Projeto de Estatística Computacional, Report 5

DIOGO RAMOS CAVALCANTI DE SOUZA, 95244
RENAN YUDI HAMADA NUNES, 95357

Este documento apresenta a abordagem do grupo para estudar o conjunto de dados fornecido pelo docente, cobrindo o conteúdo da disciplina. O *dataset* apresenta informações de indivíduos afro-americanos na *Central Virgínia*. São aplicados métodos de exploração e análise de dados, pré-processamento e regressão linear generalizada. Os resultados são apresentados e devidamente analisados.

1 INTRODUÇÃO

O conjunto de dados analisado apresenta 390 registros de pessoas que foram entrevistadas para um estudo visando a entender a prevalência de obesidade, diabetes e outros fatores de risco cardiovascular para afro-americanos na *Central Virgínia*.

Um dos objetivos do estudo era verificar a relação entre Hemoglobina glicada e os fatores de risco cardiovasculares observados. Por esse motivo, através de uma abordagem Bayesiana, foram feitos modelos lineares generalizados para analisar a influência das outras variáveis na Hemoglobina glicada.

2 ANÁLISE EXPLORATÓRIA E PREPARAÇÃO DOS DADOS

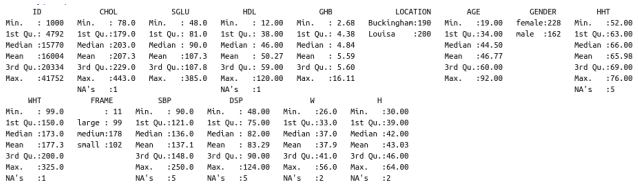
Ao deparar-se com um conjunto de dados, convém compreendê-lo melhor em aspectos gerais antes de aprofundar-se nas técnicas de pré-processamento e modelagem. Isso envolve aspectos qualitativos, quantitativos e análise estatística.

O *dataset* apresenta 390 observações, com 14 variáveis de dados e uma de ID. Sendo 11 delas com dados numéricos e 3 categóricos. As numéricas são: CHOL (Colesterol total); SGLU (Glicose estabilizada); HDL (Lipoproteína de alta densidade); GHB (Hemoglobina glicada); AGE (Idade); HHT (Altura em polegadas); WHT (Peso em libras); SBP (Pressão arterial sistólica); DSP (Pressão arterial diastólica); W (Cintura em polegadas); H (Quadril em polegada). As categóricas são: LOCATION (Localização - podendo ser Buckingham ou Louisa); GENDER (Gênero - podendo ser feminino ou masculino); FRAME (podendo ser vazio, pequeno, médio ou grande). Em especial, a variável alvo (GHB) é numérica, contínua e estritamente positiva. Essas características serão consideradas no momento da modelagem.

Algumas estatítiscas dos dados estão apresentados na figura 1. Com a figura, também verificamos que a quantidade de valores ausentes é pequena ("NA's"), apresentando, no máximo, 5 valores faltantes numa coluna. Para tratar essas informações faltantes, o grupo efetuou a inserção do valor médio da coluna quando a informação era inexistente, não comprometendo muito as distribuições dos dados.

A matriz de correlação indica o quão próximo as variáveis estão de apresentar uma relação linear entre si, então, para verificar essas correlações, a matriz está apresentada de forma gráfica na figura 2. É possível notar que a maioria das variáveis não apresentam muita correlação linear entre seus valores. As maiores correlações são entre:

Authors' addresses: Diogo Ramos Cavalcanti de Souza, 95244; Renan Yudi Hamada Nunes, 95357.



```
       ID              CHOL            SGLU            HDL              GHB          LOCATION       AGE          GENDER         HHT       
 Min.   : 1000   Min.   : 78.0   Min.   : 48.0   Min.   : 12.00   Min.   : 2.68   Buckingham:190   Min.   :19.00   female:228   Min.   :52.00  
 1st Qu.: 4792   1st Qu.:179.0   1st Qu.: 81.0   1st Qu.: 38.00   1st Qu.: 4.38   Louisa    :200   1st Qu.:34.00   male  :162   1st Qu.:63.00  
 Median :15770   Median :203.0   Median : 90.0   Median : 46.00   Median : 4.84                    Median :44.50                 Median :66.00  
 Mean   :16004   Mean   :207.3   Mean   :107.3   Mean   : 50.27   Mean   : 5.59                    Mean   :46.77                 Mean   :65.98  
 3rd Qu.:20334   3rd Qu.:229.0   3rd Qu.:107.0   3rd Qu.: 59.00   3rd Qu.: 5.60                    3rd Qu.:60.00                 3rd Qu.:69.00  
 Max.   :41752   Max.   :443.0   Max.   :385.0   Max.   :120.00   Max.   :16.11                    Max.   :92.00                 Max.   :76.00  
                 NA's   :1                       NA's   :1                                                                      NA's   :5      
      WHT           FRAME          SBP             DSP              W                H        
 Min.   : 99.0          : 11   Min.   : 90.0   Min.   : 48.00   Min.   :26.0   Min.   :30.00  
 1st Qu.:150.0   large  : 99   1st Qu.:121.0   1st Qu.: 75.00   1st Qu.:33.0   1st Qu.:39.00  
 Median :173.0   medium:178   Median :136.0   Median : 82.00   Median :37.0   Median :42.00  
 Mean   :177.3   small :102   Mean   :137.1   Mean   : 83.29   Mean   :37.9   Mean   :43.03  
 3rd Qu.:200.0                3rd Qu.:148.0   3rd Qu.: 90.00   3rd Qu.:41.0   3rd Qu.:46.00  
 Max.   :325.0                Max.   :250.0   Max.   :124.00   Max.   :56.0   Max.   :64.00  
 NA's   :1                    NA's   :5       NA's   :5        NA's   :2      NA's   :2      
```

Fig. 1. Algumas estatísticas empíricas dos dados

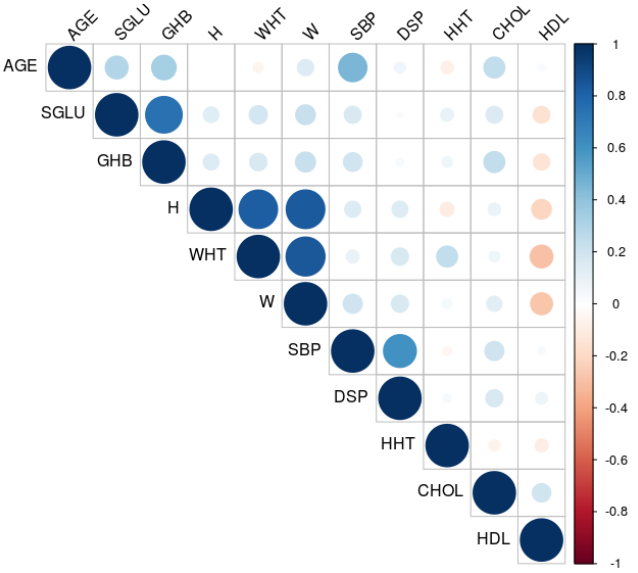


Fig. 2. Matriz de correlação entre as variáveis

Peso e quadril; Peso e cintura; Quadril e cintura; Glicose estabilizada e Hemoglobina glicada. As três primeiras não são importantes para a análise, mas a última pode ser especialmente relevante para o modelo. E essa correlação ocorre porque o número de hemoglobinas desse tipo está ligado com o nível de glicose no sangue.

Um fator de risco cardiovascular que podemos analisar é a obesidade, que pode ser verificada através de uma relação entre altura e peso. Para facilitar esse estudo, foi criada a coluna BMI (IMC - Índice de Massa Corporal), que é calculado pelo peso dividido pela altura ao quadrado. Além disso, também foi adicionado a BMI_CAT (classificação do IMC) para indicar a faixa que o IMC se encontra (exemplo: Obesidade). Outras colunas que foram adicionadas com base nas existentes são: DIABETES (indica se o nível de GHB está maior do que 7 (Positivo) ou não (Negativo)); AGE_CAT (separa as faixas de idades: 18-25, 26-35, 36-45, 46-55, 56-65, maior ou igual a 66). As estatísticas das variáveis incluídas estão apresentadas na figura 3.

```
      BMI                BMI_CAT        DIABETES    AGE_CAT
 Min.   :15.20   Underweight  :  9   Negative:330   18-25:35
 1st Qu.:24.12   Normal weight:111   Positive: 60   26-35:70
 Median :27.79   Overweight   :121                  36-45:98
 Mean   :28.74   Obesity      :149                  46-55:66
 3rd Qu.:32.23                                     56-65:65
 Max.   :55.78                                     >=66 :56
```

Fig. 3. Estatísticas empíricas das variáveis incluídas

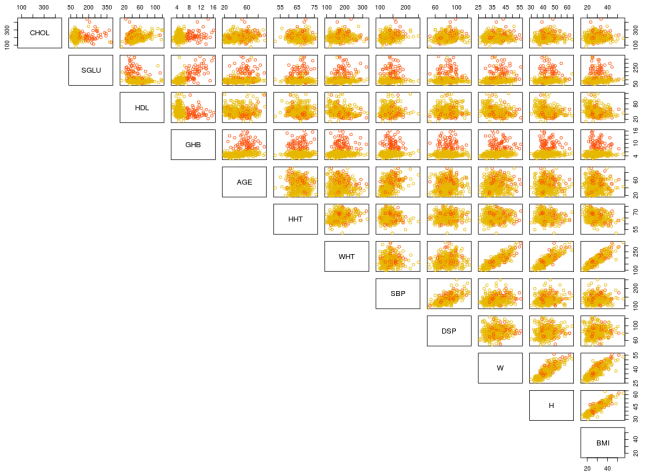


Fig. 4. Gráfico de dispersão das variáveis numéricas

Para verificar melhor o comportamento das variáveis duas a duas, utilizamos o gráfico de dispersão da figura 4. As cores laranja e amarelo foram utilizadas para distinguir os pontos com GHB maior do que 7, que são os considerados com diabetes (laranja), dos outros (amarelo). Com este gráfico, podemos ver melhor a relação entre SGLU e GHB (a figura 5 mostra o gráfico de dispersão entre os dados e sem cores). Também podemos observar a dispersão de CHOL e GHB (figura 6), porém sem poder tirar conclusões.

Para avaliar a distribuição da variável alvo, foi utilizado o histograma da figura 7, que, como já dito anteriormente, apresenta valores positivos e contínuos. Uma vez que os valores podem ser maiores do que 1, a distribuição beta não seria adequada para o modelo de regressão, já a gama poderia ser. Por esse motivo, o histograma apresenta uma linha azul, que indica a densidade de uma distribuição gama.

O próximo tópico a ser abordado é a criação do modelo linear generalizado e a aplicação do mesmo com determinadas covariáveis.

3 MODELAGEM

Como mencionado anteriormente, a distribuição gama poderia ser adequada para o modelo de regressão dadas as características da variável alvo. Logo, esta foi a distribuição utilizada para a modelagem. A função de ligação utilizada foi a canônica, ou seja, a função inversa ($\mathbf{X}\beta = \mu^{-1}$).

As informações a priori não informativas utilizadas foram: normais com média zero e variância grande para os betas (com exceção

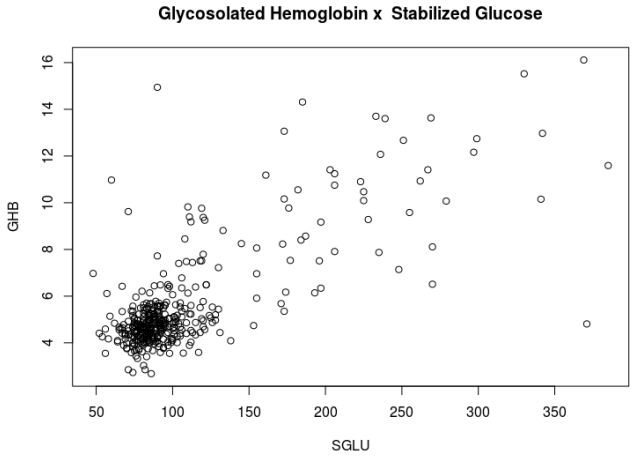


Fig. 5. Gráfico de dispersão entre SGLU e GHB

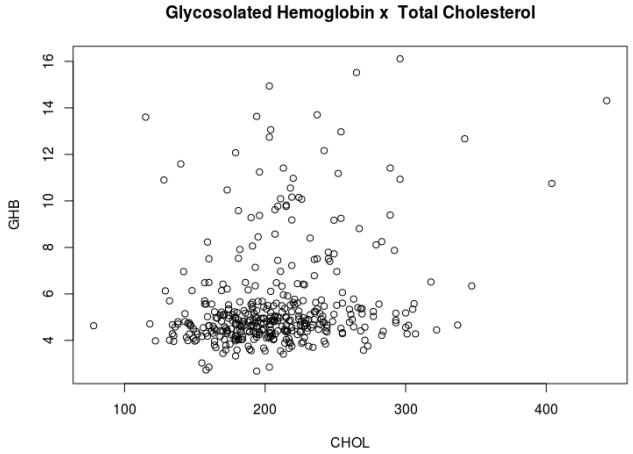


Fig. 6. Gráfico de dispersão entre CHOL e GHB

do b_0); gama com *shape* e *rate* pequenos para o *shape* da função da variável alvo. No caso do b_0, a média escolhida foi de 0.5 para evitar que o resultado da regressão fosse negativo, o que acarretaria em erro no programa. Resumindo, para n sendo o número de variáveis:

$$y_i \sim Gamma(\alpha, \beta_i)$$
$$\alpha \sim Gamma(0.001, 0.001)$$
$$\beta_i = \alpha \cdot \mu_i^{-1}$$
$$\mu_i^{-1} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_n \cdot x_n$$
$$b_0 \sim N(0.5, 10000)$$
$$b_j \sim N(0, 10000), 1 < j \leq n$$

Foi utilizado o método de amostragem Gibbs (*Gibbs Sampling*) através do pacote JAGS com 3 cadeias para as simulações, período de *burn in* igual a 1000 e 5000 iterações para cada cadeia.

A abordagem escolhida para os testes foi gerar o modelo com mais covariáveis e então retirar uma a uma de acordo com a métrica

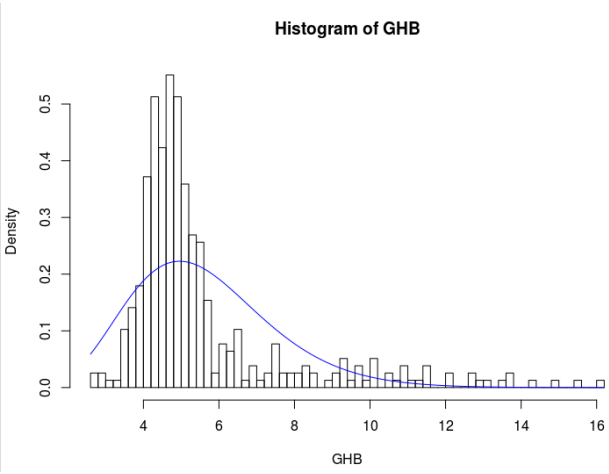


Fig. 7. Histograma de GHB com a curva de uma distribuição gama sobreposta

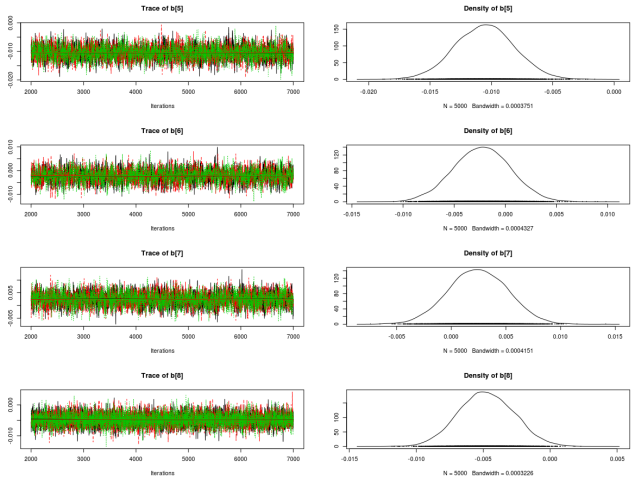


Fig. 8. Gráfico com a progressão de alguns dos parâmetros do primeiro modelo para as 3 cadeias simuladas

```
> gelman.diag(mod1_sim)
Potential scale reduction factors:

     Point est. Upper C.I.
b[1]          1          1
b[2]          1          1
b[3]          1          1
b[4]          1          1
b[5]          1          1
b[6]          1          1
b[7]          1          1
b[8]          1          1

Multivariate psrf

1
```

Fig. 9. Valores relacionados ao fator de redução de escala (Método de Gelman e Rubin)

apresentada. Portanto, o primeiro utilizou os seguintes atributos: CHOL, SGLU, HDL, AGE, SBP, DSP, e BMI normalizados (média 0 e desvio padrão 1). As variáveis H, W, WHT, HHT não foram utilizadas, pois o IMC (BMI) apresenta dados mais informativos sobre essas características dos indivíduos.

4 RESULTADOS

4.1 Modelo 1

Após simular a cadeia para o modelo descrito anteriormente, foi preciso verificar a convergência antes de analisar os resultados, por isso foi verificado os gráficos após o período de *burn in* (figura 8). Apesar do gráfico aparentar indicar a convergência, foi utilizado o método de Gelman e Rubin para coletar um segundo sinal, o mesmo se encontra na figura 9. Os valores são próximos de 1, ou seja, são indícios de que as sequências simuladas se aproximam da distribuição alvo. Por fim, foi feito o teste de Heidel, que também indicou convergência.

Outra análise relevante é a de resíduos, que se encontra na figura 10. Não existe nenhum padrão aparente no gráfico, logo não parece ter problemas em relação a isso.

Analisando o resumo do modelo, que se encontra na figura 11, podemos perceber que, de maneira geral, os valores dos b's são pequenos, em particular, o b_6 e o b_7 apresentam variação englobando números positivos e negativos. Isso indica que, por incluir o zero, a influência do atributo pode ser nula, logo será retirado um dos dois para o teste do modelo 2. O b_6, que se refere à variável SBP, foi o escolhido para remoção, pois apresentou a menor média.

A última análise do modelo 1 foi verificar a métrica DIC (Critério de informação pela desviância), que apresenta o valor de 1255, como indicado na figura 12. Esse número foi usado para comparar com os outros modelos.

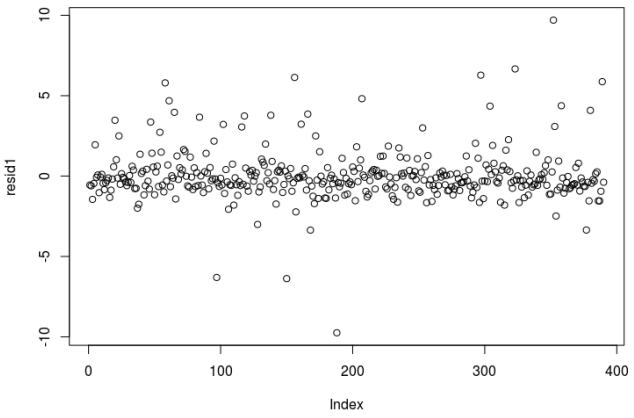


Fig. 10. Resíduos do modelo 1

4.2 Modelo 2

O modelo 2 utiliza as variáveis: CHOL, SGLU, HDL, AGE, DSP e BMI.

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.189126	0.002176	1.777e-05	2.703e-05
b[2]	-0.007032	0.001787	1.459e-05	2.119e-05
b[3]	-0.024154	0.001216	9.928e-06	1.768e-05
b[4]	0.004724	0.002093	1.709e-05	2.517e-05
b[5]	-0.010509	0.002420	1.976e-05	3.833e-05
b[6]	-0.002282	0.002830	2.311e-05	5.170e-05
b[7]	0.002452	0.002666	2.177e-05	4.368e-05
b[8]	-0.004764	0.002088	1.705e-05	2.428e-05
shape	19.897221	1.417680	1.158e-02	1.509e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.1848665	0.1876469	0.189115	0.1905647	0.1934053
b[2]	-0.0105359	-0.0082418	-0.007033	-0.0058442	-0.0034842
b[3]	-0.0265215	-0.0249796	-0.024168	-0.0233445	-0.0217254
b[4]	0.0006143	0.0033204	0.004744	0.0061350	0.0088645
b[5]	-0.0152467	-0.0121288	-0.010523	-0.0088951	-0.0057287
b[6]	-0.0076842	-0.0042211	-0.002323	-0.0003516	0.0032999
b[7]	-0.0028314	0.0006141	0.002484	0.0042848	0.0075813
b[8]	-0.0088329	-0.0061547	-0.004767	-0.0033721	-0.0006533
shape	17.2647577	18.9223878	19.861246	20.8342924	22.7851356

Fig. 11. Dados referentes ao resultado do modelo 1

```
> dic.samples(mod1, n.iter = 1e3)
  |**************************************************| 100%
Mean deviance:  1246
penalty 9.165
Penalized deviance: 1255
```

Fig. 12. Métrica DIC do modelo 1

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.189116	0.002179	1.779e-05	2.692e-05
b[2]	-0.007135	0.001784	1.457e-05	2.279e-05
b[3]	-0.024205	0.001219	9.952e-06	1.793e-05
b[4]	0.004817	0.002122	1.732e-05	2.718e-05
b[5]	-0.011390	0.002140	1.747e-05	2.796e-05
b[6]	0.001160	0.002168	1.770e-05	2.405e-05
b[7]	-0.004839	0.002051	1.674e-05	2.245e-05
shape	19.892569	1.414874	1.155e-02	1.504e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.1848680	0.1876315	0.189119	0.190594	0.193426
b[2]	-0.0105961	-0.0083533	-0.007136	-0.005932	-0.003627
b[3]	-0.0265362	-0.0250410	-0.024215	-0.023387	-0.021745
b[4]	0.0006553	0.0033862	0.004829	0.006217	0.008984
b[5]	-0.0156211	-0.0128196	-0.011390	-0.009958	-0.007153
b[6]	-0.0030942	-0.0002826	0.001134	0.002577	0.005405
b[7]	-0.0088597	-0.0062175	-0.004853	-0.003460	-0.000791
shape	17.2548865	18.9104762	19.847060	20.837544	22.777147

Fig. 13. Dados referentes ao resultado do modelo 2

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.189059	0.002189	1.787e-05	2.667e-05
b[2]	-0.006966	0.001703	1.391e-05	1.982e-05
b[3]	-0.024208	0.001245	1.016e-05	1.839e-05
b[4]	0.004846	0.002102	1.716e-05	2.539e-05
b[5]	-0.011321	0.002113	1.725e-05	2.688e-05
b[6]	-0.004633	0.002041	1.667e-05	2.207e-05
shape	19.946460	1.423926	1.163e-02	1.471e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.1847738	0.187608	0.189062	0.190509	0.1933949
b[2]	-0.0102538	-0.008119	-0.006987	-0.005821	-0.0035666
b[3]	-0.0265900	-0.025046	-0.024233	-0.023387	-0.0216673
b[4]	0.0007124	0.003435	0.004849	0.006229	0.0090730
b[5]	-0.0154590	-0.012741	-0.011319	-0.009878	-0.0071959
b[6]	-0.0086573	-0.006003	-0.004645	-0.003265	-0.0005976
shape	17.2176782	18.984294	19.913279	20.870933	22.8426726

Fig. 14. Dados referentes ao resultado do modelo 3

Fazendo os mesmos testes de convergência, os resultados foram similares, indicando não haver problemas. A análise de resíduos também apresentou resultados similares.

O resumo do modelo encontra-se na figura 13. Deste vez, apenas o b_6 apresentou o zero no seu intervalo de 95%, portanto este foi o atributo (DSP) a ser retirado para o teste do próximo modelo.

Por fim, a métrica DIC apresenta valor 1254, que é menor do que o do modelo 1, logo, por essa métrica, o modelo 2 é um pouco melhor.

4.3 Modelo 3

O modelo 3 utiliza as variáveis: CHOL, SGLU, HDL, AGE e BMI.

Novamente, testes de convergência apresentaram resultados similares, assim como a análise de resíduos.

O resumo do modelo encontra-se na figura 14. Dentre os coeficientes, o b_4 e o b_6 foram os menores, ou seja, que indicam menor influência das variáveis. Os dois encontram-se muito próximos, portanto os dois modelos seguintes foram o modelo 3 sem uma dessas duas variáveis, com o objetivo de avaliar qual é a covariável com menor importância para o modelo.

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.188917	0.002191	1.789e-05	2.797e-05
b[2]	-0.006131	0.001731	1.413e-05	1.937e-05
b[3]	-0.024996	0.001185	9.672e-06	1.694e-05
b[4]	-0.010951	0.002137	1.745e-05	2.804e-05
b[5]	-0.005689	0.002002	1.635e-05	2.096e-05
shape	19.727311	1.403819	1.146e-02	1.458e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.184674	0.187386	0.188928	0.190402	0.193224
b[2]	-0.009509	-0.007314	-0.006138	-0.004951	-0.002756
b[3]	-0.027317	-0.025798	-0.025006	-0.024197	-0.022683
b[4]	-0.015108	-0.012395	-0.010969	-0.009516	-0.006766
b[5]	-0.009530	-0.007036	-0.005713	-0.004343	-0.001798
shape	17.060633	18.757739	19.702469	20.658833	22.530099

Fig. 15. Dados referentes ao resultado do modelo 4

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.188896	0.002188	1.786e-05	2.807e-05
b[2]	-0.007369	0.001699	1.388e-05	1.961e-05
b[3]	-0.024129	0.001214	9.913e-06	1.779e-05
b[4]	0.005887	0.002087	1.704e-05	2.454e-05
b[5]	-0.011279	0.002123	1.733e-05	2.891e-05
shape	19.717538	1.408858	1.150e-02	1.489e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.184661	0.187424	0.188881	0.190354	0.193183
b[2]	-0.010668	-0.008530	-0.007378	-0.006217	-0.004098
b[3]	-0.026481	-0.024953	-0.024145	-0.023310	-0.021754
b[4]	0.001791	0.004498	0.005862	0.007267	0.010079
b[5]	-0.015380	-0.012720	-0.011292	-0.009865	-0.007123
shape	17.027770	18.750305	19.693443	20.649055	22.598942

Fig. 16. Dados referentes ao resultado do modelo 5

1. Empirical mean and standard deviation for each variable, plus standard error of the mean:

	Mean	SD	Naive SE	Time-series SE
b[1]	0.188628	0.002222	1.814e-05	2.662e-05
b[2]	-0.006525	0.001725	1.409e-05	1.922e-05
b[3]	-0.025128	0.001150	9.393e-06	1.570e-05
b[4]	-0.010681	0.002107	1.720e-05	2.833e-05
shape	19.403664	1.409312	1.151e-02	1.525e-02

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
b[1]	0.184278	0.18714	0.188638	0.190104	0.192975
b[2]	-0.009926	-0.00770	-0.006506	-0.005364	-0.003179
b[3]	-0.027331	-0.02591	-0.025142	-0.024366	-0.022838
b[4]	-0.014799	-0.01211	-0.010688	-0.009269	-0.006521
shape	16.792129	18.43696	19.360394	20.334508	22.349513

Fig. 17. Dados referentes ao resultado do modelo 6

A métrica DIC apresenta valor 1251, que é menor do que o do modelo 2, logo, por essa métrica, o modelo 3 é um pouco melhor até então.

4.4 Modelo 4

O modelo 4 foi o modelo 3 sem a variável referente ao b_4 (HDL), ou seja, apresenta: CHOL, SGLU, AGE e BMI.

Os testes citados anteriormente foram performados com sucesso. A métrica DIC foi de 1255, e o resumo do modelo encontra-se na figura 15. O modelo 5 foi analisado antes de se avaliar qual seria o próximo teste.

4.5 Modelo 5

O modelo 5 foi o modelo 3 sem a variável referente ao b_6 (BMI), ou seja, apresenta: CHOL, SGLU, HDL e AGE.

Os testes citados anteriormente foram performados com sucesso. A métrica DIC foi de 1255 também, e o resumo do modelo encontra-se na figura 16. No modelo 4, a variável referente ao coeficiente com menor média é o BMI e, no modelo 5, é o HDL, logo foi testado o modelo sem nenhuma das duas.

4.6 Modelo 6

O modelo 6 apresenta as variáveis: CHOL, SGLU e AGE.

Para passar nos testes de convergência, foi aumentado o período de *burn in* de 1000 para 2000, porque, em duas cadeias, o teste de Heidel tinha falhado para o parâmetro *shape*. O resumo do modelo, com o novo *burn in*, encontra-se na figura 17.

Ao analisar o DIC, encontramos o valor de 1261, o que, dado o DIC dos últimos modelos, pode indicar que retirar mais variáveis não melhoraria mais essa métrica. Por isso, testamos uma última

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7
DIC	1255	1254	1251	1255	1255	1261	1273

Tabela 1. DIC dos modelos testados

combinação de variáveis. Ela foi semelhante à deste modelo, apenas retirando CHOL, que apresentou o coeficiente mais próximo de zero.

4.7 Modelo 7

O modelo 7 apresenta as variáveis: SGLU e AGE.

O modelo passou nos testes de convergência, entretanto apresentou o maior DIC dentre todos, sendo ele 1273. O que reforça que retirar mais atributos não seria mais benéfico neste, portanto concluiu-se a etapa de testes de variáveis para o modelo.

5 CONCLUSÃO

Pela métrica indicada na tabela 1, o modelo 3, que contém as variáveis CHOL, SGLU, HDL, AGE e BMI, apresentou a melhor performance.

Para iniciar, faremos um breve apanhado sobre o que cada uma das substâncias analisadas representa.

-CHOL: basicamente dividido em dois; o colesterol bom (HDL) e o ruim (LDL). O LDL tende a formar placas de forma que podem colar na parede das artérias, estreitando-as, aumentando a chance de um acidente cardiovascular. Enquanto o HDL, também observado no estudo, tem o dever de evitar que esse LDL se aloje nessas paredes. Ou seja, quanto menor o LDL e maior o HDL, melhor.

-GHB: É a forma de medir a quantidade de glicose através de uma análise da hemoglobina. A alta glicose pode promover um estado inflamatório crônico nos vasos, isso predispõe danos nas paredes dos vasos, o que vai gerar formação de placas de aterosclerose, que é grande responsável por entupir o vaso e causar infarto.

-SGLU: É a glicemia em jejum. A diferença entre a GHB e a SGLU é que a GHB mostra o perfil glicêmico dos últimos 120 dias, enquanto a SGLU é mais um retrato de como está a glicemia no momento.

-AGE: A medida que se envelhece, os vasos vão perdendo a flexibilidade, assim, as chances de um infarto ocorrer são maiores.

-As medidas de HHT, WHT, W e H, no fundo, buscam ter dados sobre a obesidade, por isso, optamos por usar o BMI. Pessoas obesas têm maior acúmulo de gordura, o que aumenta a chance de entupimento das artérias. Além disso, o sobrepeso causa mudanças na estrutura e tamanho do coração e exige maior esforço para o bombeamento do sangue.

Esses são os fatores que notamos ter maior relação com as doenças cardiovasculares e diabetes. Agora vamos falar brevemente de outras variáveis visualizadas no estudo.

-Iremos aqui representar SBP e DSP apenas como a Pressão: a maior pressão pode causar danos nas paredes dos vasos, quando o corpo tenta reparar isso, atrai substâncias como o colesterol (CHOL). Além disso, o aumento da idade (AGE), pode causar uma insuficiência cardíaca, ou seja, o coração não ejetaria sangue suficiente para o corpo. Dessa forma, observamos que há bastante presença dos outros dois fatores previamente estudados, o CHOL e a idade (AGE).

6 APÊNDICE

Código do programa em R:

```
# read dataset
data = read.table('Trabalho5_EC.txt', sep = '\t',
    dec = '.', header = TRUE)

# checking some statistics of the data
summary(data)

# there is a small amount of numeric data missing.
     its going to inputed the mean for this values
for(i in 1:ncol(data)) {
  if (is.numeric(data[, i])) {
    data[, i][is.na(data[, i])] <- mean(data[, i],
        na.rm = TRUE)
  }
}
summary(data)

# checking correlation between variables
# first, we want the numerical variables
nums <- unlist(lapply(data, is.numeric))
nums[1] <- FALSE # we don't need to use id for the
     analysis
cor_data = cor(data[, nums])
cor_matrix = as.matrix(cor_data)
# install.packages("corrplot")
library(corrplot)
corrplot(cor_matrix, type = "upper", order = "
    hclust",
         tl.col = "black", tl.srt = 45)
# the most correlated variables are: SGLU and GHB;
     H and WHT; H and W (0.83455575 - biggest one)
    ; SBP and DSP

# boxplots
boxplot(data[nums], use.cols=TRUE)

# bmi calculation (with lbs and inches)
data$BMI = 703 * data$WHT / (data$HHT*data$HHT)
# adding BMI classification
data$BMI_CAT <- cut(data$BMI, breaks = c(0, 18.5,
    25, 30, 100), labels=c("Underweight", "Normal␣
    weight", "Overweight", "Obesity"))

# glycosolated hemoglobin (outcome) > 7.0 is
    usually taken as a positive diagnosis of
    diabetes
data$DIABETES <- cut(data$GHB, breaks = c(0,7,50),
     labels = c("Negative", "Positive"))

# ages' bin
```

```
data$AGE_CAT <- cut(data$AGE, breaks = c
    (17,25,35,45,55,65,200), labels=c("18-25","
    26-35","36-45","46-55","56-65",">=66"))

# summary with the created columns added
summary(data)

# chi square tests for a preliminar analysis
# diabetes x bmi_cat
test_bmi_cat = table(data$DIABETES, data$BMI_CAT)
test_bmi_cat
chisq.test(test_bmi_cat) # may indicate that
    higher BMI is related to diabetes
# fisher.test(test_bmi_cat)

# diabetes x age
test_age_cat = table(data$DIABETES, data$AGE_CAT)
test_age_cat
chisq.test(test_age_cat)

# diabetes x location
test_location = table(data$DIABETES, data$LOCATION
    )
test_location
chisq.test(test_location)

# diabetes x gender
test_gender = table(data$DIABETES, data$GENDER)
test_gender
chisq.test(test_gender)

# diabetes x frame
test_frame = table(data$DIABETES, data$FRAME)
test_frame
chisq.test(test_frame)

# some scatter plots with colors to differentiate
    HDL > 7 from HDL < 7
nums_ext <- unlist(lapply(data, is.numeric))
nums_ext[1] <- FALSE # we don't need to use id for
     the analysis
diabetes_color <- c("#E7B800", "#FC4E07")
pairs(data[, nums_ext], col = diabetes_color[data$
    DIABETES], lower.panel=NULL) # we can see that
     high values of SGLU may indicate higher HDL.
    It makes sense, because Glycosylated
    hemoglobin is the Hemoglobin to which glucose
    is bound
# zoomed chart of SGLU x GHB
plot(data$SGLU, data$GHB, main="Glycosolated␣
    Hemoglobin␣x␣␣Stabilized␣Glucose", xlab = "
    SGLU", ylab = "GHB")
# zoomed chart of CHOL x GHB
plot(data$CHOL, data$GHB, main="Glycosolated␣
    Hemoglobin␣x␣␣Total␣Cholesterol", xlab = "CHOL
    ", ylab = "GHB")

# zoomed chart of HDL x GHB
plot(data$HDL, data$GHB, main="Glycosolated␣
    Hemoglobin␣x␣␣High␣Density␣Lipoprotein", xlab
    = "HDL", ylab = "GHB")

# histogram
hist(data$GHB, xlab = "GHB", main = "Histogram␣of␣
    GHB", freq = FALSE, breaks = 50)
library(MASS)
fit.params <- fitdistr(data$GHB, "gamma", lower =
    c(0, 0))
curve(dgamma(x=x, shape=fit.params$estimate['shape
    '], rate=fit.params$estimate['rate']), col="
    blue", add=TRUE)
fit.params <- fitdistr(data$GHB, "weibull", lower
    = c(0, 0))
curve(dweibull(x=x, shape=fit.params$estimate['
    shape'], scale=fit.params$estimate['scale']),
    col="green", add=TRUE)

# Generalized linear model (without using prior)
glm_ghb <- glm(GHB ~ CHOL + SGLU + HDL + AGE + SBP
     + DSP + BMI + W + H, data=data, family=Gamma(
    link="inverse"))
summary(glm_ghb)
# removing variables with p-value > 0.05
glm_ghb_2 <- glm(GHB ~ CHOL + SGLU + HDL + AGE,
    data=data, family=Gamma(link="inverse"))
summary(glm_ghb_2)

confint(glm_ghb_2)

# anova test removing a variable that is important
     to check the result
glm_ghb_3 <- glm(GHB ~ SGLU + HDL + AGE + SBP +
    DSP + BMI + W + H, data=data, family=Gamma(
    link="inverse"))
anova(glm_ghb, glm_ghb_3, test = "Chisq")

# glm with prior
library("rjags")
set.seed(5)
# note: dgamma in JAGS = dgamma(shape, rate)
# model 1
mod1_string = "␣model␣{
␣␣␣␣for␣(i␣in␣1:length(y))␣{
␣␣␣␣␣␣y[i]␣~␣dgamma(shape,␣shape␣*␣inv_mu[i])
␣␣␣␣␣␣inv_mu[i]␣=␣(b[1]␣+␣b[2]*CHOL[i]␣+␣b[3]*SGLU
    [i]␣+␣b[4]*HDL[i]␣+␣b[5]*AGE[i]␣+␣b[6]*SBP[i]␣
    +␣b[7]*DSP[i]␣+␣b[8]*BMI[i])
␣␣␣␣}
␣␣␣␣b[1]␣~␣dnorm(0.5,␣1.0/1.0e4)
␣␣␣␣for␣(j␣in␣2:8)␣{
␣␣␣␣␣␣b[j]␣~␣dnorm(0.0,␣1.0/1.0e4)
␣␣␣␣}
␣␣␣␣shape␣~␣dgamma(0.001,␣0.001)
```

```
}␣"
params = c("b", "shape")

# normalization
library(dplyr)
columns_models <- select(data, "CHOL", "SGLU", "
    HDL", "AGE", "SBP", "DSP", "BMI")
scaled_data <- scale(columns_models)

data_jags = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], HDL=scaled_
    data[,"HDL"], AGE=scaled_data[,"AGE"], SBP=
    scaled_data[,"SBP"], DSP=scaled_data[,"DSP"],
    BMI=scaled_data[,"BMI"])
# data_jags = list(y=data$GHB, CHOL=data[,"CHOL"],
     SGLU=data[,"SGLU"], HDL=data[,"HDL"], AGE=
    data[,"AGE"], SBP=data[,"SBP"], DSP=data[,"DSP
    "], BMI=data[,"BMI"])

mod1 = jags.model(textConnection(mod1_string),
    data=data_jags, n.chains=3)
update(mod1, 1e3)

mod1_sim = coda.samples(model=mod1,
                        variable.names=params,
                        n.iter=5e3)
mod1_csim = as.mcmc(do.call(rbind, mod1_sim))

# convergence
plot(mod1_sim)
gelman.diag(mod1_sim)
heidel.diag(mod1_sim)

# residuals
X = cbind(rep(1.0, nrow(data)), scaled_data[,"CHOL
    "], scaled_data[,"SGLU"], scaled_data[,"HDL"],
     scaled_data[,"AGE"], scaled_data[,"SBP"],
    scaled_data[,"DSP"], scaled_data[,"BMI"])
(pm_params1 = colMeans(mod1_csim)) #posterior mean

# vector of predicted values from the model
yhat1 <- (drop(X %*% pm_params1[1:8]))^(-1)
# calculate residuals
resid1 <- data_jags$y - yhat1

plot(resid1)

# summary
summary(mod1_sim)

# DIC
dic.samples(mod1, n.iter = 1e3)

# model 2
mod2_string = "␣model␣{
␣␣␣␣for␣(i␣in␣1:length(y))␣{
␣␣␣␣␣␣y[i]␣~␣dgamma(shape,␣shape␣*␣inv_mu[i])
␣␣␣␣␣␣inv_mu[i]␣=␣(b[1]␣+␣b[2]*CHOL[i]␣+␣b[3]*SGLU
    [i]␣+␣b[4]*HDL[i]␣+␣b[5]*AGE[i]␣+␣b[6]*DSP[i]␣
    +␣b[7]*BMI[i])
␣␣␣␣}
␣␣␣␣b[1]␣~␣dnorm(0.5,␣1.0/1.0e4)
␣␣␣␣for␣(j␣in␣2:7)␣{
␣␣␣␣␣␣b[j]␣~␣dnorm(0.0,␣1.0/1.0e4)
␣␣␣␣}
␣␣␣␣shape␣~␣dgamma(0.001,␣0.001)
}␣"
data_jags_2 = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], HDL=scaled_
    data[,"HDL"], AGE=scaled_data[,"AGE"], DSP=
    scaled_data[,"DSP"], BMI=scaled_data[,"BMI"])

mod2 = jags.model(textConnection(mod2_string),
    data=data_jags_2, n.chains=3)
update(mod2, 1e3)

mod2_sim = coda.samples(model=mod2,
                        variable.names=params,
                        n.iter=5e3)
mod2_csim = as.mcmc(do.call(rbind, mod2_sim))

# convergence
plot(mod2_sim)
gelman.diag(mod2_sim)
heidel.diag(mod2_sim)

# residuals
X2 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    CHOL"], scaled_data[,"SGLU"], scaled_data[,"
    HDL"], scaled_data[,"AGE"], scaled_data[,"DSP"
    ], scaled_data[,"BMI"])
(pm_params2 = colMeans(mod2_csim)) #posterior mean

# vector of predicted values from the model
yhat2 <- (drop(X2 %*% pm_params2[1:7]))^(-1)
# calculate residuals
resid2 <- data_jags_2$y - yhat2

plot(resid2)

# summary
summary(mod2_sim)

# DIC
dic.samples(mod2, n.iter = 1e3)

# model 3
mod3_string = "␣model␣{
␣␣␣␣for␣(i␣in␣1:length(y))␣{
␣␣␣␣␣␣y[i]␣~␣dgamma(shape,␣shape␣*␣inv_mu[i])
␣␣␣␣␣␣inv_mu[i]␣=␣(b[1]␣+␣b[2]*CHOL[i]␣+␣b[3]*SGLU
    [i]␣+␣b[4]*HDL[i]␣+␣b[5]*AGE[i]␣+␣b[6]*BMI[i])
```

```
    }
    b[1] ~ dnorm(0.5, 1.0/1.0e4)
    for (j in 2:6) {
      b[j] ~ dnorm(0.0, 1.0/1.0e4)
    }
    shape ~ dgamma(0.001, 0.001)
}"
data_jags_3 = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], HDL=scaled_
    data[,"HDL"], AGE=scaled_data[,"AGE"], BMI=
    scaled_data[,"BMI"])

mod3 = jags.model(textConnection(mod3_string),
    data=data_jags_3, n.chains=3)
update(mod3, 1e3)

mod3_sim = coda.samples(model=mod3,
                        variable.names=params,
                        n.iter=5e3)
mod3_csim = as.mcmc(do.call(rbind, mod3_sim))

# convergence
plot(mod3_sim)
gelman.diag(mod3_sim)
heidel.diag(mod3_sim)

# residuals
X3 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    CHOL"], scaled_data[,"SGLU"], scaled_data[,"
    HDL"], scaled_data[,"AGE"], scaled_data[,"BMI"
    ])
(pm_params3 = colMeans(mod3_csim)) #posterior mean

# vector of predicted values from the model
yhat3 <- (drop(X3 %*% pm_params3[1:6]))^(-1)
# calculate residuals
resid3 <- data_jags_3$y - yhat3

plot(resid3)

# summary
summary(mod3_sim)

# DIC
dic.samples(mod3, n.iter = 1e3)

# model 4
mod4_string = " model {
    for (i in 1:length(y)) {
      y[i] ~ dgamma(shape, shape * inv_mu[i])
      inv_mu[i] = (b[1] + b[2]*CHOL[i] + b[3]*SGLU
    [i] + b[4]*AGE[i] + b[5]*BMI[i])
    }
    b[1] ~ dnorm(0.5, 1.0/1.0e4)
    for (j in 2:5) {
      b[j] ~ dnorm(0.0, 1.0/1.0e4)
    }
    shape ~ dgamma(0.001, 0.001)
}"
data_jags_4 = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], AGE=scaled_
    data[,"AGE"], BMI=scaled_data[,"BMI"])

mod4 = jags.model(textConnection(mod4_string),
    data=data_jags_4, n.chains=3)
update(mod4, 1e3)

mod4_sim = coda.samples(model=mod4,
                        variable.names=params,
                        n.iter=5e3)
mod4_csim = as.mcmc(do.call(rbind, mod4_sim))

# convergence
plot(mod4_sim)
gelman.diag(mod4_sim)
heidel.diag(mod4_sim)

# residuals
X4 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    CHOL"], scaled_data[,"SGLU"], scaled_data[,"
    AGE"], scaled_data[,"BMI"])
(pm_params4 = colMeans(mod4_csim)) #posterior mean

# vector of predicted values from the model
yhat4 <- (drop(X4 %*% pm_params4[1:5]))^(-1)
# calculate residuals
resid4 <- data_jags_4$y - yhat4

plot(resid4)

# summary
summary(mod4_sim)

# DIC
dic.samples(mod4, n.iter = 1e3)

# model 5
mod5_string = " model {
    for (i in 1:length(y)) {
      y[i] ~ dgamma(shape, shape * inv_mu[i])
      inv_mu[i] = (b[1] + b[2]*CHOL[i] + b[3]*SGLU
    [i] + b[4]*HDL[i] + b[5]*AGE[i])
    }
    b[1] ~ dnorm(0.5, 1.0/1.0e4)
    for (j in 2:5) {
      b[j] ~ dnorm(0.0, 1.0/1.0e4)
    }
    shape ~ dgamma(0.001, 0.001)
}"
data_jags_5 = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], HDL=scaled_
    data[,"HDL"], AGE=scaled_data[,"AGE"])
```

```
mod5 = jags.model(textConnection(mod5_string),
    data=data_jags_5, n.chains=3)
update(mod5, 1e3)

mod5_sim = coda.samples(model=mod5,
                        variable.names=params,
                        n.iter=5e3)
mod5_csim = as.mcmc(do.call(rbind, mod5_sim))

# convergence
plot(mod5_sim)
gelman.diag(mod5_sim)
heidel.diag(mod5_sim)

# residuals
X5 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    CHOL"], scaled_data[,"SGLU"], scaled_data[,"
    HDL"], scaled_data[,"AGE"])
(pm_params5 = colMeans(mod5_csim)) #posterior mean

# vector of predicted values from the model
yhat5 <- (drop(X5 %*% pm_params5[1:5]))^(-1)
# calculate residuals
resid5 <- data_jags_5$y - yhat5

plot(resid5)

# summary
summary(mod5_sim)

# DIC
dic.samples(mod5, n.iter = 1e3)

# model 6
mod6_string = " model {
    for (i in 1:length(y)) {
        y[i] ~ dgamma(shape, shape * inv_mu[i])
        inv_mu[i] = (b[1] + b[2]*CHOL[i] + b[3]*SGLU
    [i] + b[4]*AGE[i])
    }
    b[1] ~ dnorm(0.5, 1.0/1.0e4)
    for (j in 2:4) {
        b[j] ~ dnorm(0.0, 1.0/1.0e4)
    }
    shape ~ dgamma(0.001, 0.001)
} "
data_jags_6 = list(y=data$GHB, CHOL=scaled_data[,"
    CHOL"], SGLU=scaled_data[,"SGLU"], AGE=scaled_
    data[,"AGE"])

mod6 = jags.model(textConnection(mod6_string),
    data=data_jags_6, n.chains=3)
update(mod6, 2e3)

mod6_sim = coda.samples(model=mod6,
                        variable.names=params,
                        n.iter=5e3)
mod6_csim = as.mcmc(do.call(rbind, mod6_sim))

# convergence
plot(mod6_sim)
gelman.diag(mod6_sim)
heidel.diag(mod6_sim)

# residuals
X6 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    CHOL"], scaled_data[,"SGLU"], scaled_data[,"
    AGE"])
(pm_params6 = colMeans(mod6_csim)) #posterior mean

# vector of predicted values from the model
yhat6 <- (drop(X6 %*% pm_params6[1:4]))^(-1)
# calculate residuals
resid6 <- data_jags_6$y - yhat6

plot(resid6)

# summary
summary(mod6_sim)

# DIC
dic.samples(mod6, n.iter = 1e3)

# model 7
mod7_string = " model {
    for (i in 1:length(y)) {
        y[i] ~ dgamma(shape, shape * inv_mu[i])
        inv_mu[i] = (b[1] + b[2]*SGLU[i] + b[3]*AGE[
    i])
    }
    b[1] ~ dnorm(0.5, 1.0/1.0e4)
    for (j in 2:3) {
        b[j] ~ dnorm(0.0, 1.0/1.0e4)
    }
    shape ~ dgamma(0.001, 0.001)
} "
data_jags_7 = list(y=data$GHB, SGLU=scaled_data[,"
    SGLU"], AGE=scaled_data[,"AGE"])

mod7 = jags.model(textConnection(mod7_string),
    data=data_jags_7, n.chains=3)
update(mod7, 1e3)

mod7_sim = coda.samples(model=mod7,
                        variable.names=params,
                        n.iter=5e3)
mod7_csim = as.mcmc(do.call(rbind, mod7_sim))

# convergence
plot(mod7_sim)
gelman.diag(mod7_sim)
```

```
heidel.diag(mod7_sim)

# residuals
X7 = cbind(rep(1.0, nrow(data)), scaled_data[,"
    SGLU"], scaled_data[,"AGE"])
(pm_params7 = colMeans(mod7_csim)) #posterior mean

# vector of predicted values from the model
yhat7 <- (drop(X7 %*% pm_params7[1:3]))^(-1)
# calculate residuals
resid7 <- data_jags_7$y - yhat7

plot(resid7)

# summary
summary(mod7_sim)

# DIC
dic.samples(mod7, n.iter = 1e3)
```