Instituto Superior Técnico

Análise de Modelos Lineares

Projeto

Grupo 7

António Gouveia 92621

Filipa Costa 92626

Gustavo Caria 92633

Tomas Worm 92651

Leandro Duarte 93112

Vasco Pearson 97015

Conteúdo

1 Introdução

Este relatório foi feito no âmbito da unidade curricular de Análise de Modelos Lineares, lecionada pela Professora Conceição Amado.

1.1 Introdução ao Problema

Como as nossas horas no planeta Terra são apenas temporárias, é importante explorar a aplicação da matemática no estudo e compreensão de vários fatores eminentes ao ser humano, nomeadamente os que afetam a Esperança Média de Vida.
De maneira a encontrar os fatores que mais impactam a esperança média de vida em diferentes países foi utilizada uma amostra de dados recolhidos pelo Observatório Mundial da Saúde e disponibilizados pela OMS.
No estudo em consideração, foram registados dados entre os anos de 2000 e 2015, de 193 países diferentes, avaliando variados aspetos económico-sociais.
Este projeto inicia-se com o objetivo de encontrar o melhor *modelo de regressão linear* para justificar a variável de resposta ($Life.expectancy$) em função das variáveis explicativas e ainda, se possível, prever respostas com base no modelo escolhido.

1.2 Análise Inicial e Tratamento Preliminar dos Dados

Ao dataset inicial deu-se o nome de *life* no script do programa *R*. Este tem 2938 linhas e 22 colunas das quais se decidiu retirar as variáveis *Status* e *Year*, visto que alguns países não estão de acordo com os conformes. Assim, para cada um dos 193 países diferentes, foram medidas 20 variáveis, sendo elas: *Country, Adult.Mortality, infant.deaths, Alcohol, percentage.expenditure, Hepatitis.B, Measles, BMI, under.five.deaths, Polio, Total.expenditure, Diphtheria, HIV.AIDS, GDP, Population, thinness1.19years, thinness1.5.9years, income.composition.of.resources, Schooling e Life.expectancy.*

Analisando a variável categórica *Country*, o comando *as.factor* foi aplicado para que o programa R a reconheça como categórica e não como numérica. Note-se também que a frequência dos dados relativamente a Country não está equilibrada e esta subrepresentação em alguns países, pode também levar a problemas na modelação.

No caso da variável *BMI*, que indica o índice de massa corporal, que sabemos, por definição, que não pode ultrapassar o valor 40, verificou-se que várias observações ultrapassavam o mesmo. Assim, optou-se por substituir as observações que falhavam nesse requisito por um valor de $BMI = 40$.

Na procura de observações em falta na nossa amostra, avaliou-se o comando *any(is.na(life))*, que retornou o resultado *"True"*. Para contornar este problema, poderíamos seguir dois métodos: **Imputação** (substituindo as observações em falta pela mediana de todos os valores da variável em questão) ou **Eliminar observações em falta**. Conclui-se ao longo do projeto, com a análise e comparação de ambos, que a segunda seria a escolha mais acertada para construir um melhor modelo.

Na tabela abaixo, verifica-se que em algumas variáveis a média e a mediana diferem bastante. Esta desigualdade pode dever-se a vários erros.
Para começar, a OMS obteve os dados por região, mas como cada região organiza os seus dados de forma específica, isto levou a problemas da obtenção dos dados, pois estes foram extraídos sem levar isso em consideração (exemplo ilustrativo: Canada foi classificado como país em desenvolvimento).
Além disso, a própria natureza dos dados dificilmente seria uniforme quando são recolhidos valores de países com realidades sócio-económicas muito diferentes.
Por último, há que reconhecer que há uma subrepresentação de certos países (por exemplo: recolheram-se 16 observações da Itália, enquanto apenas se recolheu 1 observação de Nauru).

Variável	Média	Mediana
infant.deaths	30.3	3.0
percentage.expenditure	738.3	64.9
Measles	2419.6	17.0
under.five.deaths	42.04	4.00
HIV.AIDS	1.742	0.100
GDP	7483.2	1767.0

Figura 1: Tabela de dados de algumas das variáveis numéricas

É também possível identificar discrepâncias incomuns entre os valores máximo e mínimos da maioria das variáveis, o que pode indicar a presença de outliers (ex: a Variável $Measles$ apresenta mínimo 0.0 e máximo 212183.0).

Dado a elevada dimensão dos dados, fazer o diagnóstico/confirmação dos dados tornar-se-ia um processo extremamente desgastante, sendo que também seria impraticável tendo em conta o tempo disponível para realizar o projeto. Assim, para realizar uma analise à priori rápida e automática, usou-se uma rotina no programa R que com base nos resultados do comando *boxplot* eliminou 375 observações que se apresentavam como outliers.

Com as alterações anteriores, obteve-se um *dataset* com 1274 observações e 20 variáveis. Estamos assim em condições de analisar algumas estatísticas descritivas, usando os comandos *summary* do R para ter uma ideia geral do dados.

Na figura seguinte, é apresentado o gráfico obtido com o comando *ggpairs* aplicado às variáveis numéricas do *dataset*, no qual se vê os gráficos de dispersão e as correlações entre as variáveis. Assumimos que duas variáveis estariam altamente correlacionadas caso a correlação fosse acima de 80%.

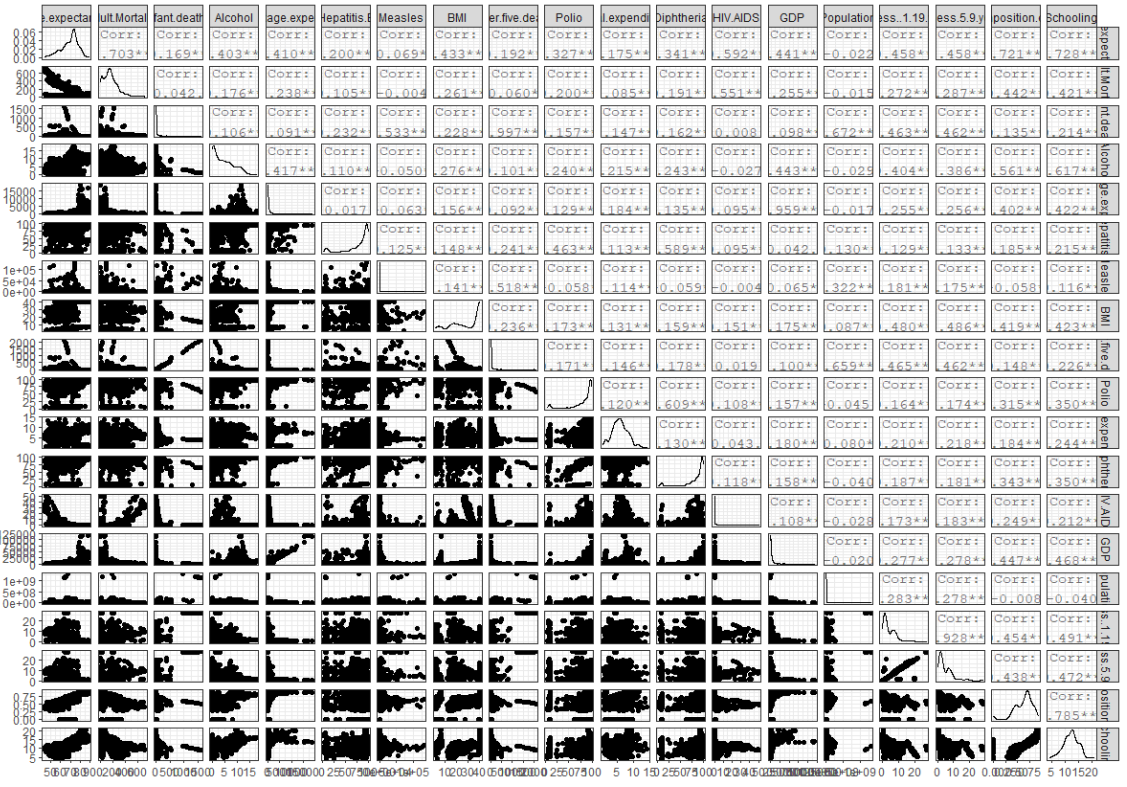


Figura 2: Gráficos de dispersão e correlação entre as variáveis quantitativas

No gráfico acima, é possível verificar que há algumas variáveis correlacionadas, nomeadamente *under.five.deaths e infant.deaths (99.7%) e GDP e percentage.expediture* (95.9%). Com estes valores elevados é de prever que em ambos os casos apenas uma das variáveis irá permanecer no modelo.
É também notória a existência de algumas variáveis com correlações acima dos 70%, um valor elevado mas que não justifica a remoção destas.

Analisando os valores da primeira linha, tornam-se visíveis as variáveis que apresentam menor correlação com a variável resposta $Life.Expectancy$, entre estas *Population*, $Measles$, *Infant Deaths, under.five.deaths e Total.expenditure.*
As variáveis *Population, Measles, Infant Deaths e under.five.deaths* têm todas uma correlação relativamente alta entre si, o que indica que a presença de uma delas no modelo poderia abdicar da presença das outras.
No caso de *Total.expenditure*, verifica-se uma baixa correlação com todas as outras variáveis, o que torna a sua influência na variável de resposta mais importante.

Inversamente, *Adult Mortality, income.composition.of.resources e Schooling* são as variáveis que apresentam maior correlação com *Life.Expectancy*. No entanto, as duas últimas estão altamente correlacionadas, e, poderia ser por este motivo que, aparecendo uma, a outra não apareceria no modelo final, dado que são explicadas mutuamente.
Adult Mortality tem baixa correlação com as restantes variáveis preditoras, o que demonstra a alta influência que esta tem na variável resposta.
É também de notar que a variável *Alcohol* está correlacionada positivamente com a variável resposta (0.403).

Analisando os gráficos de dispersão da variável resposta em função das variáveis preditoras na primeira coluna, individualmente, verificou-se que as variáveis *Adult Mortaliy, HIV-AIDS, Schooling e income.composition.of.resources* apresentam gráficos que parecem traduzir uma relação linear.

Num primeiro estudo, estas questões potencialmente problemáticas não vão ser tomadas em conta. No entanto, após serem encontrados os primeiros modelos poder-se-á verificar, recorrendo a técnicas de diagnóstico, se efetivamente existe ou não problemas.

2 Obtenção de Modelos

Neste tipo de análise estatística, é comum que se utilizem diferentes subconjuntos do *dataset* para treino e teste. Neste caso, decidiu-se repartir a amostra em 2 subconjuntos. O **conjunto de treino** (*data.train*), tem **80%** das observações, sendo que as restantes **20%** se encontram no **conjunto de teste** (*data.test*). O modelo construído irá aprender a prever a variável dependente com o conjunto de treino e posteriormente será avaliado no conjunto de teste, através da verificação da correta previsão dessa variável. Cada um dos conjuntos foi criado através da **seed(3493)** Note-se que, aquando da previsão, 2 observações estavam no *data.test*, mas não estavam no *data.train* (uma da Irlanda, e uma da Guiné Equatorial). Assim, procedeu-se a uma troca destas para o *data.train*.

2.1 Modelo Completo

Naturalmente, o primeiro modelo (m.full) que foi considerado é aquele que usa todas as variáveis preditoras que resultaram da análise exploratória inicial. Fizemos 2 modelos iniciais, um com o método imputação e outro eliminando observações em falta (NA - Not Available).
Efetuou-se o *summary* e *anova_alt* sob estes modelos, obtendo-se o seguinte:

2.1.1 Imputação

1. $R^2 = 0.9803$
2. $R^2_{adj} = 0.9778$
3. $MSE = 1.78$

Verificou-se que individualmente as variáveis são, na sua maioria, significativas, exceto, por exemplo, os casos de *Benin, Burundi, Cameroon, percentage.expenditure, Hepatitis.B e BMI.*

2.1.2 Eliminando observações NA

1. $R^2 = 0.9896$
2. $R^2_{adj} = 0.9878$
3. $MSE = 0.74$

Neste caso apresentaram-se como variáveis significativas *Adult Mortality, Infant Deaths, Alcohool, Hepatite-B, under.five.deaths, HIV-AIDS, GDP, income.composition.of.resources, Schooling*, e todos os elementos da variável *Country* exceto *Benin, Botwsuana, Equatorial-Guinea, Guinea, Guinea-Bissau, Kenia, Liberia, Mali, Nigeria, Uganda e Zimbabwe.*

2.1.3 Comparação

Para melhor comparar a forma de tratar as observações iniciais (Imputação vs NA), apresentamos a seguinte tabela:

Modelo	R^2_{adj}	$valor - p$	Mínimo	Máximo
Imputação	0.9515	$2.2e-16$	0.002801	9.581106
NA	0.9878	$2.2e-16$	0.01663	3.35731

Verificou-se que os valores mínimo e máximo pela Imputação estavam muito afastados, enquanto pela remoção dos NA, já não estão tão afastados.
Em ambos os casos o $valor - p$ é muito baixo ($2.2e-16$), o que deu a entender que a regressão é significativa.
Sendo que com o valor do R^2_{adj} pela Imputação é mais baixo que pela remoção dos NA, concluiu-se que o modelo que melhor explica a variabilidade da Esperança média de vida seria o que foi criado eliminando as observações em falta (NA).

2.2 Best Subsets

Para obter o melhor modelo, a situação ideal seria testar todos os modelos possíveis (2^p) que podemos fazer com as 19 variáveis explicativas.
Para isso, vamos começar por usar o método *Best Subset Selection*, servindo-nos dos critérios Cp, R^2_{adj} e BIC.
Para este método, vamos escolher a melhor combinação das p variáveis explicativas. Ou seja, vamos ver qual é o melhor modelo para 1 variável, para 2, e por aí em diante. Analisamos então todos os modelos obtidos, com o objectivo de identificar o melhor.

Como a *leaps*, função que determina os melhores modelos pelos critérios descritos, apenas admite variáveis quantitativas, retirou-se a variável categórica *Country*, para ser adicionada posteriormente.
O comando *leaps* do R, permite que se escolha usar como indicador o Cp de $Mallows$, o R^2_{adj} ou o R^2. Este último não foi considerado visto não ser um bom indicador para modelos com mais que 1 variável explicativa, devido ao facto de aumentar sempre que se adiciona uma variável.
Por outro lado, como o comando *leaps* não permite encontrar o *best fit* de acordo com o BIC, optamos por usar o comando $regsubsets$, da mesma *library*, para descobrir qual o melhor conjunto de variáveis segundo BIC.

2.2.1 Critério R^2_{adj}

Na figura 3, é possível observar o valor do R^2_{adj} em detrimento dos vários subconjuntos de variáveis preditoras possíveis. À primeira vista, parece que o R^2_{adj} já não sofre alterações após o número mínimo de 11 variáveis.

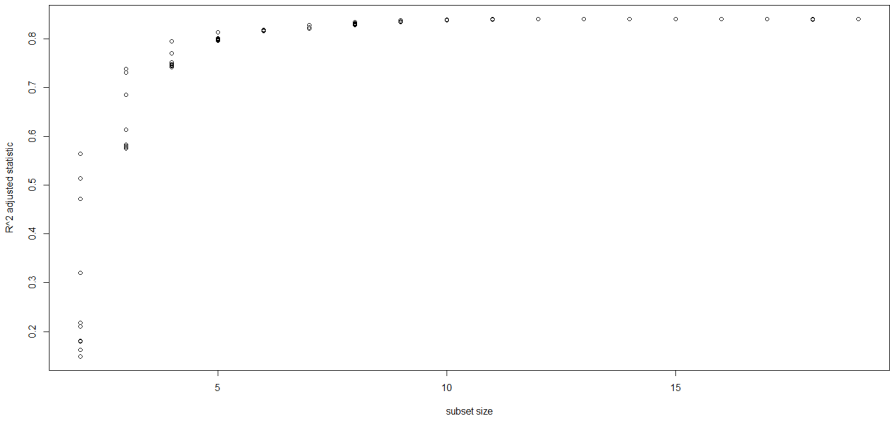


Figura 3: Valor do R^2_{adj} em função do nº de variáveis por subconjunto.

2.2.2 Critério Cp

Na figura 4, é possível observar o valor do Cp em detrimento dos vários subconjuntos de variáveis preditoras possíveis. À primeira vista, parece que o $Cp = p$ após o número mínimo de 11 variáveis.

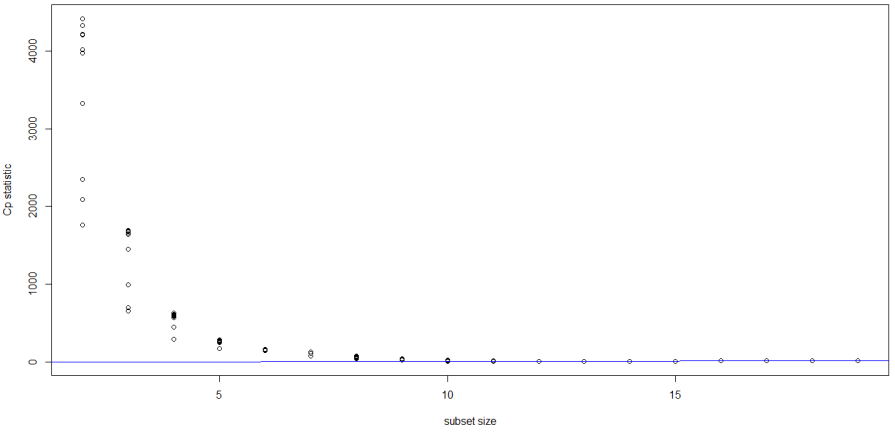


Figura 4: Valor do Cp em função do nº de variáveis por subconjunto.

Ambos os critérios apresentam resultados coerentes.

2.2.3 Seleção do melhor modelo através do método R^2_{adj}

A função $which.max()$ pode ser utilizada para identificar a localização do ponto máximo de um vector. Como estamos à procura do modelo com maior R^2_{adj}, aplicando a função $which.max()$, obtemos que o melhor subconjunto ($best.fit$) de variáveis preditoras, já com a variável categórica é: 1,3,4,6,9,10,11,12,14,18,19,20

Por sua vez, se efetuarmos o *summary* e *anova_alt* sob o novo modelo definido por estas variáveis, podemos ver que:

1. $R^2 = 0.9891$
2. $R^2_{adj} = 0.9873$
3. MSE = 0.77 (obtido pela função *anova_alt*)

Obtemos também que *valor-p* $< 2.2e-16$. Assim, não houve aumento nem diminuição significativos dos valores em relação aos resultados do modelo completo.
Por sua vez, a distância máxima em relação ao modelo completo aumentou e a mínima diminuiu, porém não foram alterações significativas.

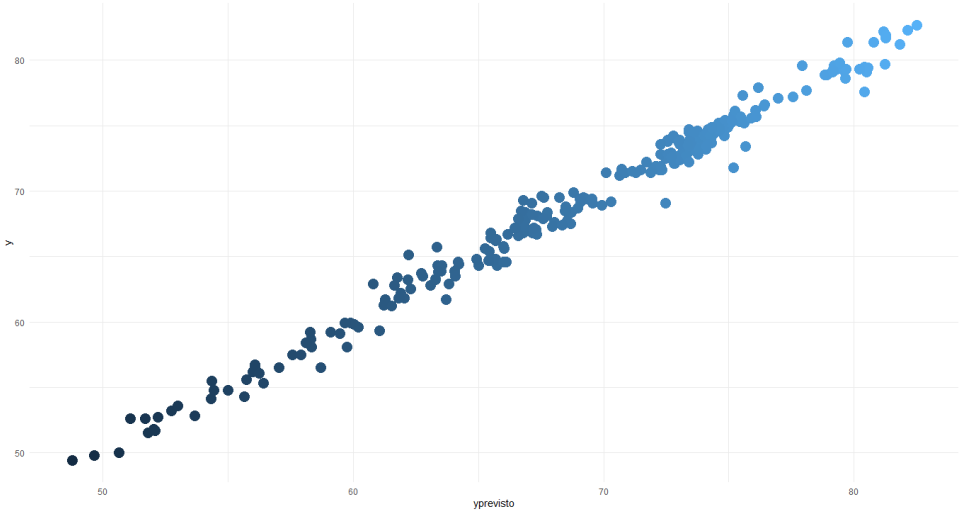


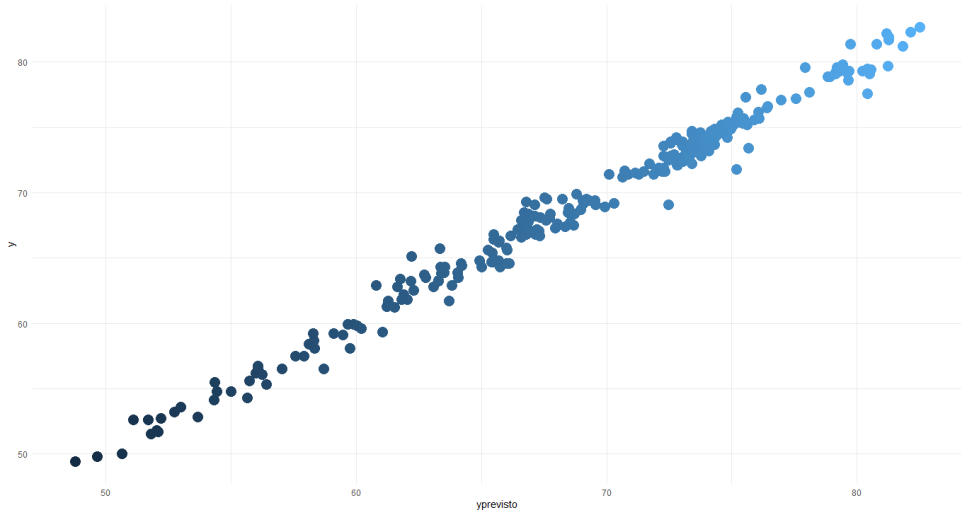
Figura 5: Comparação entre os valores previstos e os valores observados por R_{adj}

É notório que, quanto mais próximo da reta $y = x$, melhor a previsão.
Comparando este novo modelo com o modelo completo ($m.full$), é possível concluir que não existem grandes ganhos, o único é que temos 12 variáveis em vez de 19.

2.2.4 Seleção do melhor modelo através do método Cp

De modo semelhante, podemos traçar as estatísticas Cp e indicar o modelo com a estatística mais pequena, usando $which.min()$. Assim, obtemos que o melhor subconjunto ($best.fitcp$) de variáveis preditoras, já com a variável categórica é: 1 3 4 6 9 10 11 12 14 18 19 20
Por sua vez, se efetuarmos o *summary* e *anova_alt* sob o novo modelo definido por estas variáveis, podemos ver que:
1. $R^2 = 0.9891$
2. $R^2_{adj} = 0.9873$
3. $MSE = 0.77$ (obtido pela função *anova_alt*)
Obtemos também que *valor-p* $< 2.2e-16$.

Por sua vez, não houve alterações significativas nas distâncias entre os valores previstos e observados.



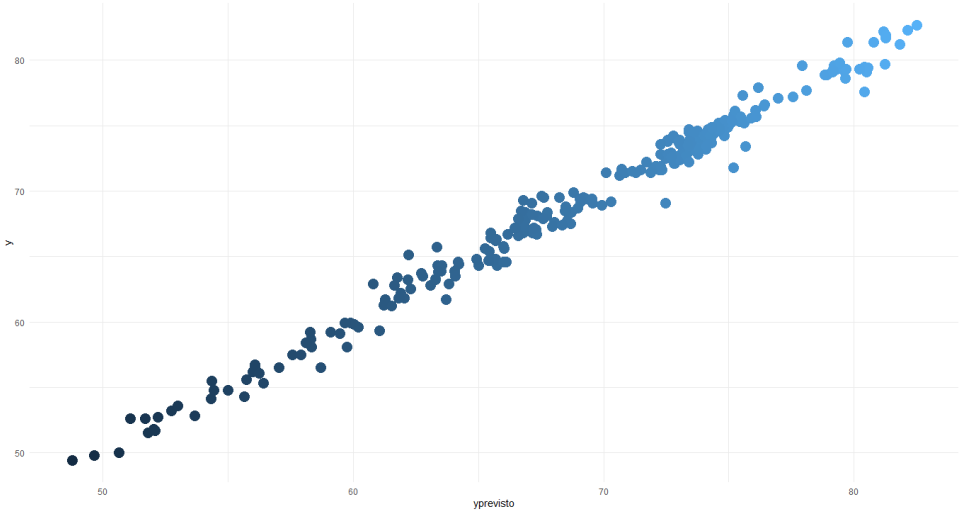

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000614 0.221287 0.457699 0.651835 0.906813 3.376822
```

Figura 6: Comparação entre os valores previstos e os valores observados por C_p

2.2.5 Seleção do melhor modelo através do método BIC

Quanto às estatísticas do BIC, vamos usar e indicar o modelo com a estatística mais pequena, usando $which.min()$. Assim, obtemos que o melhor subconjunto ($best.fitbic$) de variáveis preditoras, já com a variável categórica é: 1,3,4,5,9,10,12,14,19,20

Por sua vez, se efetuarmos o *summary* sob o novo modelo definido por estas variáveis, podemos ver que:
1. $R^2 = 0.9893$
2. $R^2_{adj} = 0.9876$
E pela tabela $anova\_alt$:
3. $MSE = 0.75$
Obtemos também que *valor-p* $< 2.2e-16$.



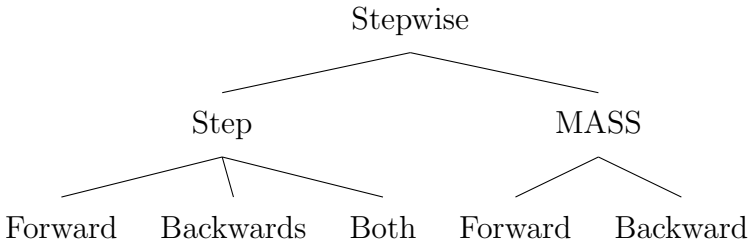

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000133 0.225124 0.456565 0.644653 0.944818 3.453163
```

Figura 7: Comparação entre os valores previstos e os valores observados por BIC

2.2.6 Conclusão

Deste modo, é possível concluir que, como o modelo obtido através do critério BIC tem uma capacidade preditiva semelhante aos restantes modelos, mas com MSE mais baixo e menos variáveis explicativas, decidimos usar este modelo como o melhor segundo o *Best Subset Selection*.

2.3 Stepwise



O objetivo do *Stepwise* é construir um modelo por acréscimo ou remoção de variáveis explicativas independentes. Em cada etapa do *Stepwise*, uma variável é considerada para acréscimo ou remoção com base num critério pré-especificado (como o teste F). As principais abordagens do *Stepwise* são:

1. **Forward selection**: Começa-se sem variáveis no modelo, depois testa-se a adição de cada uma usando um critério pré-especificado e adiciona-se a variável (se houver) cuja inclusão dá a melhoria mais significativa estatisticamente do modelo. Repete-se este processo até que nenhuma variável melhore o modelo.

2. **Backward elimination**: Começa-se com todas as variáveis, testa-se a exclusão de cada uma usando um critério pré-especificado e exclui-se a variável (se houver) cuja perda resulta no agravamento mais insignificante estatisticamente do modelo. Repete-se este processo até que nenhuma outra variável possa ser excluída.
3. **Both**: Combinação dos dois processos acima, testando em cada etapa as variáveis a serem incluídas ou excluídas.

Reparámos que ao fazer o *Stepwise* com a variável "$Country$"que as tabelas obtidas no *dropterm*, e o *summary* do modelo, iam apresentar parâmetros para cada país individualmente, o que iria dificultar a escolha das variáveis para o modelo. Decidimos então que seria melhor realizar o *Stepwise* sem a variável "$Country$"e adicioná-la no final, pois viu-se que a variável era significativa e portanto deveria ser mantida no modelo final.

O método *Stepwise* pode ser feito de forma automática (usando o comando *Step*), ou de forma manual através da biblioteca *MASS*.

Ao utilizarmos o *Step* podemos tomar a abordagem "$Backward$", "$Forward$" ou "$Both$" e as variáveis são retiradas ou acrescentadas de forma a diminuir o critério AIC.

Recorrendo à biblioteca *MASS*, executamos o comando *Stepwise manual*. Para começar, necessitamos de definir um valor-p. No nosso caso optamos pelos níveis de significância usuais (n.s.u.), 1%, 5% e 10%.
Na abordagem "$Backward$", onde começamos com o modelo $m.full$, com todas as variáveis, é necessário observar a tabela obtida por *dropterm*. Esta tabela apresenta vários parâmetros sobre o que acontece ao modelo quando eliminamos uma variável, mas focamos-nos mais no $valor-p$ da $F-statistic$.
Procuramos então o maior $valor-p$ acima dos n.s.u. e removemos a variável correspondente do modelo $m.full$. Continuamos este processo, até que todos os $valores-p$ da $F-Statistic$ sejam inferiores ao n.s.u..
No "$Forward$", começamos com o modelo vazio (temos apenas o β_0 da ordenada na origem) e adicionamos a variável com o menor $valor-p$, obtida por *addterm*. O processo continua, até os $valores-p$ das restantes variáveis, que não entraram para o modelo, serem maiores que os n.s.u..

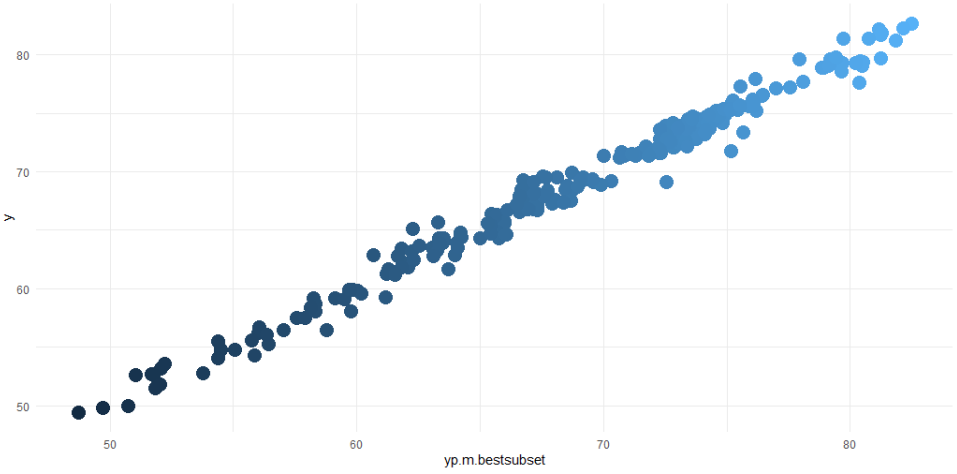
Fizemos ainda o "$Stepwise$"pela abordagem "$Backwards$"de uma terceira maneira. Através da observação dos $valores-p$ do teste t obtido pelo comando *summary*, retirou-se a variável que apresentou o $valor-p$ mais elevado. No entanto, observou-se que deste modo chegamos ao mesmo resultado do que em "$MASS\,Backwards$". Devido a isto, considerou-se apenas este último modelo na análise que se segue.

A partir da tabela seguinte escolhemos o melhor modelo dado pelo *Stepwise* (página seguinte):

Melhor modelo stepwise								
Modelo	**Nº var.**	**MSE**	R^2	R^2_{adj}	**AIC**	**BIC**	**PRESSp**	**Cp**
step.backward.NA	11	9.6	0.8427	0.841	2324.853	2386.652	10050.3	12168.9
step.backward.p.NA	10	9.6	0.8423	0.8407	2325.426	2382.076	10052.46	12200.1
step.both.NA	10	10.4	0.8294	0.8277	2405.485	2462.134	10881.06	13275.71
step.both.aic.NA	10	10.4	0.8294	0.8277	2405.485	2462.134	10881.06	13275.71
step.forward.NA	10	10.4	0.8294	0.8277	2405.485	2462.134	10881.06	13275.71
step.forward.f.NA	10	10.4	0.8294	0.8277	2405.485	2462.134	10881.06	13275.71

Concluímos então, que o melhor modelo obtido por stepwise é o *step.backwards.NA* que foi obtido usando a função *step* do *R*. De seguida, voltou-se a adicionar a variável "*Country*"e verificámos os valores dos critérios considerados.

Melhor modelo stepwise								
Modelo	**Nº var.**	**MSE**	R^2	R^2_{adj}	**AIC**	**BIC**	**PRESSp**	**Cp**
step.backward.NA	12	0.78	0.9889	0.9871	-123.276	613.1621	∞	-69.4059




```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000393 0.224255 0.451233 0.662619 0.954086 3.441092
```

Figura 8: Comparação entre os valores previstos e os valores observados por Stepwise

2.4 Best Subsets vs. Stepwise

À procura do melhor modelo possível, escolhemos fazer o diagnóstico e a análise e remoção de observações atípicas do modelo (**Secção 3**) para os melhores modelos obtidos por *Best Subsets* e por *Stepwise*.
Depois disso, fomos analisar as medidas dos dois modelos, para selecionar finalmente o melhor modelo.

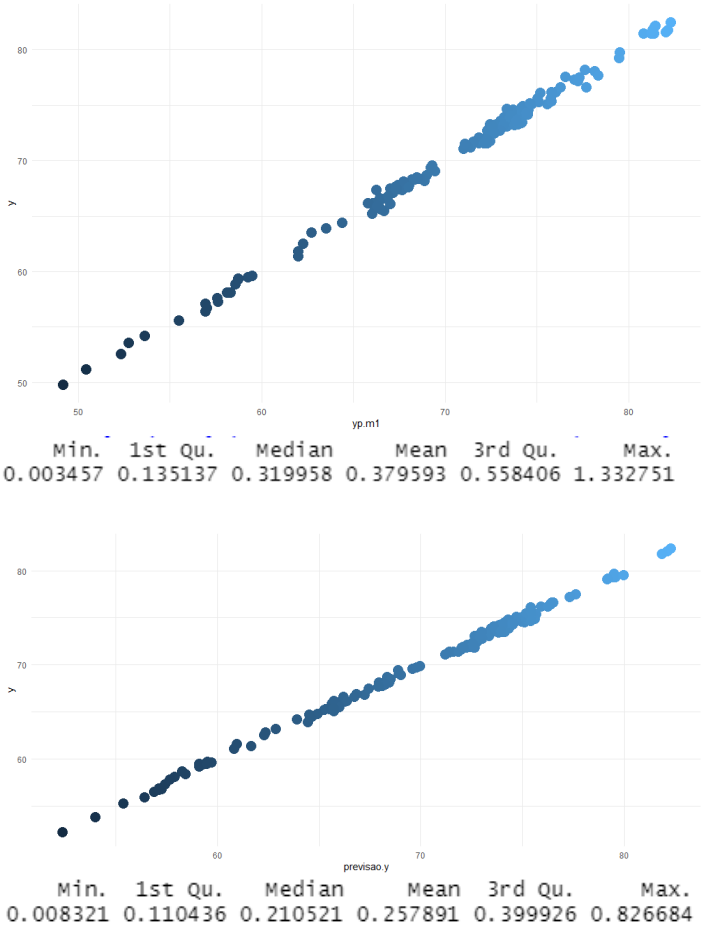


Figura 9: *Stepwise* (em cima) e *Bestsubset* (em baixo) após diagnóstico e remoção de *outliers*

Comparação de modelos								
Modelo	**Nº var.**	**MSE**	R^2	R^2_{adj}	**AIC**	**BIC**	**PRESSp**	**Cp**
Stepwise	11	0.091	0.9984	0.9981	-1307.4	-866.8631	61.40067	-551.2281
Bestsubsets	9	0.080	0.9986	0.9984	-1456.3	-992.4021	56.44894	-536.5741

Acrescentamos que **incluímos** na entrega dos ficheiros o **workspace** designado de *stepwiseSemOutliersEmEnvironmentSeparado*, que contém a variável *lifesemout* (dados para criar o modelo por *Stepwise* sem outliers). Assim, escolhemos o modelo obtido pelo *Best Subsets*, por ter menos variáveis.
Deixamos no relatório a descrição do diagnóstico e análise (e remoção) das observações atípicas para o modelo escolhido como melhor (do *Best Subsets*), e segue-se nas próximas secções.

3 Diagnóstico

3.1 Problemas no Erro

Apesar dos erros não serem observáveis, podemos observar resíduos, visto que estes são construídos com base nas suposições sobre os erros.

3.1.1 Heterogeneidade

Graficamente, interessa-nos procurar por algum tipo de padrão entre os resíduos e os valores ajustados pelo nosso modelo, como uma distribuição de pontos semelhante à de um megafone.
Pelo gráfico da figura 10, verificamos que os resíduos estão espalhados de forma relativamente simétrica em relação ao zero e o gráfico não apresenta um aspeto "megafone". No entanto, como os resíduos não estão distribuídos uniformemente ao longo da linha horizontal, este gráfico é inconclusivo perante a suposição da variância ser constante. Além disso, pelo gráfico da direita, também se pode verificar que não existe variância constante.

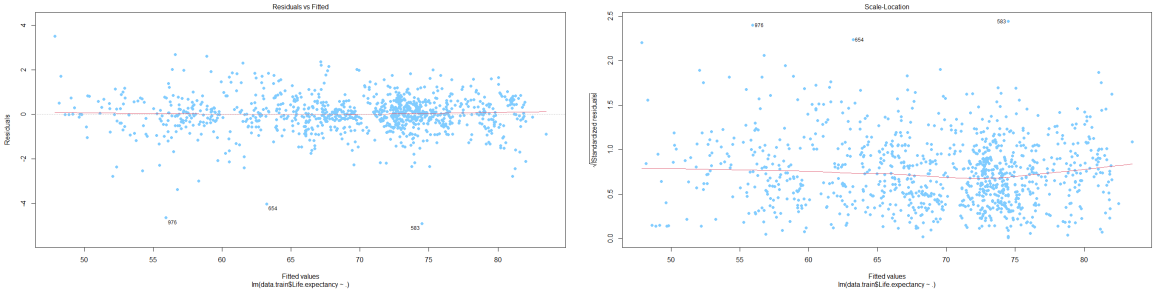


Figura 10: Valores dos resíduos vs valores ajustados (à esquerda) e Escala vs Localização para os resíduos (à direita).

Deste modo, para confirmar estas suspeitas, convém usar testes estatísticos que comprovem este facto. Para tal, recorreu-se ao teste de $Breusch - Pagan$ através do comando $ncvTest$ obtendo-se:

χ^2	Df	$Valor - p$
37.5996	1	8.6862e-10

A partir do $valor - p$, conclui-se que para os n.s.u. deve-se rejeitar H_0, que no teste de $Breush - Pagan$ refere-se à existência da homocedasticidade, pelo que se pode assumir que existe heteroscedasticidade.

Como solução a este problema, a transformação de $Box - Cox$ surge como opção para determinar qual expoente a aplicar à variável resposta. Através da rotina do R $boxCox$, foi obtido o gráfico da figura 11:

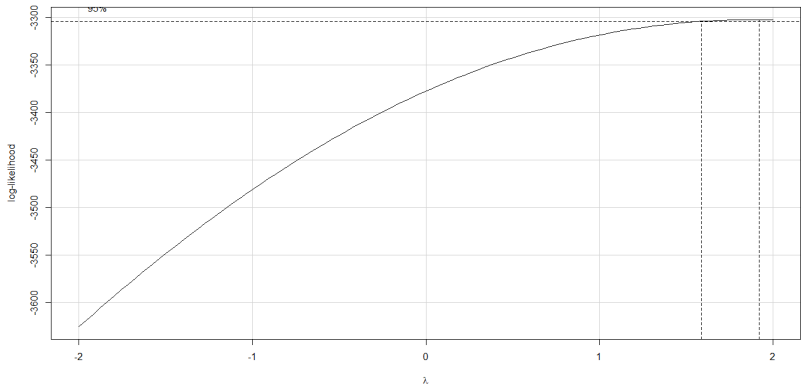


Figura 11: *Log*-Verosimilhança.

Como é possível observar, a melhor potência a elevar a variável resposta encontra-se entre os valores 1 e 2. Com a rotina $powerTransform$, obtemos que a potência à qual devemos elevar a variável resposta é 1.887818. No entanto, como este problema de variância não constante pode ser devido a outros fatores, como *outliers* influentes, voltará a ser analisado posteriormente.

3.1.2 Normalidade

Outra suposição feita na regressão linear é a normaliade dos resíduos. Para analisar esta suposição, é preciso recorrer a alguns gráficos apresentados na figura 12.

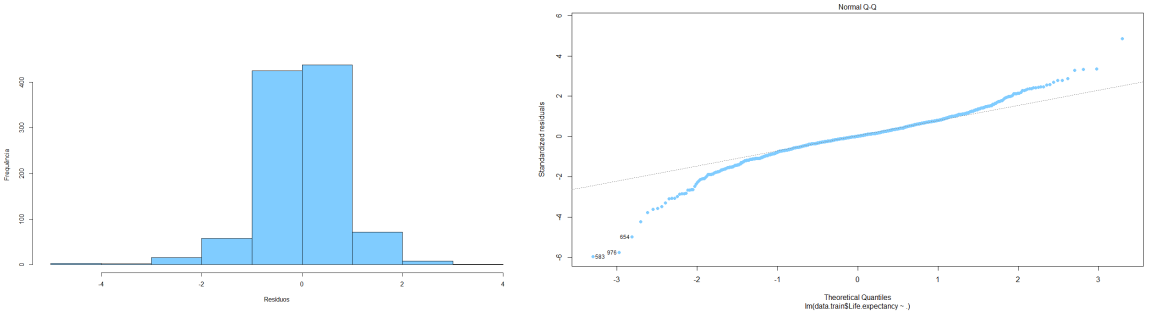


Figura 12: Histograma dos resíduos (à esquerda), *QQ Plot* dos resíduos para a distribuição normal (à direita).

Olhando rapidamente para o histograma, podemos identificar algumas informações importantes. Os valores dos resíduos distribuem-se entre -4 e 4. A concentração à direita é semelhante à concentração à esquerda, indicando que a distribuição da amostra em questão tem características simétricas, o que é característico de uma distribuição normal, visto esta ser simétrica.
Além disso, podemos ver que a maioria dos resíduos se concentram em torno do valor 0.

No entanto, pelo *QQplot*, vemos que os pontos das caudas falham na normalidade, porém os restantes não falham, por isso pode-se resumir apenas a um problema de variância não constante. Este gráfico representa uma comparação entre os quantis empíricos e os teóricos (que neste caso são os da normal). Se ambos os quantis tiverem vindo da mesma distribuição, então os pontos aproximam-se da reta a tracejado, o que não acontece nas caudas.

Para analisar melhor esta possível falta de normalidade, recorremos a testes estatísticos.
Decidimos usar 2 testes diferentes: O teste χ^2 de *Pearson* (para a normalidade) e o teste $Shapiro - Wilk$.
Ao aplicar as rotinas *pearson.test* e *shapiro.test*, obteve-se:

Pearson.test	
P	$Valor - p$
115.41	2.918e-12

Shapiro.test	
W	$Valor - p$
0.94953	2.2e-16

Como a hipótese nula é o modelo seguir uma distibuição normal e o $valor - p$ é muito pequeno, esta hipótese deve ser rejeitada aos níveis de significância usuais, pelo que se conclui que o modelo falha na normalidade.

Note-se que a variância não ser constante pode ter influência na normalidade dos resíduos, logo vamos voltar a verificar este problema quando se corrigir a heterocedasticidade.

3.1.3 Independência dos Erros/Correlação entre Resíduos

Para verificar a independência dos erros, decidimos verificar graficamente se existe ou não alguma relação entre os resíduos.

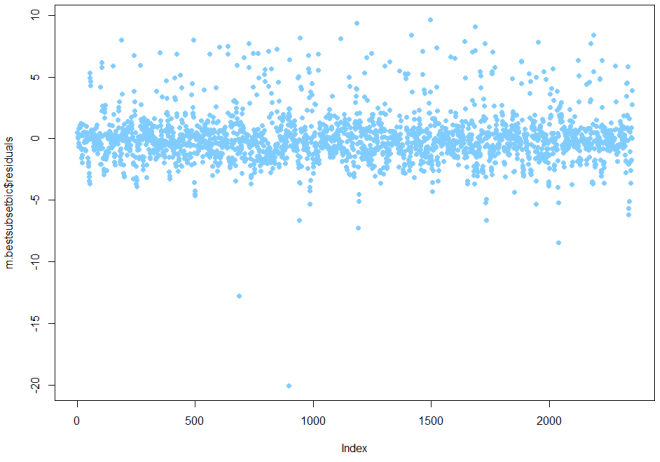


Figura 13: Resíduos para cada observação.

Pelo gráfico apresentado, podemos verificar que aparenta existir algum tipo de correlação entre os resíduos, visto os valores não se encontrarem dispersos nem aleatoriamente distribuídos.

De modo a verificar se os resíduos são independentes, realiza-se o teste de $Durbin - Watson$, e as correspondentes funções $durbinWatsonTest$ e $dwtest$ no programa R.

Obteve-se os seguintes resultados:

DW	$Valor - p$	Autocorrelação
1.21697	2.2e-16	0.3915046

Com este teste testa-se a hipótese nula $\rho = 0$ ($\rho = 0$ significa que não existe autocorrelação). O mesmo retorna valores entre 0 e 4, sendo que 2 traduz-se na não existência de autocorrelação, valores acima de 2 traduzem-se em autocorrelação positiva e abaixo de 2, negativa. Assim, conclui-se que os resíduos têm uma fraca autocorrelação (negativa).

3.2 Observações atípicas (ou $Outliers$)

3.2.1 Introdução

Antes de começar o estudo dos $outliers$ do nosso modelo $m.bestsubset$, fomos confirmar suspeita inicial de alta correlação entre as variáveis $under.five.deaths$ e $infant.deaths$.
Assim, fizemos uma análise rápida dos $VIF's$ (utilizando o comando vif) dessas variáveis e verificámos que, de facto, estavam altamente correlacionadas. $VIF_{infant.deaths} = 275.450687$, $VIF_{under.five.deaths} = 275.874261$ (enquanto as restantes não eram correlacionadas, i.e. $VIF_{restantes} < 5$).
Ao analisar qual destas duas variáveis seria a mais significativa no modelo, verificámos que por várias medidas os seus valores eram muito semelhantes:

Modelo	MSE	R^2	R^2_{adj}	AIC	BIC	$PRESS$	C_p
Sem $under.five.deaths$	0.78	0.9888	0.9871	−122.197	603.941	∞	−68.780
Sem $infant.deaths$	0.78	0.9889	0.9871	−126.431	599.708	∞	−72.63756

Inicialmente, ao retirar outliers ao modelo sem $infant.deaths$ (começámos por examinar este modelo por ter valores das medidas ligeiramente melhores), foi preciso retirar mais de 90% das observações para que passasse os testes das suposições.
No entanto, para o modelo sem $under.five.deaths$, conseguimos manter cerca de 38% das observações. Apesar de nenhum caso ser o desejado, preferimos optar pelo modelo com mais observações.

3.2.2 Análise e Remoção

Idealmente, a ideia para retirar $outliers$ seria utilizar vários critérios para obter o $outlier$ mais influente em cada um deles, e verificar qual a observação que estes

múltiplos critérios acusam como mais influente, e retirar a mais acusada (assim, esta seria o *outlier* mais influente, ditado por todos os métodos).
Por causa de existirem demasiados *outliers* influentes a retirar, não achámos pertinente agir dessa forma, tendo em conta que demoraria demasiado tempo.
Assim, agimos de forma algorítmica, tal que, servindo-nos da *Distância de Cook*, do *Teste de Bonferroni*, dos $DFFITS$ e da rotina $influencePlot$, fomos retirando *outliers* influentes acusados por todos estes testes/critérios.

Para remover os outliers, começou-se por programar um gráfico para os $DFFITS$ e colocar no gráfico duas linhas guias com base na constante $\pm 2\sqrt{\frac{k}{n}}$, (que serve como *cutoff* para saber se os outliers devem ser considerados influentes, e assim, guiar a remoção dos mesmos).
Seguidamente, implementou-se uma rotina que remove os valores do vetor *data.train* que se encontram fora das linhas guias.
Em seguida, implementou-se uma outra rotina que, com base no teste de Bonferroni (usando o comando $outlierTest$), seleciona e retira os *outliers* mais significativos (i.e. com $Bonferroni\ p < 0.05$).
Com base nestas rotinas definidas, a remoção dos outliers procedeu-se da seguinte forma, em que ilustramos cada passo apenas para a primeira aplicação do pseudo-algoritmo (i.e. aplicado ao modelo com todos os *outliers* por remover):

1. Removeu-se outliers com base nos DFFITS;

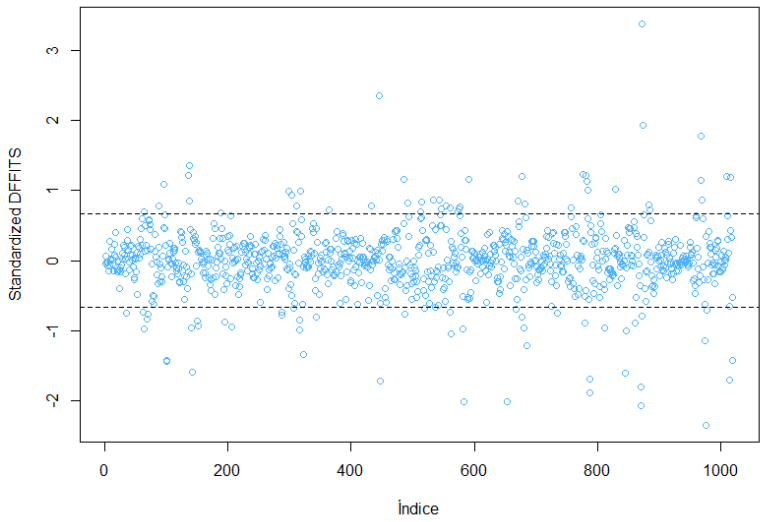


Figura 14: Outliers acusados por DFFITS (cutoff a tracejado)

2. Removeu-se outliers com base em Bonferroni;

	rstudent	unadjusted p-value	Bonferroni p
976	-6.139538	1.2511e-09	1.2761e-06
583	-5.973301	3.3721e-09	3.4396e-06
654	-5.495788	5.0955e-08	5.1975e-05
873	5.139998	3.3860e-07	3.4537e-04
968	4.557994	5.8943e-06	6.0122e-03
788	-4.383028	1.3117e-05	1.3379e-02

Figura 15: Outliers influentes acusados por Bonferroni (outlierTest)

3. Analisando o gráfico das distâncias de Cook, removeu-se as três observações com maior distância de Cook (supondo que eram maiores que o *cutoff* para a distância de Cook: $\frac{4}{n-k-1}$);

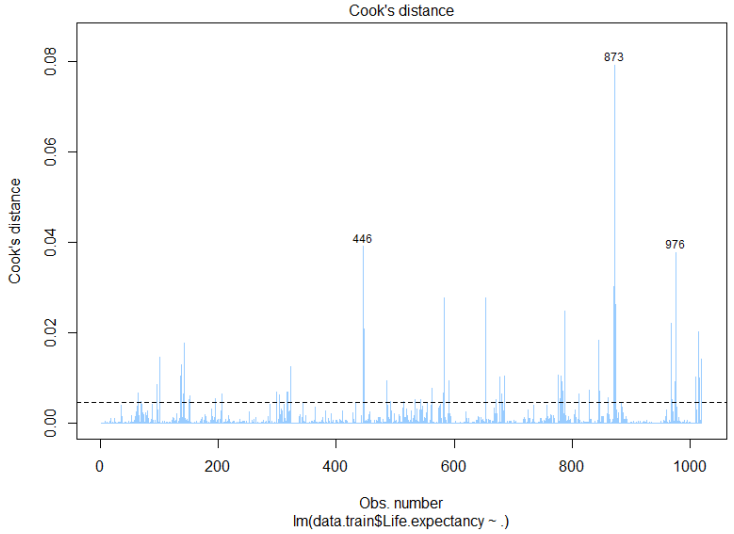


Figura 16: Outliers influentes acusados pela Distância de Cook (os três mais influentes aparecem identificados: 446, 873 e 976)

4. Analisando os valores devolvidos pelo comando $influencePlot$, retirou-se os que apresentavam $hat_{value} = 1$, pois, caso contrário, a presença destes valores implica $PRESS = \infty$.

	StudRes	Hat	CookD
446	3.293148	0.3395714	0.03910953
583	-5.973301	0.1020427	0.02766738
873	5.139998	0.3029233	0.07914172
976	-6.139538	0.1283575	0.03779317
1021	NaN	1.0000000	NaN
1022	NaN	1.0000000	NaN

Figura 17: Observações acusadas por influencePlot (só retirámos observações com $hat_{value} = 1$)

É de notar que, quando o programa *outlierTest* devolvia um $Bonferroni\ p > 0.05$, deixou-se de considerar a observação como *outlier* influente, e portanto, não foi removida.

Assim, com base nestes critérios, procedeu-se à execução deste "algoritmo" até o modelo passar as suposições exigidas para um modelo linear (designamos este modelo com menos *outliers* que passa os testes das suposições do modelo e com $PRESS$ finito de **M2**).
Depois de, finalmente, o modelo passar essas suposições, notámos que ainda se mantinha $PRESS = \infty$, pelo que iterámos apenas os passos 3 e 4, até o modelo verificar $PRESS$ finito (denominamos este modelo com ainda menos outliers e $PRESS$ finito de **M3**).
Nota: denominamos o modelo imediatamente antes de retirar outliers de **M1**.

Segue-se a seguinte tabela ilustrativa da melhoria do modelo, resultante da remoção dos outliers:

Modelo	MSE	R^2	R^2_{adj}	AIC	BIC	$PRESS$	C_p
M1	0.78	0.9888	0.9871	-122.197	603.941	∞	-68.780
M2	0.08	0.9988	0.9985	-1535.158	-981.639	∞	-571.545
M3	0.08	0.9986	0.9984	-1456.259	-992.402	56.449	-536.574

Notamos que o M2 tem medidas ligeiramente melhores que o M3, mas como a diferença não é significativa, o $PRESS$ do M3 é finito e como o M3 continua a ser muito melhor que o M1, escolhemos o modelo M3.

3.3 Multicolinearidade

Uma das suposições de um modelo linear é que todas as variáveis usadas no modelo são independentes. Sendo assim, é necessário proceder à análise do mesmo.
Como tal, considerou-se o critério $VIF's$, definido como se segue:

$$VIF(\hat{\beta_k}) = \frac{1}{1 - R_k^2} \tag{3.1}$$

Deste modo, obteve-se a seguinte tabela usando o comando *vif* do R (ao modelo sem a variável *Country*):

Variáveis	VIF
Adult.Mortality	1.777798
infant.deaths	1.140015
percentage.expenditure	1.340850
BMI	1.278058
Total.expenditure	1.105085
HIV.AIDS	1.330953
Income.composition.of.resources	2.746256
Schooling	3.217287

Como podemos ver pela tabela, os valores de *VIF* indicam que não há problemas de multicolinearidade entre as variáveis (estão todos abaixo de 5, considerado o *cutoff* a partir do qual há suspeita de problemas de multicolinearidade).

3.4 Verificação final e possível Retificação

Começamos por fazer novo diagnóstico para o modelo final, obtendo os seguintes gráficos:

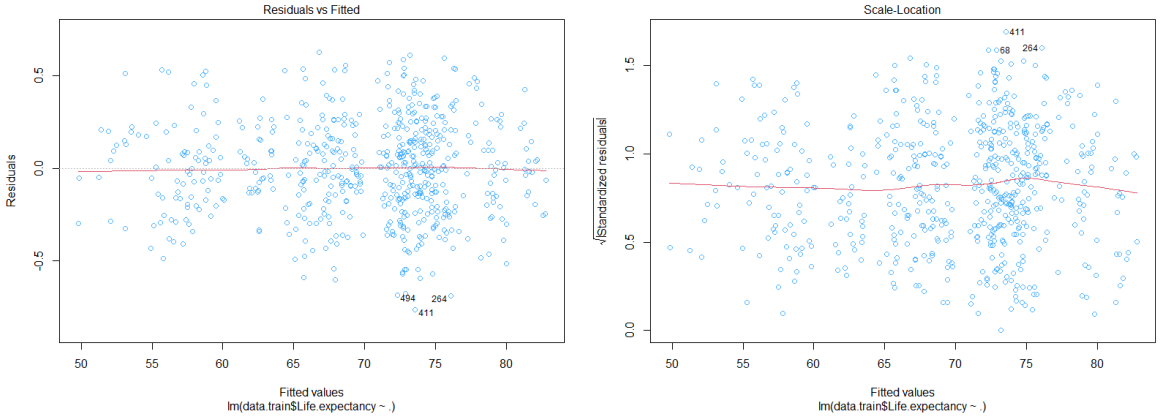


Figura 18: Valores dos resíduos vs valores ajustados (à esquerda) e Escala vs Localização para os resíduos (à direita).

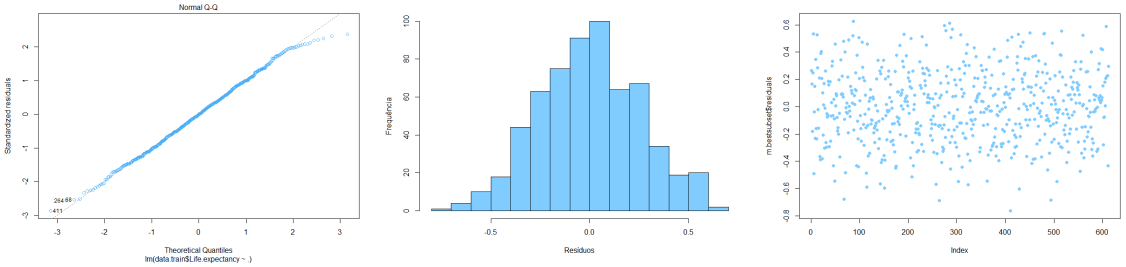


Figura 19: Histograma dos resíduos (à esquerda), *QQPlot* dos resíduos para a distribuição normal (ao centro) e Resíduos para cada observação (à direita).

Em relação aos primeiros dois gráficos (figura 18), apesar de na imagem à esquerda se notar uma distribuição aleatória dos resíduos, na imagem à direita, nota-se uma ligeira perturbação na curva a vermelho, devido ao facto de, mesmo tendo retirado tantos *outliers*, o nosso modelo continuar a ter alguns.
Nos últimos três gráficos (figura 19), nota-se que a normalidade do modelo se comprova, visto que o histograma está relativamente simétrico (ao centro) e vemos também pelo *QQPlot* que a normalidade do modelo se verifica (à esquerda). O gráfico à direita permite-nos concluir que não existe correlação entre os resíduos (porque os pontos no gráfico estão aleatoriamente distribuídos) e por isso a suposição da independência dos erros verifica-se.

Como podemos ver na tabela seguinte, o valor-p para todos os testes é superior a 10%, sendo que claramente não se rejeita as suas hipóteses para os níveis de significância usuais:

teste	valor obs. da estatística	valor-p
χ^2 Pearson	17.029	0.847
Shapiro-Wilk	0.997	0.212
Durbin-Watson	2.025	0.756
Breusch-Pagan	0.994	0.319

Para ilustrar a previsão do modelo final, apresentamos o gráfico seguinte da comparação entre os valores previstos e os observados:

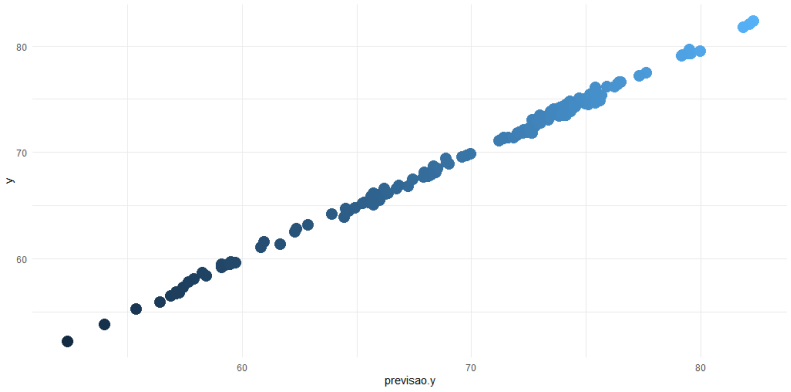


Figura 20: Previsão do modelo final

Além disso, apresentamos (de novo) as medidas do modelo final:

Modelo	MSE	R^2	R^2_{adj}	AIC	BIC	$PRESS$	C_p
Modelo final (M3)	0.08	0.9986	0.9984	-1456.259	-992.402	56.449	-536.574

Pela tabela, vemos que as medidas são muito boas. Lembramos que é preciso ter em conta que este R^2 explica 99.86% da variância da variável resposta, i.e. da esperança média de vida, ou seja, segundo os resultados obtidos, a esperança média de vida é altamente explicada pela origem das pessoas, pelas taxas de mortalidade (adulta e infantil), pelo índice de massa corporal e doenças como a sida, pela qualidade de ensino, gastos/despesas percentuais da população, gastos/despesas totais da população, e a composição do rendimento dos recursos de cada país (*Income.composition.of.resources*).

4 Conclusão

Os resultados alcançados deixam-nos globalmente satisfeitos. No entanto, é relevante mencionar que surgiram algumas limitações na construção deste modelo, nomeadamente a quantidade de dados que a OMS fornece não congruentes com a própria realidade. Houve, assim, necessidade de remover vários Outliers dos mesmos com vista à obtenção de um modelo descritivo da realidade.

Dentro do possível para este projeto, a melhor regressão encontrada para modelar a esperança média de vida tem como variáveis explicativas *Country, Adult.Mortality, infant.deaths, percentage.expenditure, BMI, Total.expenditure, HIV.AIDS, Income.composition.of.resources* e *Schooling*.

É de notar que este modelo é adequado pois é de esperar que, pessoas com doenças como a SIDA e com o índice de massa corporal fora dos conformes tenham um período de vida mais curto, que a educação escolar e os alicerces construtivos de uma pessoa tenham influência na maneira como vive a sua vida , que a despesa pública e os recursos que o estado dispõe não sejam irrelevantes, e também que a morte, como é óbvio, tanto de adultos como de crianças tenha influência nesta análise. Reparamos também que, sendo a variável *Total.expenditure* significativa no modelo e assumindo que esta engloba gastos na saúde pública, um país deve aumentar os seus gastos nesta área de modo a aumentar a esperança média de vida. No entanto, isto são tudo fatores que variam dependendo do quão desenvolvido é o país em questão, e por isso, a origem da pessoa é um fator importante que deve ter tido em consideração.

Isto, na verdade, poderá muito bem ter sido uma das razões para a presença de tantos *outliers*. Porque, como não tínhamos observações muito concordantes como um todo, mas sim, concordantes em grupos (i.e. as observações europeias têm valores muito mais parecidos entre si do que entre observações na África, por exemplo).
Portanto, para dizer a verdade, para levar possivelmente este projeto a um nível acima, poder-se-ia ter feito uma regressão para cada continente, por exemplo, ou talvez, pelo menos, uma regressão para os Países considerados Desenvolvidos e os considerados Em Desenvolvimento.

Deste modo, o modelo final que se apresenta é uma ferramenta que até poderá servir de auxílio para a OMS na sua tomada de decisões no que diz respeito a saber quais são, de facto, os fatores que afetam a esperança média de vida da população. No entanto, se aplicássemos o procedimento do projeto a um novo e melhor conjunto de dados inicial, aí sim, o modelo final obtido já poderia ser uma melhor ferramenta para a tomada de decisões da OMS (ou se se fizessem as várias regressões por região).

Para concluir, este trabalho permitiu-nos, em geral, abrir novos horizontes e descobrir o universo escondido por detrás da análise e construção de modelos lineares, através da programação em R. Assim, trabalhar com técnicas de ajustamentos de modelos lineares a conjuntos de dados foi algo absolutamente novo e também para isso este trabalho foi importante. Mais do que nunca, aqui realça-se a importância de se aprender a reformular, debater e saber questionar, de modo a decidir, em grupo, as razões que nos levam a tomar uma direção e não outra.

5 Referências

1. Documentação do R
2. Material disponibilizado da cadeira AML 2020/2021
3. Math Stack Exchange
4. University of Cambridge on collinearity
5. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning: With applications in R.