Longitudinal Data Analysis with Composite Likelihood Methods

Haocheng Li and Grace Y. Yi

Department of Statistics and Actuarial Science, University of Waterloo

The National Population Health Survey Data (NPHS)

Data Description

- The National Population Health Survey (NPHS) is a longitudinal study which collects information on health and related socio-demographic characteristics
- The study is designed to follow a group of Canadian household residents for 10 cycles
- The survey is conducted every second year from 1994/1995 and has completed nine cycles: Cycle 1 (1994/1995), Cycle 2 (1996/1997), $\cdots$, Cycle 9 (2010/2011)
- One person in each household was randomly selected as the longitudinal respondent to answer an in-depth questionnaire

Data Description

The questions for the NPHS include

- Health information
  - Health status
  - The use of health services
  - Chronic conditions
  - Activity restrictions
  - ...
- Social background information
  - Age, Gender, Education, Income level, Marital status
  - ...
- One objective: understanding how health status may be associated with variables of social background information

Health-Related Quality of Life

Health-Related Quality of Life

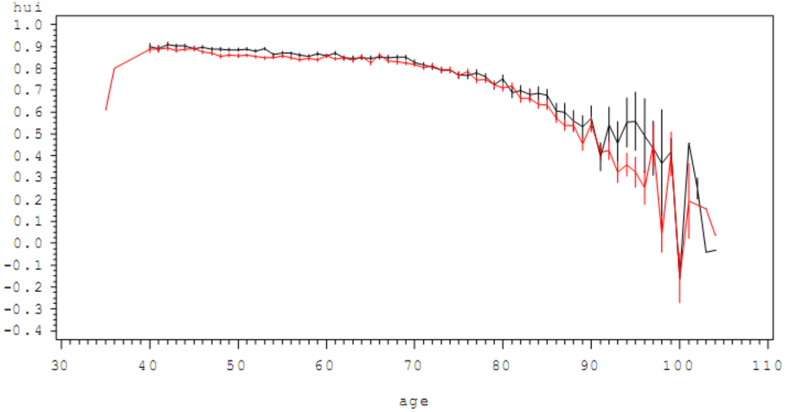
- Health-Related Quality of Life is measured by the Health Utilities Index Mark 3 (HUI)
- The HUI describes health status using eight factors
  - Vision, Hearing, Speech, Ambulation
  - Dexterity, Emotion, Cognition, Pain and Discomfort
- Each factor has 5 or 6 levels that range from severely impaired to no impairment
- HUI is obtained based on the combination of all factor levels

Health-Related Quality of Life

- HUI scores can range from -0.36 to 1.00
  - A score of 1.00 represents perfect health
  - A score of 0 represents the state of being dead
  - A score less than 0 is a state "worse than dead"
  - Scores less than 0 are possible because a health status can be considered as less preferable than being dead

Health-Related Quality of Life

The average HUI for respondents after age 40



Male: ———; Female: ———

Household Income

Household Income

The NPHS employs various of indexes to evaluate the income level of respondents

- Total household income
- Total personal income
- Food insecurity
- Distribution of household income - national level
- Distribution of household income - provincial level
- ...

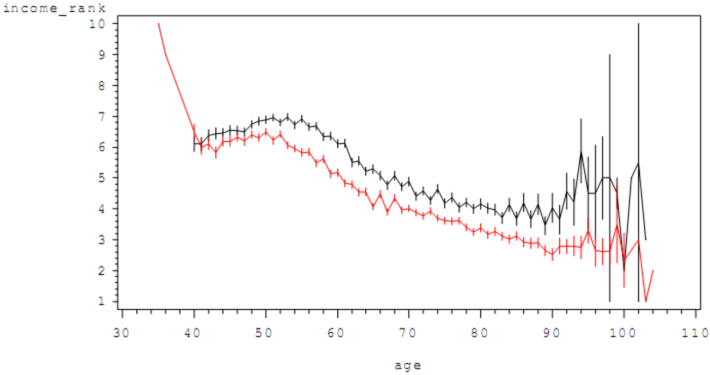
Household Income

For the variable "Distribution of household income - provincial level",

- it represents the ranking of household income
- it ranges from 1 to 10
- 10 represents the highest income decile in the entire sample
- 1 represents the lowest income decile in the entire sample

Household Income

The average household income for respondents after age 40



Male: ———; Female: ———

Marital Status

- 6 categories
  - married, living common-law, living with a partner
  - widowed, separated, divorced, single (never married)

Variable Transformation

- Marriage=1 if married, living common-law, or living with a partner
- Marriage=0 if widowed, separated, divorced, or single (never married)

Education

- 14 categories
  - no schooling, elementary school, secondary school graduation
  - bachelor degree, master degree, degree in medicine, doctorate degree
  - ...

Variable Transformation

- Two dummy variables: Education 1 (secondary school level)
  Education 2 (college level)
  - Education1=0 and Education2=0, if no schooling or elementary school
  - Education1=0 and Education2=1, if bachelor or higher degree
  - Education1=1 and Education2=0, otherwise

Missing Data in the NPHS

Missing Data in the NPHS

- The NPHS started with a sample of 17276 individuals
- The NPHS data are subject to information incompletion
- Three main possible reasons of incompleteness
  - non-tracing
  - refusal or unknown to question items
  - death

Missing Data: Non-tracing

- "Non-tracing" denotes the situation that interviewers failed to reach the respondents
- To deal with non-tracing issue, many approaches were introduced into the survey
  - workload restriction
  - interviewers training
  - tracking individuals who moved within Canada or to United States
- Despite those efforts, the non-tracing rate in all 17276 members increased over time

Cycle 2	$\longrightarrow$	Cycle 7
1.7%	$\longrightarrow$	5.4%

Missing Data: Refusal or Unknown to Question

- Respondents may refuse to participate in the survey because of personal privacy, time schedule arrangement or other concerns
- The NPHS made efforts to persuade all members to continue the study
  - persuasive letter
  - senior interviewers
- Though many strategies were applied, refusal rate in survey sample increased from 3.1% in cycle 1 to 13.2% in cycle 7
- Respondents might attend the survey but refuse to answer some questions
  - A typical example: respondents may finish other questions but refuse to report their income status
- For some questions, respondents may not be sure about the answers and just report "unknown"

Missing Data: Death

- Until cycle 7, there are 2032 (11.76%) members died before the end of the NPHS
- Death leads to another source of information loss that may not be well handled by general approaches

Simple Imputation for Missing Data

Ad Hoc Approach of Handling Missing Data:

Age

- Add 2 years from last observation

Education & Marriage

- Education: time-invariant variable
  use education record in cycle 1
- Marriage: impute missing value by the "last observation carried forward" method

A Subset of the NPHS Data

A Subset of the NPHS Data

- Longitudinal Data: 6 Cycles; 1349 subjects
- Age 50-70 at cycle 1; Still alive at cycle 6; Male
- Response: Health Utility Index
- Incomplete Covariate: Household Income
- Complete Covariates: Age, Education, Marital Status

Rate	Health Utility Index						Household Income					
	1	2	3	4	5	6	1	2	3	4	5	6
43.2%	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4.2%	✓	×	×	×	×	×	✓	×	×	×	×	×
...												
2%	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
1%	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓
1%	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

✓ Observed; × Missing

Notation (For Individual i)

Response & Covariates

Visit:	1	2	3	4	$\cdots$	m
$\mathbf{Y}_i$:	✓	✓	×	✓	$\cdots$	✓
$\mathbf{X}_i$:	×	✓	×	×	$\cdots$	×
$\mathbf{Z}_i$:	✓	✓	✓	✓	✓	✓

- Incomplete response: Y_{ij} - scalar, $\mathbf{Y}_i = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$
- Incomplete covariate: X_{ij} - scalar, $\mathbf{X}_i = (\mathbf{X}_i^o, \mathbf{X}_i^m)$
- Complete covariates: $\mathbf{Z}_{ij}$ - scalar/vector (including intercept)

Missing Data Indicator

- For Y_{ij}: $R_{ij}^y = \mathrm{I}(Y_{ij} \text{ is observed})$
- For X_{ij}: $R_{ij}^x = \mathrm{I}(X_{ij} \text{ is observed})$

Inference Strategy - Observed Likelihood

- Inference framework:
  $f(R_i^y, R_i^x, \mathbf{y}_i, \mathbf{x}_i|\mathbf{z}_i) = f(R_i^y, R_i^x|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i)f(\mathbf{x}_i|\mathbf{z}_i)$
- Strategy: observed likelihood

$$L_i = \int\int f(R_i^y, R_i^x|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i)f(\mathbf{x}_i|\mathbf{z}_i)\mathrm{d}\mathbf{y}_i^m \mathrm{d}\mathbf{x}_i^m \quad (1)$$

Missing Data Mechanism

- MCAR: $f(R_i^y, R_i^x|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i) = f(R_i^y, R_i^x|\mathbf{z}_i)$
- MAR: $f(R_i^y, R_i^x|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i) = f(R_i^y, R_i^x|\mathbf{y}_i^o, \mathbf{x}_i^o, \mathbf{z}_i)$
- **MNAR**: $f(R_i^y, R_i^x|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i) = f(R_i^y, R_i^x|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{x}_i^o, \mathbf{x}_i^m, \mathbf{z}_i)$
  typically depends on unobserved $\mathbf{y}_i^m, \mathbf{x}_i^m$

Computational Challenge: High Dimensional Integrals

- Modeling missing data process is generally required if **MNAR** holds when using likelihood-based methods
- High dimensional integrals would be involved in the observed likelihood (1)
- Other Possible Options:
  EM algorithm (e.g. Roy & Lin 2002)
  MCEM algorithm (e.g. Stubbendick & Ibrahim 2003)
- Challenges: difficult in modeling
  computationally expensive
  not robust

Proposed Methods to Address the Challenges: Composite Likelihood Method

Proposal: Composite Likelihood Method

- Composite likelihood consists of a combination of valid likelihood objects corresponding to a marginal or conditional event **in small subsets of data**
- Suppose **correlated** random variables $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$:
  Marginal uni-wise likelihood $L_{C1}(\mathbf{Y}_i) = \prod_{j=1}^{m} f(Y_{ij})$
  Marginal pairwise likelihood $L_{C2}(\mathbf{Y}_i) = \prod_{j<k} f(Y_{ij}, Y_{ik})$
- Unbiasedness: $E[S(\beta)] = E\left\{\frac{\partial log L_C}{\partial \beta}\right\} = 0$

  Remark: This ensures the resulting estimator $\hat{\beta}$ is consistent
- Asymptotic Distribution:

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_D N(0, J(\beta)^{-1}\{K(\beta)\}[J(\beta)^{-1}]^T)$$

where $\hat{J}(\beta) = \frac{1}{n}\sum_i -\left\{\frac{\partial S_i(\beta)}{\partial \beta}\right\}_{\beta=\hat{\beta}}$, $\hat{K}(\beta) = \frac{1}{n}\sum_i S_i(\hat{\beta})\{S_i(\hat{\beta})\}^T$

Analysis of NPHS Data by Pairwise Model

- Response Process
  $(HUI_{ij}, HUI_{ik}) \sim N_2(\mu_{ij}^{HUI}, \mu_{ik}^{HUI}; \Sigma_{HUI}(\sigma_y^2, \sigma_{jk}^{HUI})),$
  - $\mu_{ij}^{HUI} = \beta_0 + \beta_1 INC_{ij} + \beta_2(AGE_{ij} - 50) + \beta_3 EDU1_i + \beta_4 EDU2_i + \beta_5 MARR_{ij}$
- Covariate Process
  $(INC_{ij}, INC_{ik}) \sim N_2(\mu_{ij}^{INC}, \mu_{ik}^{INC}; \Sigma_{INC}(\sigma_x^2, \sigma_{jk}^{INC}))$
  - $\mu_{ij}^{INC} = \alpha_0 + \alpha_1(AGE_{ij} - 50) + \alpha_2 EDU1_i + \alpha_3 EDU2_i + \alpha_4 MARR_{ij}$
- Missing Process
  $f(r_{ij}^y = 1, r_{ik}^y = 1) = \Phi_2(\mu_{ij}^y, \mu_{ik}^y; \rho^y),$

$$\mu_{ij}^y = \Phi(\eta_0^y + \eta_1^y HUI_{ij} + \eta_2^y INC_{ij} + \eta_3^y Age_{ij})$$

  $f(r_{ij}^x = 1, r_{ik}^x = 1) = \Phi_2(\mu_{ij}^x, \mu_{ik}^x; \rho^x),$

$$\mu_{ij}^x = \Phi(\eta_0^x + \eta_1^x HUI_{ij} + \eta_2^x INC_{ij} + \eta_3^x r_{ij}^y + \eta_4^x Age_{ij})$$

  where
  $\Phi(\cdot)$ is standard normal distribution function
  $\Phi_2(\mu_1, \mu_2; \rho)$ is standard bivariate normal distribution function

Analysis of NPHS Data by Pairwise Model

		Composite Likelihood			Available Data		
		Est.	S.E	P-value	Est.	S.E	P-value
Intercept	β_0	0.754	0.021	< 0.001	0.795	0.016	< 0.001
INC	β_1	**0.012**	0.001	< 0.001	**0.006**	0.001	< 0.001
AGE	β_2	-0.001	0.001	0.253	-0.002	< 0.001	< 0.001
EDU1	β_3	0.026	0.015	0.085	0.030	0.013	0.021
EDU2	β_4	0.050	0.017	0.003	0.063	0.017	< 0.001
MARR	β_5	0.030	0.011	0.007	0.021	0.008	0.010

Composite Likelihood in Handling Variable Selection

Variable Selection Problem

- Response: Health Utility Index (HUI)
- Candidate Variables
  - alcohol dependence, chronic conditions, drugs
  - health care, injuries, mental health
  - nutrition, physical activities, self care
  - smoking, social support, stress
  - $\cdots$ many more
- Question: how do we know what variables should be included when building a model to explain response variable *HUI* ?

Proposal: Model Selection via Composite Likelihood

Our methods: penalized composite likelihood

$$logL_{pen}(Y) = logL_C(Y) - n\sum_{s=1}^{p} p_\lambda(|\beta_s|)$$

- $p_\lambda(|\beta_s|)$ is the penalty function for the s-th element in $\boldsymbol{\beta}$
- Choice of penalty functions is not unique. Fan and Li (2001) suggest the SCAD penalty

$$p'_\lambda(\beta_s) = \lambda\left\{I(\beta_s \leq \lambda) + \frac{(a\lambda - \beta_s)_+}{(a-1)\lambda} I(\beta_s > \lambda)\right\}$$

Asymptotic Results for Penalized Composite Likelihood

Model parameter: $\boldsymbol{\beta} = (\boldsymbol{\beta}_I^T, \boldsymbol{\beta}_{II}^T)^T$

- $\boldsymbol{\beta}_I \neq \mathbf{0}$: corresponding to "important" variables
- $\boldsymbol{\beta}_{II} = \mathbf{0}$: corresponding to "unimportant" variables

- Theorem 1:
  There exists a local maximizer of $logL_{pen}(Y)$ *such that*

$$\|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I\| = O_p(n^{-1/2})$$

- Theorem 2:
  With probability tending to 1, the root-n consistent local maximizers $\hat{\boldsymbol{\beta}}$ *satisfies:*
  (a) Sparsity: $\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}$
  (b) Asymptotic normality for $\hat{\boldsymbol{\beta}}_I$

Example

- Response: Health Utility Index (HUI)
- Candidate Variables
  - Household Income (INC), Age (Age)
  - INC^2, INC^3, Age^2, Age^3
  - Interaction terms (e.g. $\mathrm{INC} \times \mathrm{Age}$, $\mathrm{INC}^2 \times \mathrm{Age}$, etc)

The Analysis of NPHS Data

Variable	Maximum Likelihood		Composite Likelihood	
	Full Model	Selected Model	Full Model	Selected Model
Intercept	$-0.02(0.04)$	0.00(0.03)	-0.02(0.04)	0.01(0.03)
INC	0.11(0.06)	0.09(0.01)	0.16(0.07)	0.10(0.02)
INC^2	-0.01(0.03)		-0.02(0.03)	-0.01(0.01)
INC^3	-0.00(0.03)		0.01(0.04)	0.04(0.01)
Age	0.35(0.22)	0.07(0.02)	0.23(0.20)	0.08(0.02)
Age^2	-0.28(0.32)		-0.08(0.30)	
Age^3	0.03(0.12)	-0.04(0.01)	-0.03(0.11)	-0.03(0.01)
$\text{INC} \times \text{Age}$	-0.25(0.38)		-0.35(0.39)	
$\text{INC}^2 \times \text{Age}$	-0.13(0.17)		-0.04(0.16)	
$\text{INC}^3 \times \text{Age}$	0.24(0.20)		0.27(0.22)	
$\text{INC} \times \text{Age}^2$	0.29(0.55)		0.44(0.57)	
$\text{INC}^2 \times \text{Age}^2$	0.16(0.24)		0.04(0.24)	
$\text{INC}^3 \times \text{Age}^2$	-0.35(0.30)		-0.40(0.32)	
$\text{INC} \times \text{Age}^3$	-0.09(0.20)		-0.15(0.21)	
$\text{INC}^2 \times \text{Age}^3$	-0.05(0.09)		-0.01(0.09)	
$\text{INC}^3 \times \text{Age}^3$	0.13(0.11)		0.15(0.12)	

Concluding Remarks

Summary & Comments

- When data have complex features, such as missing values and a large number of covariates, standard likelihood-based methods may become infeasible in
  - Model building
  - Computation implementation
  - Robustness
- Composite likelihood serves as an attractive alternative
- We particularly discuss a composite likelihood that handles incomplete data and model selection
- Computational gain: reduction in the dimensions of integrals
- Statistical gain: ease of modeling
  robustness