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.

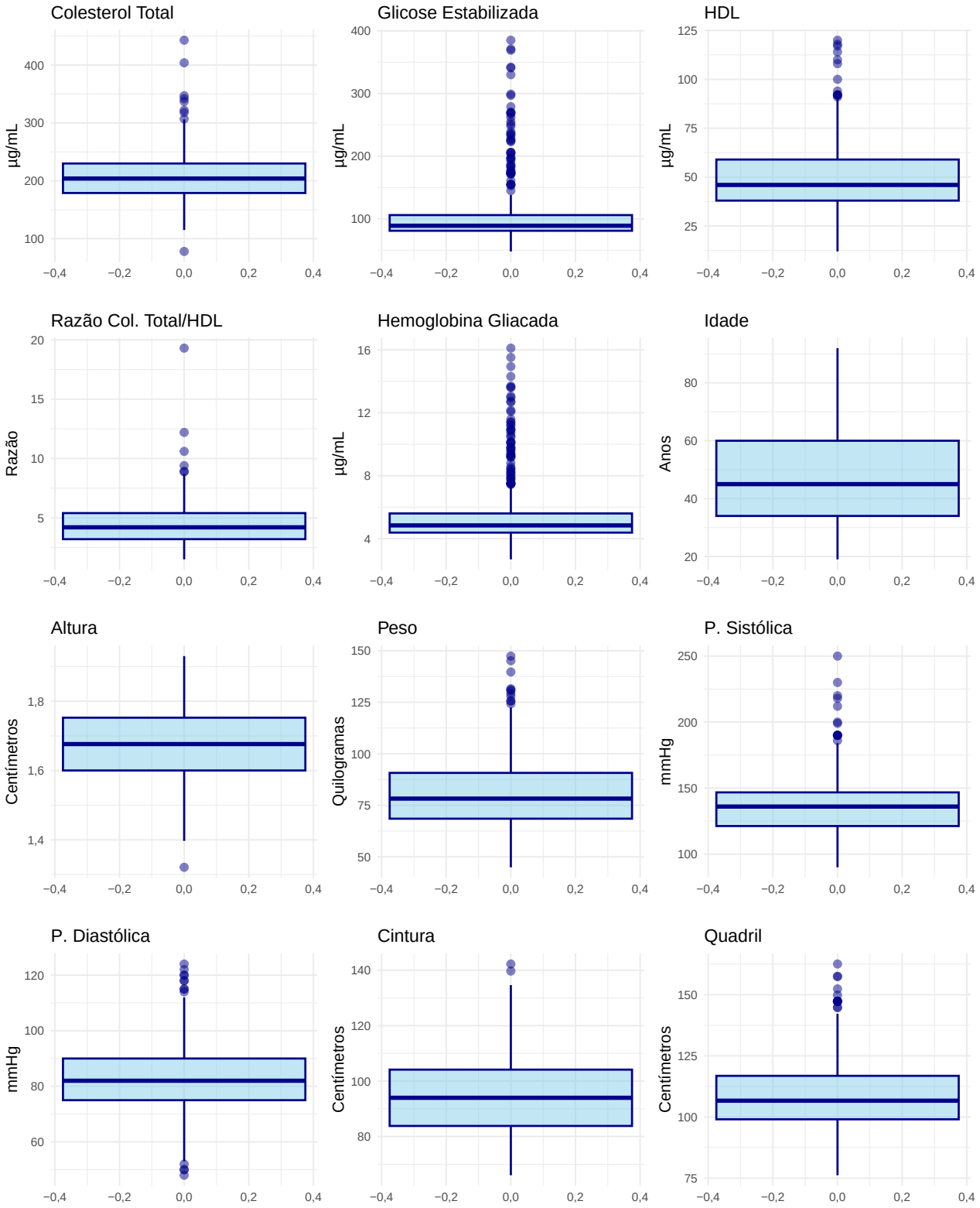
Tabela 1: Medidas Resumo dos dados

	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 = 0,015 X_{1i}+ 0,021 X_{2i} -0,187 X_{3i} -0,975 X_{4i} -0,24 X_{5i} -0,164 X_{6i}+ 1,107 X_{7i} -0,069 X_{8i}+ 0,046 X_{9i}+ 0,726 X_{10i}+ 0,295 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 = 0,783, o que denota que 78,3% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a R^2_a = 0,776.

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

	Estatística t
Colesterol total	0,780
Glicose estabilizada	1,660
Lipoproteína de alta densidade	-3,195
Razão colesterol total e colesterol bom	-1,468
Hemoglobina glicada	-0,758
Idade	-4,804
Altura	0,427
Pressão sanguínea sistólica	-2,409
Pressão sanguínea diastólica	1,025
Cintura	12,159
Quadril	5,387

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 = -0,049 X^*_{1i}+ 0,031 X^*_{2i}+ 2,955 X^*_{3i}+ 0,049 X^*_{4i}+ 33,032 X^*_{5i}+ 0,222 X^*_{6i}

Onde:
Y_i - Peso;
X^*_{1i} - Colesterol total;
X^*_{2i} - Glicose estabilizada;
X^*_{3i} - Razão colesterol total e colesterol bom;

X^*_{4i} - Hemoglobina glicada;
X^*_{5i} - Altura;
X^*_{6i} - Pressão sanguínea diastólica.

Análise de Variâncias

Realizando uma análise de variância com as variáveis significativas, 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	1,4147	0,2350
Glicose estabilizada	1	3.722,204	3.722,2039	13,0504	0,0003
Razão colesterol total e colesterol bom	1	8.113,832	8.113,8320	28,4479	<0,0001
Hemoglobina glicada	1	2,053	2,0527	0,0072	0,9324
Altura	1	5.590,041	5.590,0414	19,5992	<0,0001
Pressão sanguínea diastólica (1ª medida)	1	3.527,244	3.527,2442	12,3669	0,0005
Resíduos	370	105.530,465	285,2175		

Legenda:
[1] GL: Graus de Liberdade

A análise da Tabela 3 permite avaliar que apenas para a razão colesterol total e colesterol bom e para a altura o peso do paciente está relacionado aos resultados obtidos de forma significativa. Para todos os demais a correlação não é evidente.

Gráficos de Diagnóstico

Figura 2: Valores Ajustados e Resíduos Studentizados

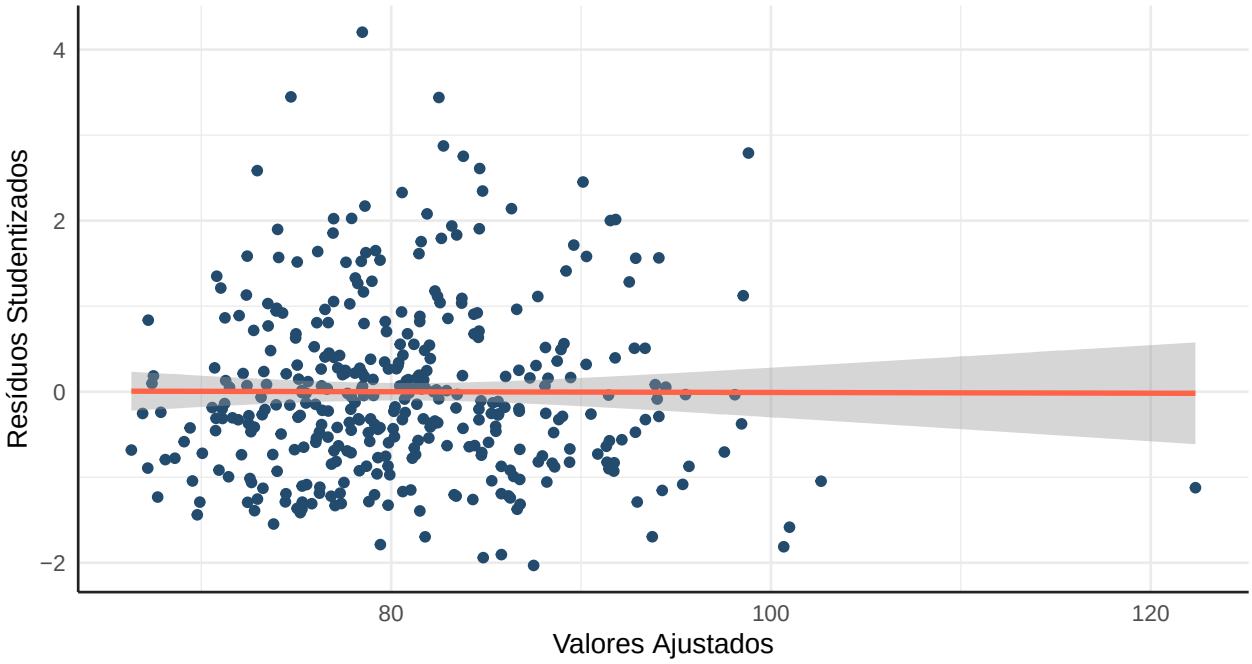
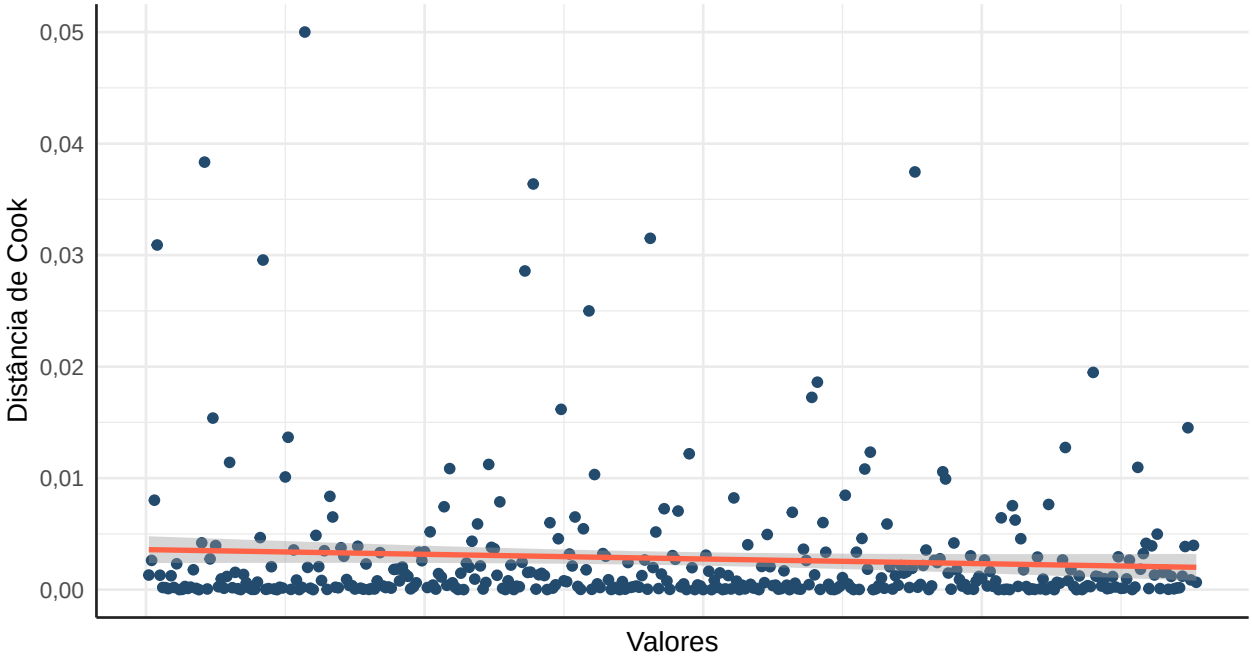


Figura 4: Distância de Cook



Conclusões