Relatório Do Trabalho Prático

Regressão linear múltipla

Estatística aplicada 2019/2020

José Pedro Ribeiro Ferreira Pinto – 201603713

1-Descrição do problema

O problema em questão tratasse da utilização de regressão linear múltipla para tentar prever os valores de uma variável objetivo com o uso de diversas variáveis explicativas. Tal foi proposto para obter experiência pratica de regressão linear em bases de dados reais e complexas, compreender as dificuldades desta tarefa, os passos a realizar e as limitações inerentes, tanto ao modelo como aos dados.

Com este objetivo em mente foi obtida uma base de dados cujas especificidades serão descritas posteriormente.

1.1-Base de dados

A base de dados obtida foi uma base de dados de doenças cardiovasculares encontrada no kaggle https://www.kaggle.com/sulianova/cardiovascular-disease-dataset/data.

Esta base de dados é constituída por 11 variáveis explicativas de interesse teórico e 1 variável alvo binaria, correspondente há existência ou inexistência de uma doença cardiovascular do individuo em causa. Há no total 70.000 entradas, um valor muito superior ao necessário para uma análise de regressão linear.

Todos os valores foram obtidos no momento de uma consulta medica e como tal poderão não representar a população em geral (pessoas geralmente saudáveis menos frequentemente procuram consultas medicas).

Nota: Após várias tentativas de encontrar a fonte dos dados, tal não foi conseguido, como tal não foi possível verificar a autenticidade dos dados e todas as conclusões tiradas a partir da análise terão de ser cuidadosamente analisadas. No entanto para o objetivo prático do trabalho esta base de dados serve perfeitamente.

1.1.1-Descrição das variáveis

As variáveis, tal como previamente mencionado, são 12 no total, com 11 dessas explicativas e uma delas alvo. Estas podem ser divididas em 3 categorias. As de natureza objetiva (valores factuais), as de exame (valores obtidos em exames médicos) e as subjetivas (providenciadas pelo individuo a ser examinado).

As variáveis em questão são:
Age: a idade do individuo em dias. Valor objetivo e inteiro.
Height: a altura do individuo em cm. Valor objetivo e inteiro.
Weight: o peso do individuo em kg. Valor objetivo e decimal.
Gender: o género do individuo. Valor objetivo e inteiro com significado categórico – 1-mulher,2-homem.
Ap-hi: valor da pressão arterial sistólica. Valor de exame e inteiro.

Ap-lo: valor da pressão arterial diastólica. Valor de exame e inteiro.
Cholesterol: valor do colesterol do individuo. Valor de exame e inteiro com significado categórico – 1-normal,2-acima do normal,3-muito acima do normal.
Gluc: valor da glucose do individuo. Valor de exame e inteiro com significado categórico – 1-normal,2-acima do normal,3-muito acima do normal.
Smoke: se o individuo fuma ou não. Valor subjetivo e booleano - 0-não fuma,1-fuma.
Alco: se o individuo consume bebidas alcoólicas ou não. Valor subjetivo e booleano – 0-não consume,1-consume.
Active: se o individuo é ou não fisicamente ativo. Valor subjetivo e booleano – 0-não é ativo,1-é ativo

Valor objetivo

Cardio: presença ou não de doença cardiovascular. Valor booleano – 0-sem doença,1-presença de doença.

1.1.2-Tabela das variáveis

Abaixo encontrasse uma pequena tabela com algumas das entradas da base de dados para se poder facilmente ter uma ideia da sua composição.

Id	Age	Gender	Height	Weight	Ap_hi	Ap_lo	Cholesterol	Gluc	Smoke	Alco	active	cardio
0	18393	2	168	62	110	80	1	1	0	0	1	0
1	20228	1	156	85	140	90	3	1	0	0	1	1
2	18857	1	165	64	130	70	3	1	0	0	0	1
3	17623	2	169	82	150	100	1	1	0	0	1	1
4	17474	1	156	56	100	60	1	1	0	0	0	0
8	21914	1	151	67	120	80	2	2	0	0	0	0

Tabela1-Visualização dos dados de doenças cardiovasculares.

2-Descrição e visualização dos dados (a)

Abaixo é apresentado um sumario da distribuição dos dados. Apresentando o mínimo, máximo, media, mediana, quartis e variância.

	Age	Gender	Height	Weight	Ap_hi	Ap_lo	Cholesterol	Gluc	Smoke	Alco	active	cardio
minimo	10798	1	55	10	-150	-70	1	1	0	0	0	0
1º Qu.	17664	1	159	65	120	0	1	1	0	0	1	0
mediana	19703	1	165	72	120	80	1	1	0	0	1	0
media	19469	1.35	164.4	74.21	128	96.63	1.367	1.23	0.088	0.054	0.80	0.49
3º Qu.	21327	1	170	82	140	90	2	1	0	0	1	1
maximo	23713	2	250	200	16020	11000	3	3	1	1	1	1
Variância	6087331	0.23	67.41	207.24	23719	35521	0.46	0.33	0.08	0.05	0.16	0.25

Tabela2-Sumario dos dados.

Na tabela acima (Tabela 2) é possível observar algumas medidas importantes. Começando pelo mínimo, é possível constatar que a idade mínima é 10798 dias, cerca de 30 anos. Há alguns valores mínimos preocupantes, a altura de 55cm, por exemplo, para pessoas de 30 anos é bastante baixa, e de facto quase impossível, uma vez que apenas um adulto no mundo tem uma altura inferior a 55cm[1]. Outro valor preocupante é o do peso, 10kg, que também é absurdo. Ap_hi e ap_lo também tem valores preocupantes, valores negativos, que estão obviamente

errados. Enquanto que os valores de altura e peso serão mantidos, os valores negativos de ap_hi e ap_lo serão removidos.

Os valores do 1º quartil, mediana e 3º quartil não fornecem muita informação sobre a distribuição e como tal não serão analisados em detalhe.

Os valores da média fornecem informação sobre as percentagens nas variáveis binarias. É possível constatar que 35% dos indivíduos são mulheres, 8.8% fumadores, 5.4% consomem álcool, 80% são ativos e 49% possuem doenças cardiovasculares. Nenhum dos valores apresentado é preocupante.

De entre os valores máximos também há alguns preocupantes, nomeadamente os de ap_hi e ap_lo. Os máximos previamente mencionados excedem os 10000, o que é simplesmente impossível. Os valores demasiado elevados foram como tal removidos.

2.1-Visualização de dados

2.1.1-Diagramas de extremos e quartis e histogramas

Nesta subsessão serão mostrados e analisados uma serie de gráficos de dispersão e localização dos dados, como por exemplo, diagramas de extremos e quartis e histogramas.

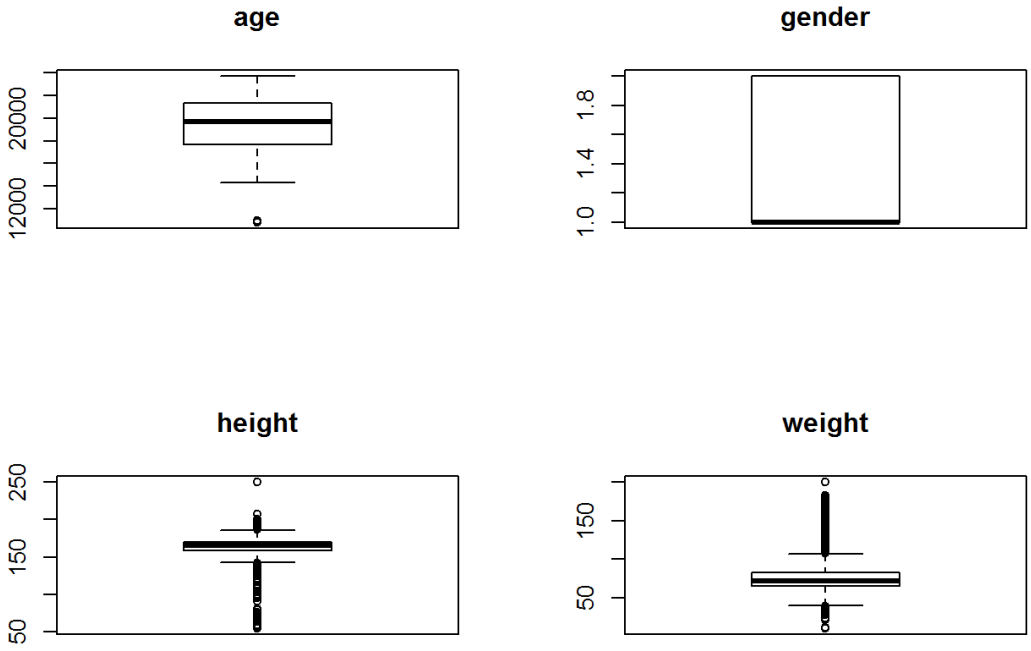


Imagem1- diagramas de extremos e quartis para age, gender, height e weight

Os diagramas acima confirmam o que já foi possível observar do sumario. Também é possível observar alguma informação nova, o facto de a distribuição da altura ser enviesada para a direita, enquanto que a do peso é enviesada para a esquerda.

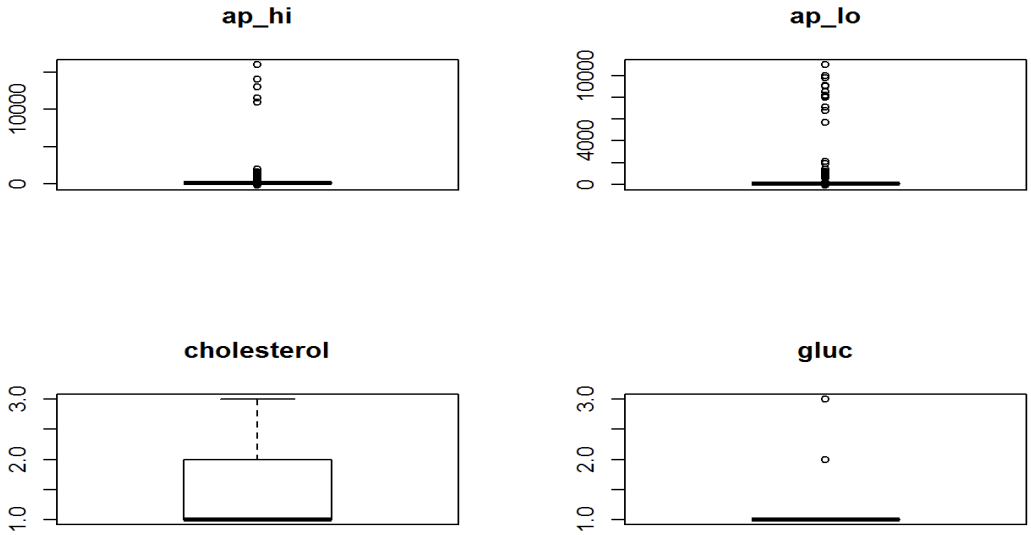


Imagem2- diagramas de extremos e quartis para ap_hi, ap_lo, cholesterol e gluc

Mais uma vez nos diagramas acima é possível observar as conclusões tiradas anteriormente. Os outliers severos do ap_hi e ap_lo são facilmente identificados. Outra observação importante é o facto de apesar de colesterol normal ser o mais comum, existem observações suficientes do elevado e muito elevado de forma a que estas não sejam consideradas outliers. O mesmo não acontece com a glucose, também o valor normal ser o mais comum, mas os outros dois terem suficientemente poucas observações para estas serem consideradas outliers.

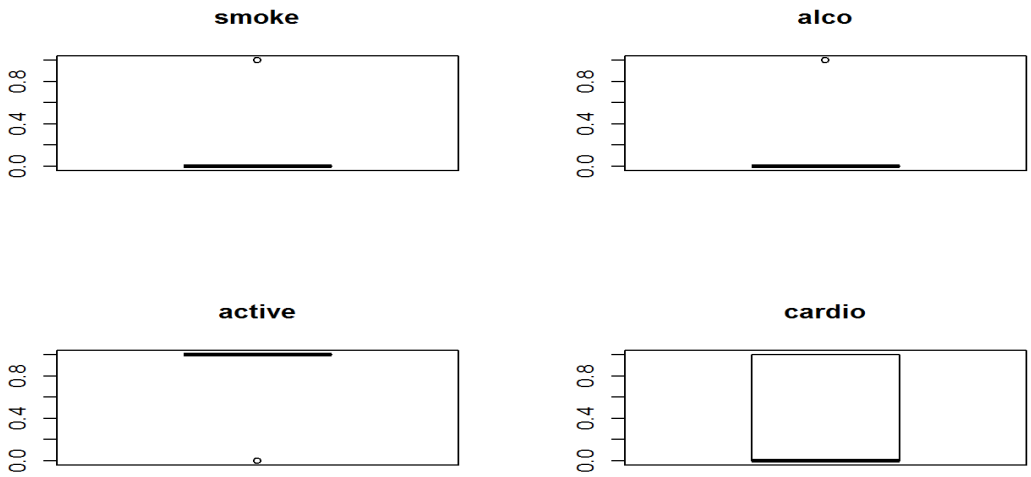


Imagem3- diagramas de extremos e quartis para smoke, alco, active e cardio

Este último conjunto de diagramas não apresenta muita informação. As três primeiras variáveis (smoke, alco e active) estão numa situação semelhante a glucose, com um valor mais comum e suficientemente poucos dos outros valores de tal forma que estes sejam considerados outliers.

O mesmo não acontece com o cardio, que não possui outliers. Tal já era de esperar dado que a percentagem de cada tipo é quase exatamente 50%.

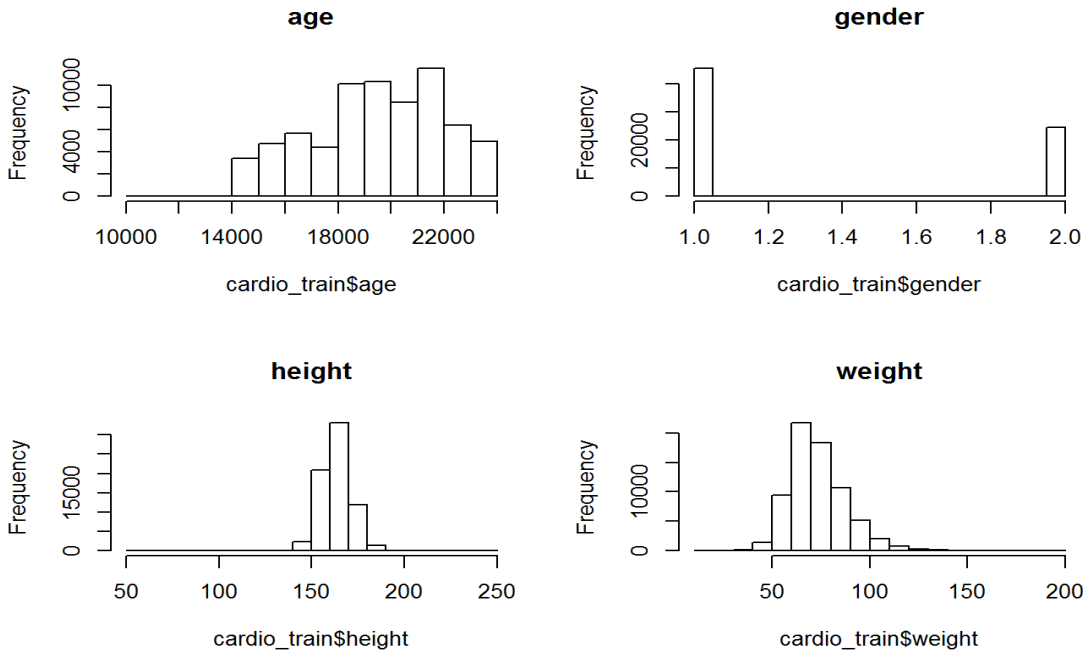


Imagem4- histogramas para age, gender, height e weight

Dos histogramas acima é possível observar que nenhuma das variáveis aparenta seguir uma distribuição normal, com a variável de idade extremamente achatada e a de peso muito enviesada para a direita.

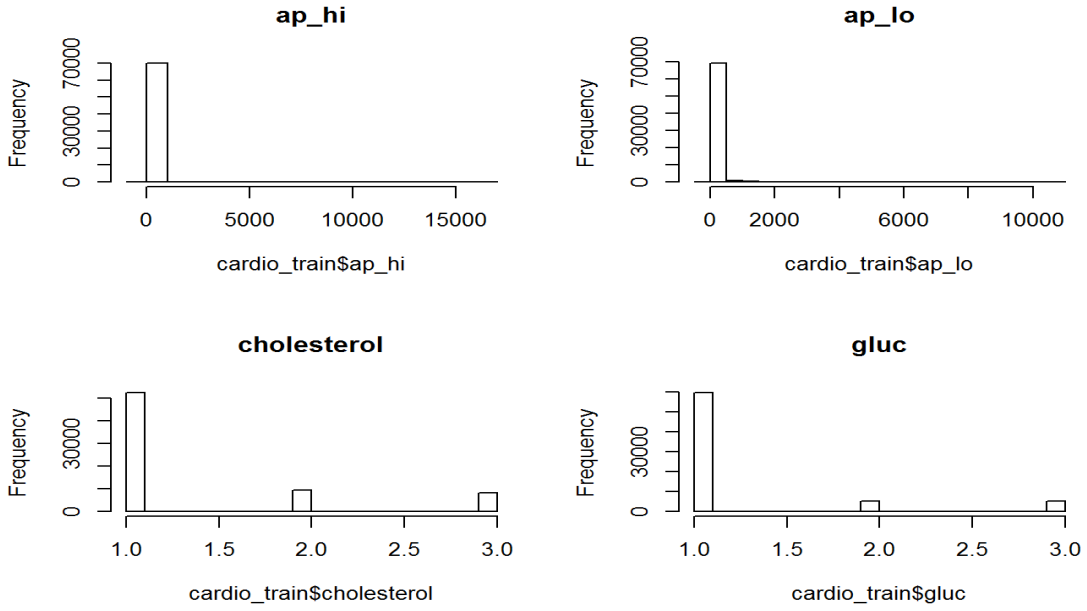


Imagem5- histogramas para ap_hi, ap_lo, cholesterol e gluc

De nenhum dos histogramas acima é possível obter novas informações. Para ap_hi e ap_lo os outliers escondem a distribuição. Enquanto que para colesterol e glucose apenas é possível observar a frequência de cada tipo de ocorrência.

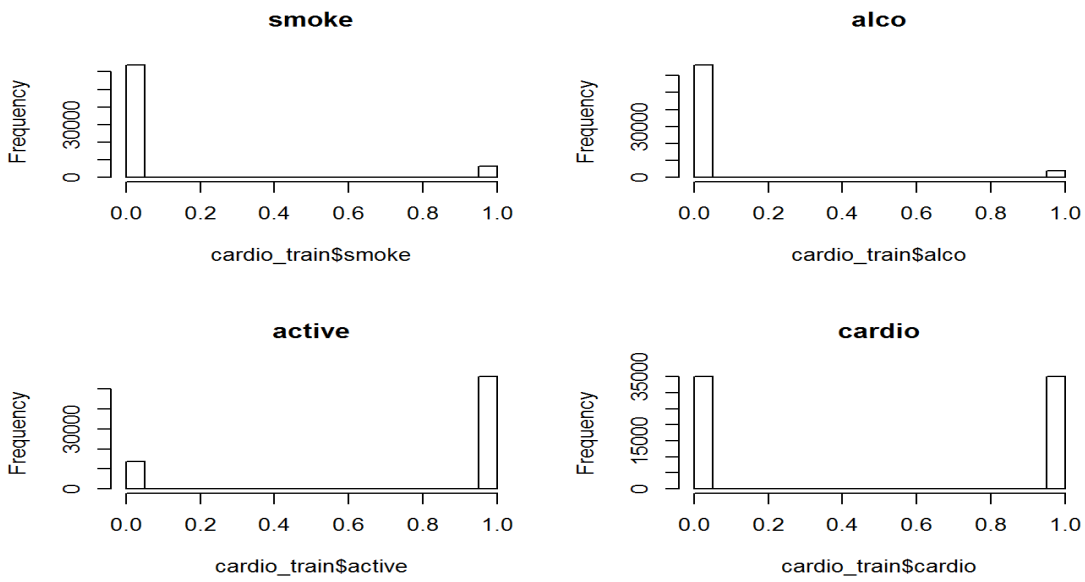


Imagem6- histogramas para smoke, alco, active e cardio

Uma vez mais os histogramas não permitem obter mais informação do que a previamente existente.

Para as variáveis categóricas e booleanas um gráfico de barras seria mais apropriado, mas uma vez que a informação apresentada é equivalente, histogramas foram utilizados para todas as variáveis para simplificar o código.

2.1.2-Diagramas e testes de normalidade

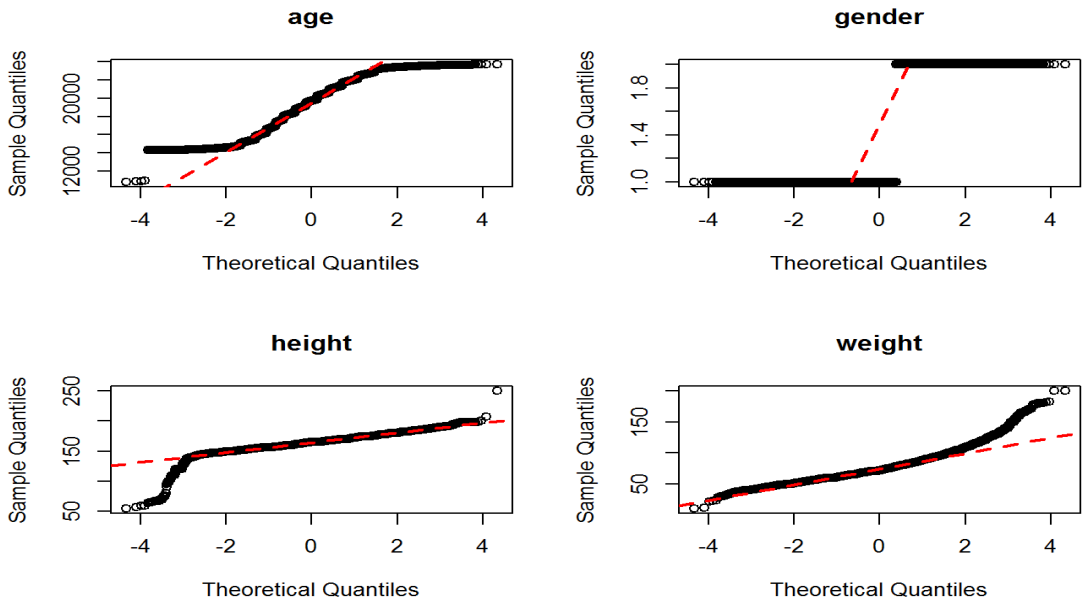


Imagem7- gráficos de normalidade para age, gender, height e weight

Todos os gráficos apresentão fortes desvios da normalidade, com o peso e altura sendo os mais perto de ser normais.

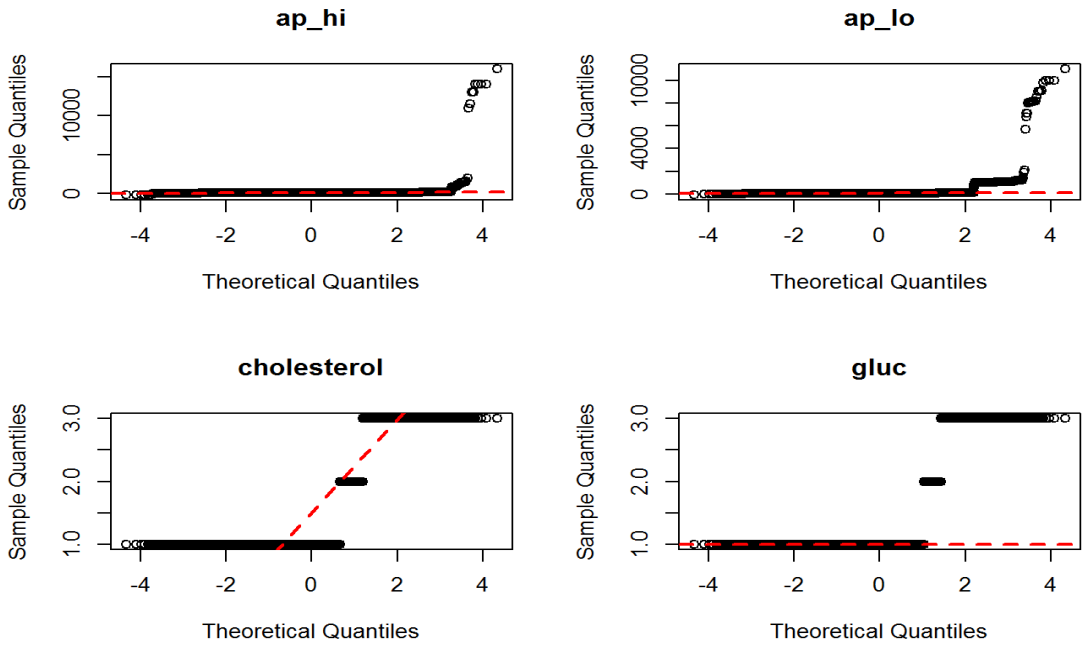


Imagem8- gráficos de normalidade para ap_hi, ap_lo, cholesterol e gluc

Mais uma vez os gráficos vão contra a ideia da normalidade das distribuições, todos eles com desvios sérios.

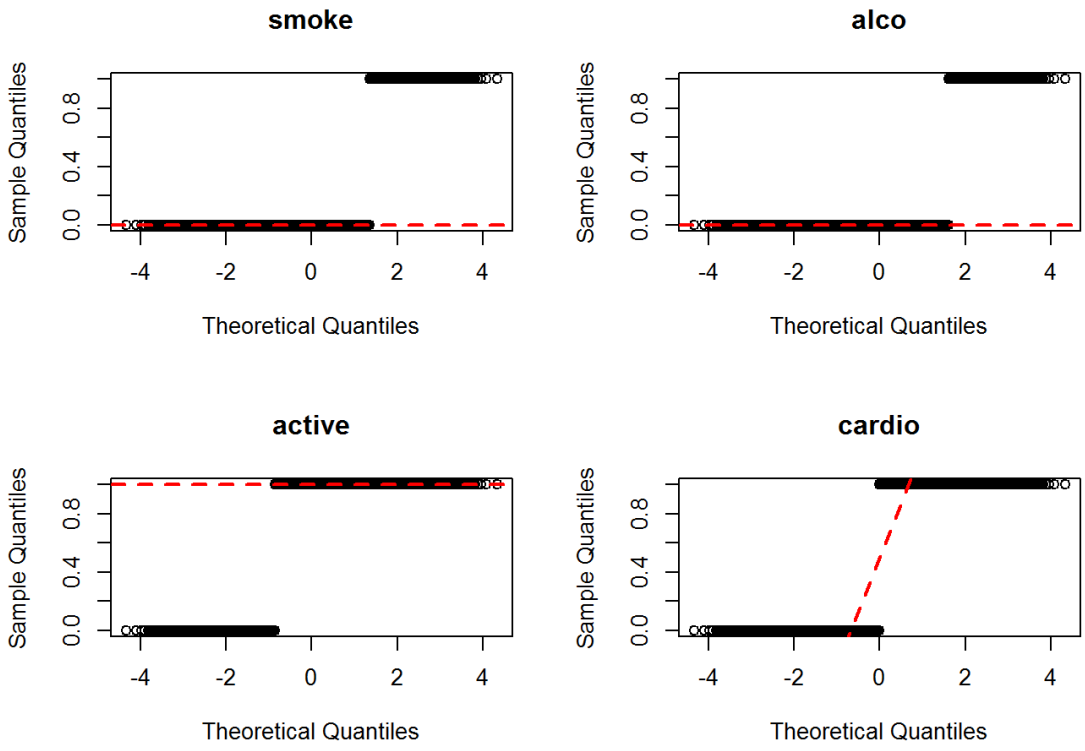


Imagem9- gráficos de normalidade para smoke, alco, active e cardio

Tal como todos os gráficos de normalidade anteriores, estes apresentam desvios fortes da normalidade.

Para confirmar as hipóteses feitas quanto à normalidade dos valores (estes não sendo normais) foram efetuados testes de normalidade. Os testes apresentados nas aulas para a normalidade foram testados, no entanto, todos retornaram erros devido ao tamanho dos dados. Como tal um teste diferente, e cujas propriedades teóricas não são totalmente conhecidas foi utilizado. As hipóteses foram confirmadas, com todos os valores p sendo inferiores a 2.2e-16(uma rejeição total da normalidade).

2.1.3-Correlações

Para verificar a existência de correlações e o seu nível entre pares de variáveis, com o objetivo de detetar colinearidade foi criada uma tabela de correlações entre todos os pares de variáveis.

Devido ao tamanho da tabela e às suas características a sua copia para o relatório seria extremamente demorada e impraticável, como tal apenas as conclusões serão apresentadas aqui. Para observar todos os valores das correlações dever-se-á ir ao ficheiro html disponibilizado.

Da análise da tabela foi possível concluir que não existe colinearidade elevada (>0.75), com o valor mais elevado sendo de 0.5(descartando correlações recíprocas), entre o género e a altura.

Por um lado, as baixas correlações diminuem os problemas causados por colinearidade, mas por outro uma vez que as correlações entre as variáveis explicativas e a variável alvo são baixas poderá ser difícil (ou mesmo impossível) obter bons resultados.

2.1.3-Grafico de dispersão

Com o objetivo de identificar as relações e padrões entre os dados foi realizado um gráfico de dispersão para cada par de variáveis.

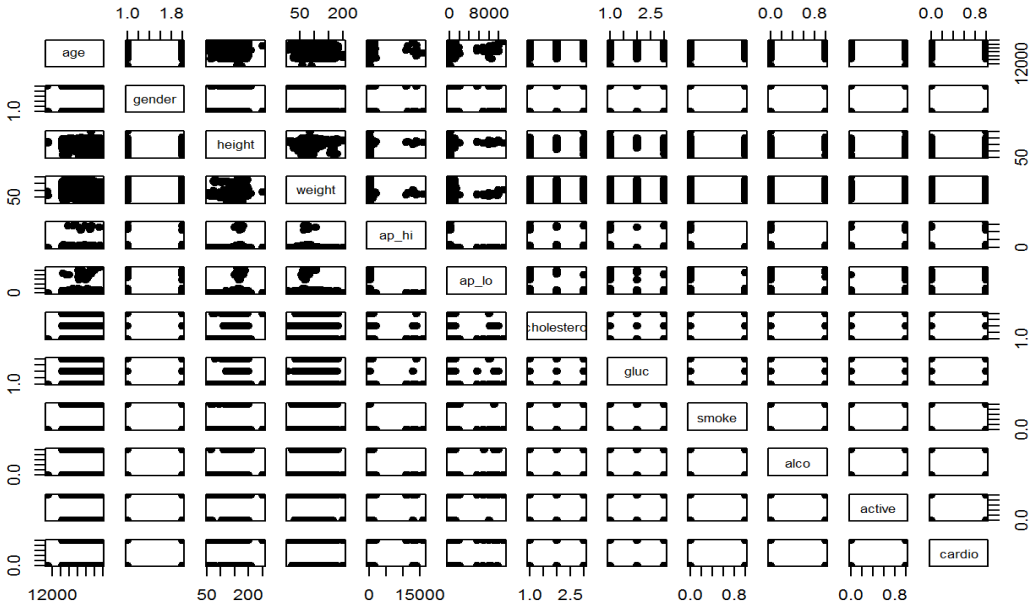


Imagem10-grafico de dispersão de todos os pares

Devido há existência de muitas variáveis categóricas e booleanas a maioria dos gráficos não apresenta muita informação. Existe, no entanto, um aspeto interessante dos dados, em grande parte deles os valores parecem distribuir-se em dois conjuntos.

2.1.4-Distribuição por categoria dos dados categóricos

Para encontrar padrões e efeitos das variáveis categóricas na variável alvo, as entradas foram divididas pelos diversos valores das variáveis categóricas e histogramas foram realizados.

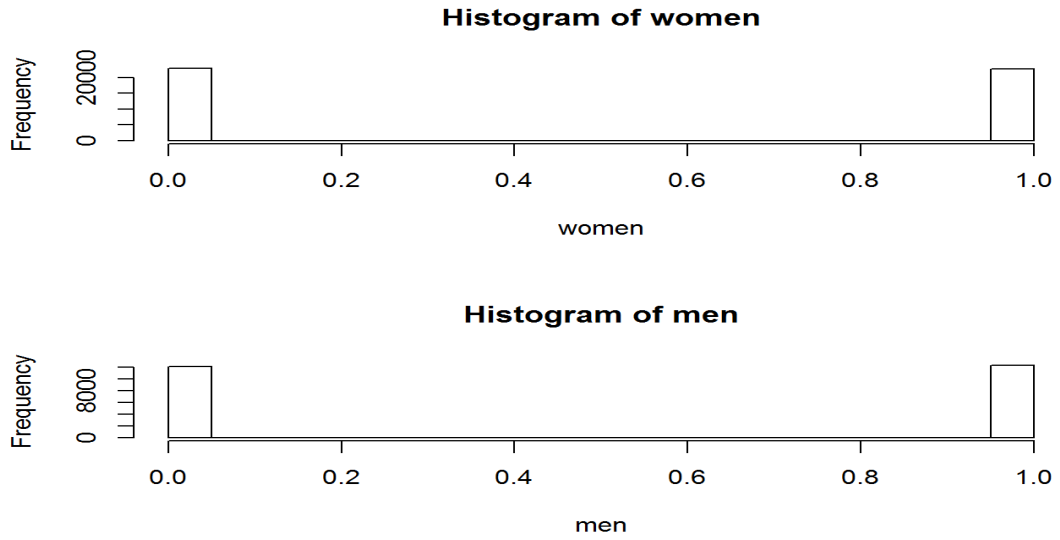


Imagem11-histogramas da resposta para diferentes valores de género.

É possível observar que, para alem da escala, as distribuições são quase idênticas, como tal não tendo muita informação.

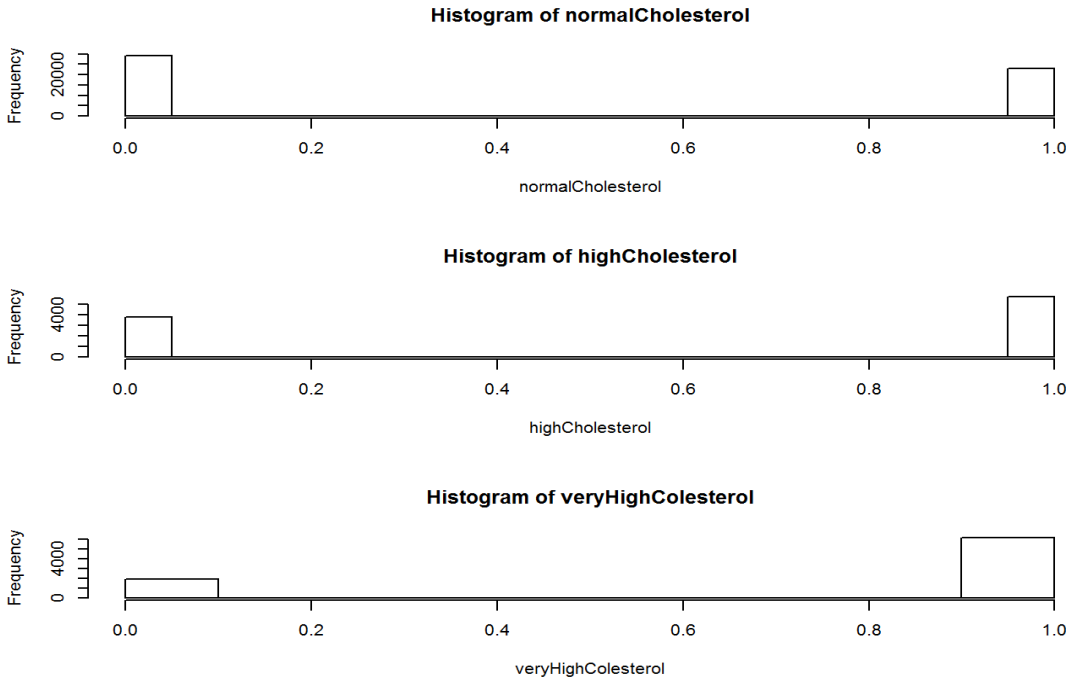


Imagem12-histogramas da resposta para diferentes valores de colesterol.

É possível observar, tal como esperado, que a presença de doenças cardiovasculares aumenta com o aumento do colesterol.

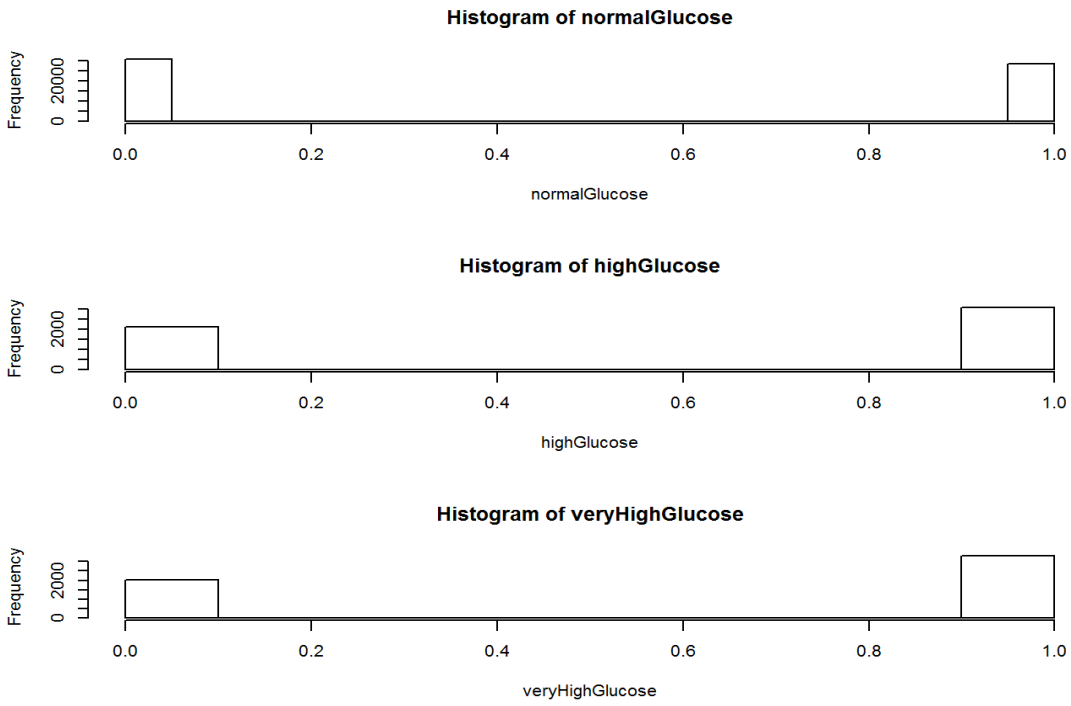


Imagem13-histogramas da resposta para diferentes valores de gluc.

Tal como para o colesterol, a presença de doenças cardiovasculares aumenta com o aumento da glucose.

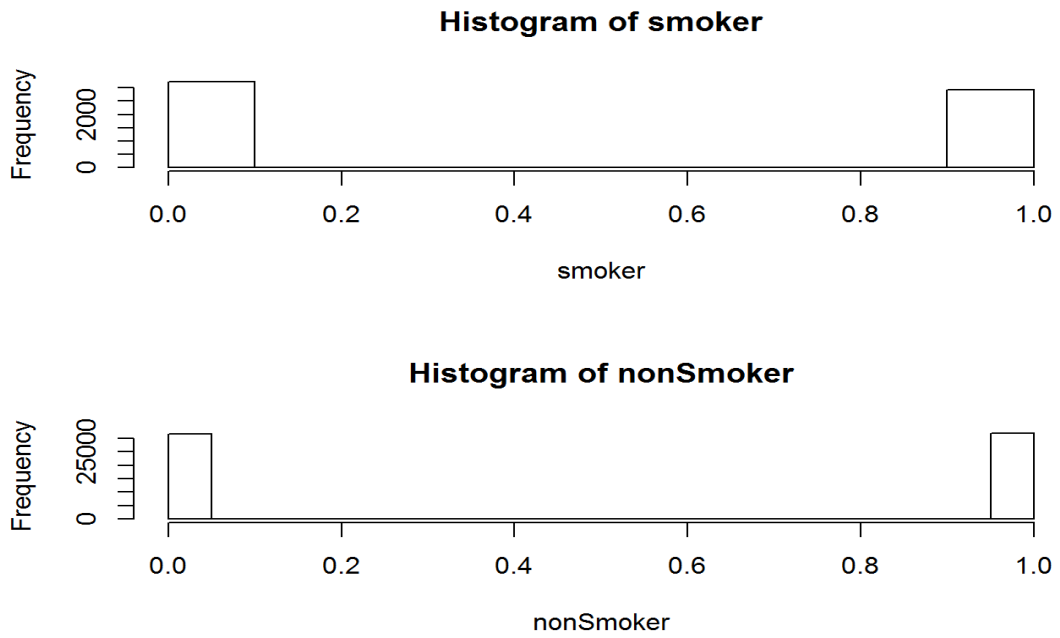
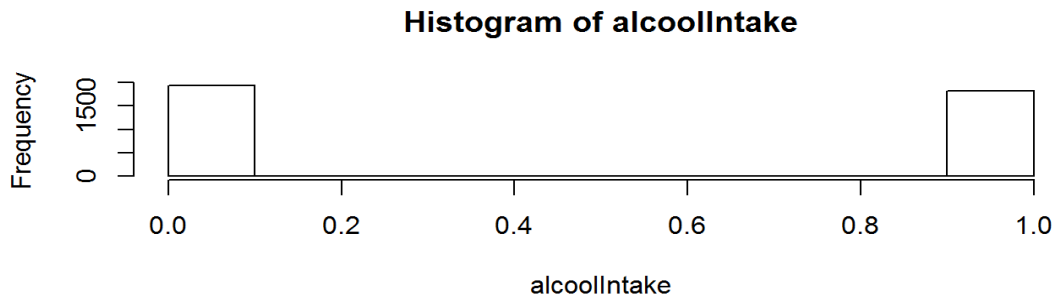


Imagem14-histogramas da resposta para diferentes valores de smoke.

Os valores são idênticos, e como tal não apresentam muita informação.



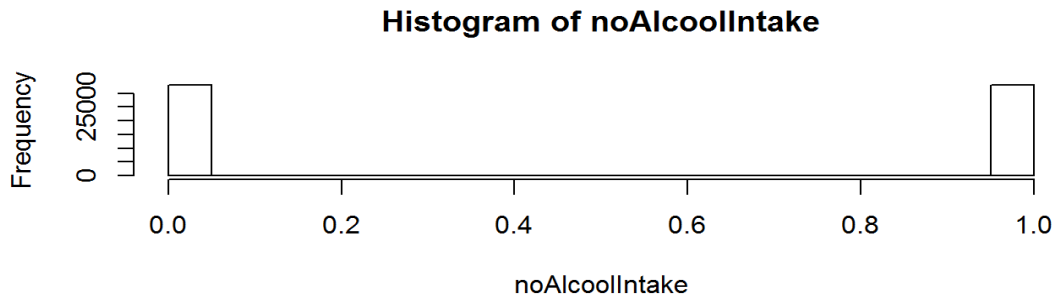
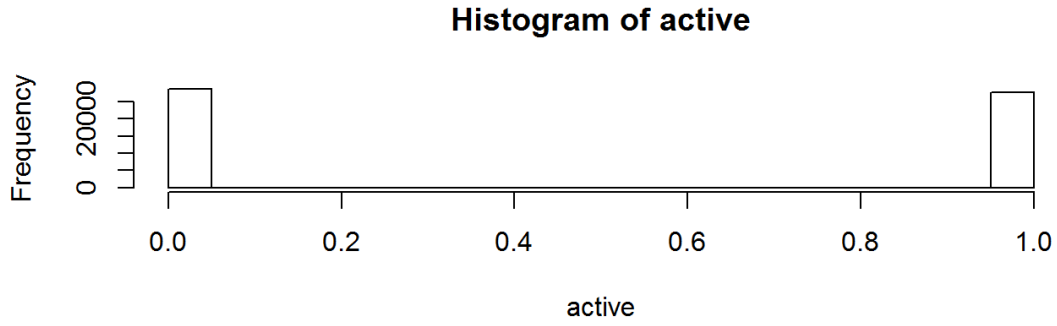


Imagem15-histogramas da resposta para diferentes valores de alco.

Mais uma vez os valores são idênticos e não contem muita informação.



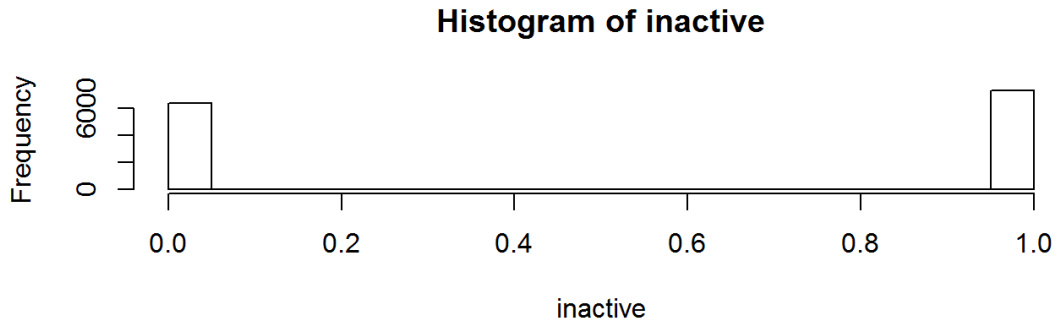


Imagem16-histogramas da resposta para diferentes valores de active.

Os valores são mais uma vez idênticos e pouca informação pode ser extraída.

3-Regressão linear

3.1-Pré processamento

Antes de proceder com a regressão linear algum pré processamento foi efetuado.

Valores extremos de ap_hi e ap_lo foram removidos (valores >500 e <0).

Os valores de género foram decrementados de 1,2 para 0,1, para estarem no mesmo formato dos outros dados.

As variáveis categóricas com 3 categorias (cholesterol e gluc) foram divididas em duas variáveis.

3.2-Modelos(b)

3.2.1-Modelo nulo

O primeiro modelo e ser criado foi o modelo nulo, sem nenhuma variável explicativa, com o qual alguns dos outros modelos serão comparados.

O coeficiente obtido foi o mesmo da percentagem de casos com doenças cardiovasculares existentes, 0.4949.

3.2.2-Modelos singulares

O segundo passo foi a criação de modelos com apenas uma variável explicativa, com o objetivo de observar o efeito de cada variável no erro e obter os efeitos brutos das variáveis.

O menor erro MSE obtido foi de 0.21 para a variável ap_hi.

Os efeitos brutos obtidos foram os seguintes:

age: 4.855*10^-5
gender: 0.007897
height: -0.0006307
weight: 0.006279
ap_hi: 0.01105
ap_lo: 0.01658
cholesterol_2: 0.1176
cholesterol_3: 0.3026
gluc_2: 0.1012
gluc_3: 0.1335
smoke: -0.02883
alco: -0.01802
active: -0.04686

3.2.3-Modelo saturado

De seguida foi criado o modelo saturado (com todas as variáveis explicativas). O erro obtido foi de 0.20, ligeiramente melhor que o melhor dos modelos simples.

3.2.4-Interação do modelo

Por fim o modelo foi iterado. Começando pelo modelo saturado as variáveis não significativas foram removidas. Apos apenas se ter variáveis significativas, as variáveis removidas foram novamente testadas para verificar se no novo modelo se tornam significativas, e nesse caso

adicionadas. De seguida as distâncias de cook e leverages foram testadas e observações com ambos os valores elevados removidos (nenhum caso foi encontrado).

Apos estes passos serem executados, o modelo final consiste de todas as variáveis exceto gluc_2(variável que identifica glucose elevada).

3.2.5-Iteração de variáveis (c - v)

Devido a restrições no tempo de elaboração do trabalho apenas um pequeno número de interações foi testado, de dentro de todas as possíveis.

De todas as interações testadas apenas a de idade contra colesterol provou ser significativa.

O gráfico abaixo mostra a diferença no declive de idade com os diversos níveis de colesterol. É possível observar uma forte interação.

O modelo com esta interação foi testado e o valor de interação foi considerado significativo.

Este modelo foi o modelo final obtido.

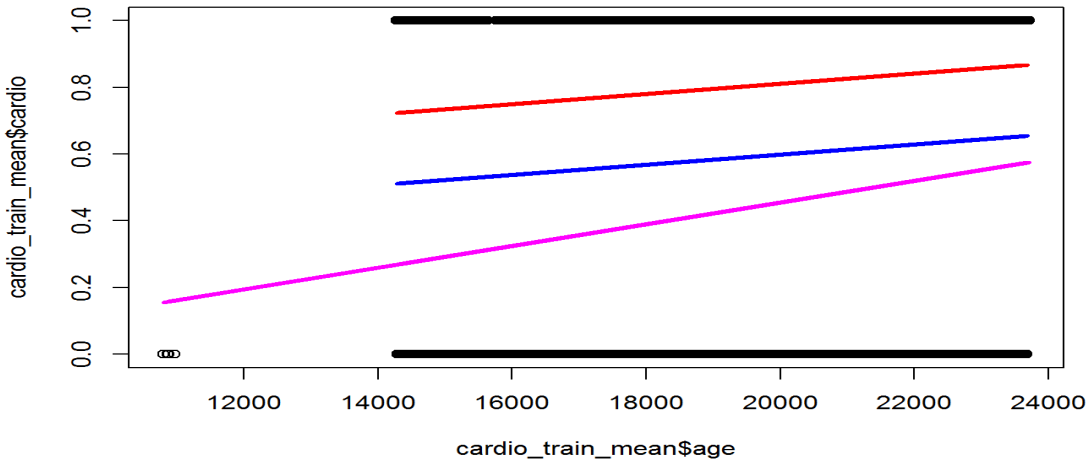


Imagem17-regressao sobre idade para os diferentes níveis de colesterol.

3.2.6-Analise dos resíduos

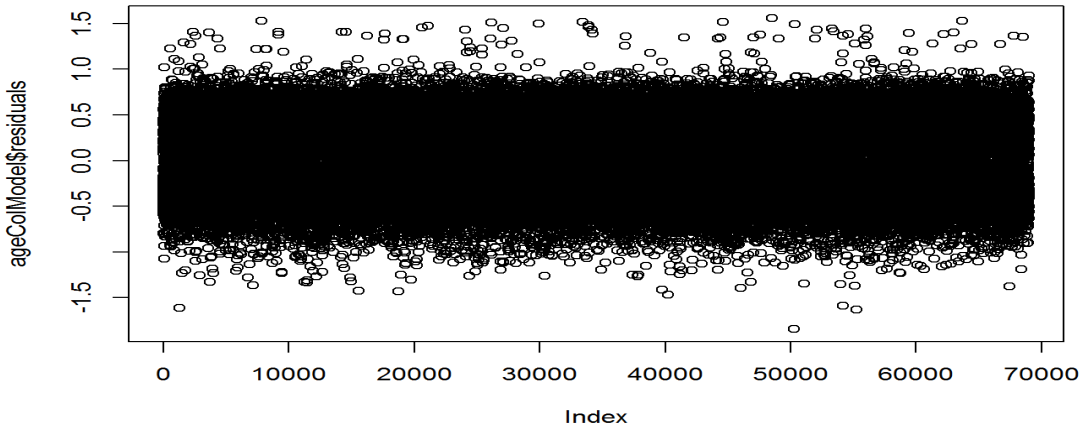


Imagem18-Grafico de dispersão dos resíduos.

A partir do gráfico acima é possível concluir que os resíduos apresentam homoestadicidade (variância constante).

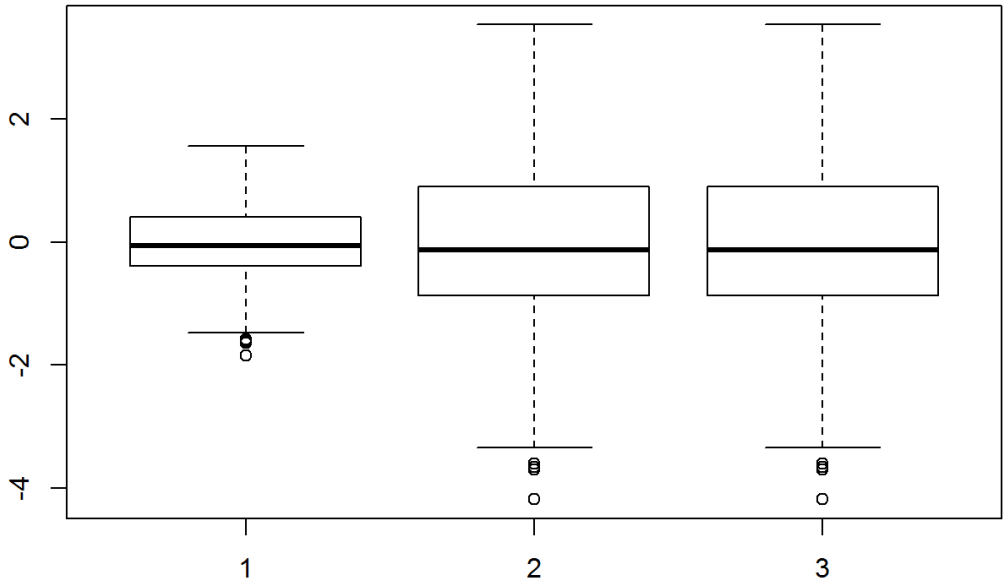


Imagem19-Diagrama de extremos e quartis dos resíduos.

É possível observar que os resíduos são aproximadamente simétricos.

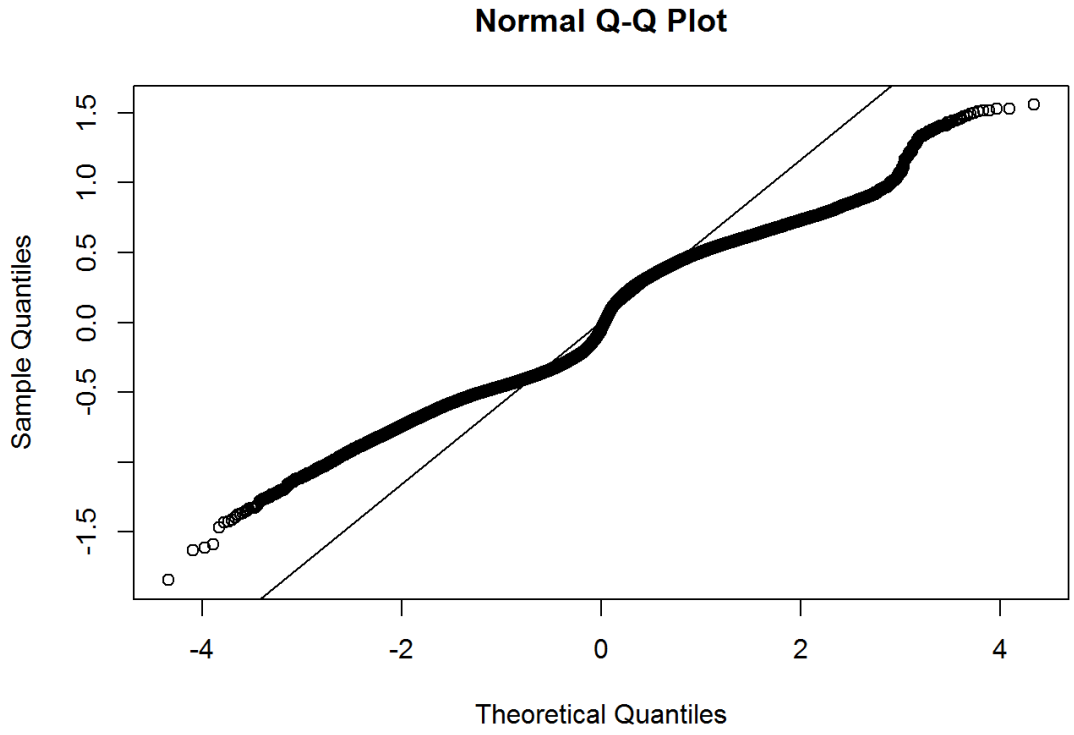


Imagem20-Diagrama de normalidade resíduos.

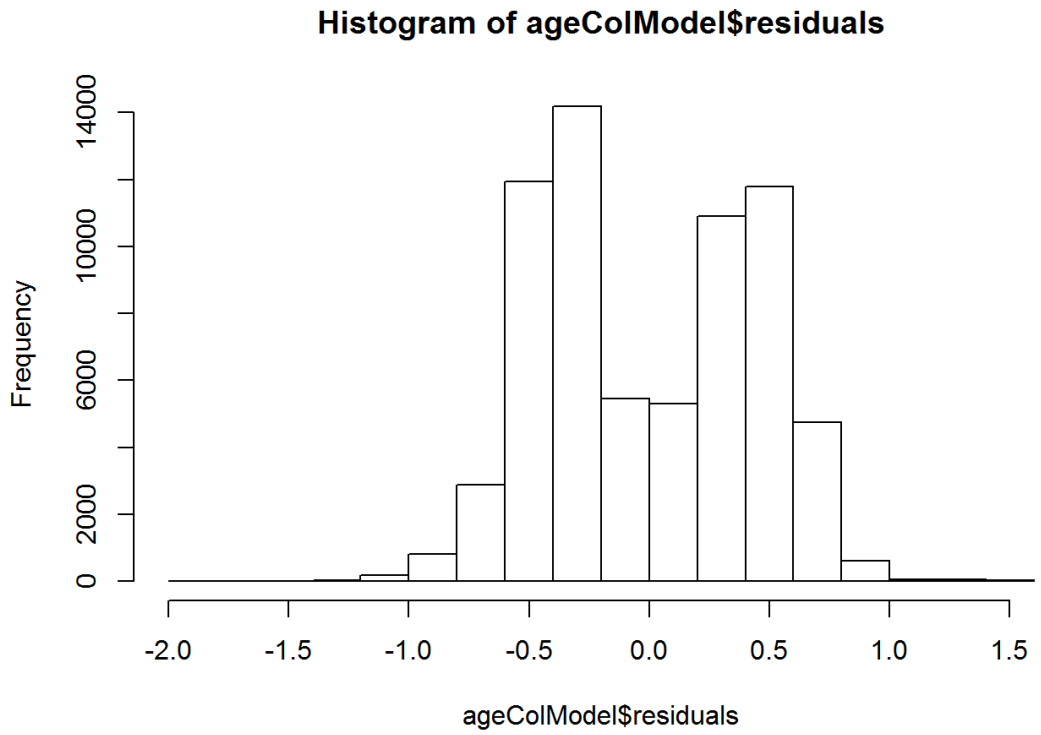


Imagem21-histograma dos resíduos.

Dos dois gráficos acima (imagem20 e imagem21) é possível concluir que os resíduos não seguem uma distribuição normal. No entanto tal não foi possível corrigir.

4-Resultados finais e previsões

Para a analise de uma variável continua e uma variável categórica com pelo menos 3 categorias foram escolhidas as variáveis age e cholesterol.

Parte da análise já se encontra acima, nomeadamente a interação entre ambas as variáveis.

4.1-Efeitos brutos contra efeitos ajustados (c - i)

Efeitos brutos:
age - 0.00004855
cholesterol_2 - 0.1176
cholesterol_3 - 0.3026
Efeitos ajustados:
age - 0.00003444
cholesterol_2 - 0.4785
cholesterol_3 - 0.6673

O coeficiente da idade diminuiu, o que significa que alguns dos efeitos causados pela idade estão agora a ser explicados pelas outras variáveis. Isto é corroborado pelo facto de quase todas as outras variáveis terem peso positivo.

Os coeficientes do colesterol aumentaram, o que significa que alguns dos efeitos relacionados com o colesterol que diminuem o risco de doenças cardiovasculares estão agora a ser tratados por outras variáveis.

4.2-Intervalos de predição e confiança (c - ii)

Mantendo todos os valores das variáveis continuas na sua média e das variáveis categóricas na sua moda (exceto para a variável de idade), o gráfico abaixo foi obtido.
A vermelho está o valor previsto, a azul as bandas de confiança, a verde as bandas de predição e a preto os valores das observações.

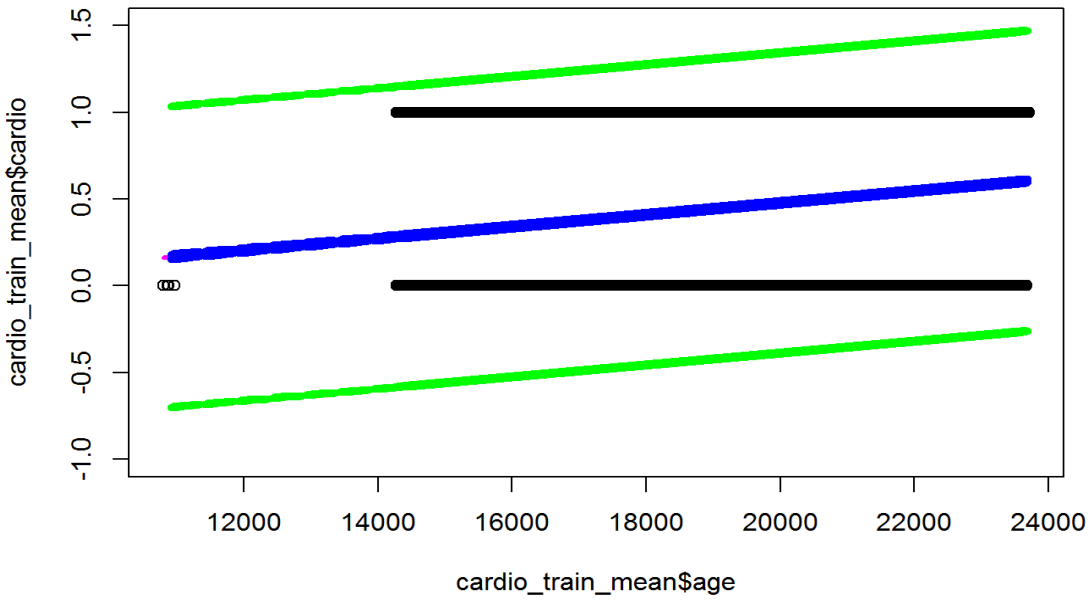


Imagem22-bandas de confiança e predição para a idade.

É possível observar que as bandas de confiança estão extremamente perto, sendo mesmo impossível distinguir as duas, tal como esperado pelo tamanho da amostra.

As bandas de predição são, no entanto, preocupantes, uma vez que todos os valores se encontram fora do intervalo dos valores possíveis.

4.3-Alteração do valor de uma variável categórica (c - iii)

A alteração do valor do colesterol de 3 para 2 tem uma mudança prevista de -0.1888, um decréscimo de 19% na probabilidade de doenças cardiovasculares.

O intervalo a 95% para esta mudança é [-0.36759020,-0.01001801], enquanto que a 90% o intervalo é [-0.3388455,-0.0387627]. O que equivale a um decréscimo de algures entre 1% e 37% no risco de doenças cardiovasculares.

4.4-Alteração do valor de uma variável continua (c - iv)

Foi analisado o resultado de aumentar o valor de uma variável continua no valor de um do seu desvio padrão.

Aumentar o valor da idade neste valor causa um aumento de 0.084, ou seja, aumenta a probabilidade de uma doença cardiovascular em 8.4%.