Doenças entre Afroamericanos

Fernado Bispo, Jeff Caponero

Sumário

Introdução

O conjunto de dados a ser analisado, 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 polegadas);
- 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 polegadas);
- Quadril (em polegadas).

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

Numa primeira análise do conjunto de dados se constatou que 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, ou seja, das 403 observações coletadas, 262 estão ausentes para estas características, diante desta grande falta de dados essas características serão descartadas.

Se constatou também que as observações das características **altura, peso, cintura** e **quadril** estão representadas em unidades do Sistema Imperial, diferentes das praticadas no Brasil, sendo necessária a conversão para o Sistema Internacional, a fim de facilitar a interoperabilidade para a nossa realidade cotidiana. Para tanto as características que possuem medidas originais em **polegadas (in), libras (lb), polegadas (in) e polegadas (in)** respectivamente, foram convertidas para **metro (m), quilograma (kg), centímetro (cm) e centímetro (cm)** respectivamente.

Sendo parte primordial para qualquer estudo, a fase exploratória dos dados está representada inicialmente na Tabela 8 com a sumarização das características em análise.

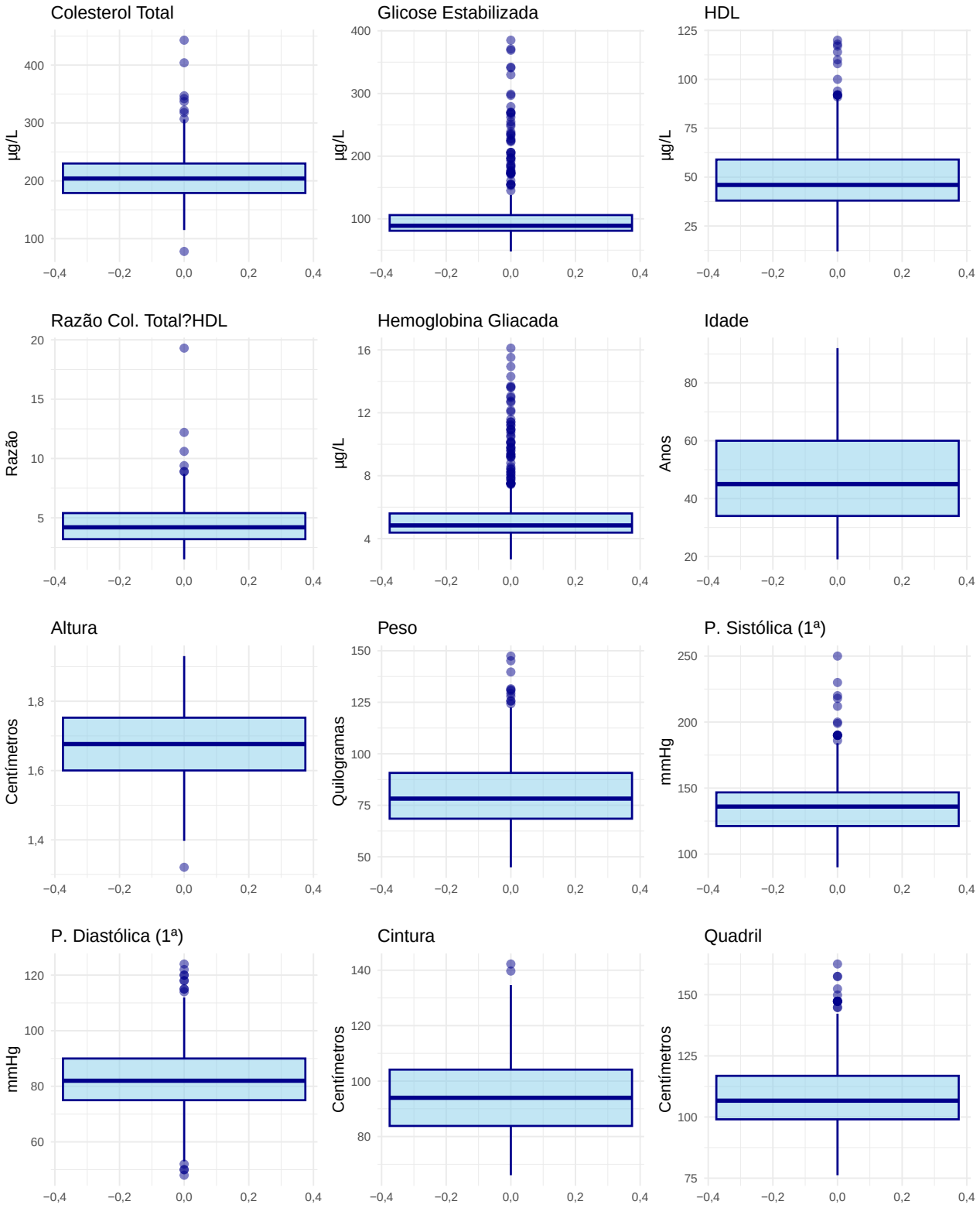
Tabela 1: Medidas Resumo dos dados

	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 (1ª medida)	48,00	75,00	82,00	83,43	92,00	124,00	13,53	0,16	0,22	0,07
Pressão sanguínea sistólica (1ª medida)	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 inicial, verifica-se que a distribuição das variáveis não apresenta fatores impeditivos da regressão linear a que nos propomos.

Pode-se realizar a 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.

Regressão Linear Múltipla

Com base na proposta da coleta de dados, a característica que apresenta melhor adequação para ser a variável resposta de um modelo de regressão é o **peso**, pois intuitivamente é a que apresenta melhor correlação. Desta forma, será feita a avaliação dessa característica com as demais a fim de se identificar as possíveis correlações intuídas.

Tabela 2: 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 (1ª medida)	1	1.873,619	1.873,6191	39,8330	<0,0001
Pressão sanguínea diastólica (1ª medida)	1	1.775,081	1.775,0806	37,7380	<0,0001
Cintura	1	76.162,631	76.162,6313	1.619,2098	<0,0001
Quadril	1	7.814,820	7.814,8199	166,1423	<0,0001
Resíduos	365	17.168,474	47,0369		

Legenda:
[1] GL: Graus de Liberdade

A análise da Tabela 2 permite avaliar que apenas para a razão colesterol total e colesterol bom e para a Hemoglobina glicada o peso do paciente está relacionado aos resultados obtidos de forma significativa. Para todos os demais a correlação não é evidente. 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 (colesterol bom);
X_{4i} - Razão colesterol total e colesterol bom;
X_{5i} - Hemoglobina glicada;
X_{6i} - Idade (em anos);
X_{7i} - Altura (em polegadas);
X_{8i} - Pressão sanguínea sistólica (1ª medida);

X_{9i} - Pressão sanguínea diastólica (1^{a} medida);
X_{10i} - Cintura (em polegadas);
X_{11i} - Quadril (em polegadas).