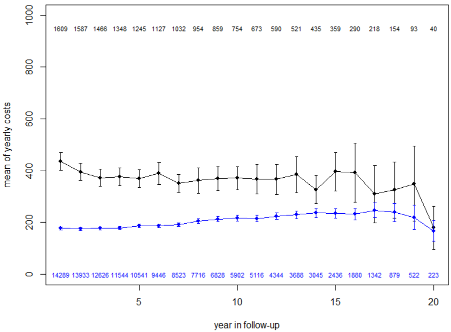
- **Results from CAYACS's previous analysis: 3-year visit counts:** (McBride et al, 2011)
  Regarding medical care demand:
  - cancer survivors > general population
    survivors often suffered the consequences of the original cancer diagnoses – mostly treatment-related (later effects)
  - females > males within survivors
    Is *gender* a risk factor?

Yearly Data Comparison: Survivor vs General

Means of Yearly Visit Counts

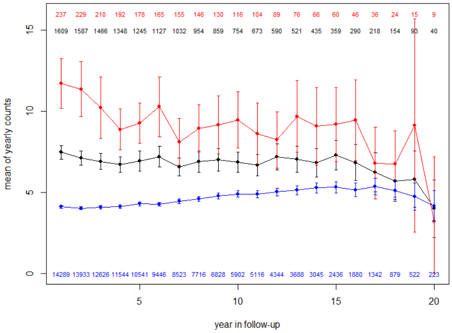


Means of Yearly Medical Costs

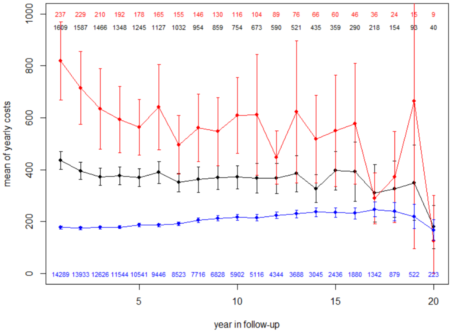


Yearly Data Comparison: survivor with RSC vs. survivor vs. general

Means of Yearly Visit Counts

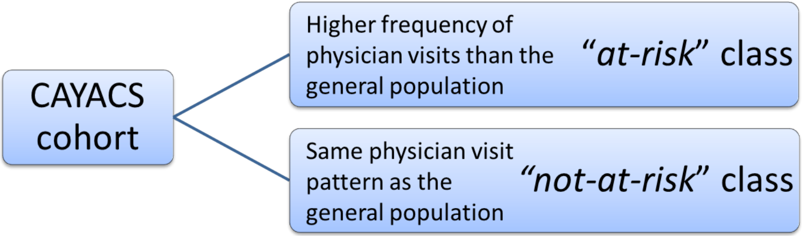


Means of Yearly Medical Costs



SC with RSC — survivor cohort — general population

1. Introduction: CAYACS Physician Claims Study



The class membership ($\eta = 1$ or 0) is not observable.

$\Longrightarrow$ to consider a latent class model

Difficulties in Analysis under Latent Class Models

- Increased number of parameters $\rightarrow$ low efficiency
- Underlying probability model specification for each latent class: since no available information directly on η, and thus on $Y|\eta$ $\rightarrow$ lack of robustness to distribution assumptions

Difficulties in Analysis under Latent Class Models

- Increased number of parameters $\rightarrow$ low efficiency
- Underlying probability model specification for each latent class: since no available information directly on η, and thus on $Y|\eta$ $\rightarrow$ lack of robustness to distribution assumptions

Additional Information

- Supplementary Information: general population (about $\eta = 0$ group?)
- Partially observed at-risk class ($\eta = 1$): a total of 168 survivors with relapse/2nd cancer ($\delta = 1$)

1. Introduction: Statistical Modelling

Model Specification

Risk model

$$P(\eta = 1|\mathbf{Z}) = p(\mathbf{Z}; \alpha)$$

Regression models for each class

$$E(\mathbf{Y}|\eta = 1, \mathbf{Z}) = \mu_1(\mathbf{Z}; \beta)$$

$$E(\mathbf{Y}|\eta = 0, \mathbf{Z}) = \mu_0(\mathbf{Z}; \theta)$$

1. Introduction: Statistical Modelling

Model Specification

Risk model

$P(\eta = 1|\mathbf{Z}) = p(\mathbf{Z}; \alpha)$ e.g. $\text{logit}\{p(\mathbf{Z}; \alpha)\} = \alpha'\mathbf{Z}$

Regression models for each class

$E(\mathbf{Y}|\eta = 1, \mathbf{Z}) = \mu_1(\mathbf{Z}; \beta)$ e.g. $\text{l}\{\mu_1(\mathbf{Z}; \beta)\} = \beta'\mathbf{Z}$

$E(\mathbf{Y}|\eta = 0, \mathbf{Z}) = \mu_0(\mathbf{Z}; \theta)$ e.g. $\text{l}\{\mu_0(\mathbf{Z}; \theta)\} = \theta'\mathbf{Z}$

1. Introduction: Statistical Problem

Estimation of (α, β, θ) based on data from the *survivor cohort* combined with the sample from the *general population*

1. Introduction: Statistical Problem

Estimation of (α, β, θ) based on data from the *survivor cohort* combined with the sample from the *general population*

Why bother? *Examples for its use:*

- by $p(\mathbf{Z}; \alpha)$, risk factor identification; risk probability estimation
- by $\mu_1(\mathbf{Z}; \beta)$, visit patterns in *"at-risk"* class
- by $\mu_0(\mathbf{Z}; \theta)$, visit patterns in *"not-at-risk"* class
- to conduct risk classification/prediction in the survivor cohort

1. Introduction: Statistical Problem

Estimation of (α, β, θ) based on data from the *survivor cohort* combined with the sample from the *general population*

Why bother? *Examples for its use:*

- by $p(\mathbf{Z}; \alpha)$, risk factor identification; risk probability estimation
- by $\mu_1(\mathbf{Z}; \beta)$, visit patterns in *"at-risk"* class
- by $\mu_0(\mathbf{Z}; \theta)$, visit patterns in *"not-at-risk"* class
- to conduct risk classification/prediction in the survivor cohort

How? *Procedures:*

- Likelihood-Based Estimation with Cross-Sectional Counts under Mixture Poisson Models (Wang et al, 2014)
- Extended GEE Procedures with Longitudinal Data

2. Likelihood-Based Estimation with Cross-Sectional Counts: Model Assumption

$\mathbf{Y} = Y$ cross-sectional visit count over $(0, T]$ with T the follow up time.

Mixture Poisson Model

$$
\begin{aligned}
& [Y|\mathbf{Z};\alpha,\beta,\theta] \\
= \ & [Y|\eta=0,\mathbf{Z};\theta][\eta=0|\mathbf{Z};\alpha]+[Y|\eta=1,\mathbf{Z};\beta][\eta=1|\mathbf{Z};\alpha]
\end{aligned}
$$

- $Y|\eta=1,\mathbf{Z} \sim \text{Poisson}\big(\mu_1(\mathbf{Z};\beta)\big)$
- $Y|\eta=0,\mathbf{Z} \sim \text{Poisson}\big(\mu_0(\mathbf{Z};\theta)\big)$
- $\eta=1|\mathbf{Z} \sim$ logistic regression model

2. Likelihood-Based Estimation with Cross-Sectional Counts: Procedures

- **Maximum Likelihood Estimation (MLE)** Likelihood function based on the data from CAYACS cohort $L(\alpha, \beta, \theta; \mathrm{Data}_{\mathcal{P}}) \propto \prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \theta]$
  - EM algorithm via the "full-data" likelihood based on $[Y_i, \eta_i | \mathbf{Z}_i]$
  - computationally intense

2. Likelihood-Based Estimation with Cross-Sectional Counts: Procedures

- **Maximum Likelihood Estimation (MLE)** Likelihood function based on the data from CAYACS cohort
  $L(\alpha, \beta, \theta; \mathrm{Data}_{\mathcal{P}}) \propto \prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \theta]$
  - EM algorithm via the "full-data" likelihood based on $[Y_i, \eta_i | \mathbf{Z}_i]$
  - computationally intense

- **Pseudo-MLE** With rich information on θ from the general population, likelihood function:

$$L(\alpha, \beta, \theta; \mathrm{Data}_{\mathcal{P}}, \mathrm{Data}_{\mathcal{Q}}) \propto \prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \theta] \prod_{i \in \mathcal{Q}} [Y_i | \eta_i = 0, \mathbf{Z}_i; \theta]$$

2. Likelihood-Based Estimation with Cross-Sectional Counts: Procedures

- **Maximum Likelihood Estimation (MLE)** Likelihood function based on the data from CAYACS cohort $L(\alpha, \beta, \theta; \mathrm{Data}_{\mathcal{P}}) \propto \prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \theta]$
  - EM algorithm via the "full-data" likelihood based on $[Y_i, \eta_i | \mathbf{Z}_i]$
  - computationally intense

- **Pseudo-MLE** With rich information on θ from the general population, likelihood function:

$$L(\alpha, \beta, \theta; \mathrm{Data}_{\mathcal{P}}, \mathrm{Data}_{\mathcal{Q}}) \propto \prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \theta] \prod_{i \in \mathcal{Q}} [Y_i | \eta_i = 0, \mathbf{Z}_i; \theta]$$

  Type AB pseudo MLE.
  - $\hat{\theta}$ from $\prod_{i \in \mathcal{Q}} [Y_i | \eta_i = 0, \mathbf{Z}_i; \theta]$
  - $(\hat{\alpha}, \hat{\beta})$ from $\prod_{i \in \mathcal{P}} [Y_i | \mathbf{Z}_i; \alpha, \beta, \hat{\theta}]$

2. Likelihood-Based Estimation with Cross-Sectional Counts: Properties

- Consistency and asymptotic normality
- MLE vs the Pseudo-MLE: efficiency?
- Extended Huber sandwich variance estimator: *e.g.* account for $\hat{\theta}$ estimated from $\mathcal{Q}$

2. Likelihood-Based Estimation with Cross-Sectional Counts: Properties

- Consistency and asymptotic normality
- MLE vs the Pseudo-MLE: efficiency?
- Extended Huber sandwich variance estimator: *e.g.* account for $\hat{\theta}$ estimated from $\mathcal{Q}$

However,

- Simulation results show that likelihood-based estimators were biased under distribution misspecification, especially for α
- CAYACS physician visit counts are highly overdispersed. Plus physician claims include costs and are longitudinal.

$\Longrightarrow$ to adapt the GEE approach

3. Extended GEE Procedures with Longitudinal Data: Modelling

Consider the Mean-Variance Models:

$$E(Y|\mathbf{Z}) = p(\mathbf{Z};\alpha)\mu_1(\mathbf{Z};\beta) + \{1 - p(\mathbf{Z};\alpha)\}\mu_0(\mathbf{Z};\theta) \equiv \Lambda$$

$$\begin{aligned} V(Y|\mathbf{Z}) = {} & p(\mathbf{Z};\alpha)\Sigma_1 + \{1 - p(\mathbf{Z};\alpha)\}\Sigma_0 \\ & + p(\mathbf{Z};\alpha)\{1 - p(\mathbf{Z};\alpha)\}\{\mu_1(\mathbf{Z};\beta) - \mu_0(\mathbf{Z};\theta)\}^2 \equiv \Sigma \end{aligned}$$

3. Extended GEE Procedures with Longitudinal Data: Modelling

Consider the Mean-Variance Models:

$$E(Y|\mathbf{Z}) = p(\mathbf{Z};\alpha)\mu_1(\mathbf{Z};\beta) + \{1 - p(\mathbf{Z};\alpha)\}\mu_0(\mathbf{Z};\theta) \equiv \Lambda$$

$$\begin{aligned} V(Y|\mathbf{Z}) = {} & p(\mathbf{Z};\alpha)\Sigma_1 + \{1 - p(\mathbf{Z};\alpha)\}\Sigma_0 \\ & + p(\mathbf{Z};\alpha)\{1 - p(\mathbf{Z};\alpha)\}\{\mu_1(\mathbf{Z};\beta) - \mu_0(\mathbf{Z};\theta)\}^2 \equiv \Sigma \end{aligned}$$

Directly applying the GEE approach:

$$\sum_{i=1}^{n} \frac{\partial \Lambda_i(\alpha,\beta,\theta)}{\partial(\alpha,\beta,\theta)} \Sigma_i^{-1}[Y_i - \Lambda_i(\alpha,\beta,\theta)] = 0$$

the evaluations of the estimator for (α,β,θ)*?*

3. Extended GEE Procedures with Longitudinal Data: Procedure

- Using the information from the general population to set the standard for "not-at-risk", the group of $\eta = 0$:

$$\sum_{i \in \mathcal{Q}} \frac{\partial \mu_0(\mathbf{Z}_i; \theta)}{\partial \theta} \Sigma_{0i}^{-1} [Y_i - \mu_0(\mathbf{Z}_i; \theta)] = 0$$

3. Extended GEE Procedures with Longitudinal Data: Procedure

- Using the information from the general population to set the standard for "not-at-risk", the group of $\eta = 0$:

$$\sum_{i \in \mathcal{Q}} \frac{\partial \mu_0(\mathbf{Z}_i; \theta)}{\partial \theta} \Sigma_{0i}^{-1} [Y_i - \mu_0(\mathbf{Z}_i; \theta)] = 0$$

- Using the information from the sub-cohort of subjects with relapse/2nd cancer to set the standard for "at-risk", the group of $\eta = 1$:

$$\sum_{i: \delta_i = 1} \frac{\partial \mu_1(\mathbf{Z}_i; \beta)}{\partial \beta} \Sigma_{1i}^{-1} [Y_i - \mu_1(\mathbf{Z}_i; \beta)] = 0$$

3. Extended GEE Procedures with Longitudinal Data: Procedure

- Using the information from the general population to set the standard for "not-at-risk", the group of $\eta = 0$:

$$\sum_{i \in \mathcal{Q}} \frac{\partial \mu_0(\mathbf{Z}_i; \theta)}{\partial \theta} \Sigma_{0i}^{-1} [Y_i - \mu_0(\mathbf{Z}_i; \theta)] = 0$$

- Using the information from the sub-cohort of subjects with relapse/2nd cancer to set the standard for "at-risk", the group of $\eta = 1$:

$$\sum_{i : \delta_i = 1} \frac{\partial \mu_1(\mathbf{Z}_i; \beta)}{\partial \beta} \Sigma_{1i}^{-1} [Y_i - \mu_1(\mathbf{Z}_i; \beta)] = 0$$

- Pulling the information together to obtain an estimator of α:

$$\sum_{i \in \mathcal{P}} \frac{\partial \Lambda_i(\alpha, \beta, \theta)}{\partial \alpha} \Sigma_i^{-1} [Y_i - \Lambda_i(\alpha, \beta, \theta)] = 0$$

3. Extended GEE Procedures with Longitudinal Data: Procedure

- Using the information from the general population to set the standard for "not-at-risk", the group of $\eta = 0$:

$$\sum_{i \in \mathcal{Q}} \frac{\partial \mu_0(\mathbf{Z}_i; \theta)}{\partial \theta} \Sigma_{0i}^{-1} [Y_i - \mu_0(\mathbf{Z}_i; \theta)] = 0$$

- Using the information from the sub-cohort of subjects with relapse/2nd cancer to set the standard for "at-risk", the group of $\eta = 1$:

$$\sum_{i : \delta_i = 1} \frac{\partial \mu_1(\mathbf{Z}_i; \beta)}{\partial \beta} \Sigma_{1i}^{-1} [Y_i - \mu_1(\mathbf{Z}_i; \beta)] = 0$$

- Pulling the information together to obtain an estimator of α:

$$\sum_{i \in \mathcal{P}} \frac{\partial \Lambda_i(\alpha, \beta, \theta)}{\partial \alpha} \Sigma_i^{-1} [Y_i - \Lambda_i(\alpha, \beta, \theta)] = 0$$

$\Longrightarrow$ extended GEE estimator with consistency and asymptotic normality: how is it compared with MLE?

3. Extended GEE Procedures with Longitudinal Data: Implementing with CAYACS Data

responses $\mathbf{Y}_i \to Y_{ij}$: $j = 1, \ldots, J_i$, $J_i \in [1, 20]$
yearly visit counts/log-trans costs

potential risk factors
- **sex:** male vs female
- **age at study entry:** 5 years after diag
- **SES:** socioeconomic status, high vs low
- **diagnosis period:** 1990s vs 1980s
- **cancer treatment:** chemotherapy no radiation, radiation no chemotherapy, both vs others

Table 3.1. LCM Analysis: intercept, sex, age at entry – effect time-varying; compound symmetric corr

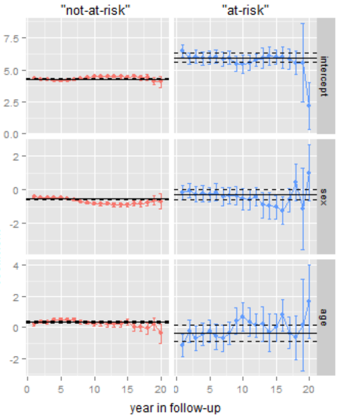
Factor	counts estimate	counts sw.se	costs estimate	costs sw.se
α estimates in the Risk Model				
intercept	0.179	(0.435)	0.196	(0.314)
male (vs female)	-0.329	(0.341)	-0.286	(0.247)
SES high (vs low)	0.365	(0.342)	0.280	(0.248)
age at diagnosis	0.097	(0.590)	-0.302	(0.389)
diag in 90s (vs 80s)	**-1.347**	(0.283)	0.017	(0.178)
treatment (vs other)				
chemo no rad	**0.474**	(0.246)	0.269	(0.181)
rad no chemo	**1.524**	(0.525)	**1.729**	(0.509)
both	**1.463**	(0.413)	**0.946**	(0.241)
β estimates in the Regression Model for the "at-risk" class				
GEE estimates based on $\delta = 1$ subgroup				
intercept	**2.360**[b]	(0.128)[c]	**5.664**[b]	(0.232)[c]
male (vs female)	**-0.293**[b]	(0.124)[c]	**-0.421**[b]	(0.201)[c]
SES high (vs low)	-0.078	(0.111)	-0.094	(0.159)
age at study entry	0.070[b]	(0.186)[c]	-0.071[b]	(0.287)[c]
dispersion/scale parameter	**10.59**	(1.302)	**2.641**[b]	(0.224)[c]
correlation parameter	**0.331**	(0.042)	**0.401**	(0.048)
θ estimates in the Regression Model for the "not-at-risk" class				
GEE estimates based on general population				
intercept	**1.537**[b]	(0.036)[c]	**4.324**[b]	(0.032)[c]
male (vs female)	**-0.546**[b]	(0.040)[c]	**-0.697**[b]	(0.030)[c]
SES high (vs low)	**-0.062**	(0.019)	**-0.049**	(0.020)
age at study entry	**0.399**[b]	(0.060)[c]	**0.235**[b]	(0.047)[c]
dispersion/scale parameter	**10.029**	(0.537)	**2.801**[b]	(0.025)[c]
correlation parameter	**0.381**	(0.013)	**0.333**	(0.005)

[a]Significant Effect with P-value $\leq$ 0.05 in **Boldface**

[b]Average values over 20 estimates

[b]se of the 20 averaged estimate

Time-varying coefficients: (yearly costs)



4. Application to Risk Classification and Prediction

Statistical Modelling II

To capture heterogeneity within individual ...

$$l\Big\{\mathrm{E}\big(Y_{ij}\big|\eta_i = 1, Z_{ij}, b_i\big)\Big\} = \beta_j' Z_{ij} + b_i' X_{ij}$$

$$l\Big\{\mathrm{E}\big(Y_{ij}\big|\eta_i = 0, Z_{ij}, c_i\big)\Big\} = \theta_j' Z_{ij} + c_i' X_{ij}$$

4. Application to Risk Classification and Prediction

Statistical Modelling II

To capture heterogeneity within individual ...

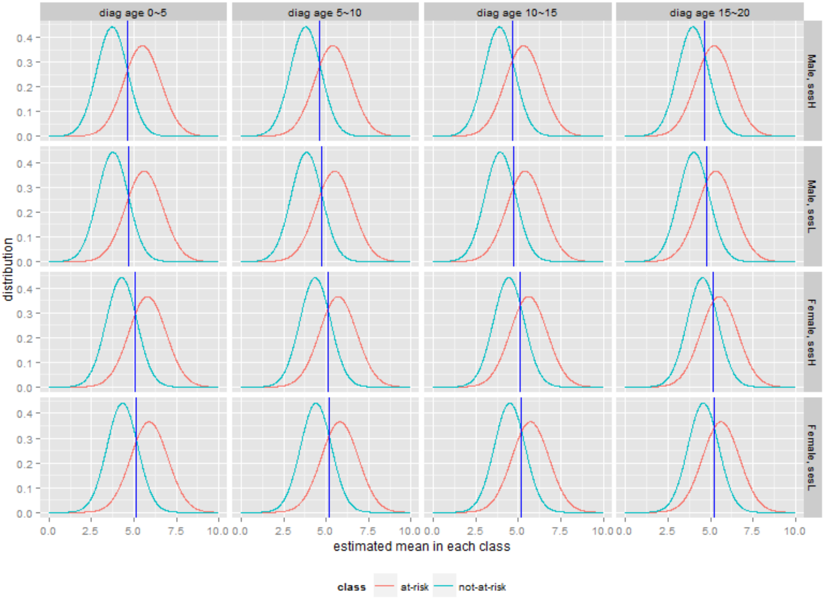
$$l\Big\{\mathrm{E}\big(Y_{ij}\big|\eta_i = 1, Z_{ij}, b_i\big)\Big\} = \beta_j' Z_{ij} + b_i' X_{ij}$$

$$l\Big\{\mathrm{E}\big(Y_{ij}\big|\eta_i = 0, Z_{ij}, c_i\big)\Big\} = \theta_j' Z_{ij} + c_i' X_{ij}$$

Special cases:

(i) $X_{ij} = 1 \Longrightarrow b_i$ and c_i are scalar ("random intercept")

(ii) for cost $\Longrightarrow l\{\} = I$; for count $\Longrightarrow l\{\} = \log$



Application A. Risk Classification by Yearly Costs Using $E(\mathbf{Y}_i|\mathbf{Z}_i, \hat{b}_i, \hat{c}_i)$

Application B. Risk Classification based on Estimated Risk Probabilities

1. $\hat{P}(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i; \hat{\alpha})$

Application B. Risk Classification based on Estimated Risk Probabilities

1. $\hat{P}(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i; \hat{\alpha})$
2. $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i]P(\eta_i = 0|\mathbf{Z}_i)}$$

Application B. Risk Classification based on Estimated Risk Probabilities

1. $\hat{P}(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i; \hat{\alpha})$
2. $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i]P(\eta_i = 0|\mathbf{Z}_i)}$$

3. $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, \hat{b}_i, \hat{c}_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, $\hat{b}_i, \hat{c}_i$ estimated by BLUP

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i, c_i]P(\eta_i = 0|\mathbf{Z}_i, b_i, c_i)}$$

Application B. Risk Classification based on Estimated Risk Probabilities

1. $\hat{P}(\eta_i = 1|\mathbf{Z}_i) = p(\mathbf{Z}_i; \hat{\alpha})$
2. $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i]P(\eta_i = 1|\mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i]P(\eta_i = 0|\mathbf{Z}_i)}$$

3. $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, \hat{b}_i, \hat{c}_i; \hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, $\hat{b}_i, \hat{c}_i$ estimated by BLUP

$$P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i) = \frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Z}_i, b_i, c_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i, c_i]P(\eta_i = 0|\mathbf{Z}_i, b_i, c_i)}$$

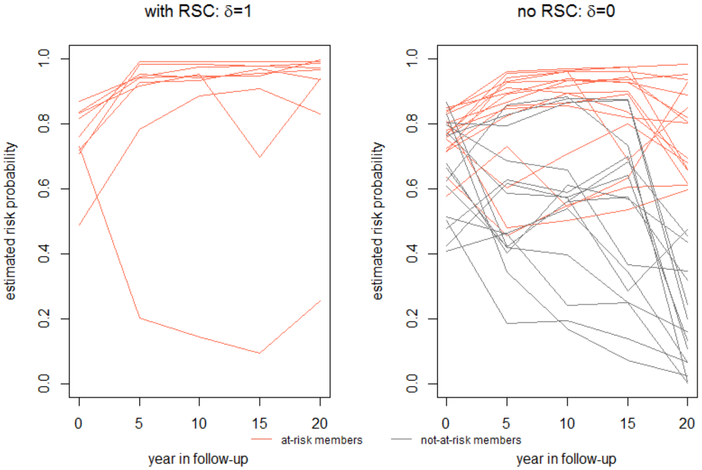
(A) jointly model $\mathbf{Y}$ and η via b and c
(B) to approximate it by

$$\frac{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)}{[\mathbf{Y}_i|\eta_i = 1, \mathbf{Z}_i, b_i]P(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i) + [\mathbf{Y}_i|\eta_i = 0, \mathbf{Z}_i, c_i]P(\eta_i = 0|\mathbf{Y}_i, \mathbf{Z}_i)}$$

Histograms of estimated risk probabilities for full survivor cohort and parametric bootstraps

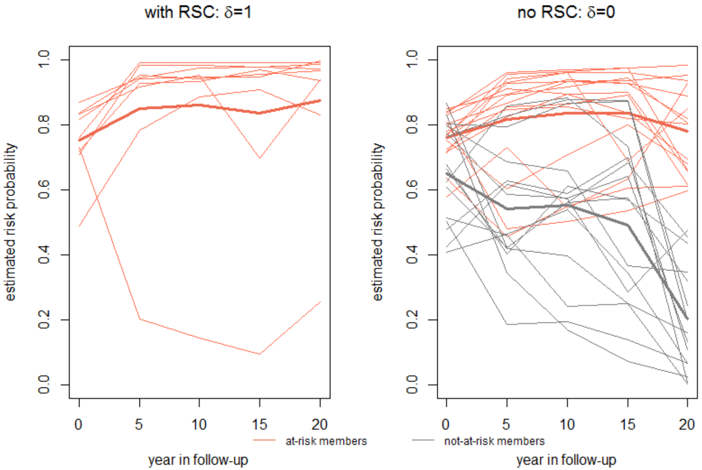
$\hat{P}(\eta_i = 1|\mathbf{Z}_i)$ $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i)$ $\hat{P}(\eta_i = 1|\mathbf{Y}_i, \mathbf{Z}_i, b_i, c_i)$

Dynamic risk probabilities:
40 survivors diagnosed in 1981 and followed until 2006



$$P(\eta = 1|\mathbf{Z}) \mapsto P(\eta = 1|\mathbf{Z}, \mathbf{Y}_5) \mapsto P(\eta = 1|\mathbf{Z}, \mathbf{Y}_{10}) \mapsto P(\eta = 1|\mathbf{Z}, \mathbf{Y}_{15}) \mapsto P(\eta = 1|\mathbf{Z}, \mathbf{Y}_{20})$$

Dynamic risk probabilities:
40 survivors diagnosed in 1981 and followed until 2006, with estimated means for the two classes



Thanks for your attention!