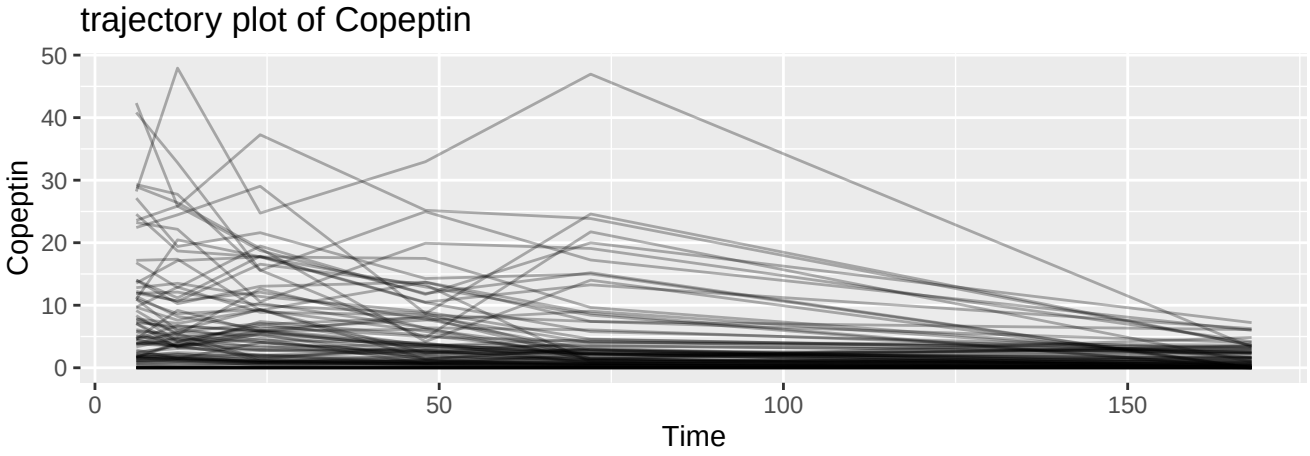
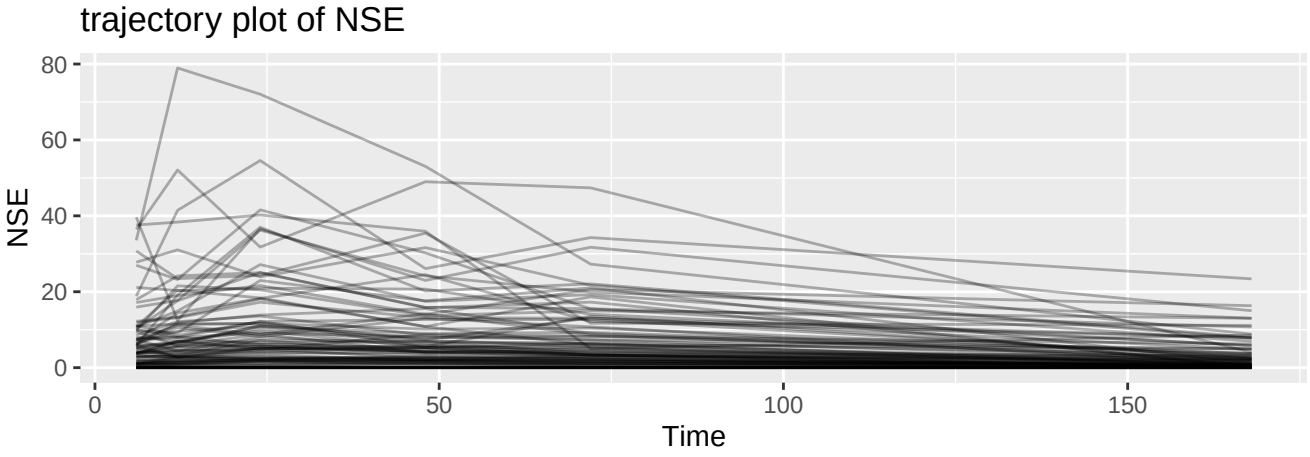
Comp exam 2019

Nam-Anh Tran

1 Aim 1

1.1 Data exploration



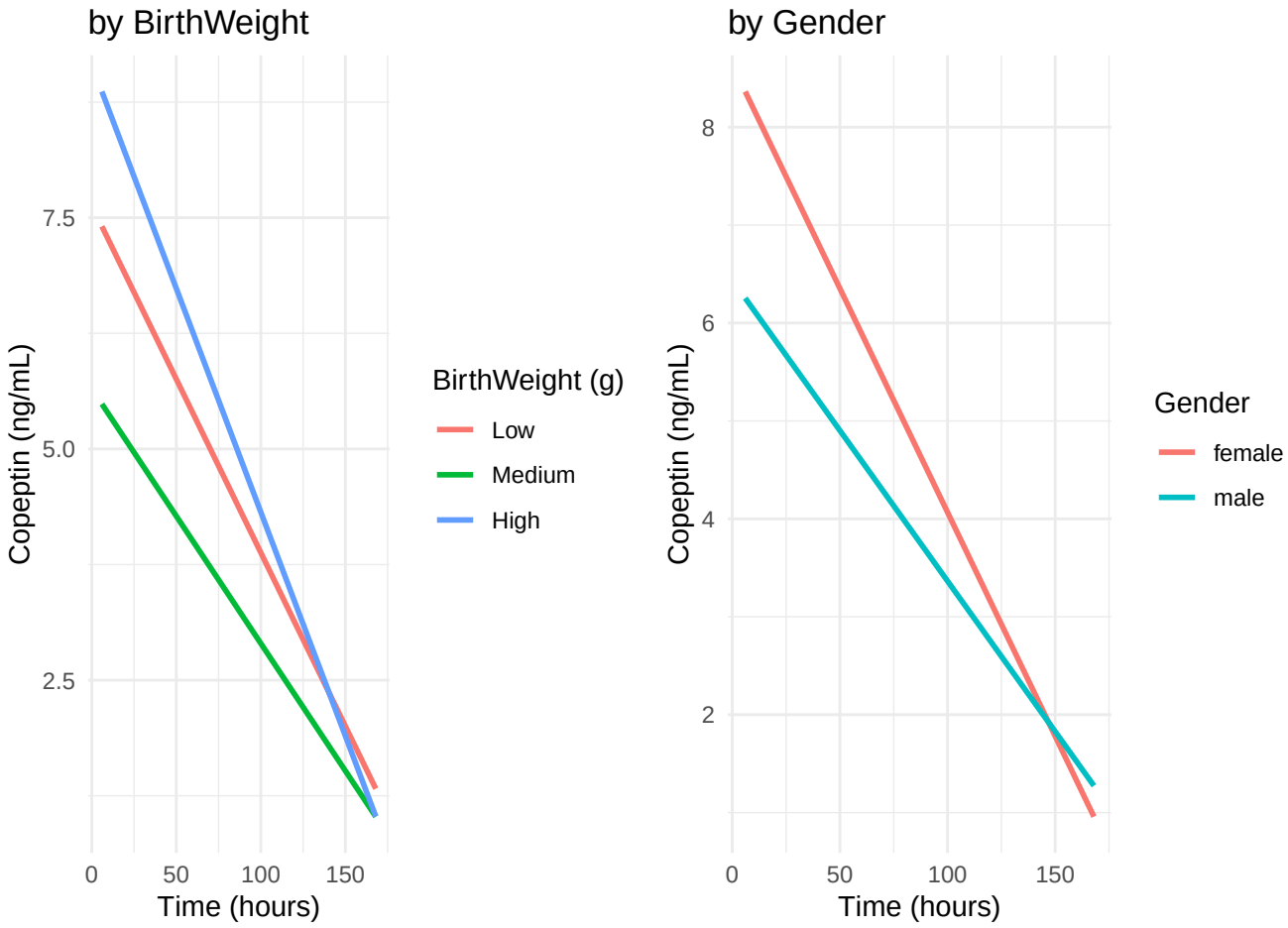


We explore the data to specify the structure of the model.

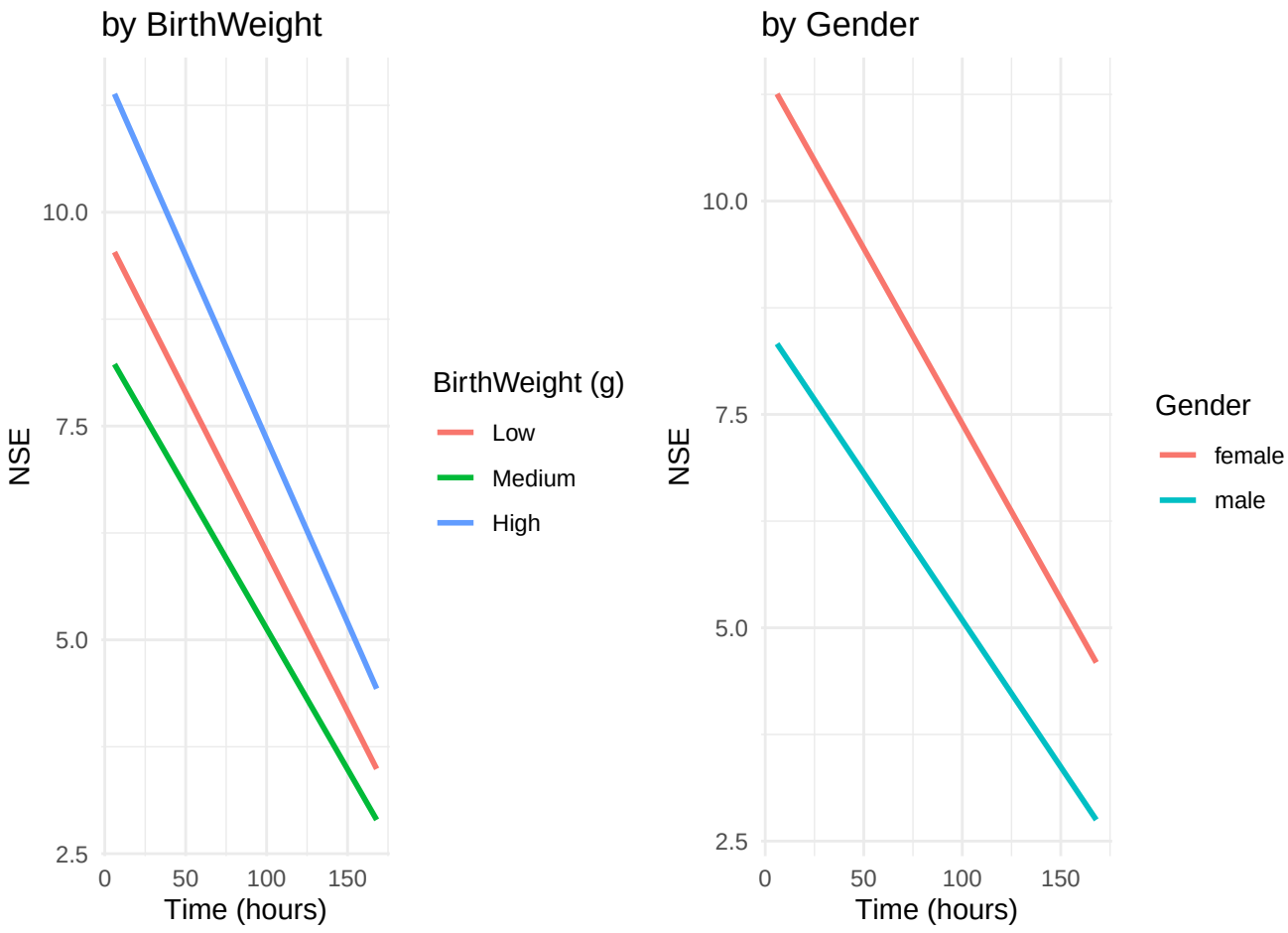
- The trajectories start from different points, which suggests a random intercept.

- The trajectories with a nonlinear relationship between the outcomes and time, specifically the curvature, suggest that the quadratic term of time should be included in the model. Also, the curvature looks similar across trajectory lines, implying a random slope is unnecessary.

We evaluate the necessity of interactions between two predictors of interest, gender and birth weight, and time, by regressing the outcomes on time for each group of predictors. As birth weight is a continuous variable, we categorize it into three groups, representing low, medium, and high weights.



The plot above shows that the decrease in copeptin over time varies across birth weight and gender. This indicates that two interactions, birth weight versus time and gender versus time, should be included in the model.



The plot above indicates that the interactions of birth weight and gender versus time are not necessary, as the lines do not intersect.

1.2 Models

We fit two linear mixed models for the two biomarkers separately:

Copeptin:

$$y_{ij} = \beta_0 + \boldsymbol{x}'_{ij1}\boldsymbol{\beta}_1 + \beta_2 \text{Time}_{ij} + \beta_3 \text{Time}_{ij}^2 + \boldsymbol{x}'_{ij2}\boldsymbol{\beta}_4 + b_i + \epsilon_{ij},$$

where

- $\boldsymbol{x}_{ij1}$ includes birthweight, gender, and other covariates.
- $\boldsymbol{x}_{ij2}$ includes interactions: birthweight $\times$ Time and gender $\times$ Time.

NSE:

Similar to the Copeptin's model but excluding the interaction vector $\boldsymbol{x}_{ij2}$. The association between biomarkers and gender and birth weight is evaluated based on the corresponding coefficients.

The models are completed after defining prior distributions.

$$\beta_0, \beta_2, \beta_3 \sim N(0, 10^2),\ \boldsymbol{\beta}_1 \sim N(0, 10^2\boldsymbol{I}),\ \boldsymbol{\beta}_4 \sim N(0, 10^2\boldsymbol{I})$$

$$b_i \sim N(0, \sigma_b^2),\ \sigma_b \sim \text{half-cauchy}(0, 2)$$

$$\epsilon_{ij} \sim N(0, \sigma^2),\ \sigma \sim \text{half-cauchy}(0, 2)$$

Both models assumes linearly additive effect of predictors. As fitting two models separately, the models do not capture correlation between two biomarkers. This assumption may not be correct in real-world data. This limitation can be overcome by jointly fitting the model where the two random intercepts are assumed correlated. We then let $\boldsymbol{b}_i = (b_i^{\text{copeptin}}, b_i^{\text{NSE}}) \sim N(0, \Sigma_b)$ and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}) \sim N(0, \Sigma)$.

2 Aim 2

While both biomarkers are time-varying covariates, the outcome, neurological outcomes at 2 years of age, is a time-invariant covariate. The direct and simple approach is to fit the logistic model for the outcome, which is a binary variable, using the aggregated measure of time-varying covariates. That is, we calculated the mean of the time-varying covariates over time and fitting the model. Specifically,

$$y_i \sim \text{Bern}(\pi_i),\ \text{logit}(\pi_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}_1,$$

where $\boldsymbol{x}_i$ includes time-invariant covariates and aggregated time-varying covariates.

The model is completed after defining priors. We employ spike-and-slab priors to shrink the coefficients toward to zero if the corresponding covariates are not predictive. As stan does not accept the discrete parameters, the traditional spike-and-slab, which includes a degenerate function with the mass at zero, is not possible. Thus, we select a "soft" spike, i.e. the spike follows a distribution about zero (spike) with a relatively small variance. The model then becomes the Gaussian mixture model.

$$\beta_0 \sim N(0, 5^2),$$

$$\boldsymbol{\beta}_1^{\text{slab}} \sim N(0, \sigma_\beta^2 I),\ \sigma_\beta \sim \text{truncated-normal}(0, 5),\ \text{truncated at 0.5}$$

$$\boldsymbol{\beta}_1^{\text{spike}} \sim N(0, 0.001^2),$$

$$\beta_{1k} \sim pN(0, \sigma_\beta^2) + (1-p)N(0, 0.001^2),\ k = 1, 2, \ldots, K$$

$$p \sim \text{Uniform}(0, 1).$$

We select a deterministic variance for spike to reflect the minimal-to-non effect of covariates as its corresponding coefficients are close to zero. In contrast, if the effect of covariates are evident, we let the data decide the level of uncertainty of the corresponding coefficients. We truncate distribution of σ_β at 0.5 to ensure a clear distinction between the spike and slab components, as our goal, choosing variance of spike small, is to enforce sparsity (coefficients of irrelevant covariates are near 0), while the slab (σ_β) should be large enough to accommodate the true coefficients of relevant covariates, enabling the model to distinguish between included and excluded variables.

With the SD of 0.001, the spike component is intended to shrink coefficients of irrelevant covariates within $[-0.002, 0.002]$ with the probability of 0.95. Truncating at 0.5 ensures $\sigma_\beta^2 \geq 0.25$, maintaining a significant gap with the ratio greater than 500, (i.e. $\sigma_\beta^{\text{slab}}/\sigma_\beta^{\text{spike}} \geq 500$), which is critical for effective variable selection.

After fitting the model, we then obtain the posterior predictive realizations of the latent binary variable γ_k based on the p, i.e. $\gamma_k \sim \text{Bern}(p)$. γ_k indicates the presence of the corresponding covariate in the model.

As this model uses the aggregated measure of time-varying covariates, it may not well capture the effect of the change in those covariates over time on the outcome. to capture the change effects, we can fit the joint model, including the longitudinal submodel and the outcome submodel. That is, we jointly fit both models in aim 1 and aim 2 using the shared random effect. For example, the longitudinal model includes both random intercept and slope of time. The random intercept and slope summarise the between-subject difference at the baseline and the rate change of time-varying covariates over time. we are then interested whether the between-individual difference has difference in predicting the outcome.