THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Marginal proportional hazards models for multivariate interval-censored data

Yangjianchen Xu

Department of Biostatistics
The University of North Carolina at Chapel Hill

November 11, 2022

Table of Contents

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Introduction

Multivariate Failure Time Data

- Multiple events: each subject can potentially experience several events
- Cluster: natural/artificial clustering of study subjects

Interval Censoring: Failure occurs within a time interval

Medical Research: Periodic monitoring of asymptomatic diseases

- diabetes and hypertension in family members
- asymptomatic SARS-Cov-2 infection

Theoretical/Computational Issues

- no exact failure time
- unknown dependence structures

Existing methods

Random-effects models for multivariate interval-censored data

- Chen et al. (2009): current-status data
- Chen et al. (2014): a common set of examination times for all subjects
- Wen & Chen (2013): asymptotic theory for bivariate interval-censored data
- Chang et al. (2007): asymptotic theory for current-status family data
- Zeng et al. (2017): transformation models for general multivariate interval-censored data

Marginal models for multivariate interval-censored data

- assume examinations occur at a common set of time points
- parameterize the event time distribution
- Goggins & Finkelstein, 2000; Kim & Xue, 2002; Chen et al., 2007; Tong et al., 2008; Chen et al., 2013; Shen, 2015; Zhang & Sun, 2010; Kor et al., 2013

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Notation

- n = number of clusters
- n_i = number of subjects in the ith cluster
- K = number of types of failures
- $X_{ijk}(\cdot)$ = (potentially time-dependent) covariates
- T_{ijk} = kth failure time for the jth subject of the ith cluster
- $U_{ijk1} < \cdots < U_{ijk,M_{ijk}}$ = random sequence of examination times on T_{ijk}
- $(L_{ijk}, R_{ijk}]$ = the shortest time interval that brackets T_{ijk}

Data, models and pseudo-likelihood

Data: $(L_{ijk}, R_{ijk}, X_{ijk})$ $(i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, K)$

Marginal Proportional Hazards Models

$$\lambda_{ijk}(t) = \lambda_k(t) \exp\left\{\beta_k^{\mathrm{T}} X_{ijk}(t)\right\}$$

- β_k = regression parameters
- $\lambda_k(t)$ = arbitrary baseline hazard function
- $\Lambda_k(t) = \int_0^t \lambda_k(s)\mathrm{d}s$

Pseudo-Likelihood

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\prod_{k=1}^{K}\left[\exp\left\{-\int_0^{L_{ijk}} e^{\beta_k^{\mathrm{T}} X_{ijk}(t)}\mathrm{d}\Lambda_k(t)\right\} - \exp\left\{-\int_0^{R_{ijk}} e^{\beta_k^{\mathrm{T}} X_{ijk}(t)}\mathrm{d}\Lambda_k(t)\right\}\right]$$

Estimation procedure

Nonparametric Maximum Pseudo-Likelihood Estimation

- $0 < t_{k0} < t_{k1} < \cdots < t_{km_k} < \infty = \{L_{ijk} > 0, R_{ijk} < \infty; i = 1, \ldots, n; j = 1, \ldots, n_i\}$
- λ_{kq} = jump size of $\Lambda_k(\cdot)$ at t_{kq}
- $X_{ijkq} = X_{ijk}(t_{kq})$
- For each k, we maximize

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left\{\exp\left(-\sum_{t_{kq}\leqslant L_{ijk}}\lambda_{kq}e^{\beta_k^{\mathrm{T}}X_{ijkq}}\right)-I(R_{ijk}<\infty)\exp\left(-\sum_{t_{kq}\leqslant R_{ijk}}\lambda_{kq}e^{\beta_k^{\mathrm{T}}X_{ijkq}}\right)\right\}$$

Estimation procedure

Implementation

- Direct maximization
  - non-concave likelihood
  - no analytic expression for λ_{kq}
  - many λ_{kq} are zero
- EM-type algorithm
  - latent Poisson variables with same observed-data likelihood
  - analytic expression for λ_{kq}
  - partial-likelihood like estimating equation for β_k
  - observed-data pseudo-likelihood increases at each iteration

EM-type algorithm

Latent variables: $W_{ijkq} \overset{\text{ind}}{\sim} \text{Poisson}\left(\lambda_{kq} e^{\beta_k^{\mathrm{T}} X_{ijkq}}\right)$

$(i = 1, \ldots, n; j = 1, \ldots, n_i; k = 1, \ldots, K; q = 1, \ldots, m_k)$

Observed data: $(L_{ijk}, R_{ijk}, X_{ijk}, A_{ijk} = 0, B_{ijk} > 0)$

- $A_{ijk} = \sum_{t_{kq} \leqslant L_{ijk}} W_{ijkq}$
- $B_{ijk} = I\left(R_{ijk} < \infty\right) \sum_{L_{ijk} < t_{kq} \leqslant R_{ijk}} W_{ijkq}$

Observed-data likelihood

$$\prod_{i=1}^{n} \prod_{j=1}^{n_i} \left\{ \prod_{t_{kq} \leqslant L_{ijk}} \Pr\left(W_{ijkq} = 0\right) \right\} \left\{ 1 - \Pr\left(\sum_{L_{ijk} < t_{kq} \leqslant R_{ijk}} W_{ijkq} = 0 \right) \right\}^{I(R_{ijk} < \infty)}$$

Complete-data log-likelihood

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{q=1}^{m_k} I(R_{ijk}^{*} \geqslant t_{kq}) \left\{ W_{ijkq} \log(\lambda_{kq} e^{\beta_k^{\mathrm{T}} X_{ijkq}}) - \lambda_{kq} e^{\beta_k^{\mathrm{T}} X_{ijkq}} - \log W_{ijkq}! \right\},$$

- $R_{ijk}^{*} = I(R_{ijk} < \infty) R_{ijk} + I(R_{ijk} = \infty) L_{ijk}$

EM-type algorithm

E-step

$$\widehat{E}(W_{ijkq}) = I(L_{ijk} < t_{kq} \leqslant R_{ijk} < \infty) \frac{\lambda_{kq} \exp(\beta_k^{\mathrm{T}} X_{ijkq})}{1 - \exp\{-\sum_{L_{ijk} < t_{kq'} \leqslant R_{ijk}} \lambda_{kq'} \exp(\beta_k^{\mathrm{T}} X_{ijkq'})\}}.$$

M-step

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{q=1}^{m_k} I(R^*_{ijk} \geqslant t_{kq}) \widehat{E}(W_{ijkq}) \times \left\{ X_{ijkq} - \frac{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(R^*_{i'j'k} \geqslant t_{kq}) \exp(\beta_k^{\mathrm{T}} X_{i'j'kq}) X_{i'j'kq}}{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(R^*_{i'j'k} \geqslant t_{kq}) \exp(\beta_k^{\mathrm{T}} X_{i'j'kq})} \right\} = 0.$$

We then update

$$\lambda_{kq} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(R^*_{ijk} \geqslant t_{kq}) \widehat{E}(W_{ijkq})}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(R^*_{ijk} \geqslant t_{kq}) \exp(\beta_k^{\mathrm{T}} X_{ijkq})}$$

Asymptotic properties

Let $\widehat{\beta} = \left(\widehat{\beta}_1^{\mathrm{T}}, \ldots, \widehat{\beta}_K^{\mathrm{T}}\right)^{\mathrm{T}}$ and $\widehat{\Lambda} = \left(\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_K\right)$.

Consistency

Theorem 1

Under some regularity conditions, $\|\widehat{\beta} - \beta_0\| + \sum_{k=1}^{K} \sup_{t \in [0, \tau_k]} |\widehat{\Lambda}_k(t) - \Lambda_{0k}(t)| \to 0$ *almost surely, where* $\|\cdot\|$ *is the Euclidean norm.*

Asymptotic distribution

Theorem 2

Under some regularity conditions, $n^{1/2}(\widehat{\beta} - \beta_0)$ *converges in distribution to a zero-mean multivariate normal random vector with covariance matrix* Ω.

Sandwich variance estimator

Profile pseudo-log-likelihood for β_k

$$\mathrm{pl}_k(\beta_k) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \log\left\{ \exp\left(- \sum_{t_{kq}\leqslant L_{ijk}} \widetilde{\lambda}_{kq} e^{\beta_k^{\mathrm{T}} X_{ijkq}} \right) - I\left(R_{ijk} < \infty\right) \exp\left(- \sum_{t_{kq}\leqslant R_{ijk}} \widetilde{\lambda}_{kq} e^{\beta_k^{\mathrm{T}} X_{ijkq}} \right) \right\}$$

- $\widetilde{\lambda}_{kq}\,(q = 1, \ldots, m_k)$ are obtained from EM with fixed β_k

Covariance matrix estimator between $\widehat{\beta}_k$ and $\widehat{\beta}_l$

$$\widehat{V}_{kl} = \left\{ D_{h_n}^2 \mathrm{pl}_k(\widehat{\beta}_k) \right\}^{-1} \sum_{i=1}^{n} D_{h_n}\mathrm{pl}_{ki}(\widehat{\beta}_k) D_{h_n}\mathrm{pl}_{li}(\widehat{\beta}_l)^{\mathrm{T}} \left\{ D_{h_n}^2 \mathrm{pl}_l(\widehat{\beta}_l) \right\}^{-1}$$

- $\mathrm{pl}_{ki}\left(\beta_k\right) =$ contribution of the i th cluster to $pl_k\left(\beta_k\right)$

Sandwich variance estimator

Theorem 3

Under some regularity conditions, $\{n(\widehat{V}_{kl}); 1 \leqslant k, l \leqslant K\}$ is a consistent estimator for the limiting covariance matrix Ω.

Statistical Inference

$$L\widehat{\beta} \sim N\left(L\beta, LVL'\right)$$

$$V = \begin{bmatrix} V_{11} & \cdots & V_{1K} \\ \vdots & \vdots & \vdots \\ V_{K1} & \cdots & V_{KK} \end{bmatrix}$$

- linear combinations (e.g., a subset of parameters, difference of two parameters)

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Simulation studies

Clustered data

- $K = 1$
- n_i varies from 1 to 5 with probabilities 0.1, 0.5, 0.2, 0.1 and 0.1
- Treatment assignment: cluster-level $X_1 \sim \mathrm{Ber}(0.5)$
- Correlated continuous covariate: subject-level $X_2 = X_1^* + X_2^*$
  - cluster-level $X_1^* \sim \mathrm{Un}(0, 1)$
  - subject-level $X_2^* = \max\{0, N(0, 0.04)\}$
- $\beta_1 = 0.5$, $\beta_2 = -0.5$
- Generate $(Z_1, \ldots, Z_{n_i}) \sim N_{n_i}(0, \Sigma)$ and $T_j^* = -\mu \log\{1 - \Phi(Z_j)\}$
  - Σ is an exchangeable correlation matrix with $\rho = 0.2$, 0.5 or 0.75
  - $\mu = 2 \exp\{-(\beta_1 X_1 + \beta_2 X_2)\}$
- Event times: $T_j = T_j^{*2}$
- Cumulative hazard function: $0.5 t^{1/2} \exp(\beta_1 X_1 + \beta_2 X_2)$

Simulation studies

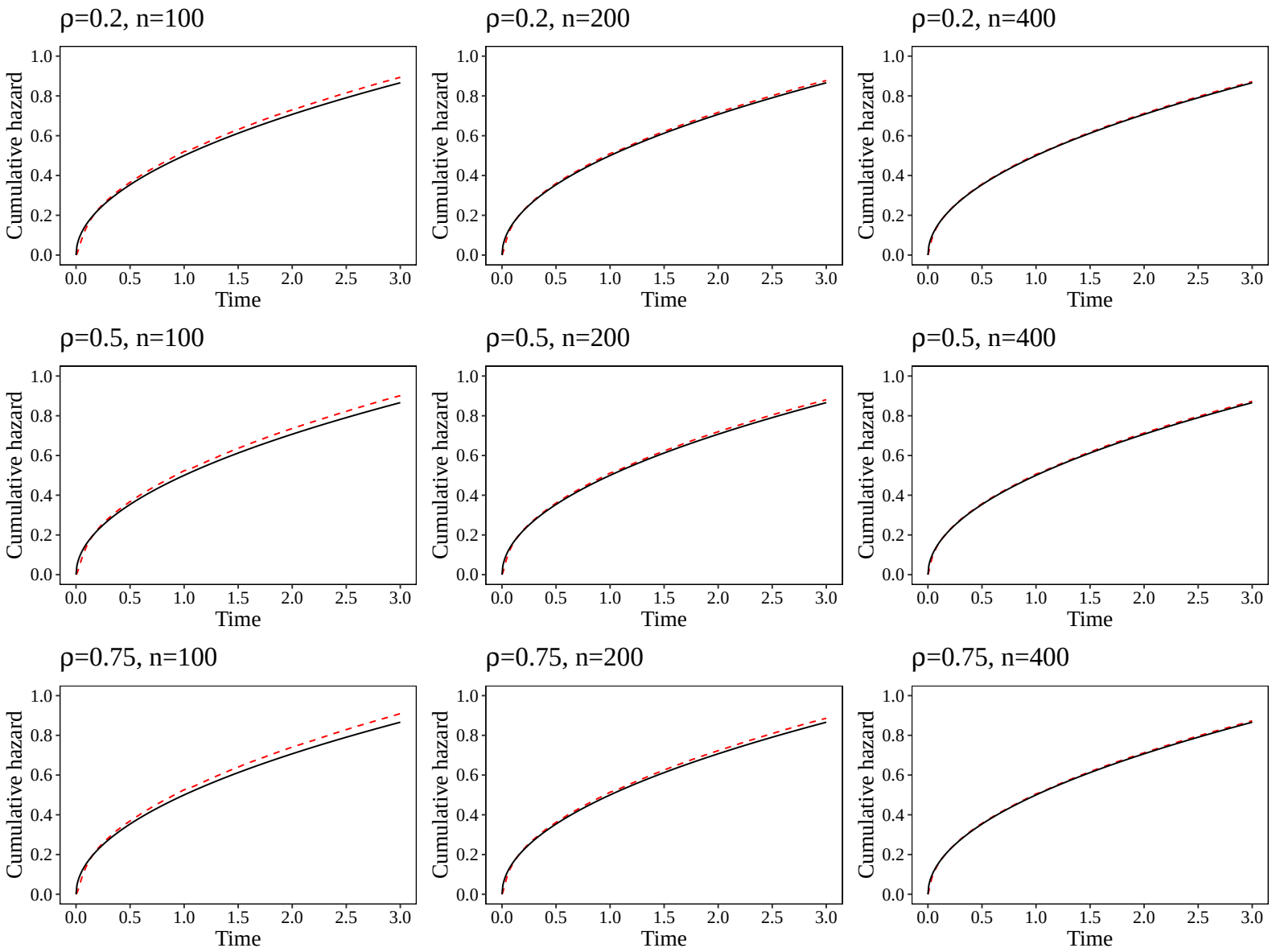
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Table 1. *Parameter estimation results for simulation studies with clustered data*

			Marginal model						Random-effects model			
ρ	No. of clusters	Parameter	Bias	SE	SEE	CP	SEEn	CPn	Bias	SE	SEE	CP
0.2	$n = 100$	$\beta_1 = 0.5$	0.005	0.202	0.211	95.5	0.182	92.0	0.060	0.222	0.220	94.2
		$\beta_2 = -0.5$	−0.012	0.314	0.314	94.8	0.279	91.9	−0.064	0.355	0.344	94.2
	$n = 200$	$\beta_1 = 0.5$	0.005	0.142	0.145	95.4	0.125	91.6	0.060	0.157	0.154	93.1
		$\beta_2 = -0.5$	−0.006	0.219	0.216	94.6	0.194	91.7	−0.069	0.244	0.241	94.2
	$n = 400$	$\beta_1 = 0.5$	0.003	0.097	0.100	95.7	0.087	92.3	0.059	0.110	0.109	91.7
		$\beta_2 = -0.5$	−0.004	0.154	0.152	94.5	0.136	91.5	−0.061	0.171	0.170	93.4
0.5	$n = 100$	$\beta_1 = 0.5$	0.007	0.238	0.247	95.7	0.182	86.8	0.224	0.335	0.326	89.4
		$\beta_2 = -0.5$	−0.013	0.369	0.359	94.1	0.279	86.5	−0.229	0.493	0.483	92.6
	$n = 200$	$\beta_1 = 0.5$	0.004	0.165	0.169	95.6	0.125	86.5	0.215	0.233	0.227	84.7
		$\beta_2 = -0.5$	−0.005	0.253	0.249	94.5	0.194	86.7	−0.210	0.340	0.336	90.7
	$n = 400$	$\beta_1 = 0.5$	0.002	0.114	0.118	95.7	0.087	86.5	0.209	0.162	0.160	74.6
		$\beta_2 = -0.5$	−0.002	0.178	0.174	94.6	0.136	87.0	−0.209	0.238	0.236	85.7
0.75	$n = 100$	$\beta_1 = 0.5$	0.013	0.266	0.278	95.9	0.182	82.2	0.515	0.521	0.493	82.2
		$\beta_2 = -0.5$	−0.013	0.410	0.399	93.9	0.279	82.2	−0.519	0.699	0.677	88.4
	$n = 200$	$\beta_1 = 0.5$	0.006	0.184	0.191	95.6	0.125	82.0	0.527	0.359	0.346	67.0
		$\beta_2 = -0.5$	−0.007	0.285	0.277	94.2	0.194	81.7	−0.524	0.491	0.473	80.0
	$n = 400$	$\beta_1 = 0.5$	0.004	0.130	0.132	95.4	0.087	81.4	0.516	0.254	0.239	42.8
		$\beta_2 = -0.5$	−0.001	0.197	0.194	94.4	0.136	82.6	−0.506	0.338	0.331	66.4

SE, standard error of the parameter estimator; SEE, mean of the proposed standard error estimator; SEEn, mean of the naive standard error estimator; CP, coverage percentage of the 95% confidence interval based on the sandwich variance estimator; CPn, coverage percentage of the 95% confidence interval based on the naive variance estimator.

Simulation studies



Simulation studies

Multiple-event data

- $K = 2$
- $n_i = 1$
- Subject-level $X_1 \sim \mathrm{Ber}(0.5)$ and $X_2 = \mathrm{Un}(0, 1)$
- $(\beta_{11}, \beta_{12}) = (0.5, -0.5)$ and $(\beta_{21}, \beta_{22}) = (0.5, 0.5)$
- Generate $T_j^* = -\mu_j \log\{1 - \Phi(Z_j)\}$ $(j = 1, 2)$
  - (Z_1, Z_2): standard normal with correlation $\rho = 0.2$, 0.5 or 0.75
  - $\mu_1 = 2 \exp\{-(\beta_{11}X_1 + \beta_{12}X_2)\}$
  - $\mu_2 = 5 \exp\{-(\beta_{21}X_1 + \beta_{22}X_2)\}$
- Event times: $(T_1, T_2) = (T_1^{*2}, T_2^*)$

Simulation studies

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Table 2. *Parameter estimation results for simulation studies with multiple-event data*

		First event				Second event				Optimal combination			
ρ	No. of subjects	Bias	SE	SEE	CP	Bias	SE	SEE	CP	Bias	SE	SEE	CP
0.2	$n = 100$	0.019	0.287	0.297	95.9	0.018	0.281	0.289	95.8	0.007	0.213	0.223	96.0
	$n = 200$	0.009	0.196	0.202	95.8	0.010	0.191	0.196	95.7	0.005	0.147	0.151	95.5
	$n = 400$	0.005	0.136	0.139	95.5	0.006	0.134	0.135	95.2	0.004	0.103	0.104	95.1
0.5	$n = 100$	0.020	0.285	0.297	96.2	0.026	0.280	0.289	96.0	0.011	0.232	0.248	96.3
	$n = 200$	0.010	0.195	0.202	95.9	0.014	0.191	0.196	95.8	0.008	0.160	0.167	96.1
	$n = 400$	0.007	0.138	0.139	95.4	0.007	0.133	0.135	95.7	0.005	0.112	0.115	95.7
0.75	$n = 100$	0.020	0.282	0.297	96.2	0.021	0.276	0.289	96.0	0.004	0.260	0.268	96.5
	$n = 200$	0.009	0.195	0.202	96.0	0.009	0.188	0.195	95.8	0.003	0.170	0.180	96.4
	$n = 400$	0.006	0.136	0.139	95.7	0.010	0.134	0.135	95.2	0.006	0.120	0.124	95.8

SE, standard error of the parameter estimator; SEE, mean of the proposed standard error estimator; CP, coverage percentage of the 95% confidence interval based on the proposed variance estimator.

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

ARIC study

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Atherosclerosis Risk in Communities Study (ARIC): cohort of 14,751 white and black individuals from 4 U.S. communities

Examinations: 1987–1989 → (3-year intervals) → 2011–2013

Diabetes

- fasting glucose $\geqslant 126 \mathrm{mg/dL}$
- non-fasting glucose $\geqslant 200 \mathrm{mg/dL}$
- self-reported physician diagnosis of diabetes
- use of diabetic medication

Hypertension:

- systolic blood pressure $\geqslant 140$
- diastolic blood pressure $\geqslant 90$
- use of anti-hypertensive medication

Analysis set: 8,735 individuals without prior diabetes or hypertension

ARIC study

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Table 3. *Regression analysis of the Atherosclerosis Risk in Communities Study*

	Diabetes			Hypertension			Overall test		Difference		
Risk factor	Est	SE	p-value	Est	SE	p-value	Test	p-value	Est	SE	95% CI
Jackson	−0.145	0.149	0.332	−0.239	0.077	0.002	10.14	0.006	0.094	0.162	(−0.234, 0.413)
Minneapolis suburbs	−0.389	0.076	$< 10^{-4}$	−0.100	0.046	0.031	29.17	$< 10^{-4}$	−0.289	0.085	(−0.455, −0.122)
Washington County	0.115	0.073	0.114	0.078	0.048	0.103	4.68	0.096	0.037	0.083	(−0.125, 0.199)
Age	−0.014	0.005	0.007	0.013	0.003	$< 10^{-4}$	26.17	$< 10^{-4}$	−0.027	0.006	(−0.038, −0.016)
Male	−0.062	0.055	0.265	−0.238	0.034	$< 10^{-4}$	49.34	$< 10^{-4}$	0.176	0.062	(0.056, 0.297)
Caucasian	−0.451	0.160	0.005	−0.480	0.081	$< 10^{-4}$	40.29	$< 10^{-4}$	0.029	0.172	(−0.307, 0.366)
Body mass index (kg/m^2)	0.075	0.005	$< 10^{-4}$	0.017	0.004	$< 10^{-4}$	236.83	$< 10^{-4}$	0.059	0.006	(0.047, 0.070)
Derived glucose value (mg/dl)	0.096	0.003	$< 10^{-4}$	0.001	0.002	0.595	961.78	$< 10^{-4}$	0.095	0.004	(0.088, 0.102)
Systolic blood pressure (mmHg)	0.005	0.003	0.096	0.058	0.002	$< 10^{-4}$	914.31	$< 10^{-4}$	−0.053	0.003	(−0.060, −0.046)
Diastolic blood pressure (mmHg)	0.005	0.004	0.310	0.011	0.003	$< 10^{-4}$	17.48	0.0002	−0.007	0.005	(−0.016, 0.003)

Est, estimate; SE, standard error; CI, confidence interval.

Outline

Future work

- Model checking
- A more flexible working dependence structure
- Intermittent missing values or measurement error of covariates

Thank you!