UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 6: INFERÊNCIA E ANÁLISE DIAGNÓSTICA

Salvador

2023

CAMILLE MENEZES PEREIRA DOS SANTOS
MICHEL MILER ROCHA DOS SANTOS

LABORATÓRIO 6: INFERÊNCIA E ANÁLISE DIAGNÓSTICA

Atividade de laboratório apresentada ao Instituto de Matemática e Estatística da Universidade Federal da Bahia como parte das exigências da disciplina Análise de Regressão ministrada pela professora Dra. Edleide de Brito.

Salvador

2023

SUMÁRIO

1 Introdução

Nesta atividade, será realizado uma análise de Regressão Linear Múltipla, com o objetivo de investigar a relação entre variáveis e a prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares em dados de 403 afro-americanos residentes no Estado da Virginia, nos Estados Unidos.

A primeira etapa consiste em uma análise descritiva e exploratória dos dados, onde as características das variáveis quantitativas serão destacadas. Essa análise permitirá uma melhor compreensão acerca da natureza dos dados e identificar possíveis padrões ou tendências.

Em seguida, será ajustado um modelo de regressão linear múltipla os dados, em que a variável resposta é o IMC, e os coeficientes estimados do modelo serão interpretados. Para determinar a significância das variáveis independentes, serão realizados testes estatísticos. Assim, será possível identificar quais variáveis têm um efeito estatisticamente significativo na prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares.

A avaliação da bondade de ajuste do modelo será realizada por meio da análise de variância (ANOVA), que fornecerá informações sobre a qualidade global do modelo.

Também será avaliado o coeficiente de determinação e o coeficiente de determinação ajustado do modelo. Por fim, será realizado uma série de gráficos de diagnóstico, esses gráficos ajudarão a identificar pontos atípicos, verificar a normalidade dos resíduos e avaliar a influência de observações individuais sobre os resultados do modelo.

Assim, será obtida uma compreensão mais aprofundada da relação entre as características dos afro-americanos residentes no Estado da Virginia e a obesidade.

2 Resultados

2.1 Manipulação dos dados

Os dados de 403 afro-americanos residentes no Estado da Virginia, apresentam as variáveis: colesterol total, glicose estabilizada, colesterol bom, razão entre colesterol total e colesterol bom, hemoglobina glicada, município de residência, idade em anos, sexo, altura, peso, pressão sanguínea sistólica e diastólica (1^{o} e 2^{o} medidas), cintura e quadril.

Nessa manipulação dos dados, algumas variáveis tiveram suas unidades convertidas: as variáveis altura, cintura e quadril foram convertidas para metros e a variável peso foi convertida para quilos.

Duas novas variáveis foram calculadas a partir das variáveis existentes. A variável IMC (Índice de Massa Corporal) foi calculada dividindo o peso em quilogramas pelo quadrado da altura em metros. A variável RCQ (Relação Cintura Quadril) foi calculada dividindo a cintura pela medida do quadril.

Por haver muitos valores ausentes, as variáveis que denotam as pressões sanguíneas sistólica e diastólica na 2^{o} medida foram retiradas dos dados. Além disso, as linhas que contêm valores ausentes foram removidas. Após a remoção, restaram 377 observações no conjunto de dados.

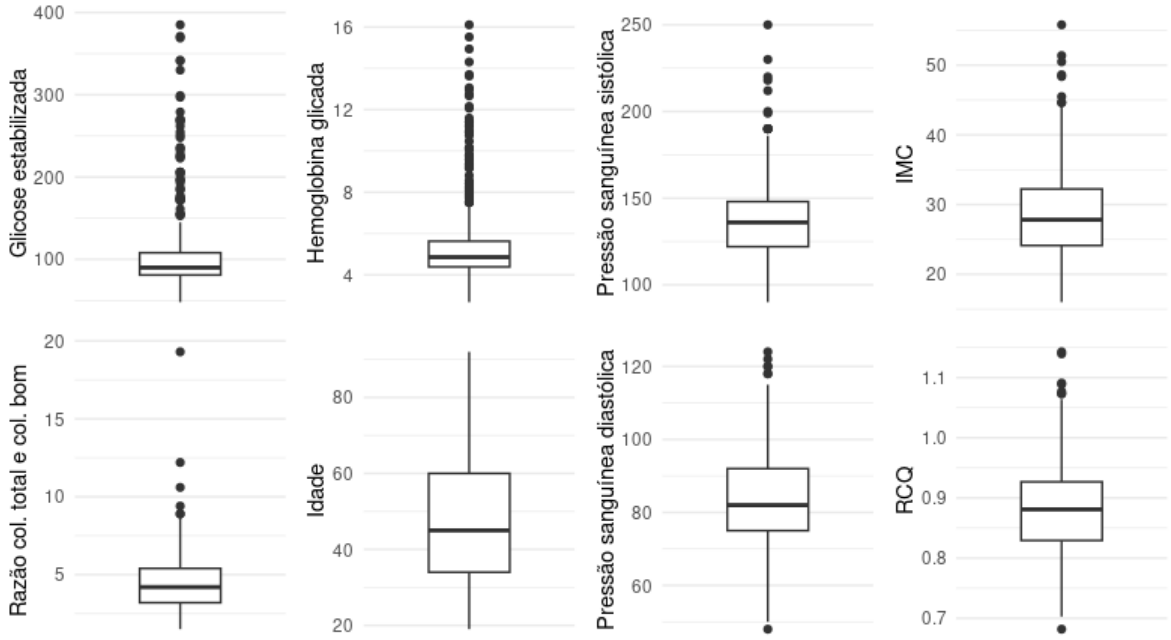
Além disso, como a razão entre o colesterol total e colesterol bom, o IMC e o RCQ são funções de variáveis que estão nos dados, essas variáveis que deram origem as funções serão desconsideradas, assim como as variáveis qualitativas. Portanto, as variáveis que serão utilizadas no modelo de regressão são: glicose estabilizada, razão entre colesterol total e colesterol bom, hemoglobina glicada, idade em anos, pressão sistólica, pressão diastólica e as novas variáveis IMC e RCQ.

2.2 Análise Descritiva

Ao observar os *boxplots* da Figura 2.1, é possível notar que há presença de muitos valores atípicos, principalmente nas variáveis glicose estabilizada e hemoglobina glicada. Esses valores atípicos não devem ser retirados da análise, pois eles fazem parte da natureza das variáveis e parecem indicar alguma informação sobre a presença de diabetes nesses indivíduos, já que ambas variáveis avaliam a glicose no sangue.

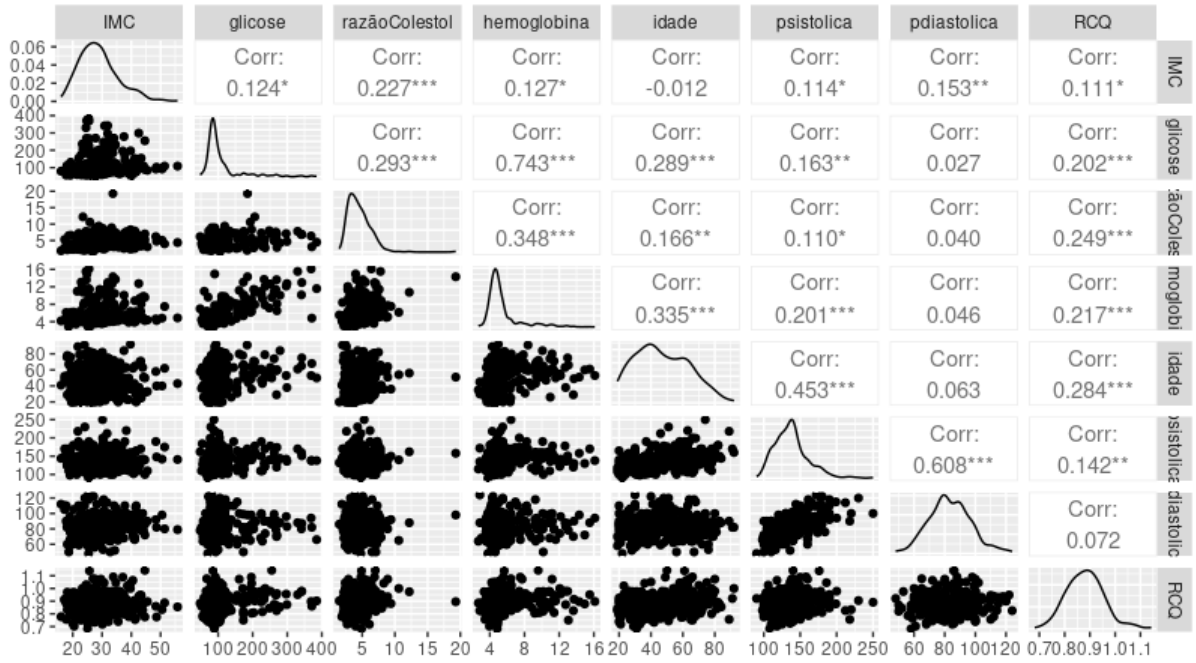
As variáveis pressão sanguínea diastólica, IMC e RCQ aparentam uma leve simetria, isso pode ser observado nos *boxplots* da Figura 2.1 e no gráfico da densidade das variáveis na Figura 2.2.

Figura 2.1: *Boxplots* das características de afro-americanos residentes no Estado da Virginia (EUA)



As variáveis, em geral, não aparentam ter uma relação linear visível entre si, quando se observa os gráficos de dispersão na Figura 2.2, com exceção das relações entre as variáveis glicose estabilizada e hemoglobina glicada, e entre pressão sanguínea diastólica e sistólica, que parecem ser linear. Esta análise é reiterada quando observamos o coeficiente de correlação linear de *Pearson*, apenas entre glicose estabilizada e hemoglobina glicada e entre pressão sanguínea diastólica e sistólica apresentaram o valor do coeficiente de correlação maior que 0,6.

Figura 2.2: Correlação e gráfico de dispersão e densidade das características de afro-americanos residentes no Estado da Virginia (EUA)



2.3 Modelo de regressão linear múltiplo

A Tabela 2.1, apresenta uma sumarização da regressão linear em que as características dos afro-americanas residentes no Estado da Virginia são utilizadas para predizer o seu IMC.

Na Tabela 2.1, é possível notar que a estimativa de β_0, apesar de ser um valor teoricamente possível do IMC, não possui interpretabilidade. Pois, o indivíduo não pode ter valores de RCQ, glicose estabilizada, pressão sistólica e diastólica iguais a zero, por exemplo.

Tabela 2.1: Sumarização da regressão linear em que as características dos afro-americanos residentes no Estado da Virginia (EUA) são utilizadas para predizer o seu IMC

Coeficientes	Estimativa	EP	Est. T	Valor-p
β_0	14,76	4,40	3,35	<0,001
Glicose est.	0,007	0,009	0,76	0,45
Razão colest.	0,74	0,21	3,61	< 0,001
Hemoglobina	0,08	0,23	0,34	0,73
Idade	-0,05	0,03	-1,93	0,06
Pressão sist.	0,02	0,02	0,75	0,45
Pres. dias.	0,06	0,03	1,70	0,09
RCQ	5,69	4,80	1,19	0,23

Apenas as variáveis razão colesterol total e colesterol bom, idade e pressão diastólica

foram significativas a um nível de 10%. Além disso, o coeficiente de determinação foi aproximadamente igual a $0,09$, ou seja, este o modelo de regressão linear múltiplo com estas variáveis independentes explicam apenas 9% da variabilidade do IMC. Portanto, parece que esse modelo de regressão linear múltiplo não foi relevante em descrever a natureza da variável IMC.

2.4 Modelo de regressão linear múltiplo reduzido

Um novo modelo de regressão linear foi estimado, considerando as variáveis que foram significativas a um nível de 10%. A Tabela 2.2, apresenta uma sumarização dos coeficientes das variáveis incluídas no modelo.

Tabela 2.2: Sumarização da regressão linear em que as características dos afro-americanos residentes no Estado da Virginia (EUA) são utilizadas para predizer o seu IMC

Coeficientes	Estimativa	EP	Est. T	Valor-p
β_0	19,95	2,31	8,65	<0,001
Razão colest.	0,85	0,19	4,58	<0,001
Idade	0,07	0,03	2,95	0,003
Pres. dias.	-0,02	0,02	-1,18	0,239

Dessa forma, a equação do modelo de regressão ajustado é

$$\begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \begin{pmatrix} 1 & 3,6 & 46 & 59 \\ 1 & 6,9 & 29 & 68 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 3,8 & 41 & 78 \end{pmatrix} \begin{pmatrix} 19,95 \\ 0,85 \\ 0,07 \\ -0,02 \end{pmatrix}.$$

E as interpretações para os coeficientes estimados do modelo de regressão linear são:

- Considerando todas as outras variáveis constantes, o aumento de uma unidade na razão entre colesterol total e bom representa, em média, um aumento de 0,85 no valor do IMC;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade na idade representa, em média, um aumento de 0,07 no valor do IMC;
- Considerando todas as outras variáveis constantes, o aumento de uma unidade na pressão diastólica representa, em média, um decréscimo de 0,02 no valor do IMC.

No modelo reduzido, a pressão diastólica deixou de ser significativa, a um nível de significância de 5%. A idade, que era não era significativa a um nível de 5%, mas apenas ao nível de 10%, passou a ser significativa a um nível de 5%.

A Tabela 2.3 apresenta a Tabela ANOVA do modelo reduzido. O valor-p do teste F obtido foi menor que 0,5. Portanto, rejeita-se a hipótese nula de igualdade de todos os coeficientes angulares do modelo a zero, a um nível de significância de 5%. Ou seja, rejeita-se a hipótese de que o modelo não tem nenhuma validade para explicar a variável resposta IMC.

Tabela 2.3: Tabela ANOVA

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrado médio	Estatística F	Valor-p
Regressão	3	1255,81	418,6	10,21	<0,001
Resíduos	373	15293,06	41		
Total	376	16548.87			

Realizando o teste exato da razão de máxima verossimilhança para testar o modelo completo contra o reduzido foi obtido um valor-p de 0,36. Ou seja, não rejeitou a hipótese nula de equivalência entre o modelo reduzido com o modelo completo, a um nível de significância de 5%. Desse modo, é possível escolher o modelo reduzido em detrimento do modelo completo.

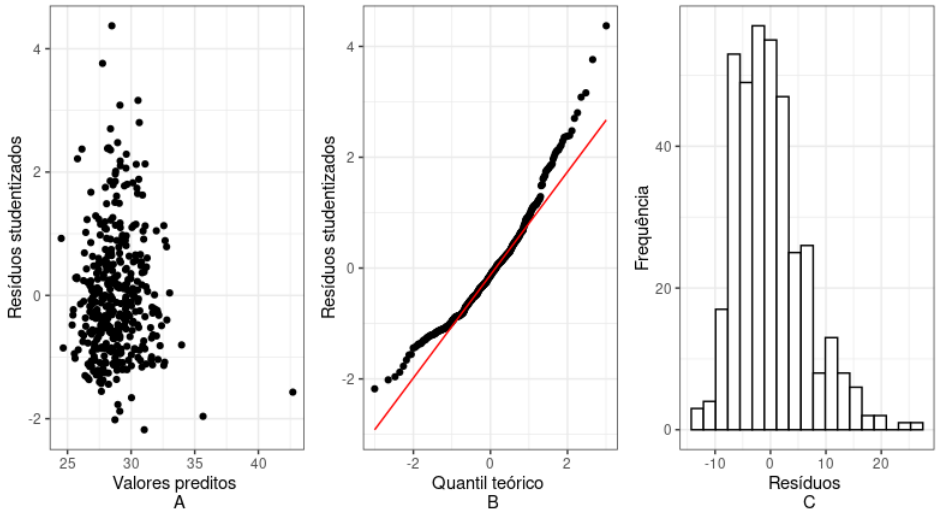
O coeficiente de determinação foi 0,076, enquanto que o coeficiente de determinação ajustado foi 0,068. Portanto, 7,6% da variabilidade da variável IMC pode ser explicada pelo modelo de regressão linear reduzido.

O coeficiente de determinação ajustado pode ser utilizado para comparar dois modelos. O coeficiente de determinação ajustado para o modelo completo foi 0,069, praticamente igual ao coeficiente de determinação ajustado do modelo reduzido. Portanto, mais uma evidência de que o modelo reduzido é equivalente ao modelo completo, mas com a vantagem de ser mais parcimonioso.

2.4.1 Análise dos resíduos

No gráfico de dispersão dos resíduos contra os valores preditos na Figura 2.3, é possível perceber que a variação dos resíduos parece ser constante, pois não parece haver nenhuma tendência de crescimento ou diminuição da variação dos resíduos ao longo dos valores preditos. Realizando o teste de *Goldfeld-Quandt* para avaliar a homoscedasticidade dos resíduos, foi obtido um valor-p de 0,81, logo, a hipótese de homocedasticidade não foi rejeitada a um nível de significância de 5%. Portanto, é possível afirmar que a suposição da variância constante dos resíduos não foi violada.

Figura 2.3: Valores Ajustados e Resíduos Studentizados, Gráfico Quantil-Quantil e histograma



No gráfico QQ-plot e no histograma dos resíduos contidos também na Figura 2.3, nota-se um grande problema no que tange à normalidade nas caudas da distribuição do resíduos, indicando uma assimetria à direita incomum a distribuição normal. Realizando o teste de *Shapiro-Wilk* para avaliar a normalidade dos resíduos, foi obtido um valor-p de aproximadamente zero, rejeitando a hipótese nula de que os resíduos seguem uma distribuição normal, a um nível de 5%. Portanto, a suposição de normalidade dos resíduos foi violada.

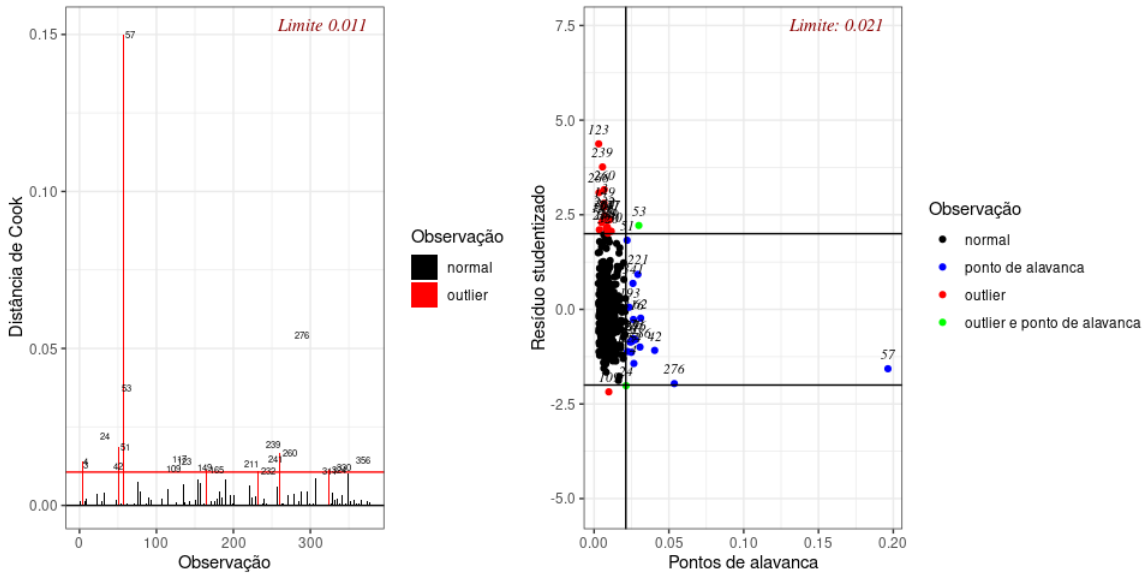
Para avaliar se os resíduos são independentes, foi realizado o teste de *Durbin-Watson*, que afere se os resíduos não têm autocorrelação serial. O valor-p obtido foi igual a 0,70, logo, a hipótese nula de não autocorrelação serial dos resíduos não é rejeitada a um nível de 5%. Desse modo, a suposição de independência dos resíduos não foi violada

A Figura 2.3 (A) indica a presença de pontos influentes no modelo, uma vez que apresentam uma distância de Cook maior que o limite considerado de 4/377, em especial a observação 57 que apresenta uma distância de Cook bem acima do limite tolerado e das demais observações. Em outras palavras, se essas observações fossem removidas da base

de dados, isso causaria uma mudança expressiva nas estimativas de mínimos quadrados obtidas para os coeficientes de modo geral.

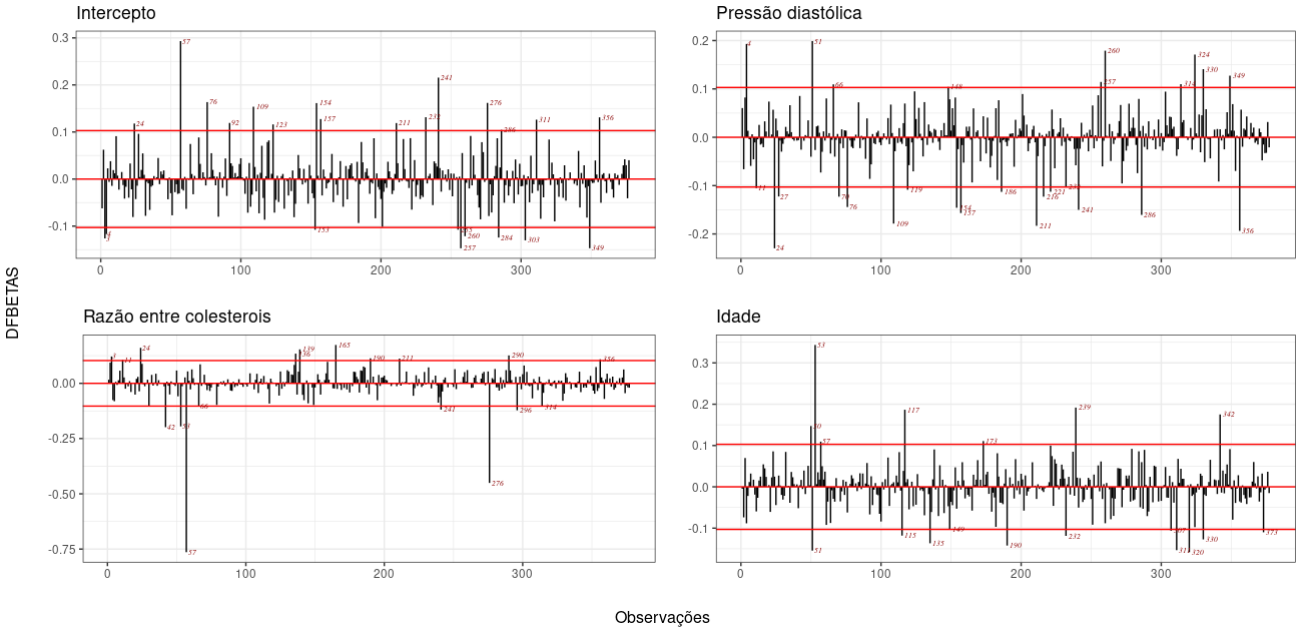
Na Figura 2.3 (B), são identificadas observações que são extremas no espaço das variáveis explicativas(ponto de alavanca), que são extremas na variável resposta (*outliers*) e observações que são tanto *outliers* quanto pontos de alavanca. Novamente a observação 57 destoa dos demais, nesse caso sendo um ponto de alavanca, e a observação 53 foi indicado tanto como *outliers* como ponto de alavanca.

Figura 2.4: Gráfico de Distância de Cook, gráfico dos pontos de Alavanca e Resíduo Studentizado



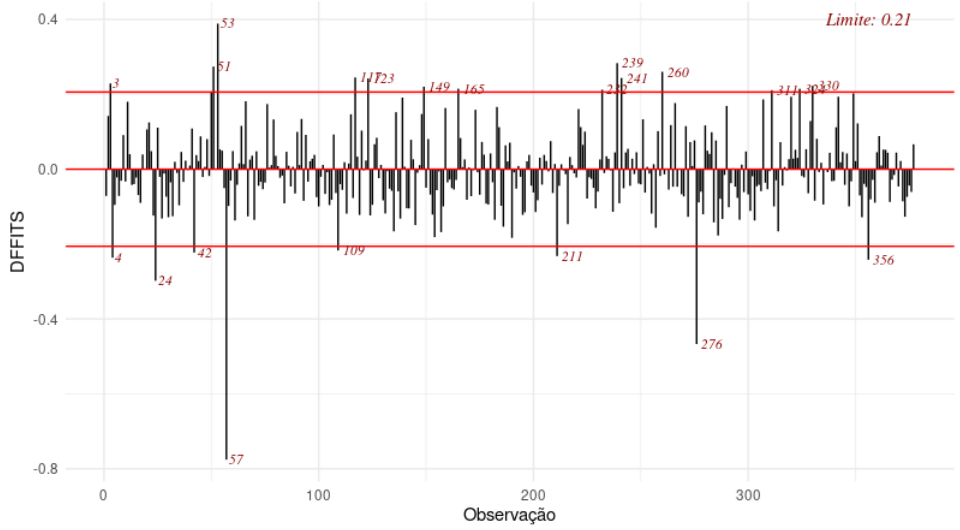
Os gráficos dos DFBetas, na Figura 2.5, também indicam a presença de pontos influentes no modelo de regressão para cada parâmetro do modelo, sendo considerados influentes aqueles com $|DFBeta_{i,j}| > 2/\sqrt{377}$. A observação 57 se apresentou como observação influente para todos os parâmetros analisados com DFBetas.

Figura 2.5: Gráfico de DfBeta



Na Figura 2.6, o gráfico dos DFFITS avalia o impacto na predição ou valor ajustado de a exclusão de determinados pontos causa um impacto significativo na predição ou valor ajustado de uma observação. A observação 57 mais uma vez é indicado como uma observação influente, dessa vez, avaliado no impacto dos valores preditos. Devido a essa observação ter sido considerado uma observação influente por diferentes métricas, seria aconselhável ajustar novamente o modelo de regressão linear múltiplo sem essa observação e verificar o impacto que a retirada dessa observação teve no modelo ao comparar com o modelo ajustado anteriormente, que considera a observação.

Figura 2.6: Gráfico de DFFITS



3 Conclusão

Foi realizada a manipulação dos dados, com o intuito de checar inconsistências, observar valores faltantes, mudar a escala das variáveis para o padrão brasileiro, criar novas variáveis a partir de variáveis já existentes no banco de dados e excluir aquelas que não são mais úteis para a nossa análise.

Na análise descritiva, foi observado que há muitos valores discrepantes nas variáveis glicose estabilizada e hemoglobina glicada. Mas, esses valores nos trazem informação sobre possíveis diabéticos, já que ambas as variáveis avaliam a glicose no sangue. Além disso, foi notado que tirando as variáveis glicose estabilizada e hemoglobina glicada, e pressão sanguínea sistólica e diastólica, as variáveis não estão muito correlacionadas e não apresentam uma relação linear visível entre si.

Um modelo de regressão linear múltiplo para predizer o IMC foi proposto. Contudo, apenas a variável explicativa razão colesterol total e colesterol bom foi significativa, ao nível significância de 5%, no modelo. Além disso, apenas 9% da variabilidade do IMC pode ser explicada pelo modelo. Portanto, o modelo linear múltiplo com estas variáveis não parece ser tão adequado.

Por isso, um modelo de regressão linear múltiplo com as variáveis razão entre colesterol total e colesterol bom, idade e pressão diastólica — significativas a um nível de 10% no modelo anterior — foi proposto. O modelo ajustado indica que a razão entre colesterol total e bom e a idade são variáveis significativas para predizer o índice de massa corporal (IMC) dos afro-americanos residentes no Estado da Virginia.

A análise dos resíduos revela que a suposição de homoscedasticidade (variância constante dos resíduos) não foi violada, mas a suposição de normalidade dos resíduos foi violada. Além disso, não há evidências de autocorrelação serial dos resíduos.

Os gráficos dos pontos influentes indicam a presença de observações que têm um impacto considerável nas estimativas dos coeficientes de regressão. Essas observações podem ser consideradas como pontos de alavanca (extremas nas variáveis explicativas), *outliers* (extremas na variável resposta) ou ambas. Um exemplo foi a observação 57, que foi considerada um ponto influente em todas as métricas utilizadas

Portanto, o modelo de regressão linear reduzido pode ser considerado válido para

explicar o IMC dos afro-americanos residentes no Estado da Virginia quando comparado com o modelo completo, embora sua capacidade de previsão seja limitada, já que apenas cerca de 7,6% da variabilidade do IMC é explicada pelo modelo. No entanto, é importante considerar as violações das suposições do modelo, como a falta de normalidade dos resíduos e a presença de pontos influentes, ao interpretar e utilizar os resultados do modelo.