Longitudinal Analysis of Cardiovascular Disease Risk Factors based on the Framingham Heart Study

Group 1:
Emily Li, 2427297
Bryan Ng, 2427348
Junyi Fang, 2427308

Changes

Dataset: unchanged

- keep original Framingham Heart Study dataset

Scientific Questions: changed

- Removed Question 2 because of its reverse causation risks.
- Outcome Variables mainly focus on blood pressure (SYSBP & DIABP)

Updated Scientific Questions:

1. Do variables such as SYSBP and DIABP, differ by demographic factors including SEX, AGE, and BMI while adjusting for EDUC and TIME?

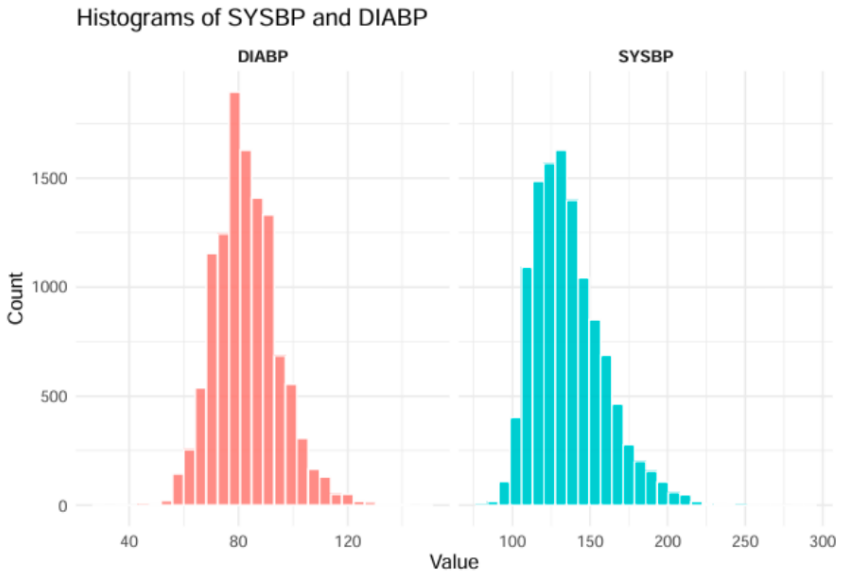
2. What is the effect of BPMEDS on SYSBP and DIABP while adjusting for SEX, AGE, BMI, EDUC, and TIME?

3. What is the effect of CURSMOKE and CIGPDAY on SYSBP and DIABP, after adjusting for SEX, AGE, BMI, EDUC, TIME, and BPMEDS?

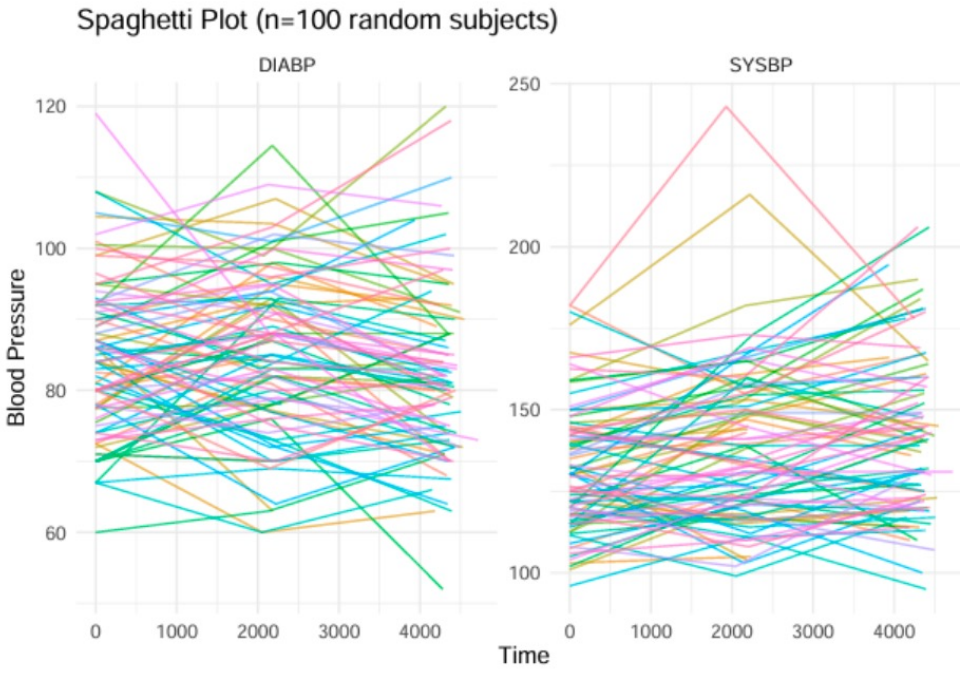
4. What is the effect of TOTCHOL, LDLC, HDLC, GLUCOSE in SYSBP and DIABP adjusting for SEX, AGE, BMI, EDUC, TIME, BPMEDS, CURSMOKE, and CIGPDAY?

Exploratory Data Analysis

Due to space constraints, we selected **representative EDA results** most relevant to the four research questions, and they are summarized below.

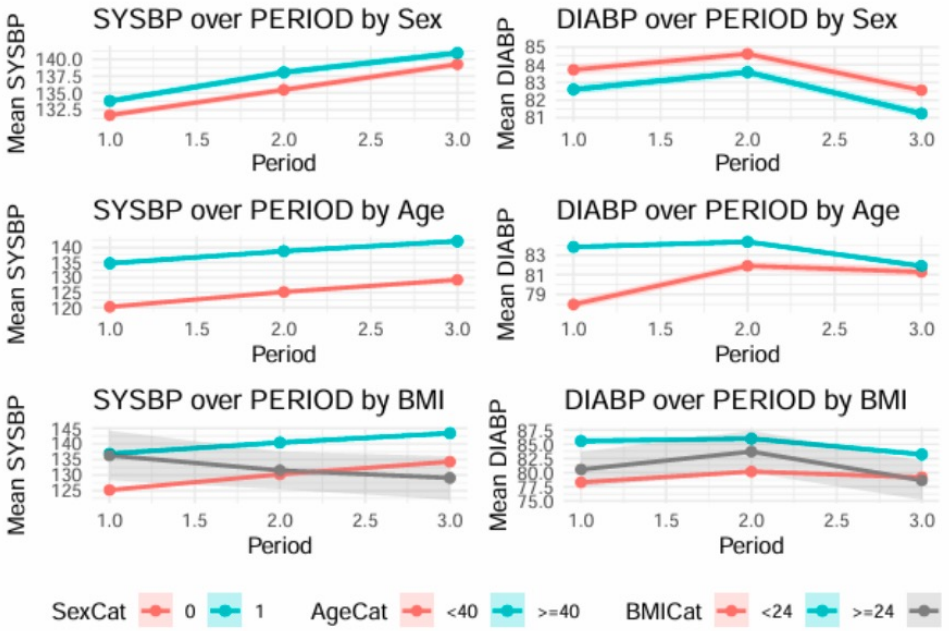


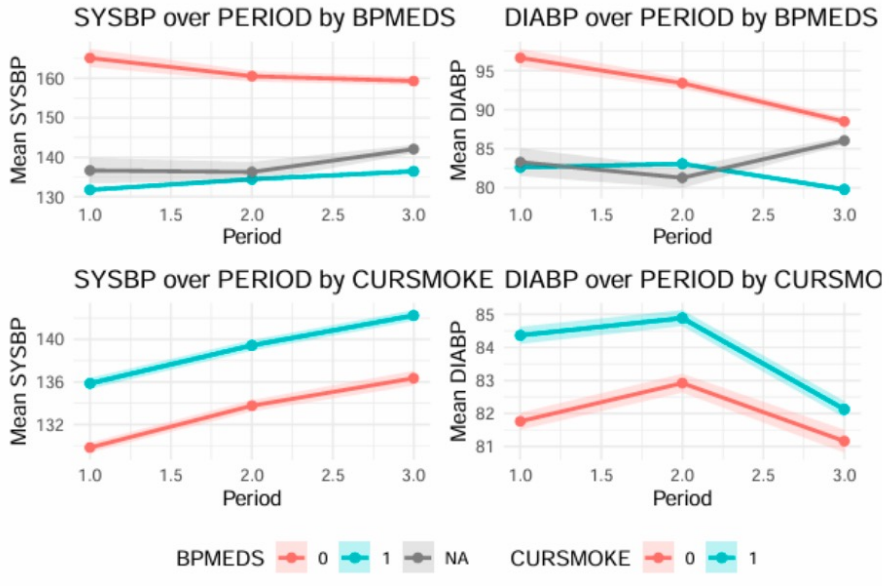
1. The **histogram** shows that DIABP is tightly clustered and roughly bell-shaped, whereas SYSBP is more variable and right-skewed with extreme outliers, suggesting the need for **robust modeling approaches** in further analysis.



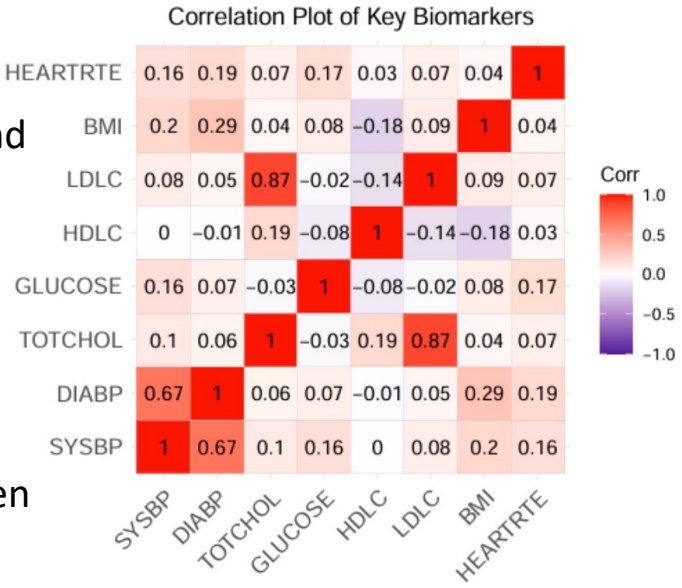
2. From the **spaghetti plot**, it is clear that each subject's blood-pressure readings are correlated over time and that individuals vary substantially in both their baseline levels and their trajectories; consequently, a **longitudinal mixed-effects model** is needed to capture the overall time trend while simultaneously allowing each person to have their own random intercept and slope.

Exploratory Data Analysis





3. We plotted mean SYSBP and DIABP trajectories by key demographic factors (age, sex, and BMI) as well as by smoking status and antihypertensive- medication use. We found that demographic groups exhibit consistent, parallel differences in both SYSBP and DIABP over time, and that treated vs. untreated subjects diverge in their blood-pressure trajectories as expected. Furthermore, smoking and medication status produced even more pronounced trends, highlighting their substantial impact on blood-pressure levels.



4. From the **correlation plot**, most variable pairs show low correlations except for total cholesterol vs. LDLC and SYSBP vs. DIABP, suggesting that if our further analyses include these pairs, we should introduce a penalty term on our regression methods to address the **multi-collinearity.**

Regression Methods

For Question 1, demographic factors vary by individual, so we're not only interested in the average effects of Sex, Age, and BMI on blood pressure but also in how each person's baseline blood pressure and its trajectory over time differ. We therefore fit a **LME** model, which estimates both the **population-level effects** and **random intercepts and slopes for each participant**, capturing subject-specific deviations from the overall trend. To address multi collinearity among predictors, we add a **ridge penalty** to the fixed-effects portion of the model.

$$\underbrace{Y_{it}}_{\text{SYSBP or DIABP}} = \underbrace{\left(\beta_0 + \beta_1 \text{SEX}_i + \beta_2 \text{AGE}_{it} + \beta_3 \text{BMI}_{it} + \beta_4 \text{EDUC}_i + \beta_5\, t\right)}_{\text{population (fixed) effects}} + \underbrace{\left(b_{0i} + b_{1i} t\right)}_{\text{random intercept \& slope}} + \varepsilon_{it}$$

For Question 2, our goal is to estimate the **population-averaged effect** of anti hypertensive medication use on blood pressure rather than individual-level trajectories. **GEE** therefore provides a simpler, more reliable interpretation of the marginal drug effect. Again, we introduce a **ridge penalty** to address multi collinearity.

$$\underbrace{\mu_{it}}_{\mathbb{E}\left[Y_{it}\right]} = \beta_0 + \beta_1 \text{BPMEDS}_{it} + \beta_2 \text{SEX}_i + \beta_3 \text{AGE}_{it} + \beta_4 \text{BMI}_{it} + \beta_5 \text{EDUC}_i + \beta_6\, t$$

For Question 3, we aim to estimate the **population-averaged effect** of current smoking status and smoking intensity on systolic and diastolic blood pressure. Thus, we again employ a **GEE** framework, clustering on RANDID and specifying an exchangeable working correlation structure. As with Question 2, we incorporate a **ridge penalty** to curb potential multicollinearity among CURSMOKE, CIGPDAY, and the additional covariates.

$$\underbrace{\mu_{it}}_{\mathbb{E}[Y_{it}]} = \beta_0 + \beta_1\, \text{CURSMOKE}_i + \beta_2\, \text{CIGPDAY}_{it} + \beta_3\, \text{SEX}_i + \beta_4\, \text{AGE}_{it} + \beta_5\, \text{BMI}_{it} + \beta_6\, \text{EDUC}_i + \beta_7\, \text{BPMEDS}_{it} + \beta_8\, t$$

Regression Methods

For Question 4, we prefer **LME** model because it directly captures within-subject correlation, naturally handles unbalanced repeated measurements, and lets us quantify **both the average effect of these biomarkers** and **each individual's deviation from that average**. Moreover, EDA revealed a **curved relationship** between blood pressure and time, so we model **TIME with a natural cubic spline** to flexibly fit that nonlinearity while preserving linear tails. Finally, to address multicollinearity among TOTCHOL, LDLC, SYSBP and DIABP, we add a **ridge penalty** to the fixed-effects portion of the model—stabilizing coefficient estimates without sacrificing the mixed-model framework.

$$\underbrace{Y_{it}}_{\text{SYSBP or DIABP}} = \underbrace{\Big(\beta_0 + \beta_1\,\mathrm{TOTCHOL}_{it} + \beta_2\,\mathrm{LDLC}_{it} + \beta_3\,\mathrm{HDLC}_{it} + \beta_4\,\mathrm{GLUCOSE}_{it} + \beta_5\,\mathrm{SEX}_i + \beta_6\,\mathrm{AGE}_{it} + \beta_7\,\mathrm{BMI}_{it} + \beta_8\,\mathrm{EDUC}_i + \beta_9\,\mathrm{BPMEDS}_{it} + \beta_{10}\,\mathrm{CURSMOKE}_i + \beta_{11}\,\mathrm{CIGPDAY}_{it} + f_{\mathrm{ncs}}(t;\boldsymbol{\gamma})\Big)}_{\text{fixed (population-level) effects}} + \underbrace{(b_{0i} + b_{1i}t)}_{\text{random intercept \& slope}} + \varepsilon$$

Overall

Question 1 & 4: Linear Mixed Effects Model

Question 2 & 3: Generalized Estimating Equations

Compared to the LME, GEE directly estimate population-averaged effects: their sandwich estimator remains consistent even if the working correlation structure is mis-specified, yielding robust standard errors. With enough clusters, GEE delivers more **stable variance estimates for fixed effects**, whereas LME's variance-component estimates can be highly variable in small or complex samples.