PROJECT

BLAL BLA BLA

anonymous

1 Introduction

- Background
- Problem formulation/scope
- Main modeling idea
- Some picture of the data?

2 Data Desctription

The "cath" dataset used in this report is obtained from Duke University Cardiovascular Disease Databank. It encapsulates a collection of 6 variables (@Data-table) that are closely related to cardiovascular health.

Table 1: TBD

sex	age	cad_dur	choleste	sigdz	tvdlm
0	73	132	268	1	1
0	68	85	120	1	1
0	54	45	NA	1	0
1	58	86	245	0	0
1	56	7	269	0	0
0	64	0	NA	1	0

The dataset consists of four explanatory variables (*sex*, *age*, *cad_dur*, *choleste*) and two response variables (*sigdz*, *tvdlm*) that provide an overview on patient demographics, clinical indicators, and critical outcomes related to coronary artery disease:

- **Sex** (*sex*): Categorized as 0 for male and 1 for female, this variable represents the gender distribution within our dataset.

- **Age** (*age*): Representing the age of patients in years, this variable serves as a demographic feature.
- **Chest Pain Duration** (*cad_dur*): The duration of chest pain symptoms in days.
- **Serum Cholesterol Level** (*choleste*): Measured in milligrams per deciliter, serum cholesterol levels are indicative of lipid metabolism and play a crucial role in cardiovascular health.
- **Significant Coronary Disease** (*sigdz*): A binary variable that captures the presence (1) or absence (0) of at least 75% blockage in one of the major coronary arteries.
- **Three Vessel Disease or Left Main Disease** (*tvdlm*): Denoting the presence (1) or absence (0) of blockage in either all three coronary vessels or in the left main coronary artery.

The univariate distributions of these variables are visualized in @Univariate-analysis.

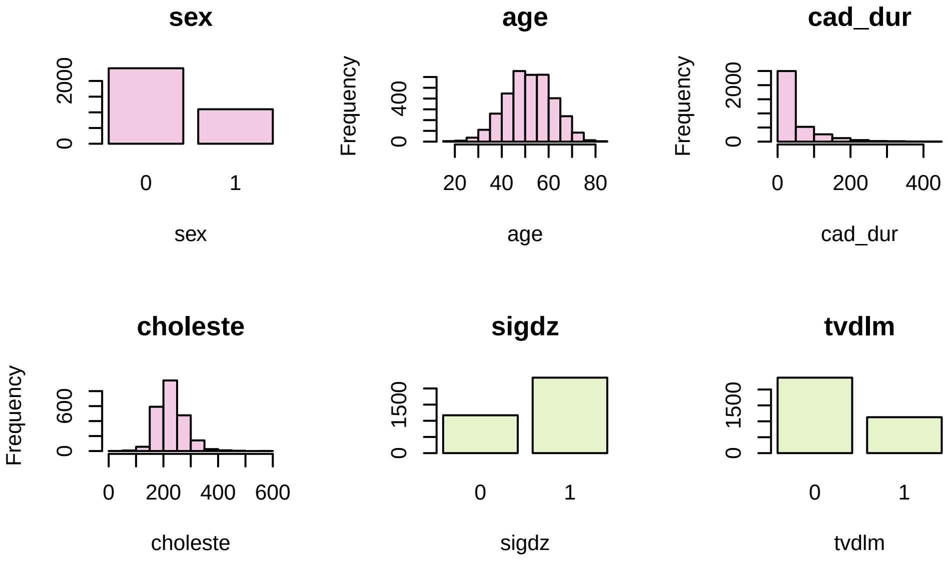

Figure 1: TBD 1

While constructing the Bayesian models to predict the probaility of significant coronary disease, the report strives to utilize the correlation between the explanatory variables (*sex*, *age*, *cad_dur*, *choleste*) and the desired response variable (*sigdz*). The *tvdlm* variable is not relevant in this report as the main focus is to predict the probability of significant coronary disease, independent of the type of the blockade.

Before the analysis, the data is preprocessed by removing *tvdlm* column and all rows that contain missing values, as well as by scaling the continuous variables to zero mean and unit variance. After this, we are left with $n = 2258$ observations. The pairwise correlations of variables are visualized in @Bivariate-analysis. We can see that variables *sex* and *age* have the most significant bivariate correlation to the responsive variable *sigdz*.

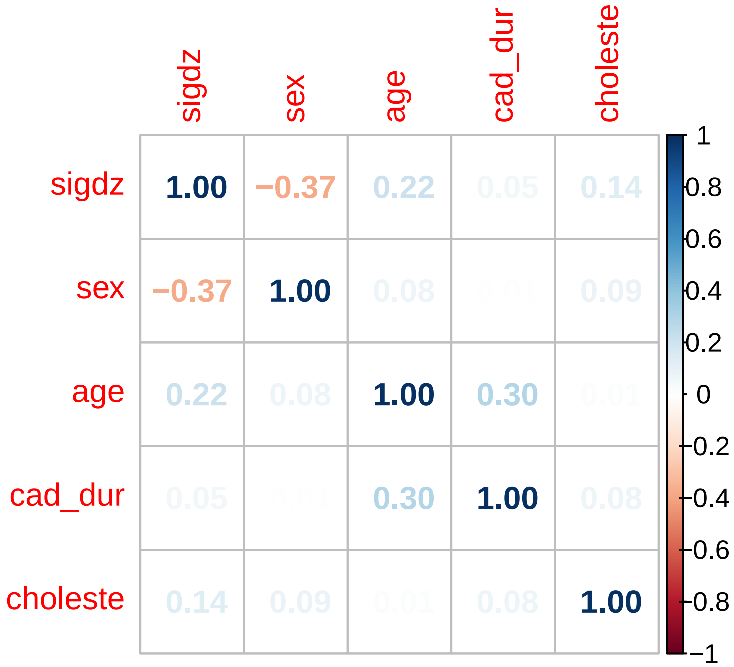

Figure 2: TBD 3

3 Mathematical Model

In this analysis, we will construct two models for inferring the binary response variable, *sigdz* based on input explanatory variables. The first model is a generalized linear model (GLM), namely Bayesian logistic regression. The other model is a generalized additive mixed model (GAMM), which implements Bayesian logistic regression with nonlinear transformations on the input variables. These models will be referred as linear and nonlinear model, respectively.

(((To-be-done still: - Check likelihood notation for nonlinear model - Prior justification - Check prior notations - Include priors with own values ? - **Posteriors** ?)))

3.1 The generalized linear model and priors

Let y be the number of times the variable *sigdz* is realized to be 1 for one individual in the dataset, and let x be the explanatory variables for this outcome. Then, this number of successes for one individual follows a Binomial distribution

$$y \sim \binom{n}{y} \theta^{y} (1-\theta)^{n-y},$$

where n is the number of observations for that specific individual and $\theta = g^{-1}(\eta)$ $(\eta = \alpha + x^{T}\beta)$ is the probability of success (patient presenting with significant coronary disease). The inverse link function g^{-1} maps the output of the linear predictor η to a probability interval between 0 and 1.

For the binomial GLM, this project utilizes logit $g(x) = \ln(\frac{x}{1-x})$ as a link function, which makes it a logistic regression model. As each individual occurs only once in the data, y can be directly presented as the binary response variable. Therefore, the likelihood of the response variable of one individual is reduced to Bernoulli distribution

$$y \sim \text{ logit}^{-1}(\eta)^{y}(1 - \text{ logit}^{-1}(\eta))^{1-y}.$$

The complete data likelihood is then a product of $n = 2258$ likelihoods, with unshared probability of success.

As in Bayesian logistic regression the scope is to infer the distribution of the regression weights, namely the intercept α and coefficients $\beta = [\beta_1, \beta_2, \beta_3, \beta_4]^{T}$, we define the prior to be Student's t-distribution

$$\begin{aligned} \alpha &\sim t_v(\mu, \sigma) \\ \beta_k &\sim t_v(\mu, \sigma),\ k = 1, .., 4 \end{aligned}$$

where v is the degrees of freedom, μ is the location and σ is the scale.

The selection of prior distribution was done based on the nature of the data. Due to correlations, there is reason to believe that the parameters are not very close to zero, but most are still rather small than large. Therefore, as Student's t-distribution has heavy tails and larger scale compared to, for example, Gaussian distribution, t-distribution is a suitable choice of prior for this purpose. The parameters of the prior were defined to be

$$v = 3, \quad \mu = 0, \quad \sigma = 2.5,$$

as the coefficients can be positive or negative.