HUDSON
COLLEGE OF PUBLIC HEALTH
The UNIVERSITY of OKLAHOMA HEALTH SCIENCES

Statistical Analysis of Family Studies with Known Kinship Matrices: Applications to the Strong Heart Family Study

Daniel Zhao, PhD
Department of Biostatistics and Epidemiology

March 26, 2025

Outline

- Introduction
  - Strong Heart Study
  - Strong Heart Family Study
- Motivation and Research Questions
- Three Aims
  - Statistical Models
  - Simulation Studies
  - Real Data Example
- Overall Summary

Introduction

Strong Heart Study (strongheartstudy.org)

- Largest study of cardiovascular disease (CVD) and its risk factors in Native Americans (NA)
- Included 13 tribes, 4549 NAs in OK, AZ, SD, and ND
- Since 1989, six study phases have been completed, and the 7th phase is ongoing
- Since Phase III, a family study was conducted
- Over 600 publications and over 90,000 citations

Strong Heart Family Study (SHFS)

- Family studies are studies of whether a disease run in a family
- Started in 1998, SHFS enrolled 3,776 individuals from 94 families
- Family sizes ranged from 1 to 113, with a median of 31, Q1 16 and Q3 39
- Goal: Investigate the heritability of CVD
- Kinship coefficients were directly obtained by genetic test and interview

Kinship Coefficients and Kinship Matrix

- Kinship coefficient: probability that alleles randomly selected from two individuals are identical by descent
- Kinship coefficient is a measure of relatedness, ranges from 0 to 0.5
  - 0: unrelated two individuals
  - .5: identical twins
  - .25: two full siblings
- Twice of the kinship coefficient is a correlation
- Kinship matrix is a symmetric matrix that stores kinship coefficients between any two individuals

Motivation of Our Research

- Research Questions
  - Is Generalize Estimating Equations (GEE) Model a proper statistical model for SHFS?
  - Can we utilize the kinship coefficients in the statistical analysis of SHFS data
- Sophia Chen’s Dissertation Topics
  - Aim 1: Continuous Outcomes
  - Aim 2: Binary Outcomes
  - Aim 3: Survival Outcomes

Aim 1: Continuous Outcomes

Epidemiology, Biostatistics, and Public Health, 2023. 18(1): 61-67

GEE Model

- GEE model is a popular statistical model for correlated data
- It has two components for making inference on the population level
- Marginal mean model: $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$
- Working correlation matrix:
  - Independent, exchangeable
  - Allows for making valid inference even if mis-specified
- Potential drawbacks for application on family studies
  - Huge variation of family sizes
  - Does not incorporate kinship coefficients

Bayesian Model

- Conditional model: $\boldsymbol{y_i} = \boldsymbol{X_i \beta} + \boldsymbol{b_i} + \boldsymbol{\epsilon}_i$
  - $\boldsymbol{y}$ is the vector of outcomes from the ith family
  - $\boldsymbol{\beta}$ is the population regression coefficients
  - $\boldsymbol{b_i}$ is the vector of random effects from the ith family
    $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{A_i})$,
    $\boldsymbol{A_i}$ is twice the kinship matrix, σ_g^2 is the genetic variance
  - $\boldsymbol{\epsilon_i} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$,
- Derived marginal model is the same as the mean model of the GEE, making the comparison between two models straightforward

Simulation Setup (1)

- True model: $\boldsymbol{y_i} = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ gender} + \boldsymbol{b_i} + \boldsymbol{\epsilon_i}$
  - Fixed effects: $\beta_0 = 1, \beta_1 = .08,\ \beta_2 = -0.5$
  - Random effects: $\boldsymbol{b_i} \sim N\left(\mathbf{0}, \sigma_g^2 \boldsymbol{A_i}\right)$,
  - Random error $\boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \boldsymbol{I})$
  - Age and gender were obtained from the SHFS
- Sample sizes: similar to SHFS
- 1000 Simulations

Simulation Setup (2)

- Kinship matrix
  - The one from SHFS
  - Singleton family: only one member
  - Nuclear family: father, mother, two children
  - Two-trios: two families with single child, and mothers are siblings
- Genetic variance $\sigma_g^2 = 1$

Simulation Results

- Both models had similar biases and coverage probabilities
- Biases were close to zero
- Coverage probabilities were close to 95%

SHFS Data Analysis

- Outcome: systolic blood pressure
- Covariates: age, sex, body mass index (BMI), diabetes status, smoking, and alcohol consumption.
- GEE (independent, exchangeable) and Bayesian model
- Point estimates and confidence intervals were compared

Results on Point Estimates and SE

	GEE (Independent)	GEE (Exchangeable)	Bayesian Model
Point Estimates			
Intercept	96.95	98.626	96.344
Age	0.41	0.41	0.416
Sex	-6.23	-6.328	-6.43
BMI	0.368	0.373	0.382
Diabetic	1.847	1.716	1.623
Current smoke	-0.113	-0.617	-0.334
Current drink	1.487	2.267	2.204
Standard Error			
Intercept	1.717	1.529	1.483
Age	0.023	0.023	0.018
Sex	0.681	0.666	0.549
BMI	0.05	0.044	0.039
Diabetic	0.89	0.886	0.637
Current smoke	0.719	0.698	0.589
Current drink	0.786	0.734	0.61

Results on 95% CI

	GEE (Independent)	GEE (Exchangeable)	Bayesian Model
95% CI			
Intercept	(93.584, 100.316)	(93.629, 99.623)	(93.62, 99.31)
Age	(0.364, 0.456)	(0.364, 0.455)	(0.381, 0.449)
Sex	(-7.563, -4.895)	(-7.633, -5.024)	(-7.464, -5.406)
BMI	(0.27, 0.466)	(0.287, 0.459)	(0.31, 0.454)
Diabetic	(0.103, 3.59)	(-0.02, 3.452)	(0.382, 2.837)
Current smoke	(-1.523, 1.3)	(-1.985, 0.752)	(-1.525, 0.84)
Current drink	(-0.053, 3.028)	(0.828, 3.705)	(0.99, 3.351)

Conclusion

For the analysis of continuous outcomes in family studies with a known kinship matrix

- Both the GEE model and the Bayesian model work well
- The choice depends on your need
  - Inference on the population level: GEE
  - Inference on the Individual level: Bayesian model

Aim 2: Binary Outcomes

In Press: *Journal of Biopharmaceutical Statistics, https://doi.org/10.1080/10543406.2024.2333516*

GEE Model and Bayesian Model

- GEE Model
  - Marginal mean model: $\text{logit}\,(\boldsymbol{p_i}) = \boldsymbol{X_i\beta}$
  - $\boldsymbol{p_i}$ is the vector of event rates from the ith family
  - $\boldsymbol{\beta}$ is the population regression coefficients
  - Working correlation matrix: Independent, exchangeable
- Bayesian Model
  - $\text{logit}\,(\boldsymbol{p_i}) = \boldsymbol{X_i\beta} + \boldsymbol{b_i}$
  - $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{A_i})$,
    $\boldsymbol{A_i}$ is twice the kinship matrix, σ_g^2 is the genetic variance

The UNIVERSITY of OKLAHOMA
HEALTH SCIENCES

GEE and Bayesian Model Comparison

- The derived marginal mean model from the Bayesian model $\neq$ the marginal mean model of GEE
- Direct comparison between GEE and Bayesian model is not straightforward
- We derived an approximate marginal mean model from the Bayesian model
- We also proposed C-statistics as a measure of performance for model comparison

Simulation Setup

- True model: $\text{logit}\,(\boldsymbol{p_i}) = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ gender} + \boldsymbol{b_i}$
  - Fixed effects: $\beta_0 = 1, \beta_1 = -.1,\ \beta_2 = 3$
  - Random effects: $\boldsymbol{b_i} \sim N\left(\mathbf{0}, \sigma_g^2 \boldsymbol{A_i}\right)$,
  - Random error $\boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \boldsymbol{I})$
  - Age and gender were obtained from the SHFS
- Kinship matrices and genetic variances were similar to those in Aim 1
- Sample sizes: similar to SHFS
- 1000 Simulations

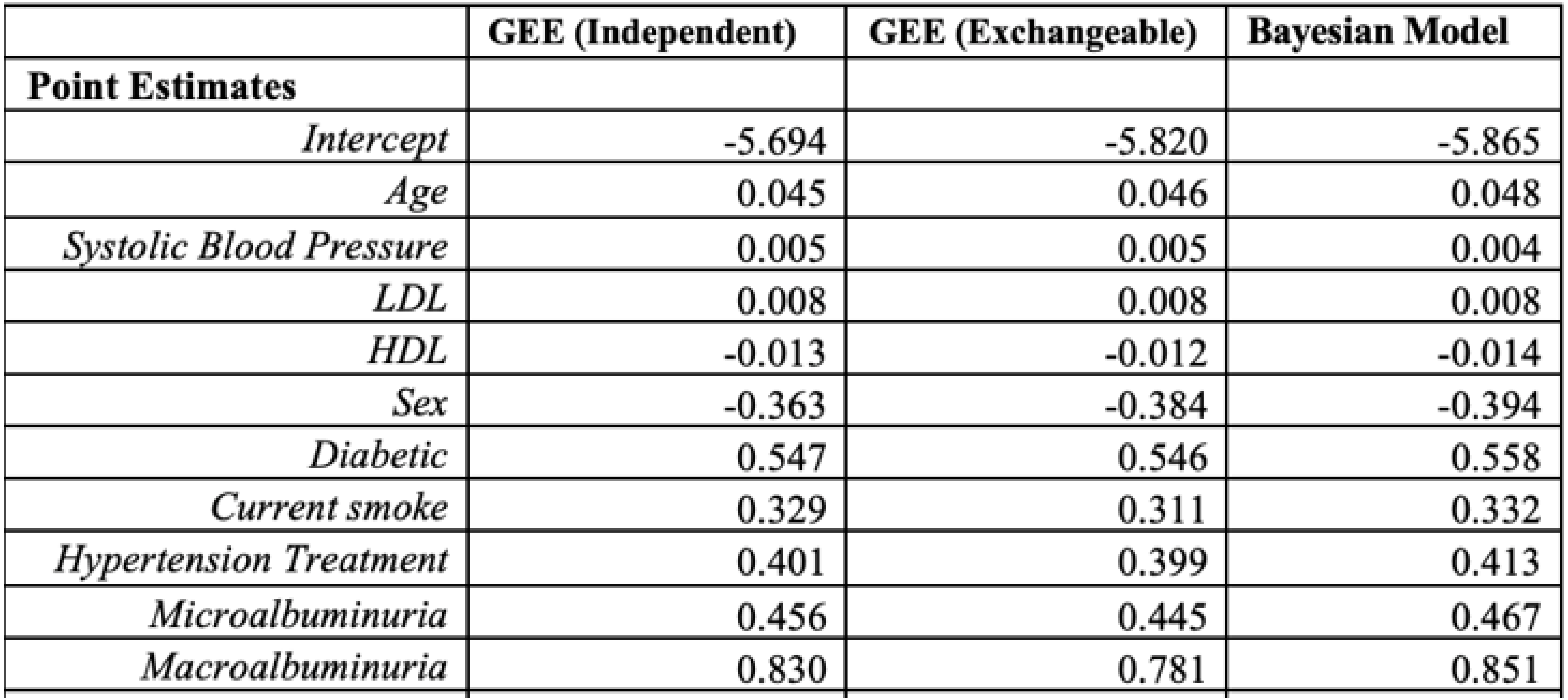
Simulation Results and Conclusion

- GEE performs well for simple family structures and small genetic variances in analyzing binary outcomes.
- However, its performance can be negatively affected by the complexity of the kinship matrix and the magnitude of the genetic variances
- If unsure, then simulation studies may be conducted

SHFS Data Analysis

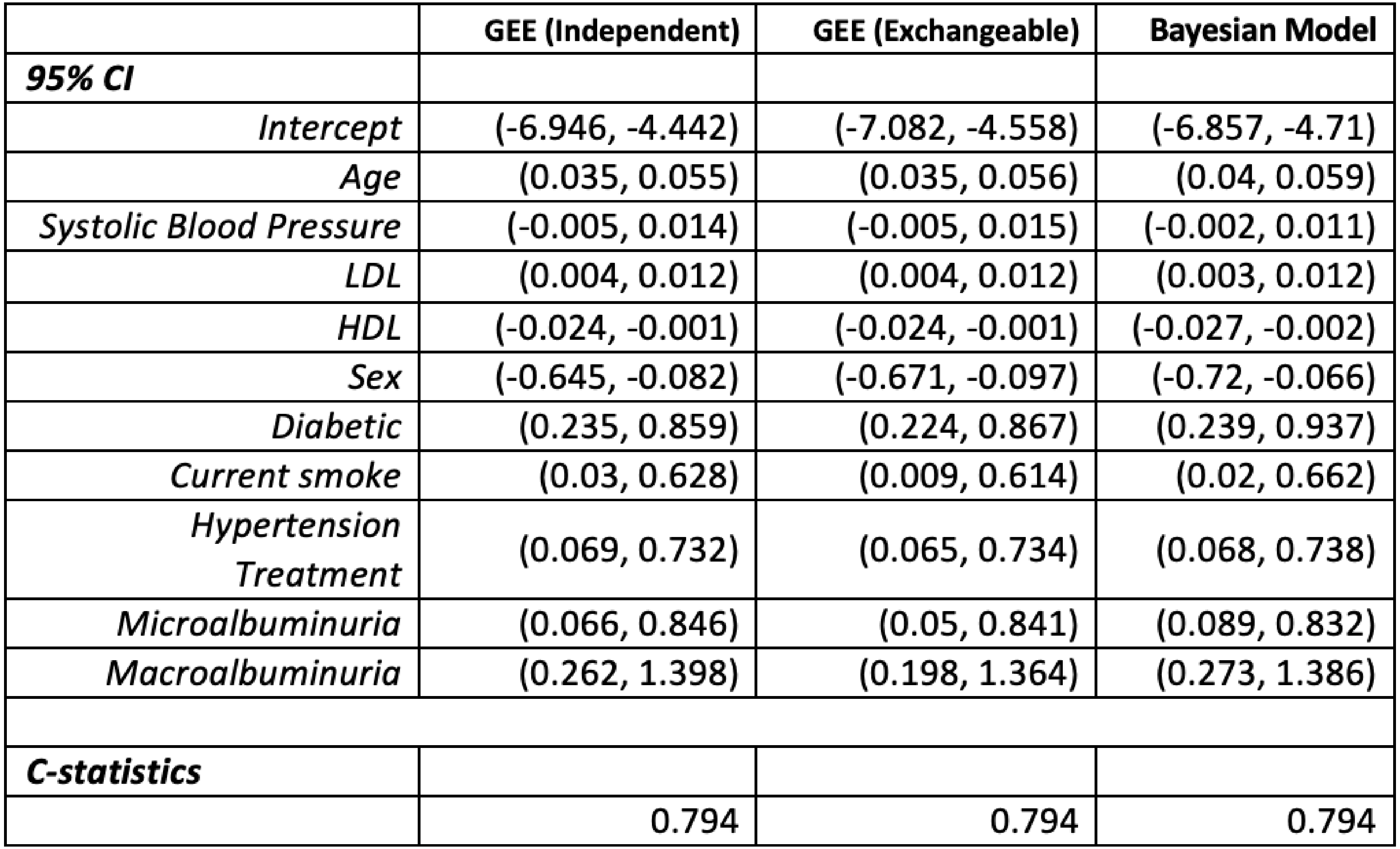
- Outcome: Coronary Heart Disease (CHD) event
- Covariates: age, sex, systolic blood pressure (SBP), LDL cholesterol, HDL cholesterol, diabetes status, current smoking, hypertension treatment, microalbuminuria, and macroalbuminuria.
- GEE (independent, exchangeable) and Bayesian model
- Point estimates and confidence intervals were compared
- C-statistic was calculated using a 5-fold cross validation

Results on Point Estimates



	GEE (Independent)	GEE (Exchangeable)	Bayesian Model
Point Estimates			
Intercept	-5.694	-5.820	-5.865
Age	0.045	0.046	0.048
Systolic Blood Pressure	0.005	0.005	0.004
LDL	0.008	0.008	0.008
HDL	-0.013	-0.012	-0.014
Sex	-0.363	-0.384	-0.394
Diabetic	0.547	0.546	0.558
Current smoke	0.329	0.311	0.332
Hypertension Treatment	0.401	0.399	0.413
Microalbuminuria	0.456	0.445	0.467
Macroalbuminuria	0.830	0.781	0.851

Results on CI and C-statistics



	GEE (Independent)	GEE (Exchangeable)	Bayesian Model
95% CI			
Intercept	(-6.946, -4.442)	(-7.082, -4.558)	(-6.857, -4.71)
Age	(0.035, 0.055)	(0.035, 0.056)	(0.04, 0.059)
Systolic Blood Pressure	(-0.005, 0.014)	(-0.005, 0.015)	(-0.002, 0.011)
LDL	(0.004, 0.012)	(0.004, 0.012)	(0.003, 0.012)
HDL	(-0.024, -0.001)	(-0.024, -0.001)	(-0.027, -0.002)
Sex	(-0.645, -0.082)	(-0.671, -0.097)	(-0.72, -0.066)
Diabetic	(0.235, 0.859)	(0.224, 0.867)	(0.239, 0.937)
Current smoke	(0.03, 0.628)	(0.009, 0.614)	(0.02, 0.662)
Hypertension Treatment	(0.069, 0.732)	(0.065, 0.734)	(0.068, 0.738)
Microalbuminuria	(0.066, 0.846)	(0.05, 0.841)	(0.089, 0.832)
Macroalbuminuria	(0.262, 1.398)	(0.198, 1.364)	(0.273, 1.386)
C-statistics			
	0.794	0.794	0.794

Aim 3: Survival Outcomes

Under review, *Journal of Biopharmaceutical Statistics*

Background

- Survival outcome is defined as the time from enrollment to date of event or the last contact date (censored)
- GEE model is not appropriate for survival outcomes
- There is no well-accepted method that can fully incorporate the kinship matrix
- We aim to develop a model with
  - Population effects similar to the Cox proportional hazard model
  - Individual effects that can incorporate the kinship matrix

Bayesian Proportional Hazard Model

- Model: $h_i(t) = h_0(t)\exp(\boldsymbol{X_i\beta} + b_i)$
  - $h_i(t)$: hazard function for the individual i at time t
  - $h_0(t)$: baseline hazard function
  - $\boldsymbol{\beta}$ is the population regression coefficients
  - $\boldsymbol{b_i} \sim N\left(\boldsymbol{0}, \sigma_g^2 \boldsymbol{A_i}\right)$
  - $\boldsymbol{A_i}$ is twice the kinship matrix, σ_g^2 is the genetic variance

Special Features of the BPHM

- Model: $h_i(t) = h_0(t)\exp(\boldsymbol{X_i}\boldsymbol{\beta} + b_i)$
- Allows for
  - Flexible specification of the baseline hazard function h_0 using mixture of piecewise constants
  - Capturing correlation defined by the Kinship Matrix using the individual random effects
  - Interpreting $\exp(\boldsymbol{\beta})$ as conditional hazard ratios

Algorithms to draw posterior samples

- Model: $h_i(t) = h_0(t)\exp(\boldsymbol{X_i}\boldsymbol{\beta} + b_i)$
- Priors: non-informative proper priors
- Due to the large dimension of the Kinship Matrix, we propose to do a **Singular Value Decomposition** of the Kinship Matrix
- Because the likelihood function is not a recognizable one, we used the **well-know “zero trick” with a Poisson distribution** to specify the likelihood function
- Finally, posterior samples can be drawn using JAGS

Simulation Setup

- For individual $i = 1, \dots n$, the survival outcome, t_i, was generated from exponential (λ_i),
$$\lambda_i = \beta_1 \text{age} + \beta_2 \text{sex} + b_i$$
- Random effects b_i were simulated by family such that
  - In the j^{th} family , $\boldsymbol{b_j} \sim N(\mathbf{0}, \sigma_g^2 \boldsymbol{A_j})$. σ_g^2 is the genetic variance, and $\boldsymbol{A_j}$ was twice the kinship matrix
- $\beta_1 = 5$ and $\beta_2 = -0.5$; $\sigma_g^2 = 0.2$
- 25% censoring rate
- Kinship matrices were chosen similarly as in Aims 1 &2

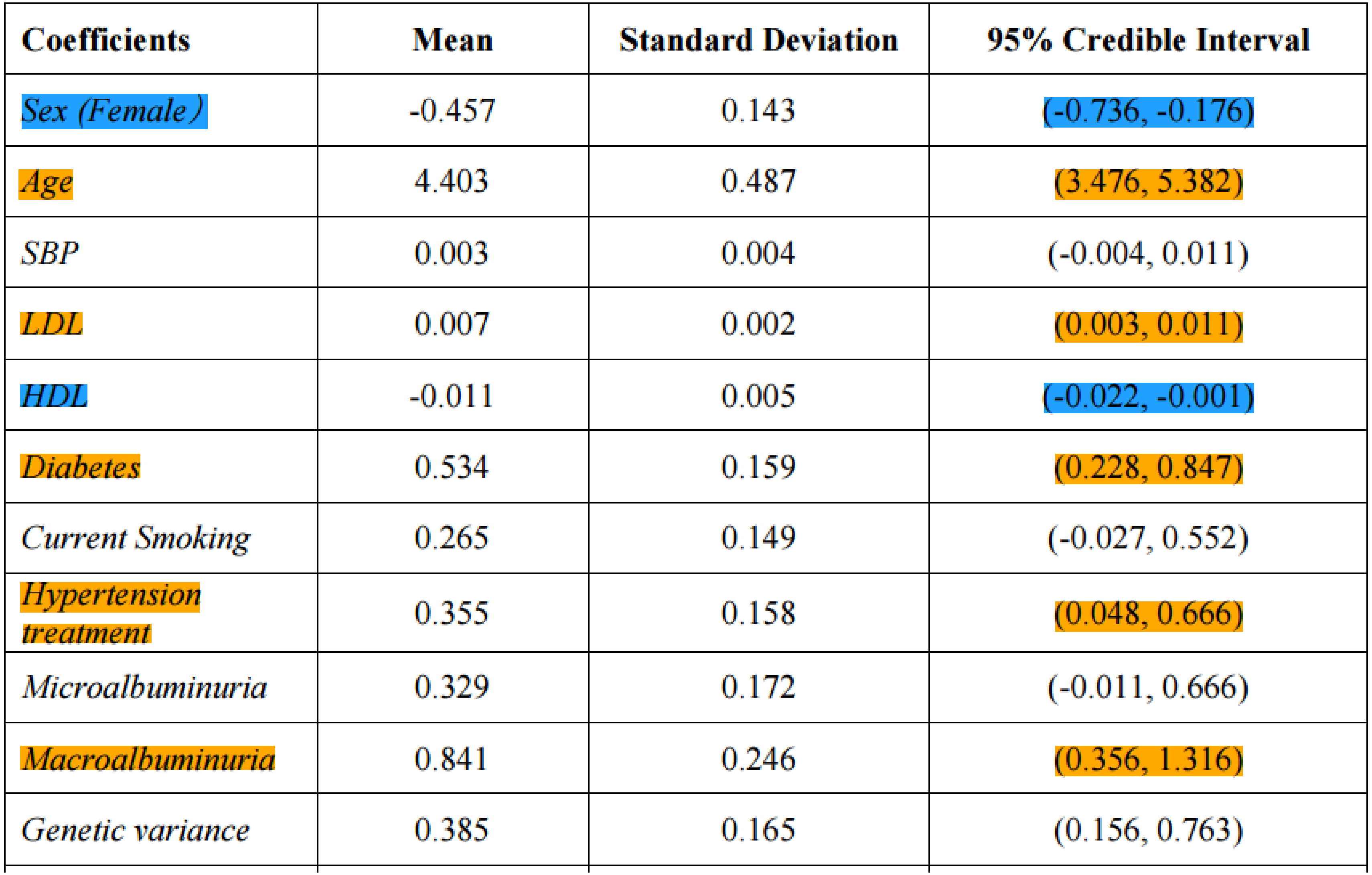
Results and Conclusions

- Relative biases are close to zero
- 95% credible intervals have an average Coverage Probabilities close to 95%

SHFS Data Analysis

- Outcome: time to CHD
- Covariates: age, sex, systolic blood pressure (SBP), LDL cholesterol, HDL cholesterol, diabetes status, current smoking, hypertension treatment, microalbuminuria, and macroalbuminuria

Results



Coefficients	Mean	Standard Deviation	95% Credible Interval
Sex (Female)	-0.457	0.143	(-0.736, -0.176)
Age	4.403	0.487	(3.476, 5.382)
SBP	0.003	0.004	(-0.004, 0.011)
LDL	0.007	0.002	(0.003, 0.011)
HDL	-0.011	0.005	(-0.022, -0.001)
Diabetes	0.534	0.159	(0.228, 0.847)
Current Smoking	0.265	0.149	(-0.027, 0.552)
Hypertension treatment	0.355	0.158	(0.048, 0.666)
Microalbuminuria	0.329	0.172	(-0.011, 0.666)
Macroalbuminuria	0.841	0.246	(0.356, 1.316)
Genetic variance	0.385	0.165	(0.156, 0.763)

Overall Summary

- Aim 1
  - Either GEE or Bayesian model works
  - Choice depends on personal preference
- Aim 2
  - Similar to Aim 1
  - GEE may be problematic for data with complex kinship matrix and large genetic variance
- Aim 3
  - Developed a model for survival outcome utilizing kinship matrix

The UNIVERSITY *of* OKLAHOMA
HEALTH SCIENCES