Doenças entre Afroamericanos

Fernado Bispo, Jeff Caponero

Sumário

Introdução

O conjunto de dados a ser analisado, o mesmo trabalhado na parte 1 deste relatório, contém informações de 403 afro-americanos residentes no Estado da Virginia (EUA), entrevistados em um estudo referente à prevalência de obesidade, diabetes e outros fatores de risco cardiovasculares. As características apresentadas são:

- Colesterol total;
- Glicose estabilizada;
- Lipoproteína de alta densidade (colesterol bom);
- Razão colesterol total e colesterol bom;
- Hemoglobina glicada;
- Município de residência (Buckingham ou Louisa);
- Idade (em anos);
- Sexo;
- Altura (em cemtimetros);
- Peso (em libras);
- Pressão sanguínea sistólica (1ª medida);
- Pressão sanguínea diastólica (1ª medida);
- Pressão sanguínea sistólica (2ª medida);
- Pressão sanguínea diastólica (2ª medida);
- Cintura (em cemtimetros);
- Quadril (em cemtimetros).

Com base nestes dados se desenvolverá nesta segunda parte:

1. Nova análise descritiva e exploratória dos dados (apenas das variáveis quantitativas), incluindo visualização de dados.

2. Determinação da equação do modelo ajustado e interpretação os seus coeficientes.

3. Condução de testes para determinar quais variáveis são estatisticamente significantes ao nível de significância de 5%.

4. Obtenção de um quadro da análise de variância e de resultado do teste F a fim de avaliar a bondade do ajuste do modelo.

5. Obtenção do coeficiente de determinação e do coeficiente de determinação ajustado do modelo.

6. Apresentação dos gráficos de diagnóstico para:

(a) Valores Ajustados e Resíduos Studentizado;

(b) Gráfico Quantil-Quantil;

(c) Gráfico de Distância de Cook;

(d) Gráfico dos pontos de Alavanca e Resíduo Studentizado;

(e) Gráfico de DfBeta;

(f) Gráfico de DfFit;

(g) Gráfico do COVRatio.

Resultados

Análise descritiva dos dados

As análises prévias (primeira parte deste relatório) permitiram determinar que:
1- As características **Pressão sanguínea sistólica (2ª medida)** e **Pressão sanguínea diastólica (2ª medida)** possuem uma quantidade muito grande de dados ausentes, cerca de 65% de ausência de dados, portanto essas características foram descartadas.

2- Se constatou também que as observações das características **altura, peso, cintura** e **quadril** estão representadas em unidades do Sistema Imperial, que foram convertidas o Sistema Internacional.

Nesta etapa, foi realizada outra análise exploratória dos dados, levando-se em conta apenas as variáveis quantitativas, que está representada na Tabela 1.

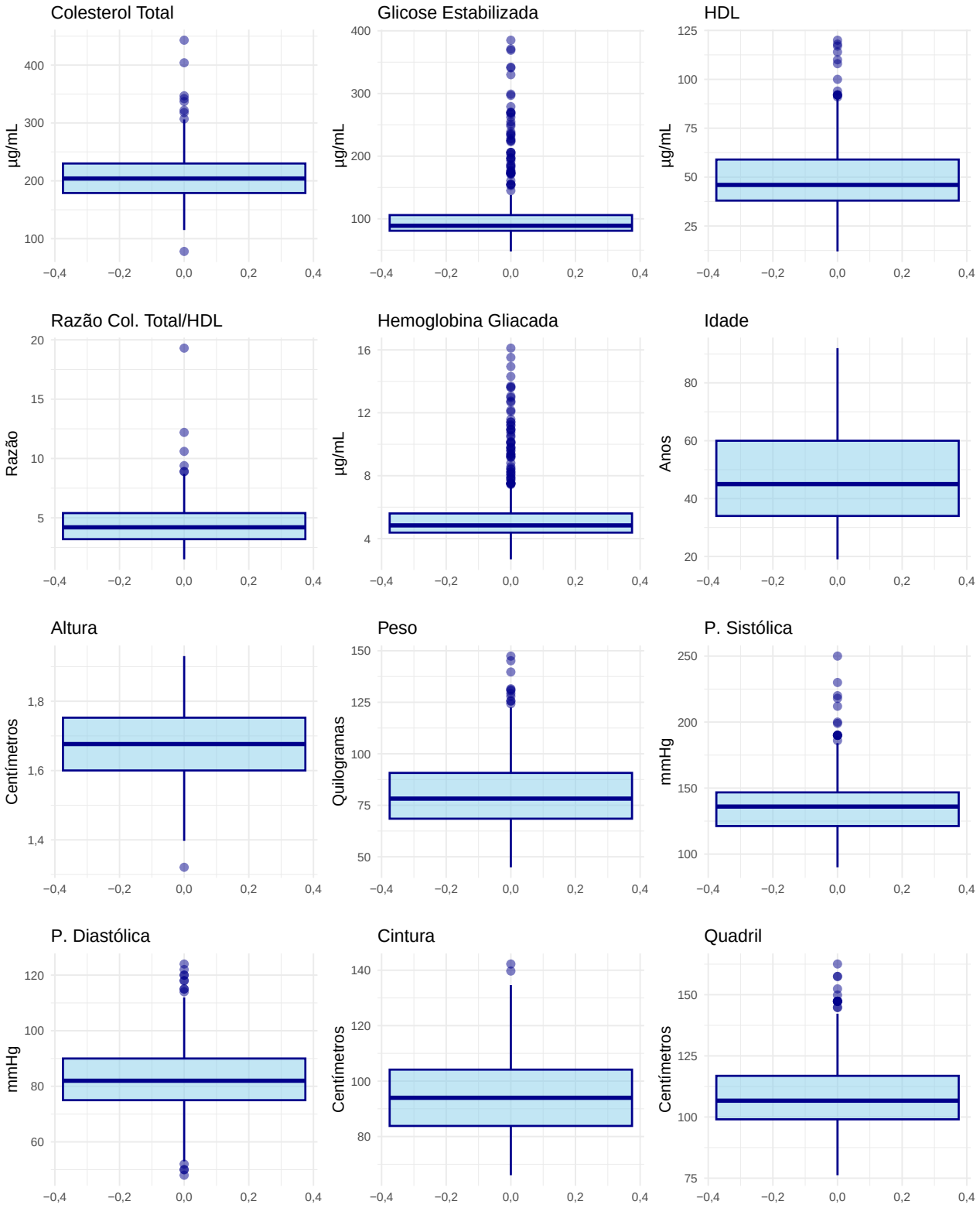
Tabela 1: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Altura	1,32	1,60	1,68	1,68	1,75	1,93	0,10	0,06	0,02	-0,20
Cintura	66,04	83,82	93,98	96,41	104,14	142,24	14,64	0,15	0,47	-0,18
Colesterol total	78,00	179,00	204,00	207,74	230,00	443,00	44,64	0,21	0,97	2,65
Glicose estabilizada	48,00	81,00	90,00	107,52	108,00	385,00	53,96	0,50	2,71	7,83
Hemoglobina glicada	2,68	4,39	4,86	5,60	5,63	16,11	2,21	0,40	2,25	5,15
Idade	19,00	34,00	45,00	46,92	60,00	92,00	16,64	0,35	0,31	-0,73
Lipoproteína de alta densidade	12,00	38,00	46,00	50,41	59,00	120,00	17,41	0,35	1,22	2,01
Peso	44,91	68,49	78,93	80,74	90,72	147,42	18,37	0,23	0,74	0,71
Pressão sanguínea diastólica	48,00	75,00	82,00	83,43	92,00	124,00	13,53	0,16	0,22	0,07
Pressão sanguínea sistólica	90,00	122,00	136,00	137,40	148,00	250,00	23,13	0,17	1,06	2,19
Quadril	76,20	99,06	106,68	109,43	116,84	162,56	14,30	0,13	0,81	0,89
Razão colesterol total e colesterol bom	1,50	3,20	4,20	4,53	5,40	19,30	1,75	0,39	2,23	13,12

Desta análise, verifica-se que a distribuição das variáveis não apresenta fatores impeditivos da regressão linear a que nos propomos.

Pode-se ainda complementar este estudo por meio de uma análise de disperção dos dados por meio de gráficos do tipo BoxPlot, como se vê na Figura 1.

Figura 1: BoxPlot das variáveis em análise.



Pode-se verificar pela Figura 1 que há diversos valores atipicos (*outlayers*), entretanto sem conhecimento especializado da fisiologia é temerário prescindir destas observações. Por outro

lado, é possivel realizar uma análise estatística destes valores de forma a indicar aqueles tem maior influência sobre o modelo proposto e assim viabilizar um tratamento mais adequado a cada um deles. Este tratamento será realizado por meio de gráficos diagnósticos ao final deste estudo.

Modelo de Regressão Linear Multipla

O modelo obtido pode ser representado por:

Y_i =-119,445 + 0 X_{1i}+ 0,012 X_{2i} -0,01 X_{3i}+ 0,386 X_{4i} -0,251 X_{5i} -0,1 X_{6i}+ 49,317 X_{7i} -0,041 X_{8i}+ 0,068 X_{9i}+ 0,545 X_{10i}+ 0,627 X_{11i}

Onde:
Y_i - Peso;
X_{1i} - Colesterol total;
X_{2i} - Glicose estabilizada;
X_{3i} - Lipoproteína de alta densidade;
X_{4i} - Razão colesterol total e colesterol bom;
X_{5i} - Hemoglobina glicada;
X_{6i} - Idade;
X_{7i} - Altura;
X_{8i} - Pressão sanguínea sistólica;
X_{9i} - Pressão sanguínea diastólica;
X_{10i} - Cintura;
X_{11i} - Quadril.

Interpretando-se o modelo pode-se dizer que para cada variável, fixadas as demais condições (*Ceteris Paribus*), temos que o peso dos indivíduos aumenta 15g a cada 1 µg/mL de colesterol total; aumenta 21g a cada 1 µg/mL de glicose estabilizada; reduz 187g a cada 1 µg/mL de lipoproteína de alta densidade; reduz 975g a cada unidade da razão colesterol total e colesterol bom; reduz 240g a cada 1 µg/mL Hemoglobina glicada; reduz 164g a cada ano de idade do indivíduo; aumenta 1.107g a cada centimetro da altura do indivíduo; reduz 69g a cada 1 mmHg de pressão sanguínea sistólica; aumenta 46g a cada 1 mmHg de pressão sanguínea diastólica; aumenta 726g a cada centimetro no perímetro da cintura e aumenta 295g a cada cemtimetro no perímetro do quadril do indivíduo.
Neste modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}865$, o que denota que 86,5% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}861$.

Da equação do modelo já se identifica que há fortes indícios de que a variável X_{1i} (Colesterol Total) não apresenta qualquer significância para o ajuste do modelo. Desta forma, é conveniente avaliar a significância estatística de cada uma das variáveis a um nível de significância de 5%.

Variáveis Estatisticamente Significantes

Considerando um teste de hipótese para os parâmetros individuais do modelo podemos avaliar se:

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$

Utilizando a estatística teste dada por:

$$t = \frac{\hat{\beta}_j - \beta_j}{ep(\hat{\beta}_j)}$$

Com base no valor tabelado de $t_{(2,5\%,365)}$ = -1,966 e realizados os calculos verificou-se os seguintes valores da estatística t:

Tabela 2: Análise de Significância

Exame	Estatística t
Colesterol total	-0,024
Glicose estabilizada	1,215
Lipoproteína de alta densidade	-0,218
Razão colesterol total e colesterol bom	0,725
Hemoglobina glicada	-1,006
Idade	-3,687
Altura	12,861
Pressão sanguínea sistólica	-1,794
Pressão sanguínea diastólica	1,924
Cintura	11,194
Quadril	12,890

Nota-se, desta forma, que as seguintes variáveis não se mostraram estatisticamente significantes ao nível de significancia de 5%: Lipoproteína de alta densidade; Idade; Pressão sanguínea sistólica; Cintura e Quadril.

Um novo modelo sem essas variáveis pode ser representado por:

Y_i^* = 69,224+ 0,028 X_{1i}^*+ 0,038 X_{2i}^*+ -0,327 X_{3i}^*+ -0,059 X_{4i}^*+ 0,134 X_{5i}^*+ -0,06 X_{6i}^*+ 0,307 X_{7i}^*

Onde:
Y_i^* - Peso;
X_{1i}^* - Colesterol total;
X_{2i}^* - Glicose estabilizada;

X^*_{3i} - Lipoproteína de alta densidade; X^*_{4i} - Razão colesterol total e colesterol bom; X^*_{5i} - Hemoglobina glicada; X^*_{6i} - Pressão sanguínea sistólica; X^*_{7i} - Pressão sanguínea diastólica.

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}15$, o que denota que 15% da variância dos dados é explicada pelo modelo. A redução em relação ao modelo anterior se deve a retirada de cinco variáveis. A redução é muito expressiva indicando que a significância individual das variáveis não pode ser usada como único critério para sua eliminação do rol de variáveis explicaivas. Pode-se calcular o coeficiente de determinação ajustado igual a $R^2_a = 0{,}125$

Análise de Variâncias

Uma outra forma de avaliar a importância de uma variável é realizando o teste ANOVA que verifica a plausibilidade de introduzir uma nova variável ao modelo de regressão. Realizando uma análise de variância com as variáveis iniciais é possível apresentar a tabela abaixo.

Tabela 3: Análise de Variância (ANOVA).

	GL^1	Soma de Quadrados	Quadrado Médio	Estatística F-Snedecor	p-valor
Colesterol total	1	403,501	403,5011	8,5784	0,0036
Glicose estabilizada	1	3.722,204	3.722,2039	79,1337	<0,0001
Lipoproteína de alta densidade	1	10.435,440	10.435,4396	221,8564	<0,0001
Razão colesterol total e colesterol bom	1	5,385	5,3850	0,1145	0,7353
Hemoglobina glicada	1	4,542	4,5418	0,0966	0,7562
Idade	1	2.012,321	2.012,3210	42,7817	<0,0001
Altura	1	5.511,324	5.511,3239	117,1702	<0,0001
Pressão sanguínea sistólica	1	1.873,619	1.873,6191	39,8330	0,0000
Pressão sanguínea diastólica	1	1.775,081	1.775,0806	37,7380	0,0000
Cintura	1	76.162,631	76.162,6313	1.619,2098	0,0000
Quadril	1	7.814,820	7.814,8199	166,1423	0,0000
Resíduos	365	17.168,474	47,0369		

Legenda:
[1] GL: Graus de Liberdade

A análise da Tabela 3 permite avaliar que apenas a introgução das variáveis "razão colesterol total e colesterol bom" e "homoglobina glicada" não seriam convenientes para o modelo de regressão para o peso do paciente.

Um novo modelo sem essas variáveis pode ser representado por:

$Y^\#_i$ = -118,502+ 0,007 $X^\#_{1i}$+ 0,006 $X^\#_{2i}$ -0,039 $X^\#_{3i}$ -0,103 $X^\#_{4i}$+ 49,535 $X^\#_{5i}$ -0,041 $X^\#_{6i}$+ 0,068 $X^\#_{7i}$+ 0,546 $X^\#_{8i}$+ 0,624 $X^\#_{9i}$

Onde:
$Y^\#_i$ - Peso;
$X^\#_{1i}$ - Colesterol total;
$X^\#_{2i}$ - Glicose estabilizada;

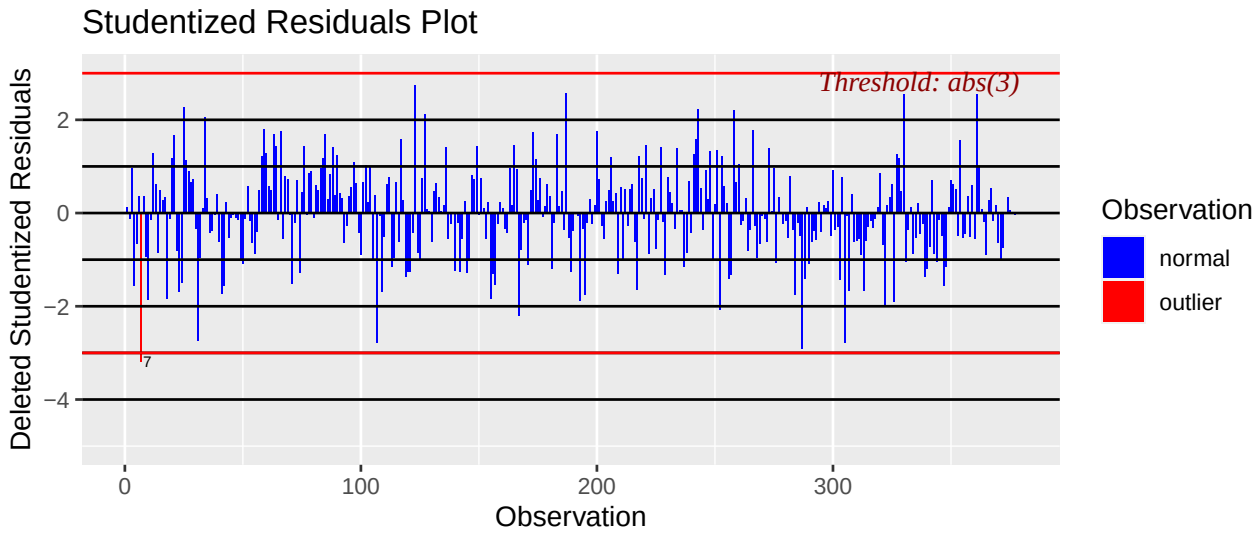
$X_{3i}^{\#}$ - Lipoproteína de alta densidade;
$X_{4i}^{\#}$ - Idade;
$X_{5i}^{\#}$ - Altura;
$X_{6i}^{\#}$ - Pressão sanguínea sistólica;
$X_{7i}^{\#}$ - Pressão sanguínea diastólica;
$X_{8i}^{\#}$ - Cintura;
$X_{9i}^{\#}$ - Quadril.

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}864$, o que denota que 86,4% da variância dos dados é explicada pelo modelo. A redução em relação ao modelo inicial é despresível, logo a retiradas das variáveis não afetou o modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}86$

Gráficos de Diagnóstico

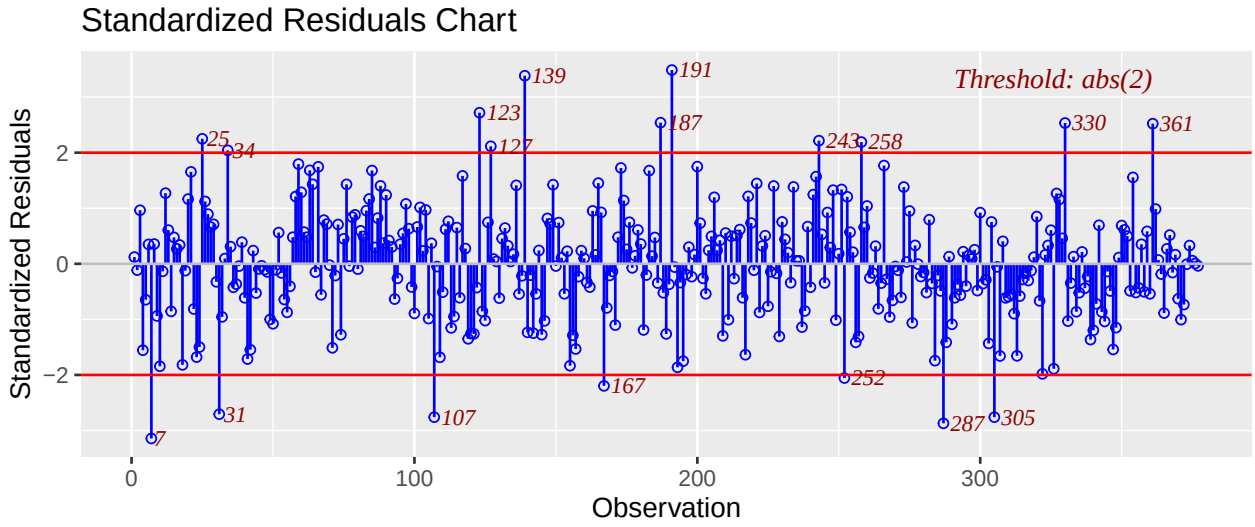
A análise dos gráficos de diagnóstico permite avaliar as observações realizadas e conhecer a influência de cada uma delas para o madelo de regressão proposto. Assim, com base no último modelo, é possível fazer as seguintes análises:

Figura 2: Valores Ajustados e Resíduos Studentizados



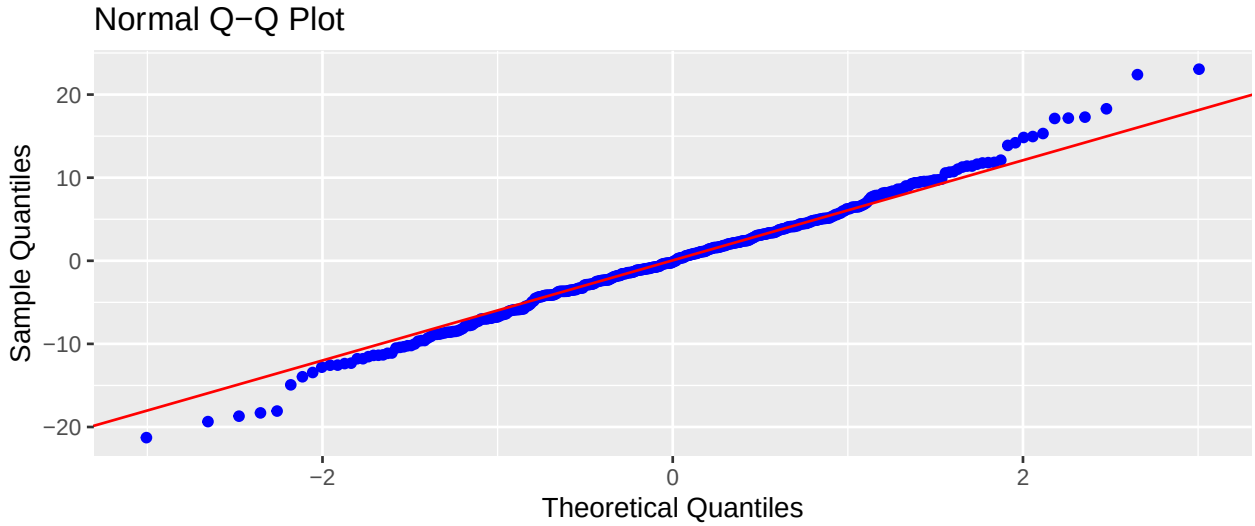
A Figura 2 demonstra que os resíduos estão todos dentro dos limites esperados, com exceção da observação 7 que por pouco ultrapassou o limite inferior. Não parece ser o caso de nenhuma intervenção por conta deste valor.

Figura 3: Valores Ajustados e Resíduos Padronizados.



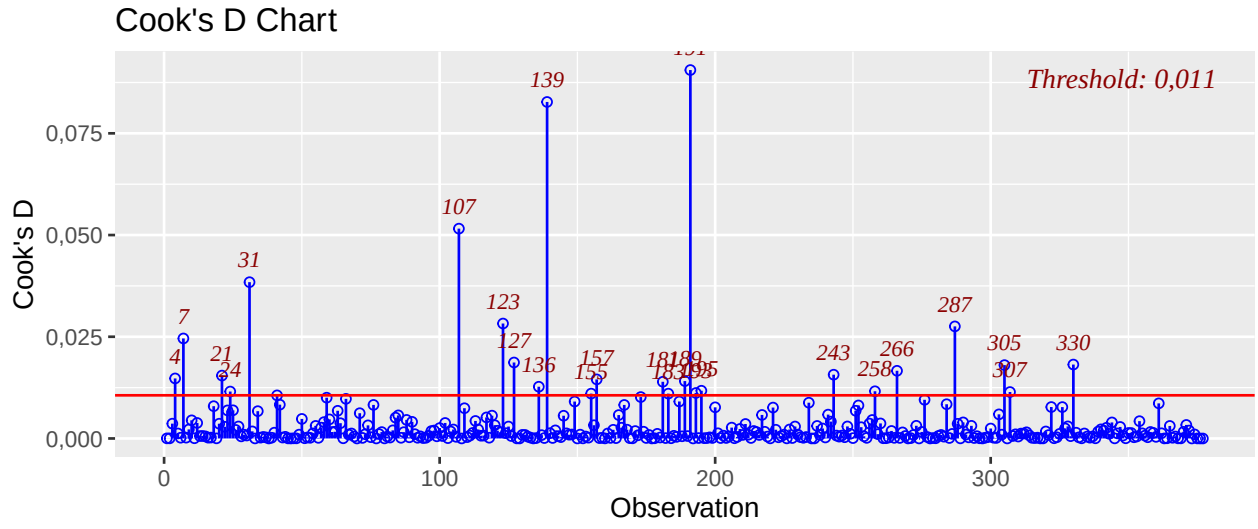
Já a análise da Figura 3, onde os resíduos foram padronizados, o número de observações que ultrapassaram os limites chegou a 4,8% do total o que é condizente com uma confiança de 95%.

Figura 4: Análise dos quantis teóricos e amostrais



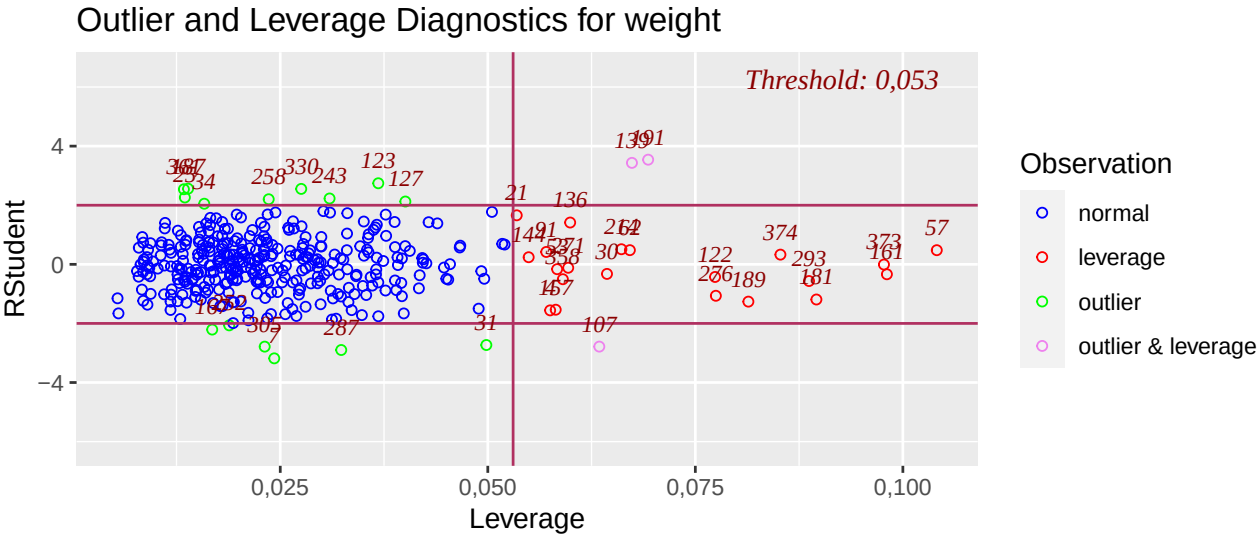
A Figura 4 apresentou um bom ajuste dos resíduos à distribuição normal, sendo um pouco pior nas caudas da distribuição por uma qauntidade pequena de pontos.

Figura 5: Distância de Cook.



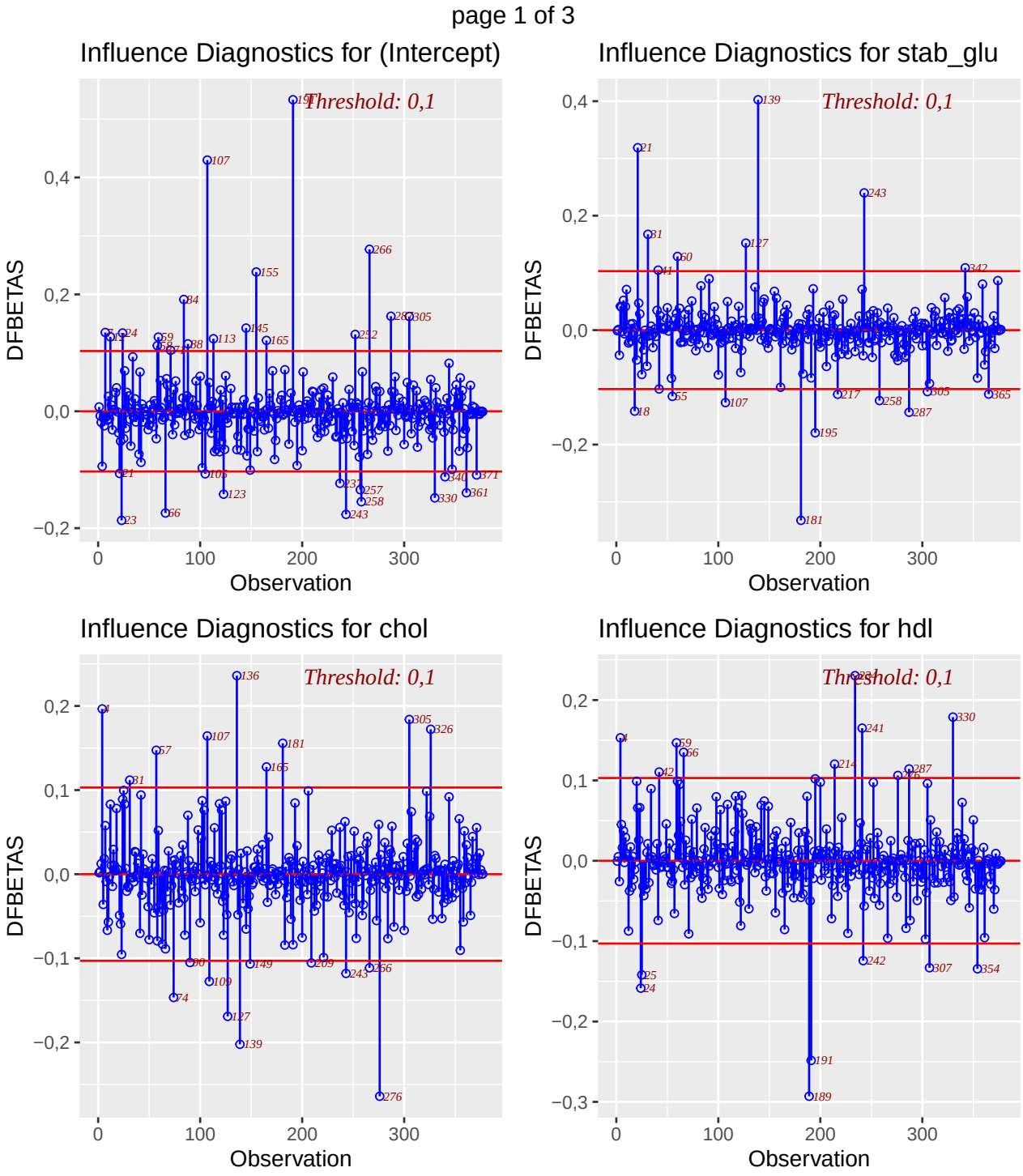
A analise da distância de Cook apresentada na Figura 5 demonstra que 25 (6,6%) observações tem uma distância expressiva, mas apenas sete deles estão acima da distância de 0,025. O tretamento destes pontos pode manter os resíduos dentro do esperado com uma confiança de 95%.

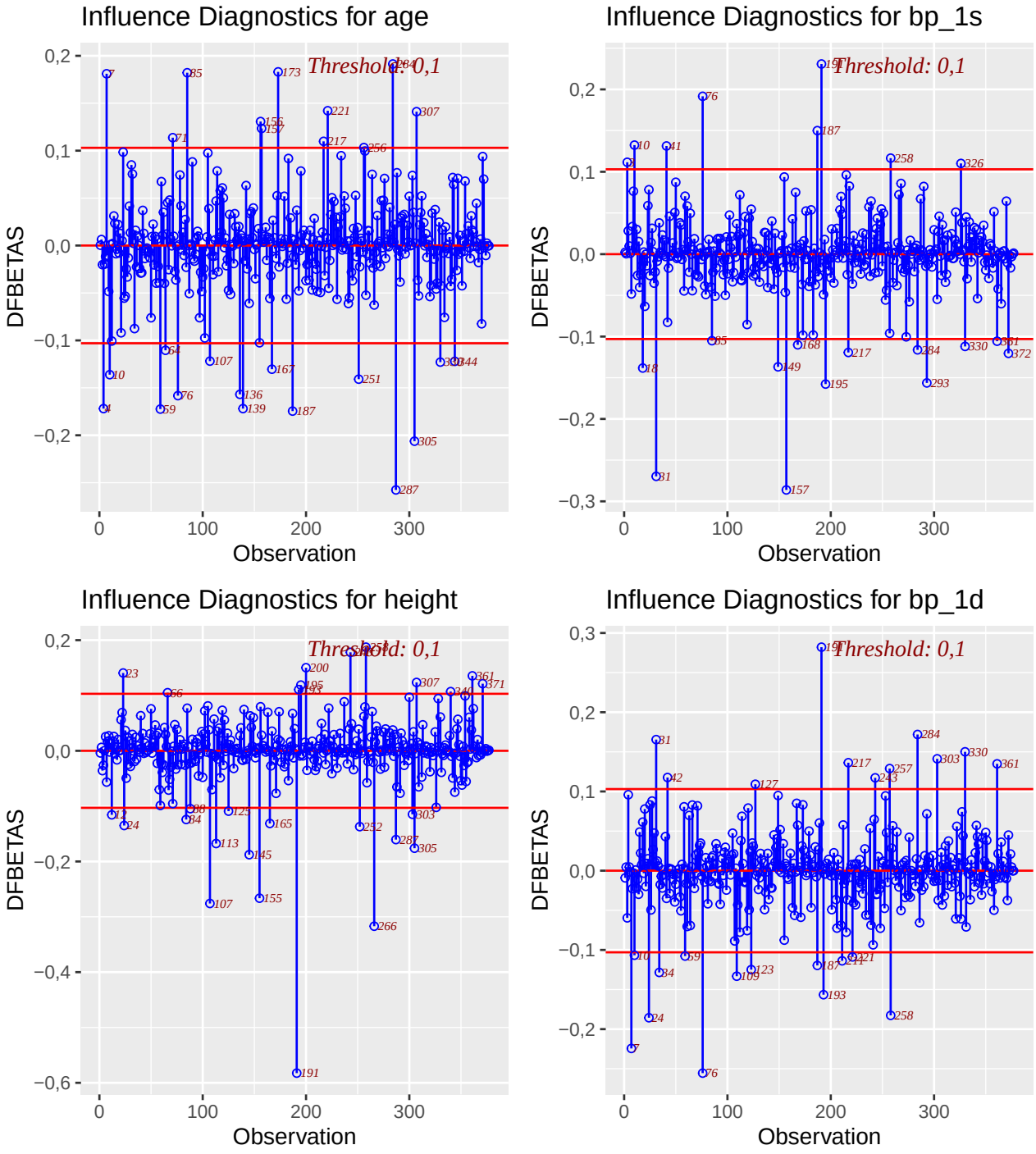
Figura 6: Análise dos pontos de Alavanca e Resíduo Studentizado.

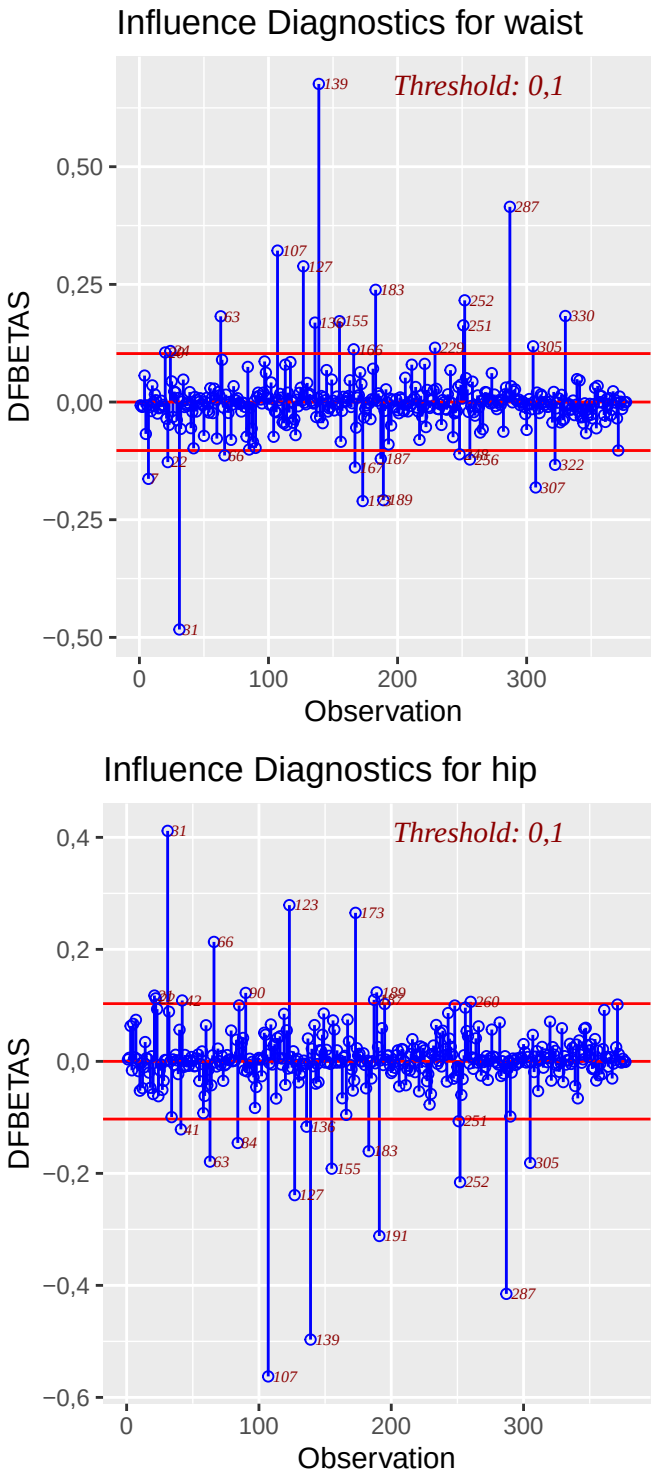


Pela Figura 6, observamos 15 observações que podem ser consideradas como *Outliers* e 21 como observações de alavanca, além de 3 com as duas características, o que representa um total de 10,3% das observações. Uma quantidade tão expressiva de dados não pode ser descartada sem o amparo de um especialista na área.

Figura 7: DFBetas para as variáveis do modelo.

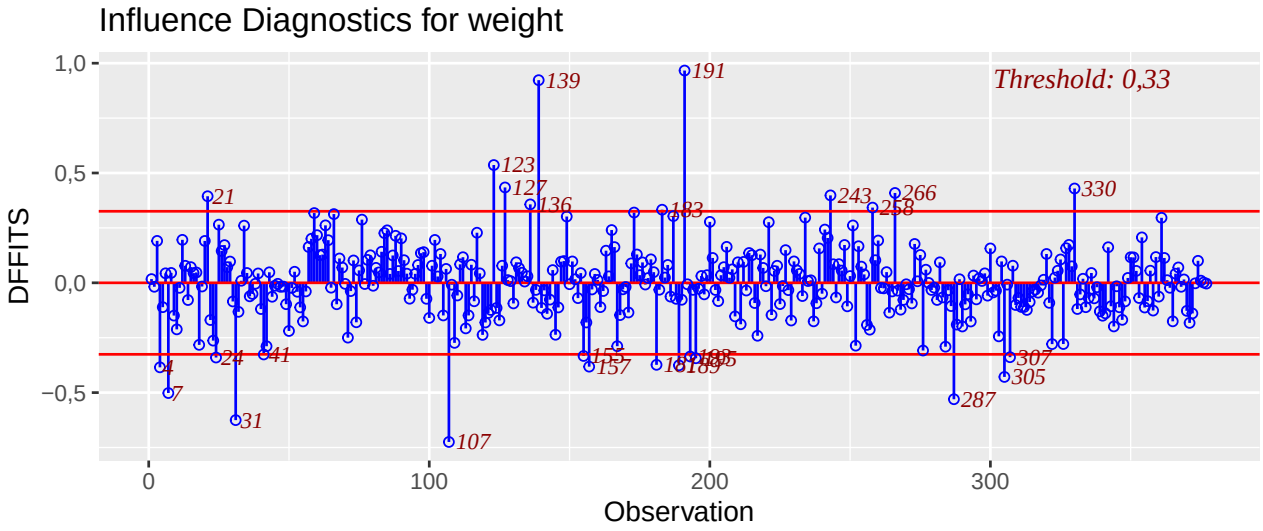



Influence Diagnostics for age
Threshold: 0,1
DFBETAS
Observation
Influence Diagnostics for bp_1s
Threshold: 0,1
DFBETAS
Observation
Influence Diagnostics for height
Threshold: 0,1
DFBETAS
Observation
Influence Diagnostics for bp_1d
Threshold: 0,1
DFBETAS
Observation



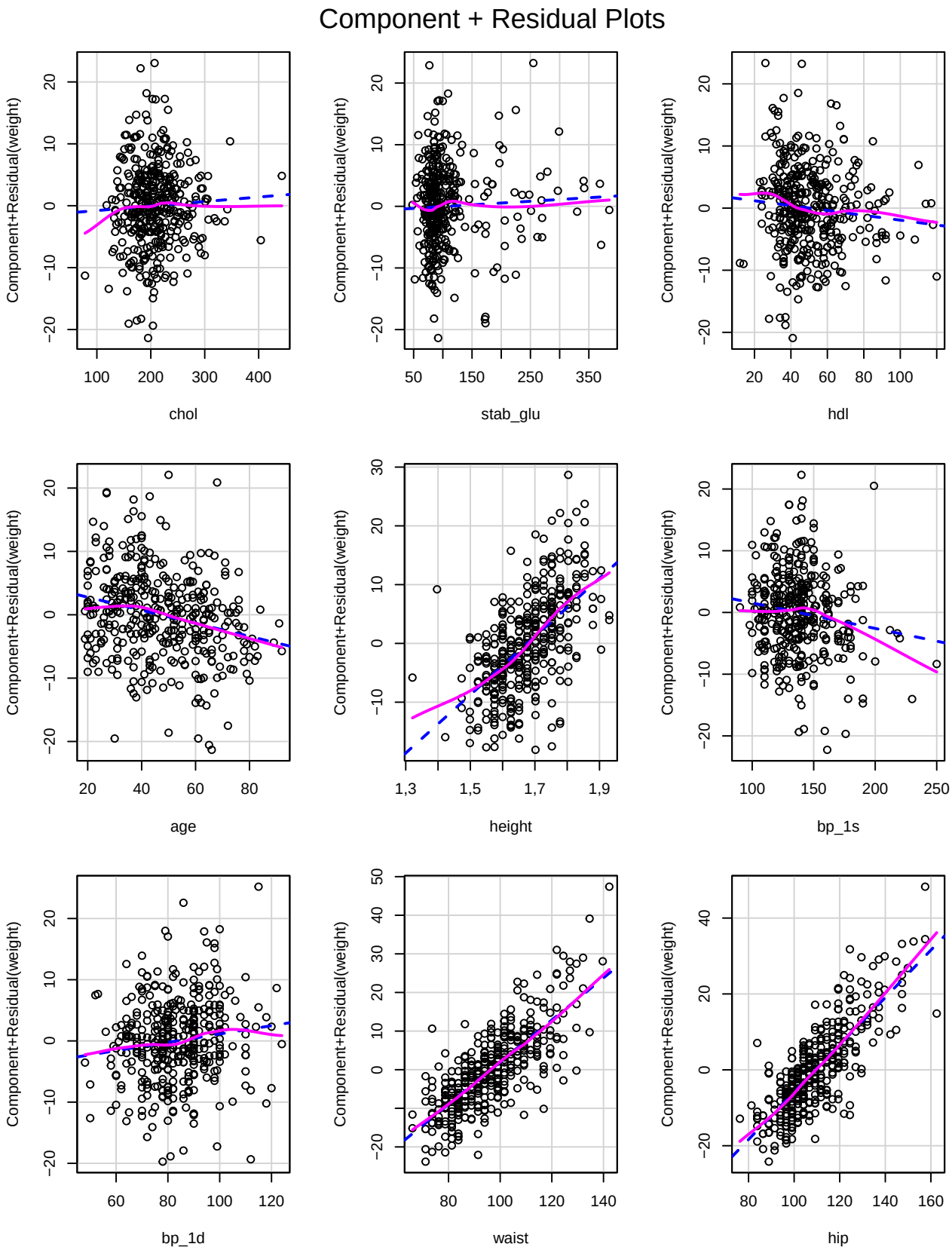
A Figura 7 apresenta os DFBetas para cada uma das variáveis utilizadas no modelo, com uma média de 6,1% de observações discrepantes com cerca de 4 observações críticas, isto é, valores mais extemos. O tratamento destas observações podem trazer o modelo para uma situação mais compatível com a confiança estabelecida.

Figura 8: DfFit para as variáveis do modelo.



A Figura 8 acompanha os gráficos anteriores apresentando 6,9% de observações discrepantes, mas apenas seis valores são extremos, desta forma pode-se da mesma forma tratá-los e manter a confiança do modelo.

Figura 9: COVRatio para as variáveis do modelo.



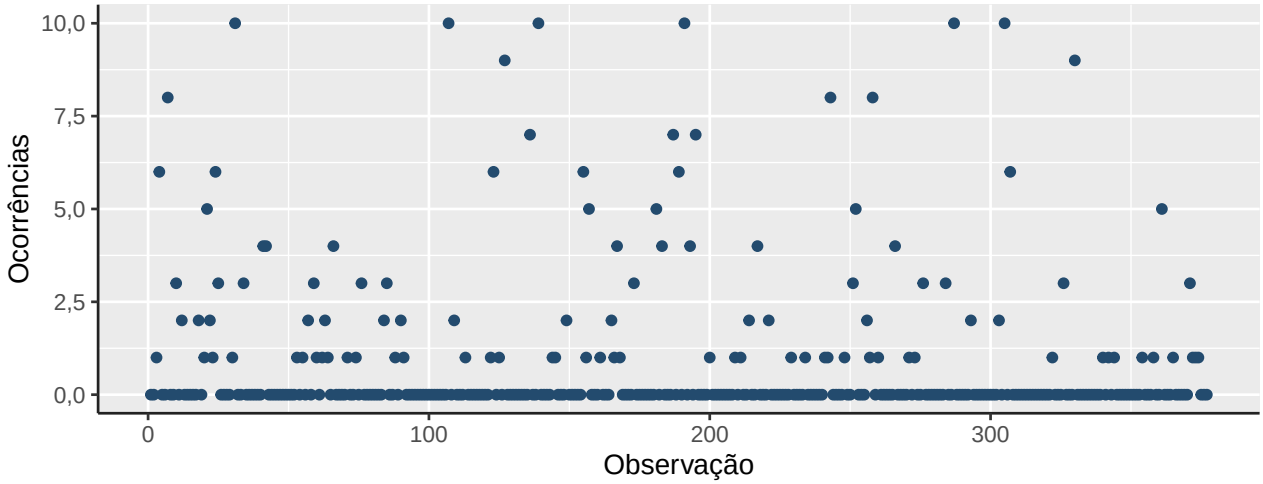
Da Figura 9 verifica-se que as variáveis “Altura”, “Cintura” e “Quadil” estão mais diretamente

correlacionadas com os resíduos do “Peso”, o que por sua vez indica que a inclusão de observações destas variáveis apresentam maior impacto ao modelo.

Eliminação de observações anômalas

Avaliando as observações que apresentaram comportamento anômalo nos diagnósticos dos valores ajustados e resíduos studentizados, valores ajustados e resíduos padronizados, distância de CooK, pontos de alavanca e *outliers*, análise de DfFit e todas as análises de BFBetas, chegamos as frequências de observações anômalas apresentadas na Figura 10.

Figura 10: Número de ocorrências para cada observação



Considerando apenas as observações com mais de 4 ocorrências temos a Tabela a seguir.

Tabela 4: Observações com maior número de ocorrências.

Observação	Ocorrências
4	6
7	8
21	5
24	6
31	10
107	10
123	6
127	9
136	7
139	10
155	6
157	5
181	5
187	7
189	6
191	10
195	7
243	8
252	5
258	8
287	10
305	10
307	6
330	9
361	5

Podemos intuir que essas são as observações com maior impacto negativo no modelo. Logo, eliminando-as do conjunto de dados analizados chegamos a um novo modelo dado por:

$Y_i^{¤}$ = -124,335+ 0,001 $X_{1i}^{¤}$+ 0,003 $X_{2i}^{¤}$ -0,042 $X_{3i}^{¤}$ -0,067 $X_{4i}^{¤}$+ 52,788 $X_{5i}^{¤}$ -0,04 $X_{6i}^{¤}$+ 0,071 $X_{7i}^{¤}$+ 0,458 $X_{8i}^{¤}$+ 0,701 $X_{9i}^{¤}$

Onde:
$Y_i^{¤}$ - Peso;
$X_{1i}^{¤}$ - Colesterol total;
$X_{2i}^{¤}$ - Glicose estabilizada;
$X_{3i}^{¤}$ - Lipoproteína de alta densidade;
$X_{4i}^{¤}$ - Idade;
$X_{5i}^{¤}$ - Altura;
$X_{6i}^{¤}$ - Pressão sanguínea sistólica;

$X_{7i}^{¤}$ - Pressão sanguínea diastólica;
$X_{8i}^{¤}$ - Cintura;
$X_{9i}^{¤}$ - Quadril.

Neste novo modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}889$, o que denota que 88,9% da variância dos dados é explicada pelo modelo. O valor deste novo coeficiente permite concluir que a eliminação das observações com maior impacto no modelo foi benéfica. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}885$

Conclusões

Verificou-se que a análise de variâncias foi um teste mais poderoso para identificar variáveis desnecessárias ao modelo que a analise individual das significancias das variáveis ao modelo.

Embora se não se tenha um conhecimento específico da área estudada, foi possível realizar uma avaliação dos dados recebidos e propor um tratamento que efetivamente melhorou o modelo de regressão linear multipla realizado.

As anomalias relatadas em cada um dos gráficos de diagnóstico elaborados foram tratadas de igual maneira contabilizando para cada observação o número de ocorrências observadas. Por este método se elencou as observações com maior potencial de prejuízo ao modelo e ao descartá-las do rol de dados avaliados obteve-se uma expressiva melhora no modelo.