Computational Statistics

Bologna Master Degree in Mathematics and Applications

Report 5 - Gamma model for glycosolated hemoglobin with bayesian approach

Authors:

Ana Marija Belingar (108827) ab30811@student.uni-lj.si
Vito Rozman (108734) vito.rozman@tecnico.ulisboa.pt

2023/2024 – 1st Semester

Contents

List of Figures

List of Tables

1 Introduction

Based on a health study conducted among 390 African Americans in central Virginia, where cardiovascular risk factors were measured, we aim to determine the most effective model, particularly a Gamma model, for predicting glycosylated hemoglobin.

To commence, we conducted explanatory data analysis to gain insights into the distribution and basic characteristics of measured risk factors and the target variable. The analysis revealed missing values, and we appropriately encoded the data.

In the second part, we focused on a general linear model without a Bayesian approach, intending to demonstrate that prediction with some prior knowledge (even minimal) about prior distributions is superior to a frequentist approach. The purpose of this section is also to explore whether it makes sense to remove certain variables to simplify the model. We utilized log link functions.

In the main part of the task, two gamma models are explored. We assume that shape parameter is constant for all observations and distributed with InverseGamma, while the rate parameter is different for each observation. This is taken into account via the regression model with log link function for the mean, which is a function of shape and rate.

The analysis was conducted in R, using R2jags software designed for Bayesian inference through Gibbs sampling. The Bayesian approach, facilitated by R2jags, allows us to incorporate prior knowledge, quantify uncertainty, and derive posterior distributions for model parameters. Alongside R2jags, the coda library plays a crucial role in analyzing Markov Chain Monte Carlo (MCMC) convergence and evaluating posterior chain diagnostics. The outline used for analysis is published on Github in this link.

2 Explanatory data analysis

Our original dataset consists of 390 records and 15 features. Each record refers to a different person, so we assume that the records are independent. Table 1 lists the feature names, short descriptions and the number of missing values. Three variables are categorical: LOCATION, GENDER, and FRAME. The other variables are numerical.

Variable name	Description	# of "NA"
ID	Subject identification	0
CHOL	Total cholesterol	1
SGLU	Stabilized glucose	0
HDL	High density lipoprotein	1
GHB	Glycosolated hemoglobin	0
LOCATION	(Buckingham, Louisa)	0
AGE	Age(years)	0
GENDER	(male, female)	0
HHT	Height (inches)	5
WHT	Weight (pounds)	1
FRAME	(large, medium, small)	11
SBP	First systolic blood pressure	5
DSP	First diastolic blood pressure	5
W	Waist (inches)	2
H	Hip (inches)	2

Table 1: Numerical features, their short descriptions and number of NA values

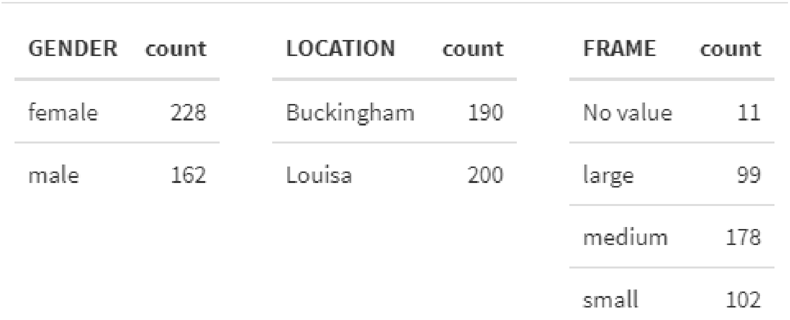
In further analysis, we omitted the ID column since it is unique for each individual, and individuals are assumed to be independent. Its removal simplifies the model and minimize potential collinearity issues, aligning with assumptions for valid regression models.

2.1 Categorical features

Figure 1 presents the categorical variables, their values and the number in each value. It can be seen that 58.5% of the participants in the study were females (48.2% from Buckingham, 51.8% from Louise) and 41.5% were males (49.4% from Buckingham, 50.6% from Louise). 48.7% of all participants were from Buckingham, and 51.3% were from Louise. The variable FRAME represents the body frame or body type of the individual. It categorize individuals as having a small, medium, or large frame, which can influence factors like body composition and metabolism. We see that most participants have a medium frame.

2.2 Numerical features

First, we plotted histograms with normalized counts of observations for numeric variables, which are presented in Figure 2. For better visualization and understanding, we added density curves. From the graphs, it would be challenging to tell that any of our numeric variables follow a normal distribution. However, there might be a suspicion that the variable SGLU is



GENDER	count
female	228
male	162

LOCATION	count
Buckingham	190
Louisa	200

FRAME	count
No value	11
large	99
medium	178
small	102

Figure 1: EDA for categorical features

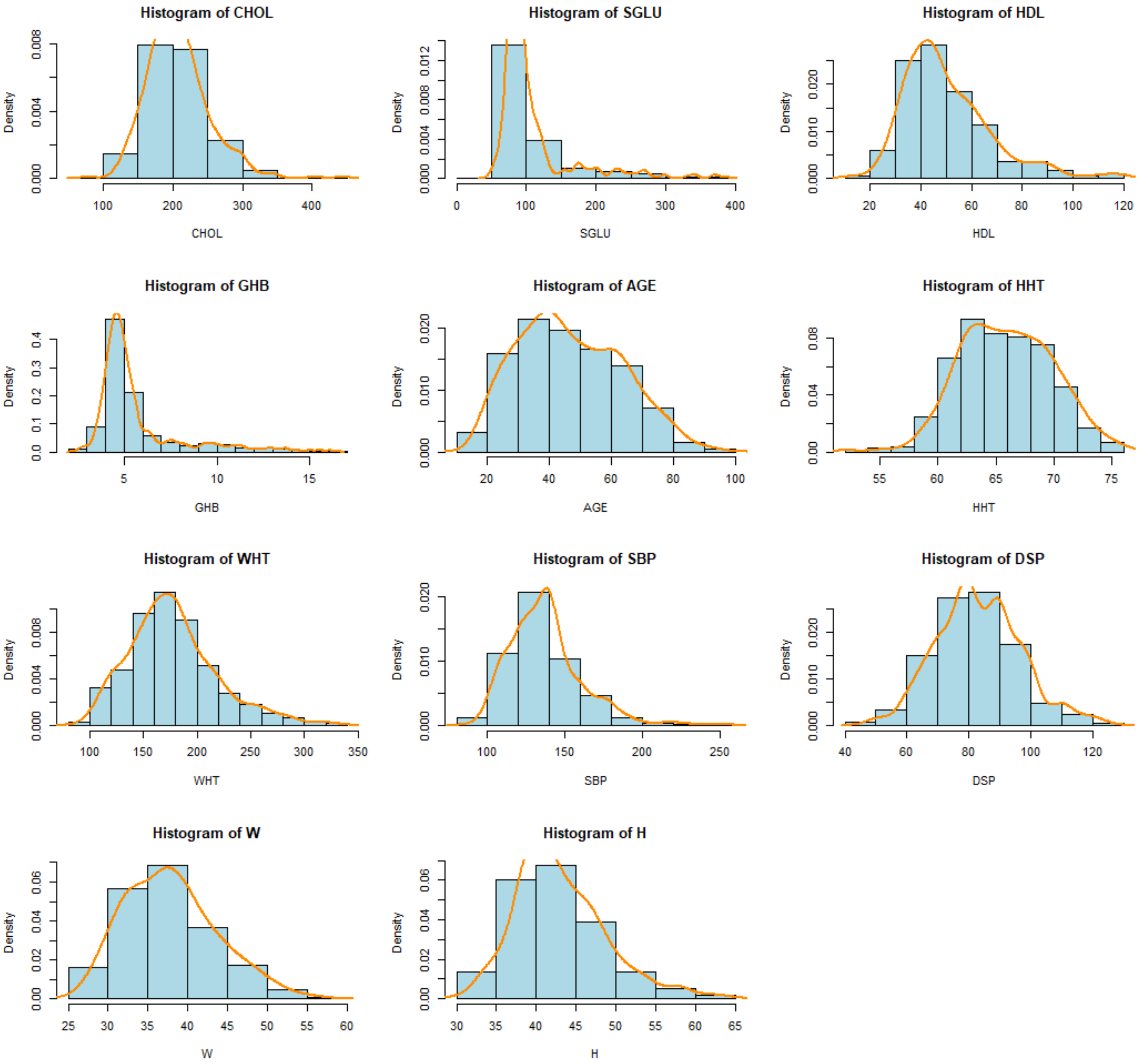


Figure 2: Histograms of numerical features

gamma distributed. In any case, in the following problem, the Gamma distribution appears appropriate for the target variable GHB, as can be seen from the graph. Due to the narrow distribution, we can suspect that the scale parameter is small.

2.2.1 Outliers

To see if we have many outliers in a variable, we have the boxplots of numeric variables in Figure 3.
For two variables, our target and SGLU, which are most likely gamma-distributed, we see many outliers. However, gamma-distributed data naturally have a heavier right tail, so this is not considered an anomaly.
Other variables, except AGE, also have some outliers. For the following models, we will keep data as they are and remove the outliers later if necessary.

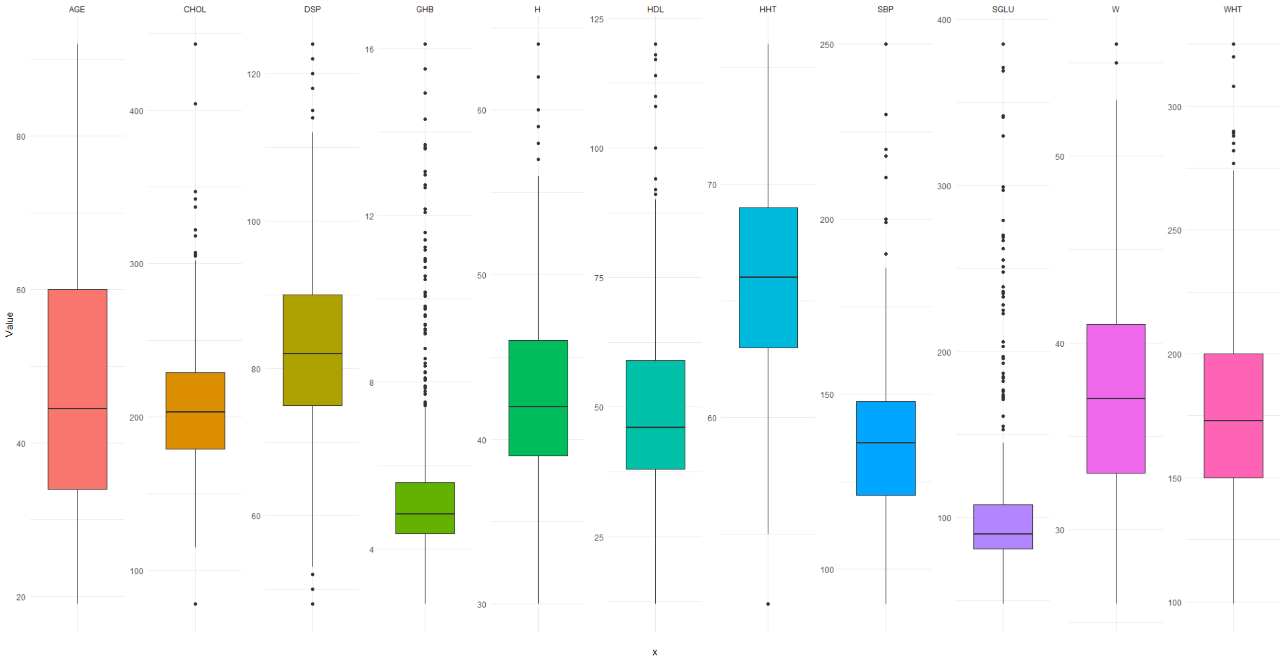


Figure 3: Boxplots of numerical features

2.3 Missing values

The number of missing values in each variable is recorded in the third column of Table 1.
We can see that the biggest part of missing values is in column FRAME. For this variable, we replaced missing values with the value 'medium' because it represents the middle value and is also the value with the highest percentage among participants.
Al rows without measurements in column W also had no data in column H. The same with columns SBP and DSP. For all numeric variables, we replaced missing values with the mean value of the respective variable.

2.4 Data encoding

For further analysis we needed entirely numerical data that all algorithms worked. Therefore, as a first step, we converted all categorical variables into numerical ones. The variables GENDER and LOCATION are binary, so we assigned them values 0 and 1. For the first one, 'female' was assigned 1, and 'male' was assigned 0. For the second one, 'Buckingham' was assigned 1, and 'Louisa' was assigned 0. From the third categorical variable, FRAME, we created three new columns using dummy variables. One of these columns was removed since it could be predicted from the other two.

Having read in several articles that BMI (body mass index) is an important predictor for predicting Glycated Hemoglobin ([2], [3]), and as we have all the data for its calculation, we decided to add a BMI column. We calculated it by formula

$$BMI = weight(lb)/(height(inches))^2 * 703.$$

We would like to be able to show later that the H and W columns are not relevant for the regression because they are already contained in the BMI variable and can be deleted, thereby reducing the dimensionality of the variables used to predict our target variable and consequently simplifying the model.

2.5 Correlation

Figure 4 respresents correlations between the variables in our encoded database. Focusing on the variable GHB that we want to predict with other variables we see a strong connection between GHB and CHOL, SGLU, AGE and W, the higher value these variables have the higher the GHB is going to be. The biggest dependency can be seen between GHB and SGLU.

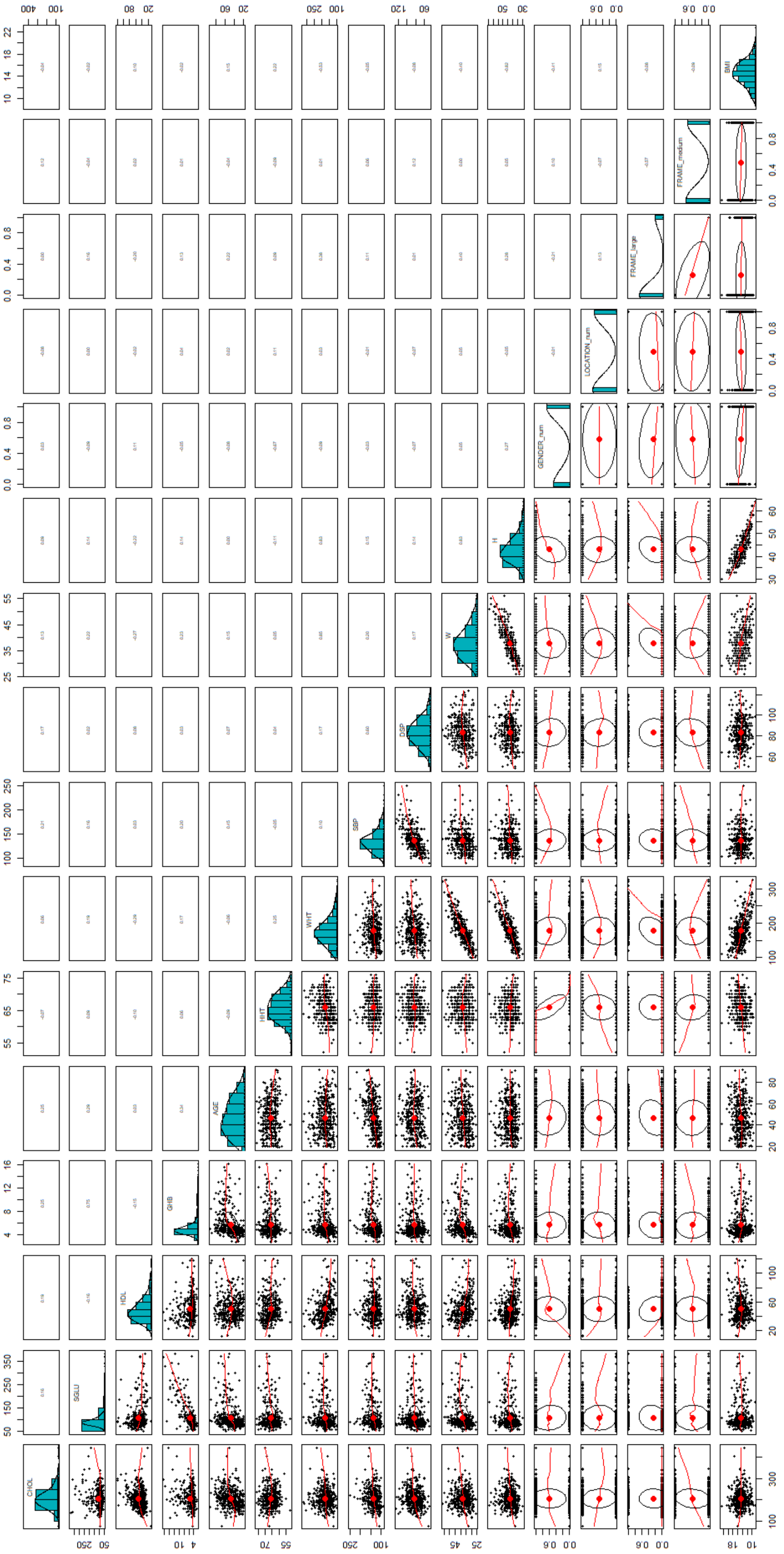


Figure 4: Scatterplots

3 General linear model

Generalized Linear Models (GLM), introduced by Nelder and Wedderburn, synthesize the normal linear model that has a linear regression structure and have in common that the response variable belongs to the exponential distribution family.
The r.v. Y is said to have distribution belonging to the *exponential family* if its p.d.f. or p.m.f. can be written in the form

$$f(y \mid \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}, \tag{1}$$

where θ and ϕ are scalar parameters, $a(\cdot)$, $b(\cdot)$, and $\delta(\cdot, \cdot)$ are known real functions. [5]

3.1 Gamma Distribution

The gamma distribution is a continuous probability distribution that is often used to model positive continuous data with a skewed distribution, such as durations, waiting times, or financial data. It is characterized by two parameters: shape ($\alpha > 0$) and rate ($\beta > 0$). The shape parameter determines the shape of the distribution, while the rate parameter determines the rate at which the distribution decays.
The probability density function (PDF) of a variable y that follows a gamma distribution, parameterized in terms of α and β, is given by:

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot y^{(\alpha-1)} \cdot e^{-\beta y}. \tag{2}$$

In this case, $E(y) = \frac{\alpha}{\beta} =: \mu$ and $Var(y) = \frac{\alpha}{\beta^2}$. If y has density function (2), we write $Y \sim \text{Gamma}(\alpha, \beta)$.

3.1.1 Gamma Distribution as Exponential Family

In this section, we prove that the gamma distribution belongs to the (dispersion) exponential family by representing its density function in the form of equation (1):

$$\begin{aligned} f(x; \alpha, \beta) =& \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{(\alpha-1)} e^{-\beta x} = \\ =& \exp\{-\beta x + \alpha \log(\beta) + (\alpha - 1)\log(x) - \log(\Gamma(\alpha))\} = \\ =& \exp\left\{\frac{-\frac{\beta}{\alpha} x + \log(\beta)}{\frac{1}{\alpha}} + (\alpha - 1)\log(x) - \log(\Gamma(\alpha))\right\}. \end{aligned}$$

Hence, we have $\theta = -\frac{\alpha}{\beta}$, $\phi = \alpha$, and $\beta = -\theta\alpha = \frac{-\theta}{\phi}$. Then

$$\log(\beta) = \log(-\theta) - \log(\phi)$$

and the density function has a form

$$f(x; \alpha, \beta) = \exp\left\{\frac{\theta x - (-\log(-\theta))}{\phi} - \frac{\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(x) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right)\right\}.$$

It shows that the gamma distribution belongs to the exponential family with

$$
\begin{aligned}
a(\phi) &= \phi = \frac{1}{\alpha}, \\
b(\theta) &= -\log(-\theta) = -\log\left(\frac{\beta}{\alpha}\right), \\
c(x,\phi) &= \frac{-\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(x) - \log\left(\Gamma\left(\frac{1}{\phi}\right)\right) = \\
&= \alpha\log(\alpha) + (\alpha - 1)\log(x) - \log(\Gamma(\alpha)).
\end{aligned}
$$

3.2 Gamma Generalized Regression Model

The gamma generalized regression model is a statistical model used to analyze data that follows a gamma distribution. It is an extension of the generalized linear model (GLM) framework, which allows for modeling of response variables that have non-normal error distributions.
In the gamma regression model, the mean of the gamma distribution is related to the covariates through a link function. The most commonly used link function is the logarithmic link, which we will use throughout the project, which takes the form:

$$\log(\mathrm{E}(Y)) = X\beta \tag{3}$$

where $\mathrm{E}(Y)$ is the expected value of the response variable Y, X is the matrix of covariates, β is the vector of regression coefficients, and log denotes the natural logarithm. The model assumes that the response variable has a gamma distribution with a mean equal to $e^{X\beta}$.

The gamma generalized regression model assumes that the response variable follows a gamma distribution with a mean and a dispersion parameter. The dispersion parameter captures the variability of the response variable around the mean. The model assumes that the logarithm of the mean is a linear combination of the covariates.

It is important to note that the gamma generalized regression model assumes that the response variable is strictly positive, as the gamma distribution is only defined for positive values. If the response variable includes zeros or negative values, alternative models may be more appropriate. In our data, the GHB target variable has only positive values, with a minimum value of 2.68 and maximum value of 16.11. As mentioned in Section 2, from the histogram in Figure 2, we find the gamma model to be a good predictor.

In the following, we aim to simplify the model by reducing the set of predictor variables to preserve key information, while eliminating irrelevant variables that could introduce noise. We have tested which model is the best using the libraries in R. In the second part of this chapter we did model evaluation.

3.3 Reducing Dataset

We started this procedure in the following way: Firstly, we identified methods to assess the importance of the variables. The methods we tested are:

- Coefficient Magnitude: We chose the top 5 variables that had the largest (in absolute terms) coefficient, because larger magnitude coefficients generally indicate a stronger impact on the response variable. This approach was not so effective for our data, because is not standardized, which was also shown in the model result.
- P-values: We chose the top 5 variables that had the lowest p-value, because a lower p-value indicates that the variable is likely to be important.
- Confidence Intervals: We chose those variables whose confidence interval with a significant level of 5% did not contain zero. Because the presence of zero suggests that the variable is statistically not significant.
- Deviance & Chi-squared tests: We selected the top 5 variables with the best deviance score obtained from the Chi-squared statistical test.
- EDA based: We selected a few the variables that capture strong relationship between glycosylated hemoglobin and the corresponding set of observed risk factors.

For each data set reduction, we ran the GLM gamma model and checked the value of three metrics to judge which of the models is better. We used:

- MSE (*mean squares of error*) on test data: was used to determine the accuracy of the model. We divided our data into train and test data in a ratio of 4:1, we used the train data to fit a GLM model and the test data to check how well the model performed. Formula for MSE is:

  $$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2,$$

  where n is number of observations, Y_i true value and $\hat{Y_i}$ predicted value. From formula we can see that smaller MSE means better performance of the model.
- AIC (*Akaike information criterion*): is a criterion for selection of regression models. It is based on the log-likelihood function plus a correction factor as penalty of the model complexity, whose statistic is given by

  $$AIC = -2\log(L(\theta \mid D)) + 2p,$$

  where $L(\hat{\theta} \mid D)$ is the likelihood function, D is the dataset (n observations) and θ is the maximum likelihood estimator of the parameter θ of dimension p, under the current statistical model. A low value for AIC indicates a better fit. [4]
- BIC (*Bayesian Information Criterion*): balances the goodness of fit of a model with the complexity of the model by penalizing the number of parameters. BIC aims to identify the model that best explains the data while avoiding overfitting. Formula for BIC is:

  $$BIC = -2\log(L) + k \cdot \log(n)$$

  where L is the likelihood of the model, k is the number of parameters, and n is the number of observations. In the context of BIC, lower values indicate better models. [1]

Results of each model reduction and metrics are in table 2. As we can see, we have the best results for Confidence intervals and EDA based choice of variables. The first choice has variables CHOL, SGLU and AGE. In the EDA based choice we added just W. Given that 2 of the 3 criteria have a lower value for the EDA based choice, we have chosen the variables CHOL, SGLU, AGE and W as the final model to which we will apply the Bayesian approach later. Thus we have also confirmed the claims made in the section 2.5.

Importance Method	**MSE**	**AIC**	**BIC**
Full Model(all data)	1.719	1228.706	1296.130
Coefficient Magnitude (top 5)	1.872	1019.708	1045.954
P-values (top 5)	1.741	988.716	1014.961
CI (not having 0)	1.636	992.488	1011.235
Deviance/Chi-squared (top 5)	4.453	1254.605	1280.850
EDA	1.635	990.642	1013.138

Table 2: Results of models with reduction dataset

3.4 Final model evaluation

For the purpose of evaluating the final model, let's examine Figure 5, which includes 4 diagnostic plots.

From the **Residuals vs Fitted Values** plot, we can observe that the majority of residuals are randomly distributed around zero. This suggests that the model effectively captures the linear relationship between the covariates and the response variable, indicating unbiased predictions with no systematic patterns left unexplained. This is a positive indication of a good fit to the data.

The **QQ** plot of residuals helps us assess whether the residuals follow a normal distribution. The points are distributed along the diagonal, confirming that the residuals exhibit a normal distribution, as assumed by the model.

The **Scale-Location** plot assesses whether residuals are evenly spread along the ranges of predictors, checking for homoscedasticity. The plot shows a horizontal line with randomly spread points, indicating that the assumption of constant variance is met, which is desirable.

The **Cook's distance** plot is used to detect influential observations in a regression model. We identified a few observations with Cook's distance greater than 0.05. Upon closer analysis, we did not observe any specialities in these observations, so we decided to proceed with the analysis using the entire dataset.

In Figure 6, the distribution of the fitted model is compared with the original data histogram, showing a good match, suggesting that our model has a reasonable fit.

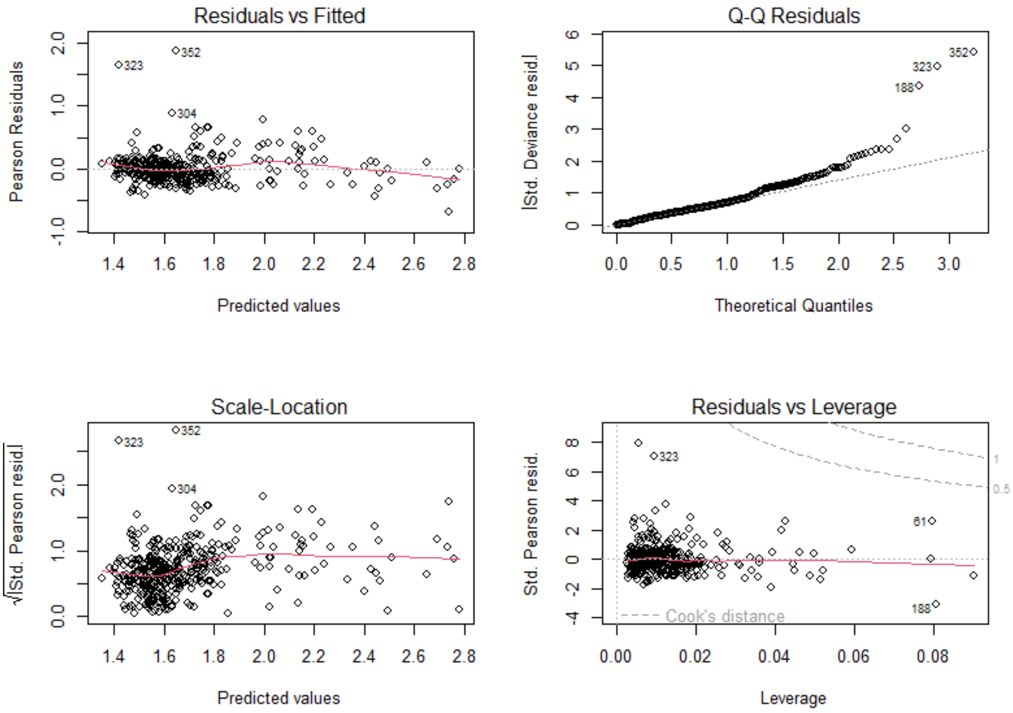


Figure 5: Model evaluation plots

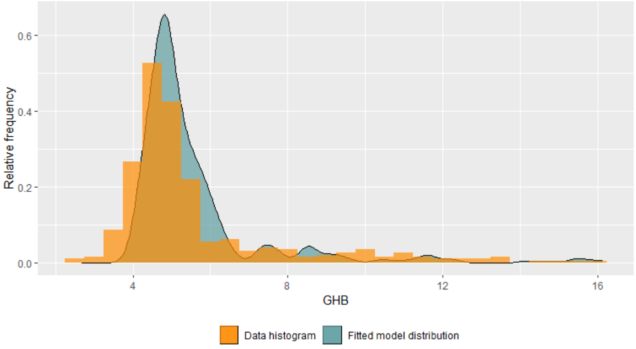
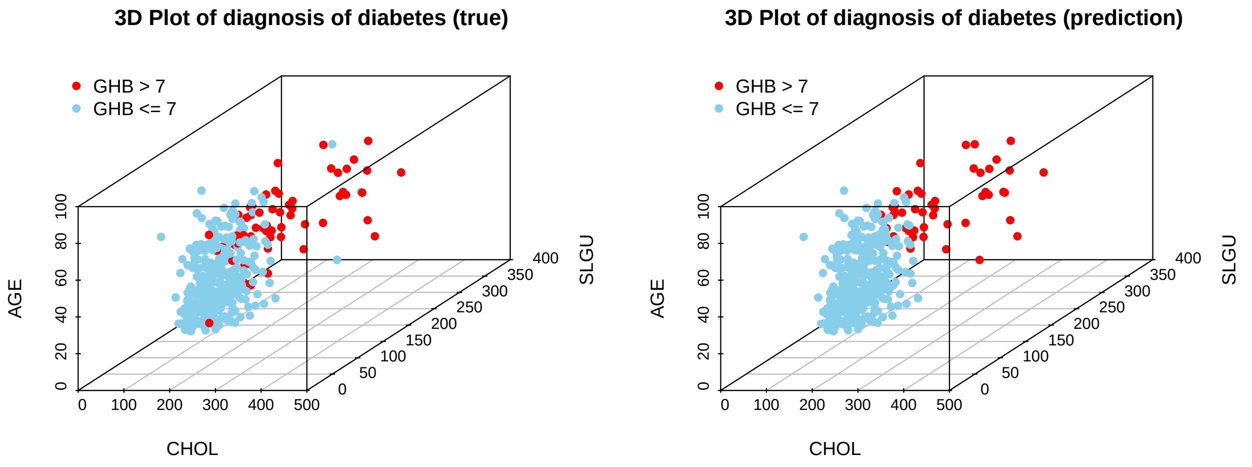


Figure 6: Fitted model distribution vs data histogram

3.5 Interpretation

From the above results, we can infer that the variables carrying the most information about Glycosolated hemoglobin (GHB), based on the tests we conducted, are Total Cholesterol, Stabilized Glucose, Age, and Waist. Our predictions were that the BMI index would be one of these variables, but it turned out that the model works much better without it. Our goal is also to assess whether a person is diabetic based on GHB since individuals with a GHB value larger than 7 are classified as diabetics. For the interpretation of which of the selected variables has a significant weight in predicting GHB, we standardized the data and observed excessively large coefficients. We found that they have a relatively proportional impact on the GHB value of the individual.

In the Figures 7a and 7b, we aimed to represent individuals with diabetes in space based on three components: Total Cholesterol, Stabilized Glucose, and Age. Figure 7a plots the marked instances for which GHB ¿ 7, based on actual data, while Figure 7b identifies positive diabetics predicted by the GLM using input data. We can observe that the graphs are very similar, hence we can assert that the GLM is well-fitted.



(a) True

(b) Prediction

Figure 7: Diagnosis of diabetes

In the next section, let's explore if we can improve the regression results by assuming that the parameters of the gamma distribution also have a distribution dependent on our four selected variables: CHOL, SGLU, AGE, and W.

4 A Bayesian approach to GLM

Let $y_1, ..., y_n$ be independent random variables such that $y_i \sim \text{Gamma}(\alpha, \beta_i)$. The gamma regression model is defined assuming that

$$\eta_i = g(\mu_i) = x_i'\delta,$$

where $\delta = (\delta_1, ..., \delta_p)'$ is a vector of unknown regression parameters $(p < n)$, $x_i = (x_{i1}, ..., x_{ip})'$ is vector of covariates of the i-th observation, η_i is a linear predictor and $\mu_i = \frac{\alpha}{\beta_i}$. Here we also use log link function, so $g(\mu_i) = \log(\mu_i)$.

In order to apply Bayesian methods to fit the gamma Bayesian model, we assume multivariate normal prior distributions for δ, that is

$$\delta \sim N_p(0, cI),$$

where $c = 0.001$.

In following, we compare the two models. For both, we assume a constant shape parameter (α), which has a prior distribution $\alpha \sim \text{InverseGamma}(a, b)$ where in Model 1 we have $a = b = 0.001$ and in Model 2 we have $a = 2, b = 1$.

4.1 Model

Assuming that the parameters α and δ are independents, the joint posterior distribution is given by:

$$f(\delta, \alpha, \eta \mid y) \propto \left[\prod_{i=1}^{n} f(y_i \mid \eta_i, \alpha)\right] \left[\prod_{i=1}^{n} f(\eta_i \mid \delta)\right] f(\alpha) f(\delta)$$

Thus, samples of $f(\delta, \alpha, \eta \mid y)$ are obtained by iterative process from the full conditional distributions:

$$f(y_i \mid \eta_i, \alpha), f(\eta_i \mid \delta), f(\alpha), f(\delta), i = 1, ..., n$$

The algorithm can be implemented using R2jags and this software can also be used to obtain posterior parameter inferences.

4.2 Application on our data

In our chosen model, where we predict GHB, we have chosen the variables in X=(SGLU, CHOL, AGE, W) based on the GLM in section 3.3, so we have a location regression structure given by: $\eta_i = \log(\mu_i) = \delta_0 + \delta_1 \text{CHOL}_i + \delta_2 \text{SGLU}_i + \delta_3 \text{AGE}_i + \delta_4 \text{W}_i$

Both models were fitted using the R2jags program given in the Appendix. The corresponding DIC values for the fitted models are given by: Model 1, DIC = 1321.1; Model 2, DIC = 1298.9. Table 3 gives the posterior estimates of the parameters associated with Model 2, which provide the least DIC value.

We considered 50000 Monte Carlo iterations (to secure convergence) and our results were obtained with the posterior samples obtained from last 45000 iterations.

4.3 Convergence diagnosis

After using the MCMC, we wanted to check the convergence of our parameters. The best known instrument for monitoring convergence for stationary distribution is the graphical representation for each scalar quantity of simulated chain values over successive iterations, connected by a

Parameter	Mean	SD	95% CI
δ_1	-0.006	0.008	$(-0.018, 0.009)$
δ_2	0.002	0.000	$(0.001, 0.002)$
δ_3	0.004	0.000	$(0.003, 0.004)$
δ_4	0.004	0.001	$(0.003, 0.006)$
δ_5	0.019	0.002	$(0.016, 0.022)$
α	17.602	1.287	$(15.256, 20.099)$

Table 3: Estimated posterior means, SD and 95% credibility intervals (CI) for the parameters in Model 2

continuous line. The Figure 8 shows the trace plot, for each of the parameters, for the selected model. We can see that the Markov chain has no trend with the exception of intercept δ_1.

For the next convergence check, we used Geweke method which provides Z-scores for each parameter in the model, based on the comparison of means between the early and late parts of the MCMC chain. Z-scores for all parameters are within a reasonable range (around -1.5 to 1.5), suggesting convergence. It's a good sign that the Z-scores are not too extreme, indicating that the means of the early and late parts of the chain are consistent.

For all parameters the stationarity test has passed, which is another criteria for convergence. This suggests that the chain has reached a stationary distribution at a specific iteration. The p-values for the stationarity test are generally high, indicating good stationarity. The halfwidth tests have also passed for all parameters, indicating reasonable precision in estimating the mean.

Finally we checked Highest Posterior Density for each parameter in the MCMC output. The HPD interval is a Bayesian statistical measure that represents the range of values containing the most credible parameter values. Each point of 95% of interval is presented in table 4. We can see that HPD of intercept δ_1 contains 0, which indicate that we could not include it in terms of improving the model.

Parameter	Lower	Upper
δ_1	-0.017	0.0092
δ_2	0.0012	0.0022
δ_3	0.0033	0.0045
δ_4	0.0024	0.0056
δ_5	0.0157	0.0217
α	15.226	20.009

Table 4: HPD of parameters

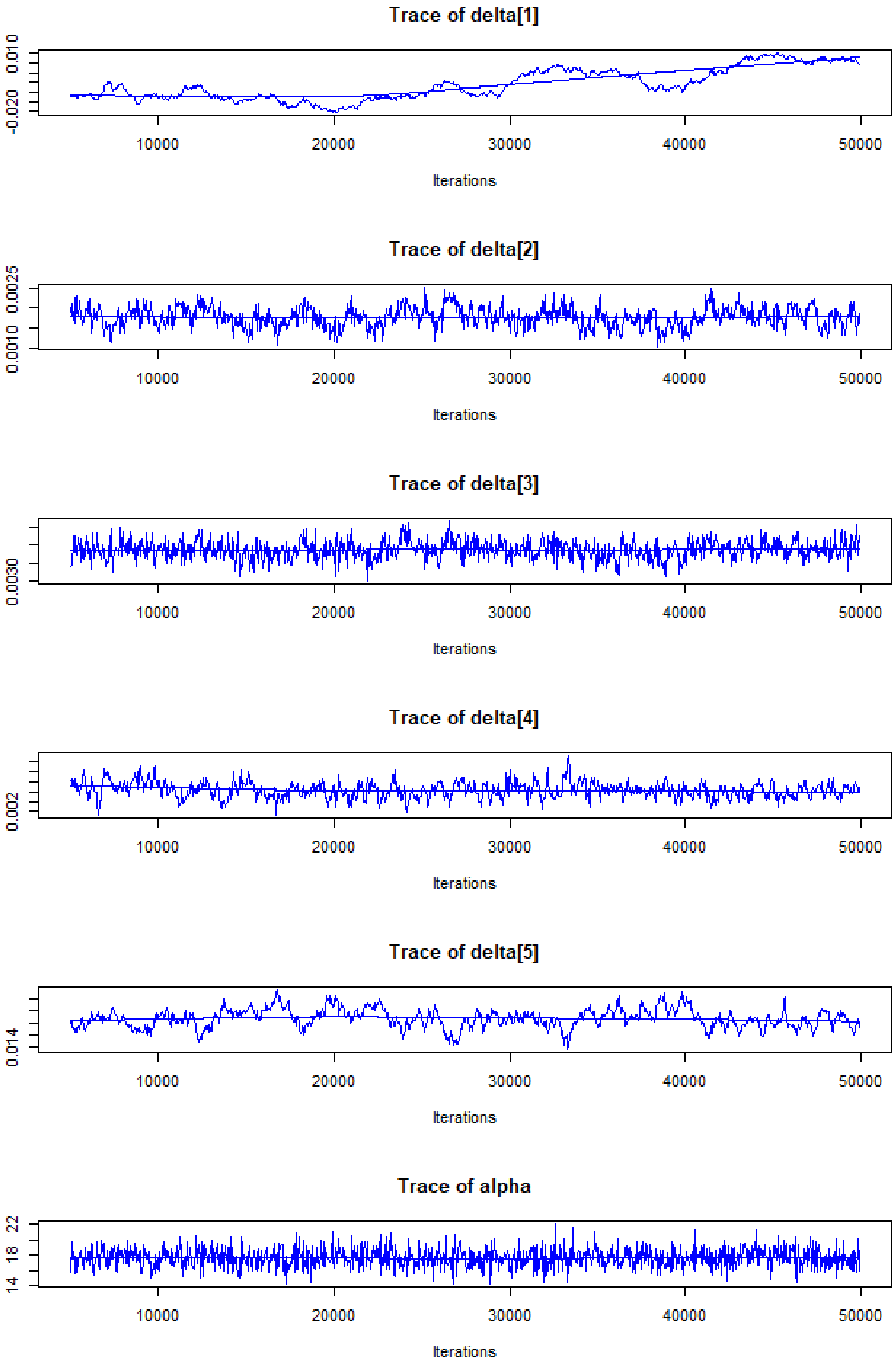


Figure 8: Trace Plot for parameters

5 Conclusion

In conclusion, in this report we perform statistical analysis of the data with the goal to find the relationship between the glycosolated hemoglobin and the associated set of the observed risk factors.

In particular we use Exploratory data analysis to get good representation of the dataset. Then, by use of GLM model we abstract important risk factors to reduce size of dataset and finaly use MCMC for Bayesian approach. Specifically we focused on a Bayesian approach to Generalized Linear Models (GLM) for predicting glycated hemoglobin levels (GHB). The result of our analysis centers on a dataset with variables such as cholesterol (CHOL), glucose (SGLU), age, and weight (W), aiming to model their correlation with GHB.

The study employs the gamma distribution within the GLM framework, using a logarithmic link function. The choice of variables for the model is refined through various reduction techniques, including coefficient magnitude, p-values, confidence intervals, and deviance tests. The final model, comprising CHOL, SGLU, AGE, and W, is evaluated using metrics like mean squares of error (MSE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

The Bayesian approach is then applied to the GLM, assuming a multivariate normal prior distribution for regression parameters and constante inverse gamma distribution for shape parameter. The resulting model is compared with an alternative, considering a distribution for parameters dependent on selected variables. The Bayesian model's performance is assessed through Deviance Information Criterion (DIC), convergence diagnosis, and posterior parameter estimates.

The report provides a comprehensive exploration of the statistical methodology, ensuring the model's reliability through convergence checks and diagnostic plots. The Bayesian approach adds a layer of complexity by introducing distributional assumptions for parameters, enhancing the model's flexibility.

References

[1] Bayesian information criterion on sciencedirect.

[2] Budgen D. Alshammari R. Al Moubayed N. Alhassan, Z. Predicting current glycated hemoglobin levels in adults from electronic health records: Validation of multiple logistic regression algorithm. *JMIR medical informatics*, 8(7), e18963, 2020.

[3] Watson M. Budgen D. Alshammari R. Alessa A. Al Moubayed N. Alhassan, Z. Improving current glycated hemoglobin prediction in adults: Use of machine learning algorithms with electronic health records. *JMIR medical informatics*, 9(5), e25237, 2021.

[4] G.L. Silva. Lecture notes of linear model analysis. *Lisbon, Instituto Superior Tecnico*, 2017.

[5] G.L. Silva. Lecture notes of computational statistics. *Lisbon, Instituto Superior Tecnico*, 2023.

Appendices

A Model 1

```
model
{
  for (i in 1:n) {
    log(mu[i]) <- inprod(x[i, ], beta[])  # Assuming beta[] is a row vector
    Y[i] ~ dgamma(a1[i], a2[i])
    a1[i] <- phi
    a2[i] <- phi / mu[i]
  }

  precision_matrix[1:p, 1:p] <- inverse(0.001 * I)  # Assuming p is the
   ↪  dimension of beta
  beta[1:p] ~ dmnorm(rep(0, p), precision_matrix)

  phiinv ~ dgamma(0.001, 0.001)
  phi <- 1 / phiinv
}
```

B Model 2

```
model
{
  for (i in 1:n) {
    log(mu[i]) <- inprod(x[i, ], beta[])  # Assuming beta[] is a row vector
    Y[i] ~ dgamma(a1[i], a2[i])
    a1[i] <- phi
    a2[i] <- phi / mu[i]
  }

  precision_matrix[1:p, 1:p] <- inverse(0.001 * I)  # Assuming p is the
   ↪  dimension of beta
  beta[1:p] ~ dmnorm(rep(0, p), precision_matrix)

  phiinv ~ dgamma(2, 1)
  phi <- 1 / phiinv
}
```