Fatores de risco para doenças cardíacas

Beatriz Rodrigues Pinna

30 de Março de 2021

Introdução

Doenças cardiovasculares são consistentemente as maiores causas de morte no mundo desde a segunda metade do século XX e não apresentam qualquer sinal de que deixarão de ocupar essa posição no futuro próximo.

A insuficiência cardíaca é a *causa mortis* mais usual em decorrência de cardiopatias. Dado o alto custo e risco de vida, a política de saúde pública tem se voltado para a detecção precoce e o controle dos principais fatores comportamentais e de riscos relacionados a mortalidade por insuficiência cardíaca.

O objetivo do trabalho é estudar a ocorrência de falecimentos em pessoas com doenças cardiovasculares ou que apresentam alto risco cardiovascular. Tal estudo é interessante, pois identificando os fatores de risco, torna-se possível auxiliar o juízo médico no prognóstico dos pacientes.

Neste trabalho, o conjunto de dados utilizados foi o *Heart Failure Prediction* extraído do site Kaggle. Os dados foram coletados em 2015, entre Abril e Dezembro, no Hospital Aliado em Faisalabad (Paquistão) de pacientes admitidos por disfunção sistólica do ventrículo esquerdo.

Descrição das Variáveis

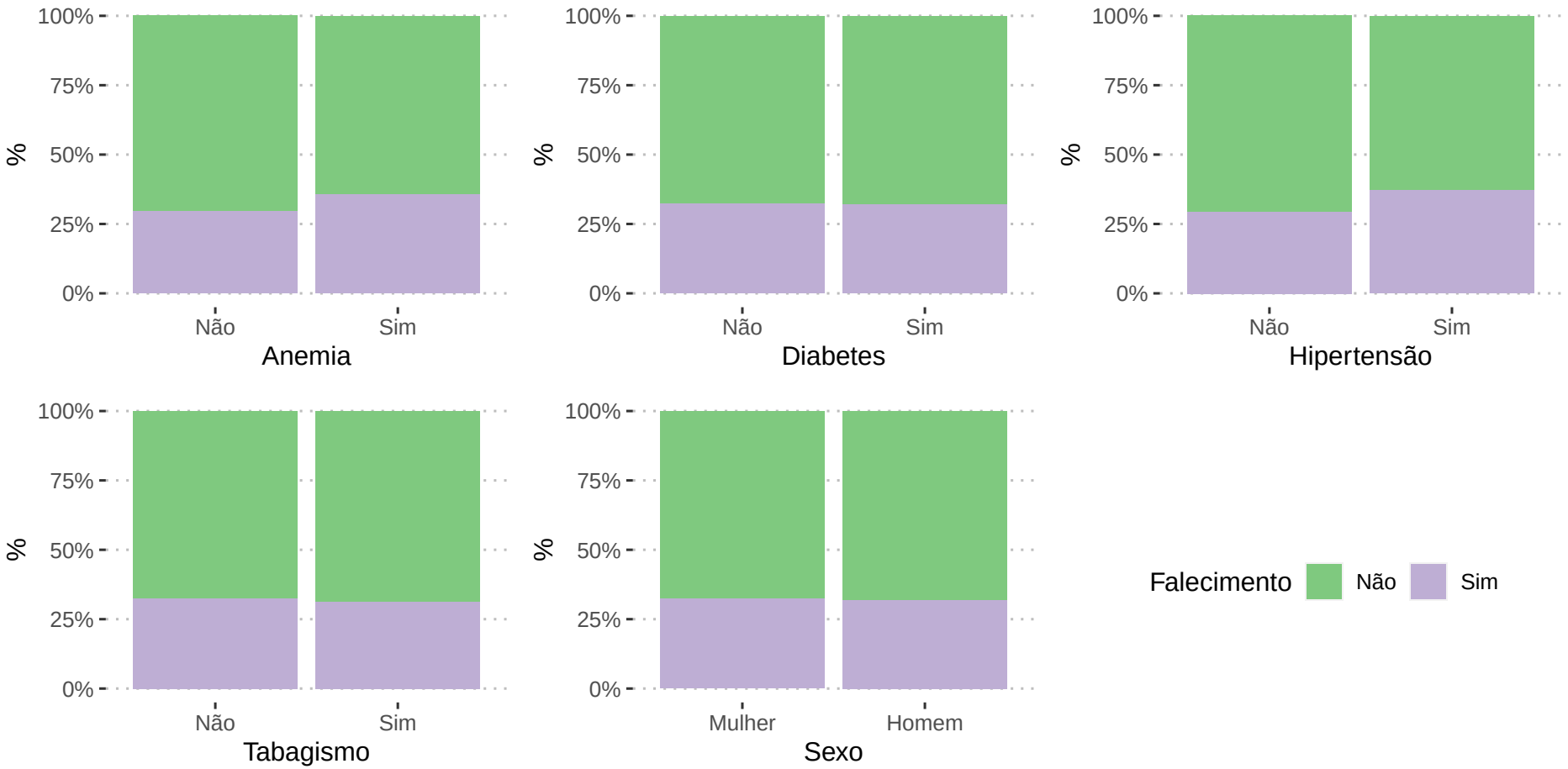
O conjunto de dados possui 299 observações e 13 variáveis:

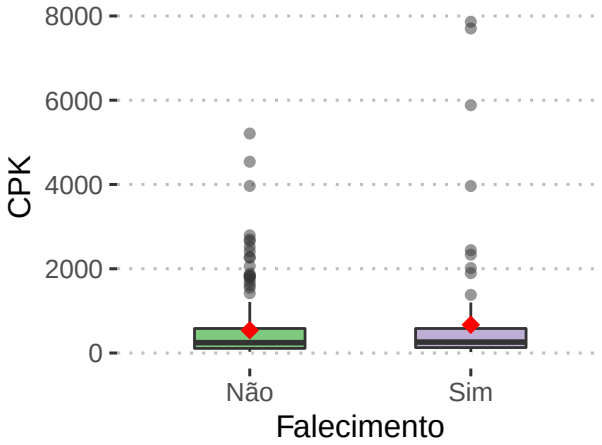
Variável	Descrição	Tipo
Idade	Idade em anos no tempo de admissão	Contínua
Anemia	Diminuição do número de hemácias ou hemoglobina	Binária
CPK	Nível sanguíneo de creatinofosfoquinase em mcg/L	Contínua
Diabetes	Níveis elevados de glicose no sangue por um longo intervalode tempo	Binária
FEVE	Porcentagem média do sangue coletado no fim do enchimento diastólico após ejeção do ventrículo esquerdo	Contínua
Hipertensão	Pressão arterial constantemente elevada	Binária
Plaquetas	Nível sanguíneo de plaquetas em quiloplaquetas/mL	Contínua
Creatinina	Nível sanguíneo de creatinina em mg/dL	Contínua
Sódio	Nível sanguíneo de sódio em mEq/L	Contínua
Sexo	Sexo biológico	Binária
Tabagismo	Consumo consistente de tabaco	Binária
Tempo	Tempo sob observação em dias	Contínua
Falecimento	Falecimento durante o tempo sob observação	Binária

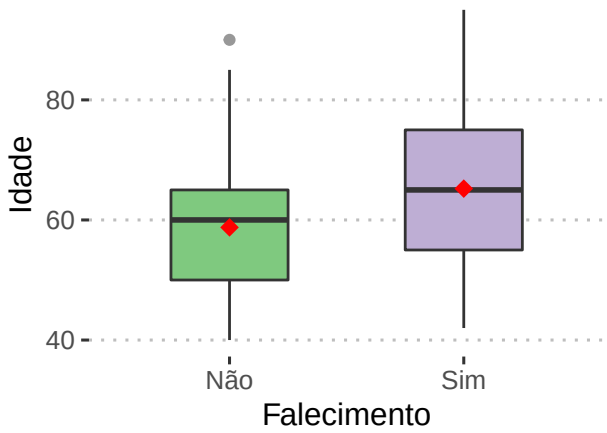
Análise Exploratória

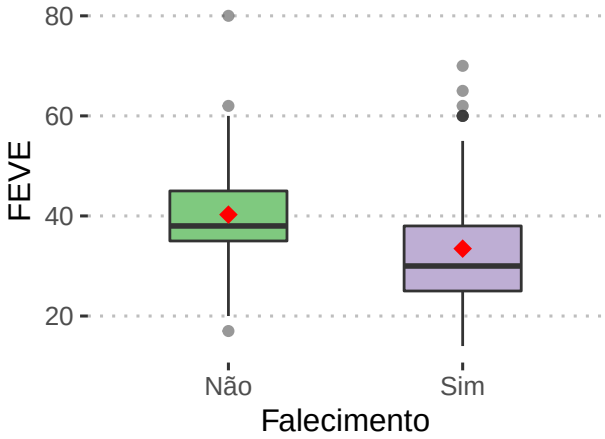
Variável resposta: **Falecimento**

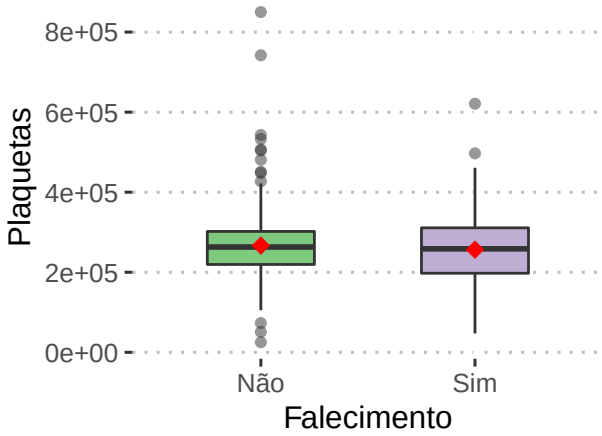
Falecimento	Frequência	Proporção
Não	203	67.89
Sim	96	32.11

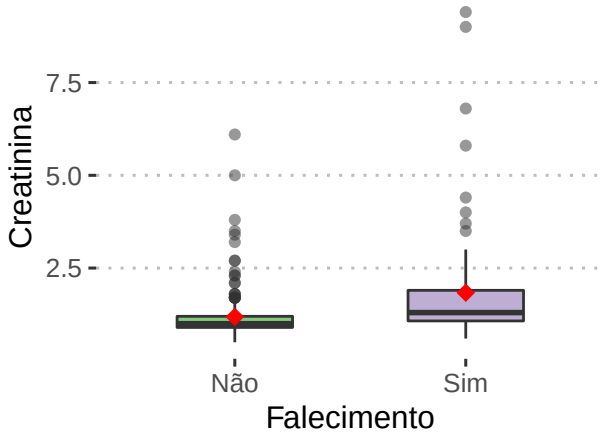

100%
75%
50%
25%
0%
%
Não
Sim
Anemia
Diabetes
Hipertensão
Tabagismo
Mulher
Homem
Sexo
Falecimento
Não
Sim

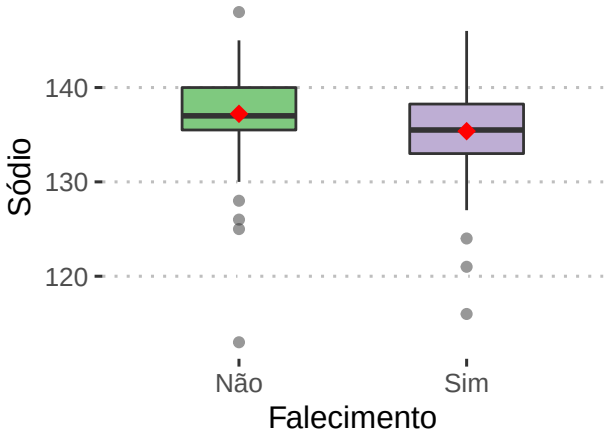

8000
6000
4000
2000
0
CPK
Não
Sim
Falecimento


80
60
40
Idade
Não
Sim
Falecimento


80
60
40
20
FEVE
Não
Sim
Falecimento


8e+05
6e+05
4e+05
2e+05
0e+00
Plaquetas
Não
Sim
Falecimento


7.5
5.0
2.5
Creatinina
Não
Sim
Falecimento


140
130
120
Sódio
Não
Sim
Falecimento

Regressão Logística

Como a variável resposta desse estudo é o falecimento ou não dos pacientes deste hospital, ou seja, uma variável binária, optou-se pela escolha de um modelo Bernoulli que modela o fracasso ou sucesso de um evento. Em nosso caso, o sucesso é a morte do paciente e o fracasso, consequentemente, o não falecimento do paciente

Os dados foram separados em uma base de treino e outra de teste/validação. Desse modo, foi possível obter as métricas relativas a acurácia do modelo proposto. Dito isso, optou-se por classificar 70% dos dados como treino e os 30% restantes como teste/validação.

Falecimento	Treino	Teste/Validação
Não	0.678	0.682
Sim	0.322	0.318

Resultados dos modelos ajustados

Para selecionar as variáveis que irão compor o modelo de regressão com as funções de ligação probit, logit e log-log complementar (C log-log) foi utilizado o algoritmo *Stepwise* considerando como critério de seleção o AIC.

	Dependent variable:		
	Falecimento		
	logistic	*probit*	*glm: binomial link = cloglog*
	(1)	(2)	(3)
Idade	0.075^{***} (0.043, 0.107)	0.044^{***} (0.026, 0.061)	0.057^{***} (0.035, 0.080)
CPK	0.0004^{**} (0.00005, 0.001)	0.0002^{**} (0.00003, 0.0004)	0.0003^{***} (0.0001, 0.0005)
FEVE	-0.080^{***} (-0.116, -0.044)	-0.043^{***} (-0.062, -0.023)	-0.062^{***} (-0.088, -0.035)
HipertensãoSim			0.493^{*} (-0.026, 1.013)
Creatinina	0.517^{***} (0.198, 0.836)	0.291^{***} (0.103, 0.479)	0.364^{***} (0.145, 0.583)
SexoHomem	-0.641^{*} (-1.363, 0.080)	-0.358^{*} (-0.777, 0.061)	
Constant	-3.124^{***} (-5.237, -1.011)	-1.944^{***} (-3.166, -0.723)	-3.291^{***} (-4.893, -1.689)
Observations	211	211	211
Akaike Inf. Crit.	218.811	220.212	218.031
Note:			*p<0.1; **p<0.05; ***p<0.01

Comparação das Ligações

As funções de ligação logit, probit e log-log complementar são as principais funções usadas para dados binários, que garantem que as probabilidades estimadas fiquem entre 0 e 1.

Tipo de Ligação	Estatística Deviance	Estatística χ^2 de Pearson	df	χ^2_{N-p}
Logit	206.81	237.39	205	239.4
Probit	208.21	232.21	205	239.4
C log-log	206.03	224.47	205	239.4

Teste Hosmer-Lemeshow

O teste de Hosmer-Lemeshow tem como objetivo atestar a qualidade de ajuste do modelo, ou seja, o teste comprova se o modelo obtido explica adequadamente os dados observados da variável resposta.

O teste de Hosmer-Lemeshow também indicou que os modelos obtidos explicam adequadamente os dados observados, ao nível de 5% de significância.

Tipo de Ligação	Estatística	df	p-valor
Logit	13.699	8	0.090
Probit	15.000	8	0.059
C log-log	7.842	8	0.449

Interpretação dos Parâmetros

Optou-se por utilizar a função de ligação logit porque mostrou ser mais adequada ao problema proposto inicialmente e pela maior facilidade de interpretação dos resultados.

O preditor linear do modelo ajustado é dado pela expressão:

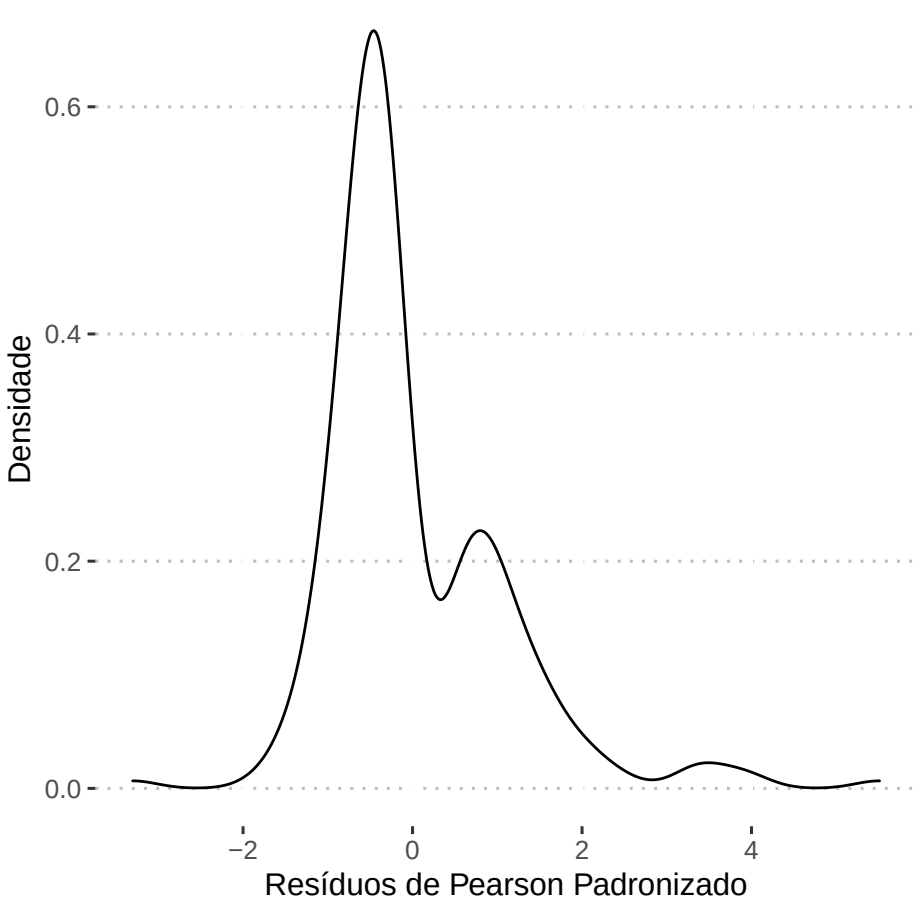
$$\text{logit}(\hat{p}_i) = ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -3{,}124 + 0{,}075\text{Idade} + 0{,}0004\text{CPK} - 0{,}08\text{FEVE} + 0{,}517\text{Creatinina} - 0{,}641\text{Sexo}$$

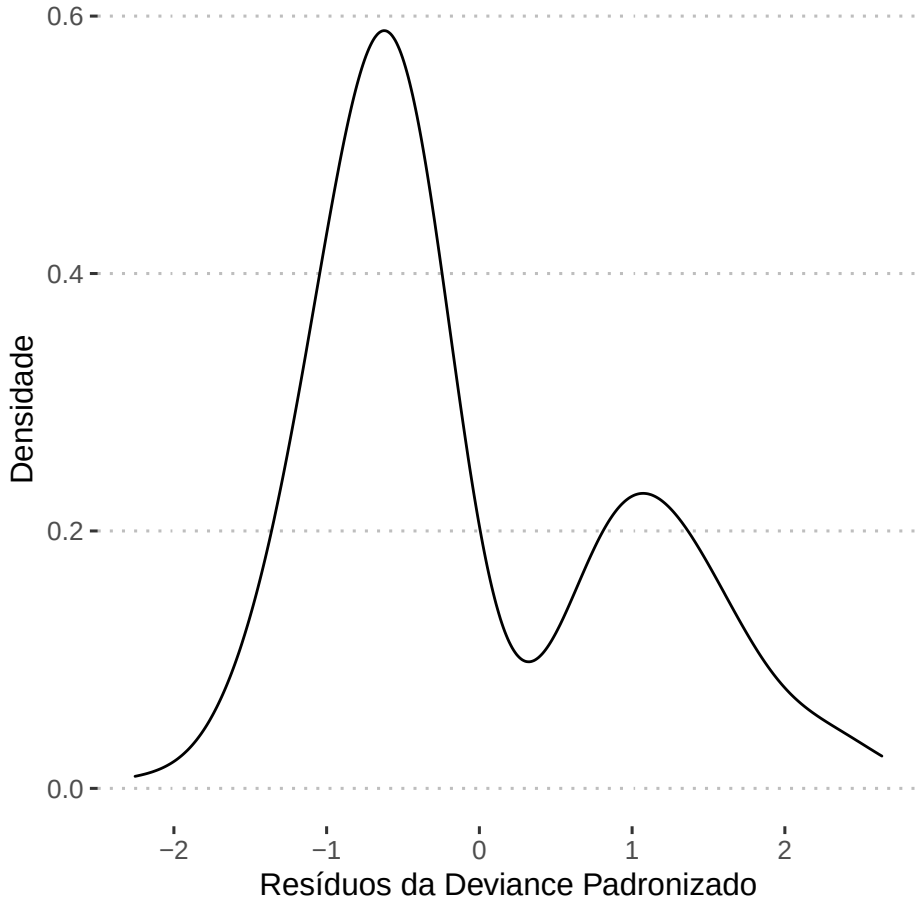
Em regressão logística, podemos interpretar os parâmetros estimados do modelo através da Razão de chances (*Odds ratio*).

		IC de 95%	
Variável	Razão de Chance	Lim inf	Lim sup
(Intercept)	0.044	0.005	0.348
Idade	1.078	1.046	1.115
CPK	1.000	1.000	1.001
FEVE	0.923	0.888	0.955
Creatinina	1.676	1.223	2.362
SexoHomem	0.527	0.254	1.081

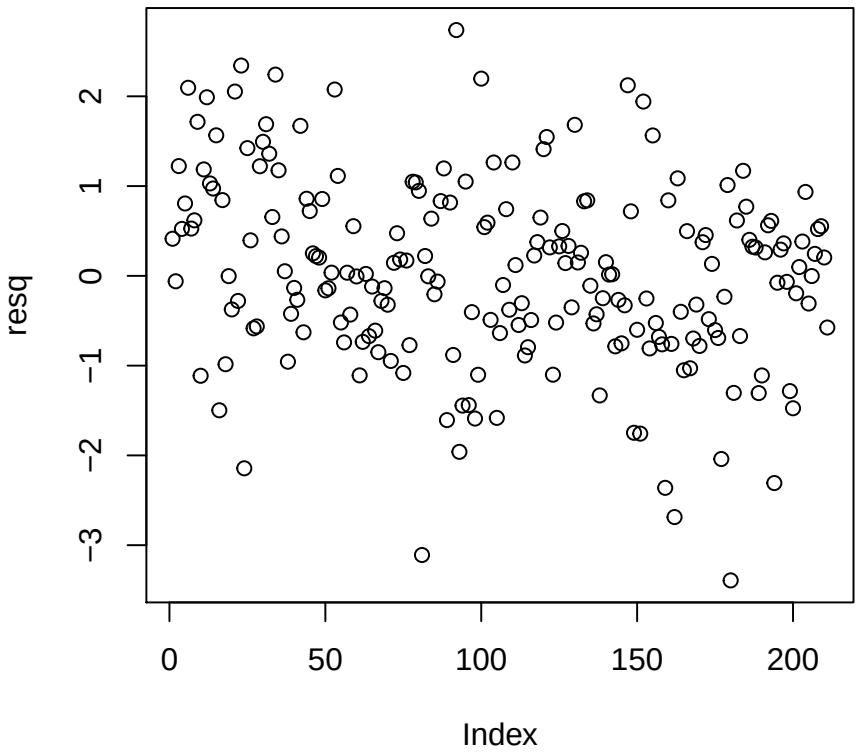
Resíduos

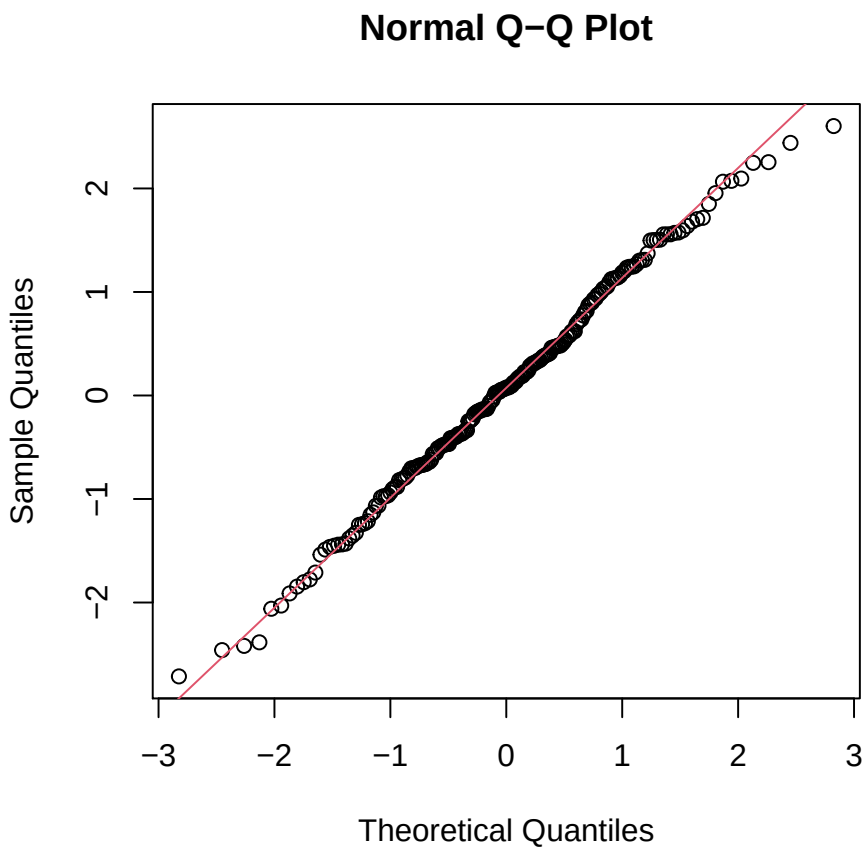
A análise de resíduos é uma importante etapa do ajuste dos modelos para dados binários.



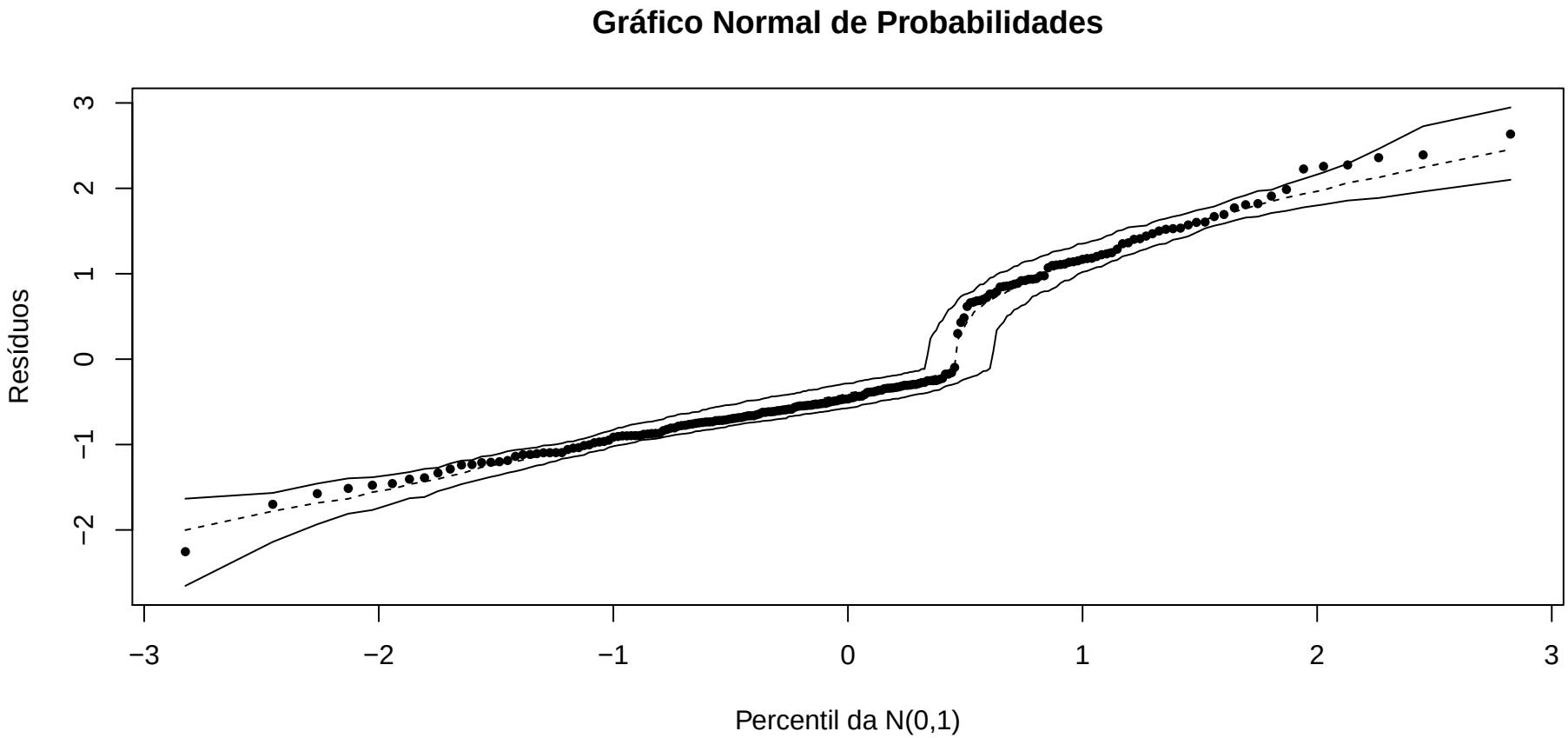


Outra alternativa é avaliar a qualidade do ajuste com base nos resíduos quantílicos aleatorizados.



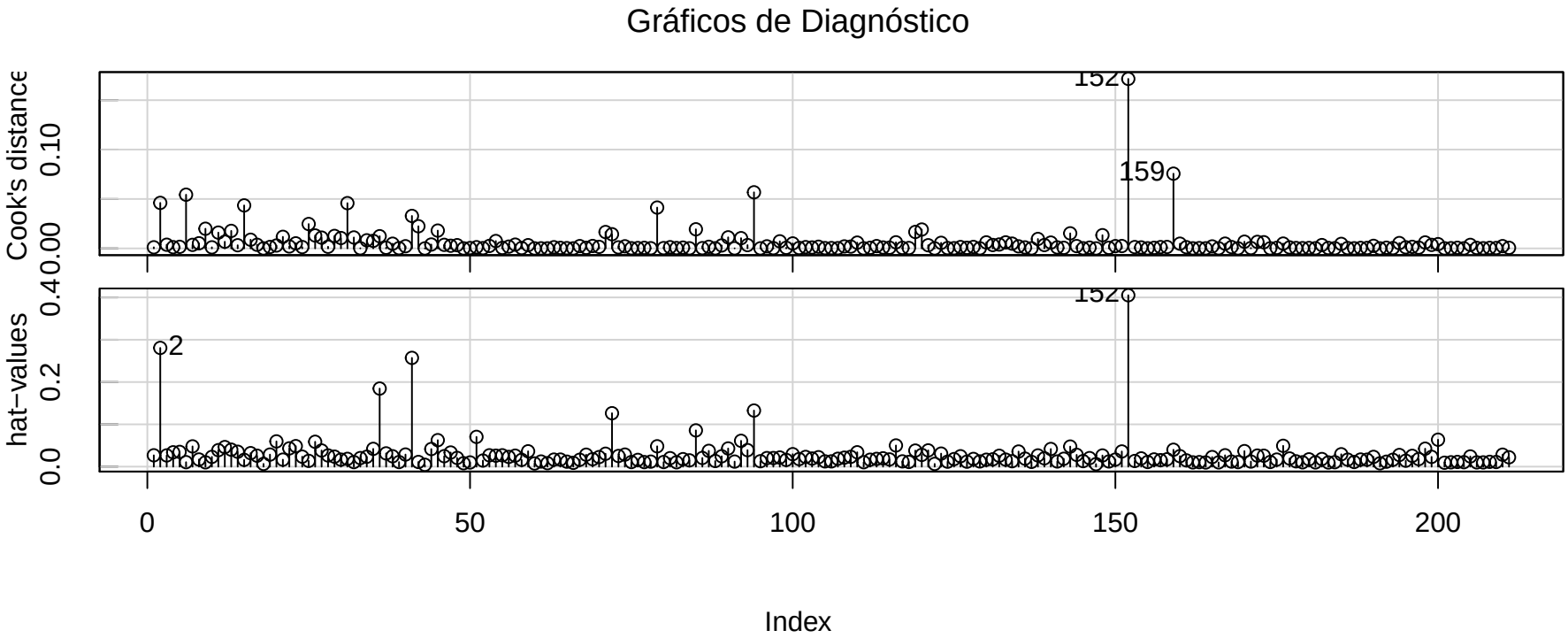


O gráfico de resíduos simulados permite verificar a adequação do modelo ajustado mesmo que os resíduos não tenham uma aproximação adequada com a distribuição Normal.



Diagnóstico de dados influentes

A distância de Cook considera a influência da i-ésima observação em todos os valores ajustados. Já a matriz H estimada para MLGs pode ser usada para avaliar a leverage (alavancagem) para cada observação.

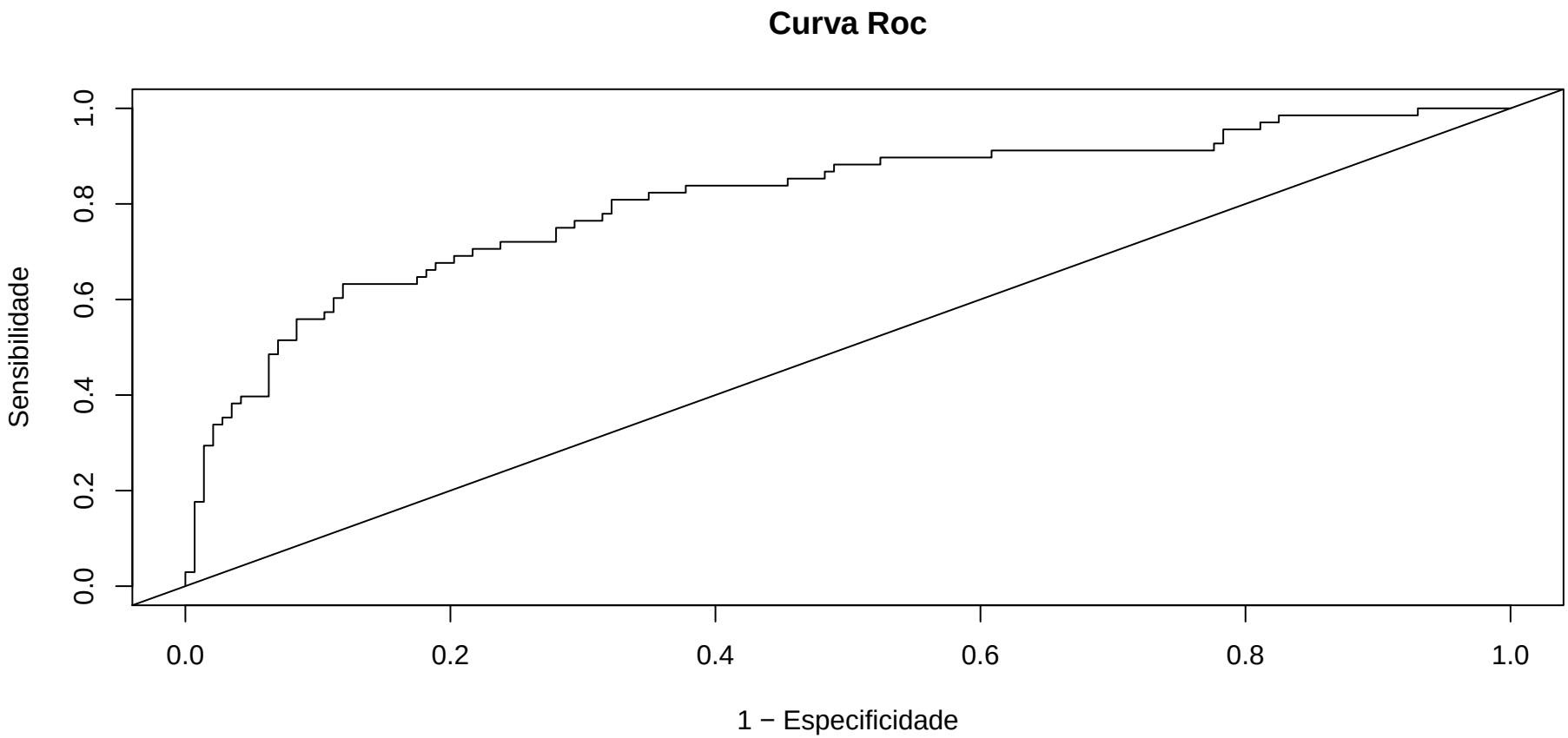


Desempenho do modelo

A curva ROC (Receiver Operating Characteristic) está entre as métricas mais utilizadas para avaliação de um modelo de classificação e mostra o quão bem um modelo criado pode diferenciar duas classificações, no caso deste trabalho, distinguir o falecimento ou não dos pacientes do hospital no Paquistão.

Para o modelo desenvolvido neste trabalho obteve-se um AUC de 0,8099, tal valor é considerado bom para um modelo de classificação. Pode-se dizer então, que o modelo logístico acerta corretamente 81% das predições feitas com a base de dados de validação.

Quando o valor predito for maior que 0,46, considera-se a predição de falecimento do paciente e, caso contrário, a predição é de não falecimento.



O poder preditivo do modelo logístico estimado foi avaliado por meio de algumas métricas de desempenho que são baseadas na Matriz de confusão.

		Valor Observado	
Valor Estimado		Não	Sim
	Não	53	17
	Sim	7	11

A acurácia das predições realizadas é de 72,73%, um valor considerado aceitável para o modelo preditivo estimado neste trabalho, visto que é difícil prever eventos deste tipo e que a amostra utilizada para teste é pequena.

Métrica	Resultado (%)
Acurácia	72.73
Sensibilidade	39.29
Especificidade	88.33

Modelo Bayesiano

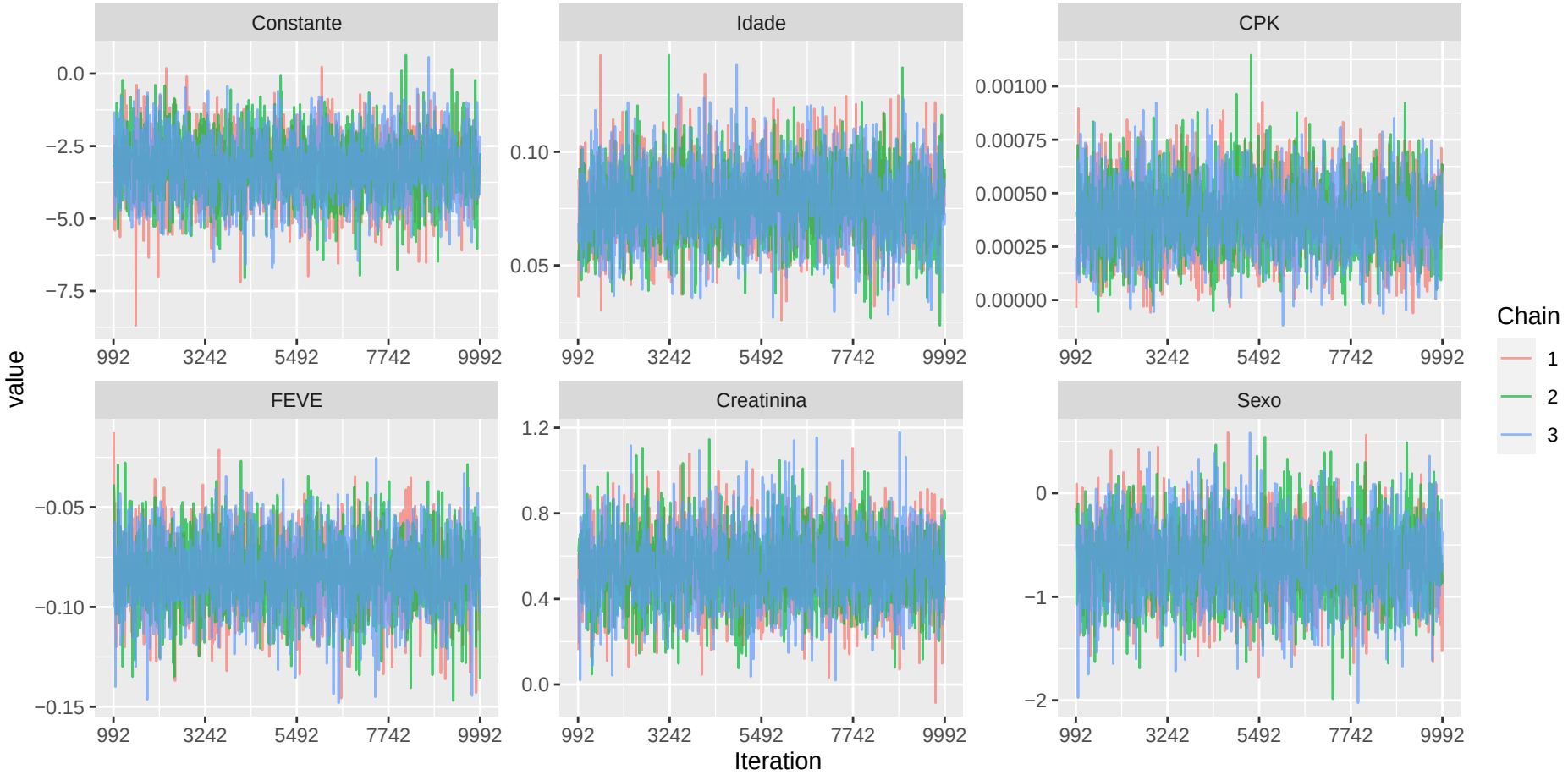
Ajustou-se o modelo logístico bayesiano para este dados, onde a estimação dos parâmetros foi feita via inferência bayesiana por métodos MCMC utilizando o software R com o pacote `R2jags`.

Além disso, sob o enfoque bayesiano, é necessário elicitar a priori para o vetor paramétrico β. Como não há nenhum conhecimento prévio sobre a influência das covariáveis, considerou-se para cada parâmetro β_k a priori não informativa $N(0, 1000)$.

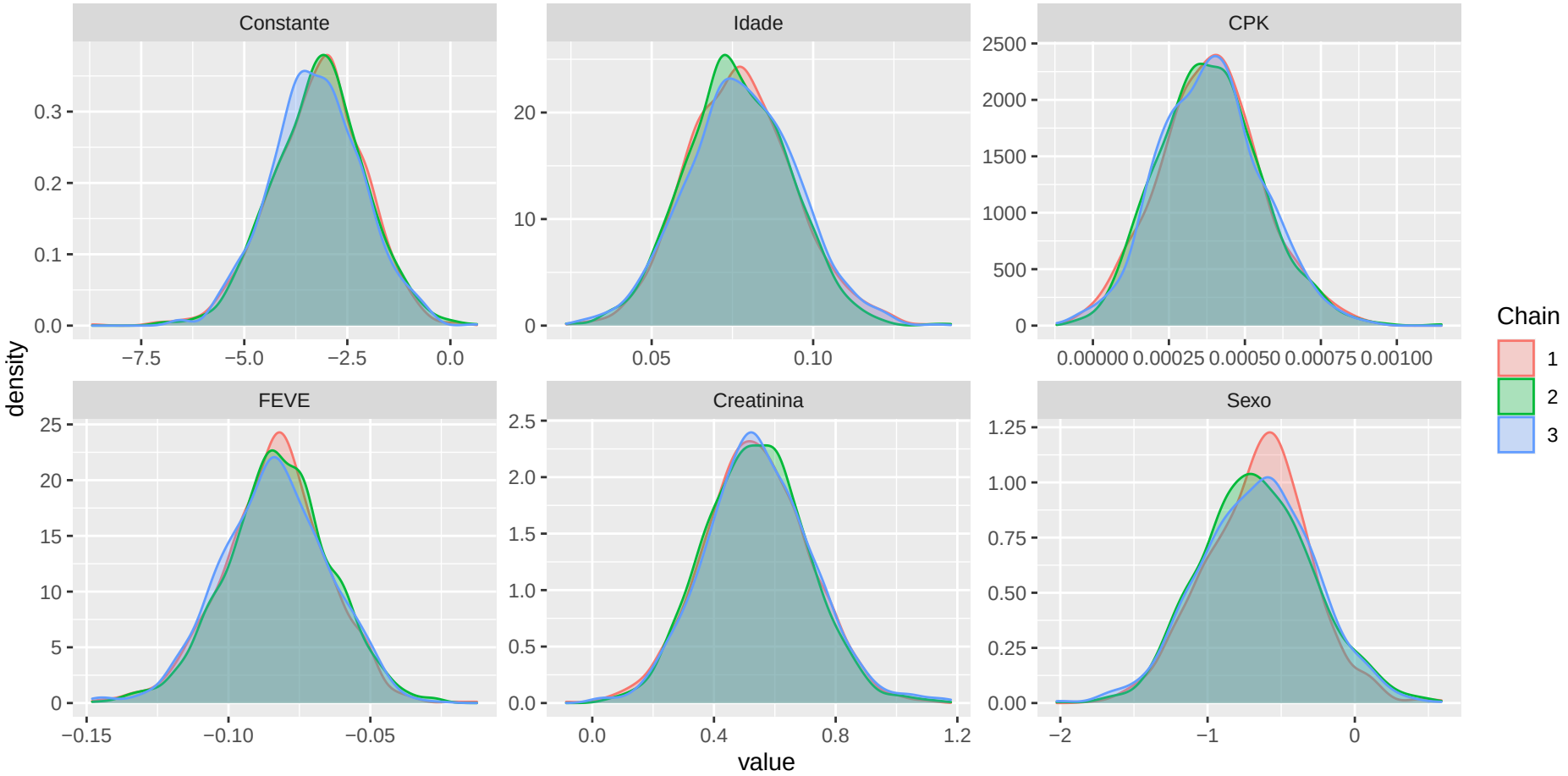
Para implementação do método MCMC, gerou-se três cadeias para cada parâmetro. Após 10.000 iterações, obteve-se os traços das cadeias. Estipulou-se um burn in de 1000 iterações e thin de 90 resultando em uma amostra de 1000 observações para cada cadeia.

Análise gráfica

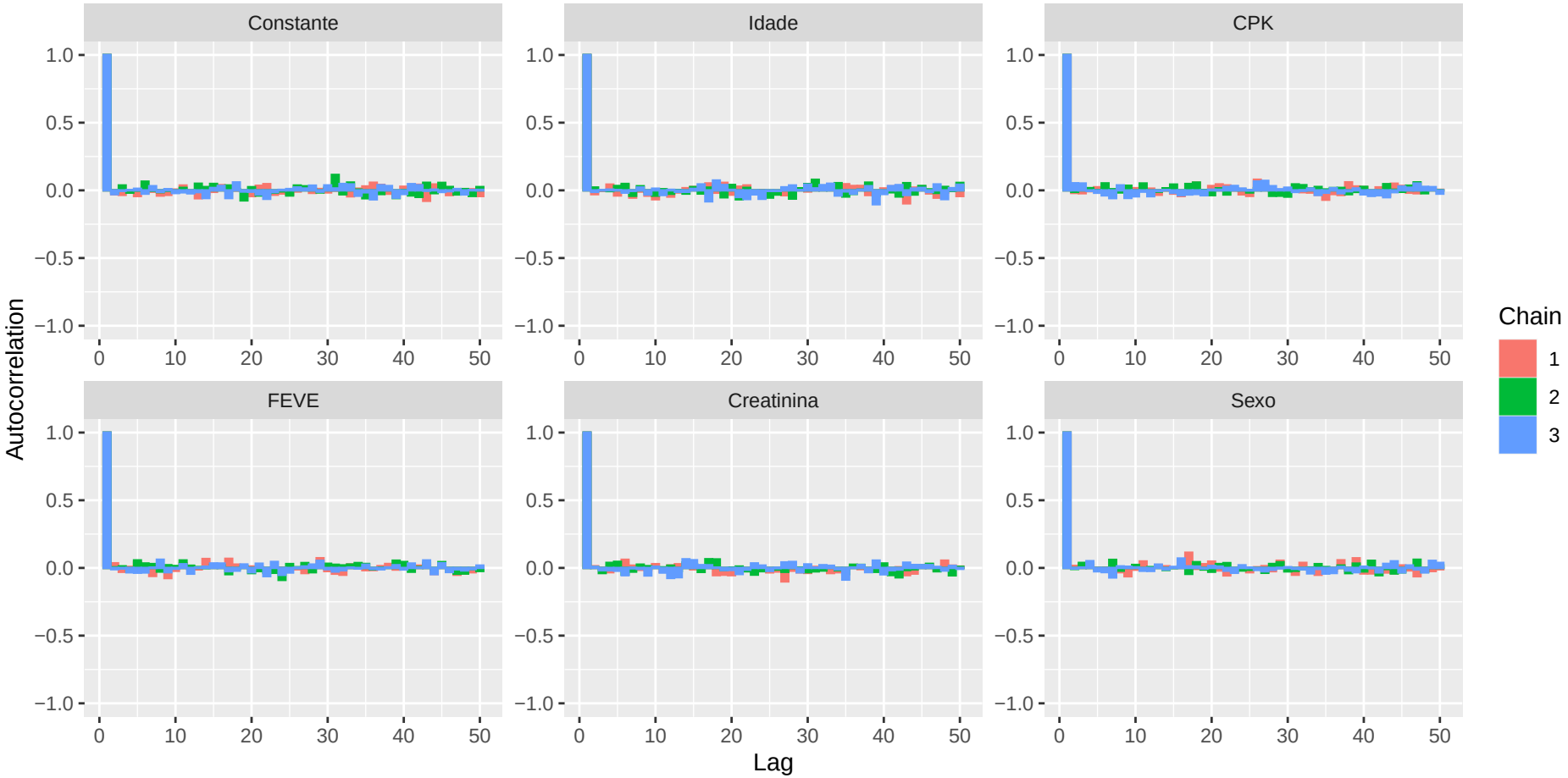
Gráficos para convergência das cadeias para os parâmetros de β_0 a β_5:



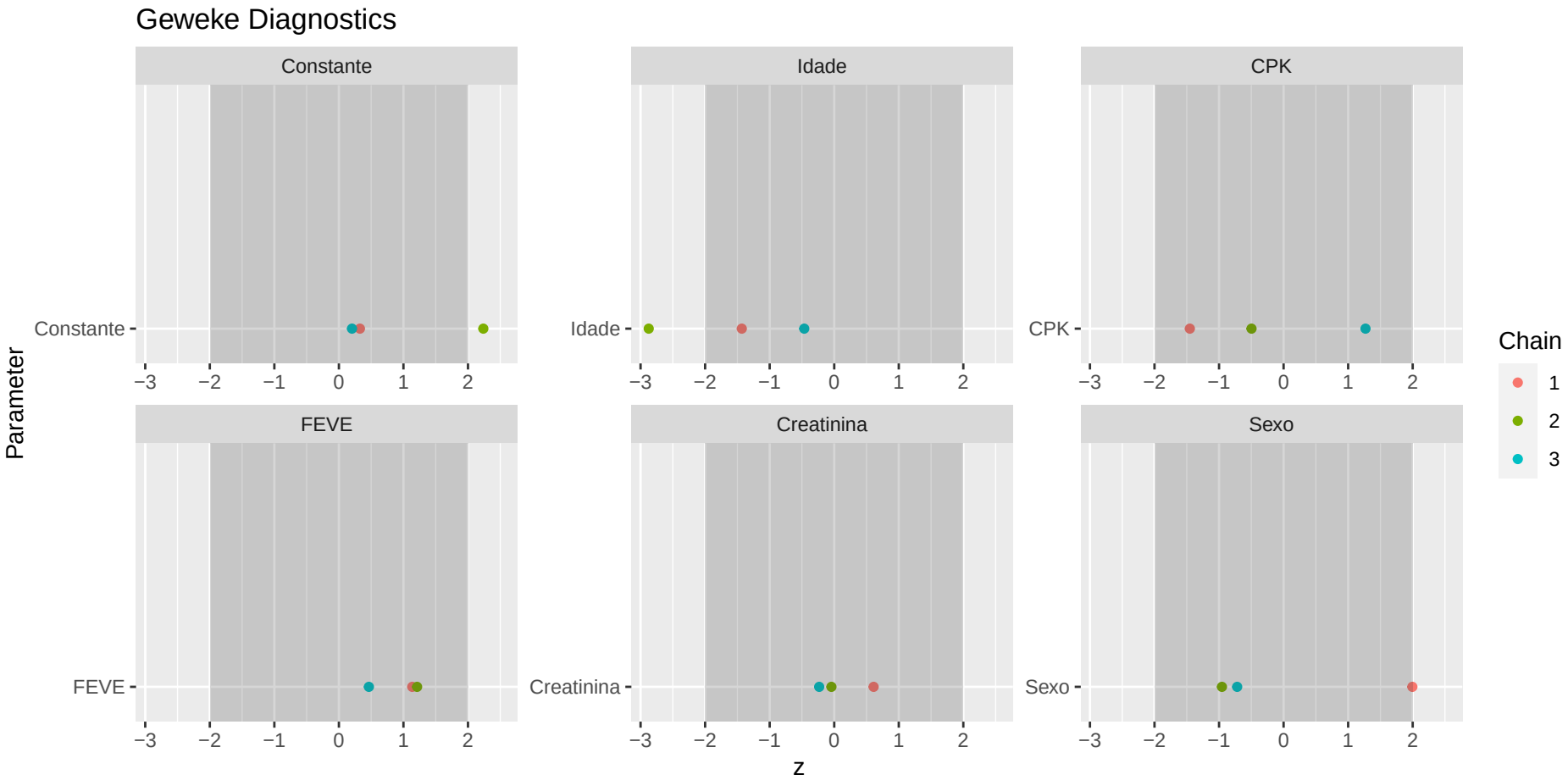
Densidade das cadeias para os parâmetros de β_0 a β_5:



Autocorrelação das cadeias para os parâmetros de β_0 a β_5:



Estatísticas do teste de Geweke das cadeias para cada parâmetro:



Estimativas dos parâmetros:

Variável	Média	Std. Dev.	Lim inf 95% IC	Lim sup 95% IC
Constante	-3.189	1.114	-5.402	-1.060
Idade	0.077	0.017	0.046	0.111
CPK	0.000	0.000	0.000	0.001
FEVE	-0.083	0.018	-0.120	-0.048
Creatinina	0.537	0.171	0.218	0.876
Sexo	-0.654	0.371	-1.376	0.083

Matriz de Confusão:

	Valor Observado	
Valor Estimado	Não	Sim
Não	53	17
Sim	7	11

Os valores da matriz de confusão foram iguais aos obtidos no modelo clássico e por isso teremos os mesmos valores das métricas de desempenho do modelo: acurácia, sensibilidade e especificidade.

Conclusão

- As três funções de ligação usadas para casos em que a variável resposta é do tipo binária ficaram bem ajustadas aos modelos propostos. Optou-se pela função logit, pois é mais intuitiva em relação a interpretação de seus resultados.
- Os resultados demonstrados pelas métricas alcançadas pelo modelo denotam uma maior capacidade do modelo em prever casos de sobrevivência.
- Outro ponto importante é a acurácia do modelo adotado, prevendo corretamente cerca de 72,73% dos eventos. As métricas alcançadas ao final do estudo poderiam ter resultados mais proveitosos caso tivéssemos uma base mais igualitária na variável resposta.
- As estimativas do modelo sob ponto de vista Bayesiano ficaram muito próximas do modelo clássico.

Referências

Tanvir Ahmad et al. “Survival analysis of heart failure patients: A case study”. Em:PLOS ONE 12.7 (jul. de 2017), pp. 1–8.

Dobson, Annette J., and Adrian G. Barnett. An introduction to generalized linear models. CRC press, 2018.

Dunn, Peter K., and Gordon K. Smyth. Generalized linear models with examples in R. New York: Springer, 2018.

Obrigada!