Análise de Modelos Lineares
Instituto Superior Técnico
Janeiro 2023

Projeto Computacional

Clara Pereira (99405)
Marta Sereno (99432)

Conteúdo

1 Introdução

1.1 Objetivo

No âmbito da cadeira de Análise de Modelos Lineares, o projeto computacional tem como objetivo a construção e o estudo de um modelo linear adequado ao problema em questão.

Neste estudo, pretende-se determinar se os programas de vigilância e controlo de infeções reduziram as taxas de infeção nosocomial nos hospitais dos Estados Unidos da América. Dado um conjunto de dados correspondente a uma amostra aleatória de 113 hospitais, será construído um modelo de regressão linear com base num *dataset* cujas variáveis explicativas são as seguintes:

- *Identification Number,* x_1 - número de identificação do paciente
- *Length of stay,* x_2 - permanência hospitalar (em dias)
- *Age,* x_3 - idade (em anos)
- *Routine culturing ratio,* x_4 - rácio entre o número de procedimentos efetuados e o número de pacientes sem sinais ou sintomas de infeção (em percentagem)
- *Routine chest X-ray ratio,* x_5 - rácio entre o número de raios-X efetuados e o número de pacientes sem sinais ou sintomas de pneumonia (em percentagem)
- *Number of beds,* x_6 - média de camas no hospital
- *Medical school affiliation,* x_7 - afiliação com medicina (1=Sim, 2=Não)
- *Region,* x_8 - região (1=NE, 2=NC, 3=S, 4=W)
- *Average daily census,* x_9 - média de pacientes por dia no hospital
- *Number of nurses,* x_{10} - número de enfermeiros
- *Available facilities and services,* x_{11} - percentagem de 35 potenciais intalações e serviços fornecidos pelo hospital

A variável resposta do modelo será *Infection risk* - risco de infeção (em percentagem).

O desenvolvimento do estudo que visa obter o modelo que melhor se adequa ao problema descrito acima está dividido em três partes. Em primeiro lugar, é necessário realizar um tratamento das covariáveis dadas, bem como da variável resposta. De seguida, através de métodos estudados, procede-se à redução do número de covariáveis que irão definir o modelo que será estudado como proposta para a resolução do problema. Por último, o modelo é analisado e avaliado, verificando-se se é ainda possível ser melhorado, e tiram-se possíveis conclusões tendo em consideração os métodos escolhidos, os resultados obtidos e o contexto do problema.

1.2 Introdução Teórica

Previamente à iniciação do estudo e da escolha do modelo linear mais adequado, deve ser feita uma introdução das noções e conceitos que serão necessários na determinação do mesmo e na interpretação dos resultados.

Modelo Linear Geral

Em forma matricial, o modelo é dado por

$$Y_{n\times 1} = X_{n\times p}\,\beta_{p\times 1} + \epsilon_{n\times 1}$$

onde $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ é o vetor de observações da variável resposta, $\mathbf{X}$ é a matriz de delineamento, com a linha $x_i^\top = (x_{i1}, \ldots, x_{ip})$ a representar a observação das p covariáveis do indivíduo i, $\beta = (\beta_1, \ldots, \beta_p)^\top$ é o vetor de parâmetros de regressão, e $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ é o vetor de componentes aleatórias.

O modelo possui as seguintes suposições

- $\epsilon_i \sim N(0, \sigma^2)$ independentes, $i = 1, \ldots, n$
- $Y \sim N(X\beta, \sigma^2 I_n)$
- *Ausência de Correlação entre as covariáveis* - as colunas de X são linearmente independentes

Modelo Completo e Modelo Reduzido

- M_C: $Y = X\beta + \epsilon$
- M_R: $Y = X\beta + \epsilon$ onde $\beta_i = 0$, para alguns $i \in \{1, \ldots, p\}$

Soma de Quadrados e Quadrado Médio

- $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ onde $\hat{Y} = \hat{E}(Y) = X\hat{\beta}$
- $MSE = \frac{SSE}{n-p}$
- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$
- $MSR = \frac{SSR}{p-1}$
- $SST = SSR + SSE$
- $MSPE = \sum_{i=1}^{n_*} \frac{(Y_i - X\hat{\beta})^2}{n_*}$ onde n_* designa o número de observações no conjunto de teste

Coeficientes

- *Coeficiente de Determinação:* $R^2 = \frac{SSR}{SST}$

  Avalia a qualidade do ajustamento do modelo, indicando a proporção da variável resposta Y que é explicada pelas covariáveis. Tanto maior quanto maior o número de variáveis explicativas.

- *Coeficiente de Determinação Ajustado:* $R^2_{aj} = 1 - \frac{(n-1)SSE}{(n-p)SST}$

  Onde $n-1$ corresponde ao total de graus de liberdade do modelo, e $n-p$ ao número de graus de liberdade do resíduo/erro. Mede o mesmo parâmetro que R^2, mas fornece informação mais precisa quando trabalhamos com um elevado número de covariáveis.

Note-se que R^2 não permite tirar as melhores conclusões sobre o modelo que se pretende estudar, já que é influenciado pela quantidade de variáveis explicativas, podendo originar conclusões enviesadas. Assim, R^2_{aj} será uma melhor alternativa para se chegar a uma conclusão, uma vez que tem em consideração o número de variáveis explicativas presentes.

Critérios

- *Critério de Informação de Akaike:* $AIC = -2\, log(L(\hat{\theta}|\mathcal{D})) + 2p$
- *Critério de Informação Bayesiano:* $BIC = -2\, log(L(\hat{\theta}|\mathcal{D})) + p\, ln(n)$

onde $L(\hat{\theta}|\mathcal{D})$ é a função de verosimilhança, $\mathcal{D}$ o conjunto de dados, e $\hat{\theta}$ o estimador de máxima verosimilhança do parâmetro θ com dimensão p.

Quanto melhor o modelo, menor será o seu valor de AIC ou BIC. Ao contrário do coeficiente de determinação, os critérios acima não são influenciados pelo número de variáveis explicativas, mas visam encontrar o valor equilíbrio do número de variáveis consideradas.

Conjunto de Treino e de Teste

Através de uma divisão aleatória dos dados, são formados o conjunto de treino, composto por 80% das observações, que é utilizado na construção do modelo, e o conjunto de teste, composto pelos restantes 20%, que é posteriormente usado na validação do modelo e no estudo da sua qualidade de previsão.

Transfromação Box-Cox

Transformação potência da variável resposta Y, dada por $Y' = Y^\lambda$, de modo a garantir a suposição de normalidade do modelo linear geral - pretende-se encontrar o valor λ que melhor aproxima as observações à distribuição normal.

Teste Wilk-Shapiro

Teste de hipóteses que visa determinar se a variável resposta Y tem uma distribuição aproximadamente normal.

H_0: *os dados seguem distribuição normal*
H_1: *os dados não seguem distribuição normal*

Métodos de Eliminação de covariáveis

- *stepwise* - método iniciado com um modelo de regressão sem as variáveis explicativas que queremos adicionar. A cada passo, as covariáveis serão incluídas/excluídas do modelo com critério num certo valor de entrada e de saída.
- *forward selection* - método iniciado com um modelo de regressão sem as variáveis explicativas que queremos adicionar. A cada passo, uma covariável é potencialmente adicionada, desde que cumpra o critério de entrada previamente estabelecido.
- *backward elimination* - método iniciado com o modelo completo, com todas as variáveis explicativas. A cada passo, uma covariável é potencialmente excluída, se verificar o critério de saída previamente estabelecido.

Todos os métodos descritos acima têm como critério de entrada/saída de covariáveis um certo valor da medida AIC, como $default$ do programa R. No entanto é possivel recorrer a outros critérios, nomeadamente testes F parciais, para a escolha.

Teste F parcial

Teste de hipóteses que pretende, neste estudo, averiguar se as covariáveis eliminadas com o método *stepwise* escolhido não explicam cabalmente a variável resposta. Tal traduz-se nas seguintes hipótestes

H_0: $\forall \beta_i \in M_C$ e $\beta_i \notin M_R$, $\beta_i = 0$
H_1: $\exists \beta_i \in M_C$ e $\beta_i \notin M_R : \beta_i \neq 0$

Estatística de Teste: $F = \frac{Q/q}{MSE} \sim F_{(q,n-p)}$

Distância de Cook

Medida que testa a influência de uma observação em todos os coeficientes de regressão β. Considera-se que a observação i é influente no modelo quando a sua distância de Cook é superior a 1 ($CD_i > 1$).

Resíduos

- *resíduo habitual:* $r_i = y_i - \hat{E}(Y_i)$

- *resíduo padronizado:* $r_i^S = \frac{r_i}{\sqrt{MSE}}$

  Se $r_i^S > 3\sqrt{MSE} ou > 3\sqrt{MSE}$, a *i-ésima* observação é tida como atípica

- *resíduo eliminação:* $d_i = y_i - \hat{y}_{-i}$

  onde $\hat{y}_{-1}$ designa o valor predito pela i-ésima observação com resposta y_i e covariáveis x_i, tendo em conta que o modelo foi ajustado sem esta observação.

  Se o valor de $|r_i - d_i|$ for elevado, a *i-ésima* observação é tida como atípica

- *resíduo internamente estudantizado:* $r_i^{IS} = \frac{r_i}{\sqrt{MSE(1-h_{ii})}}$
- *resíduo externamente estudantizado:* $r_i^{ES} = \frac{r_i}{\sqrt{MSE_{-i}(1-h_{ii})}}$

  onde h_{ii} representa a entrada i da diagonal da matriz $H = X(X^\top X)^{-1}X^\top$

Filtro de Hampel

Método que tem como objetivo detetar *outliers*. Os valores que se encontram fora do intervalo

$$I = [mediana - 3\ MAD; mediana + 3\ MAD]$$

são considerados *outliers*, onde $MAD = mediana(|Y_i - mediana(Y)|)$ designa o desvio absoluto mediano da observação Y_i.

Fator de Inflação na Variância

Medida que deteta a presença de multicolinearidade, dada por $\overline{VIF} = \frac{1}{p-1}\sum_{k=1}^{p-1}(1 - R_k^2)^{-1}$
onde R_k^2 é o coeficiente de determinação múltipla da regressão de x_k nas restantes $p-2$ covariáveis. Se o valor de $\overline{VIF}$ é muito superior a 1, diz-se que existem sérios problemas de multicolinearidade no modelo.

Teste Breusch-Pagan

Teste de hipóteses que pretende averiguar a homocedasticidade dos dados, ou seja, se a sua variância é constante. Neste estudo em particular, este teste será utilizado para verificar a condição de homocedasticidade relativamente aos resíduos associados ao modelo linear.

H_0: *Os dados apresentam homocedasticidade*
H_1: *Os dados não apresentam homocedasticidade*

Teste Durbin-Watson

Teste de hipóteses utilizado para detetar a presença de autocorrelação nos resíduos das variáveis preditoras.

H_0: *Os resíduos não estão autocorrelacionados*
H_1: *Os resíduos estão autocorrelacionados*

Nota

As estatísticas de teste e respetivas distribuições de alguns testes de hipóteses utilizados serão omitidas deste relatório pela sua complexidade, e uma vez que constituem testes automáticos realizados pelo *R*, não será necessário ter a fórmula da estatística presente aquando do teste.

2 Tratamento de Dados

2.1 Tratamento Preliminar dos Dados

A amostra fornecida para este estudo tem dimensão 113 - correspondentes aos 113 hospitais onde foram recolhidos os dados. Após a divisão dos dados em *conjunto de treino* e *conjunto de teste*, dispomos de 91 observações para a construção de uma regressão linear apropriada. Idealmente, o número de observações deve ser mais elevado, de forma a termos resultados mais exatos.

Deve-se também aplicar o comando $as.factor$ às covariáveis qualitativas - *Medical school affiliation* e *Region* - uma vez que o seu tratamento será distinto das demais covariáveis.

2.2 Normalidade da Variável Resposta

De seguida, foi necessário determinar se a hipótese de normalidade da variável de resposta Y se verifica. Para tal, procedeu-se a anásile analítica, através do comando *gvlma*, que verifica as suposições gerais do modelo, e do teste de hipóteses de Shapiro-Wilk.

Os resultados podem ser consultados em 1 e em 2

Pelo *output* do comando *gvmla*, podemos observar que as suposições gerais do modelo de regressão não são todas verificadas.
Por outro lado, o teste de Shapiro-Wilk devolve um *valor-p* de 0.03769, pelo que não dispomos de evidência suficiente para aceitar a hipótese de normalidade - para os níveis de significância 5% e 10%, H_0 não é aceite.

Para corroborar a suspeita de falta de normalidade de Y, recorreu-se também a análise gráfica:

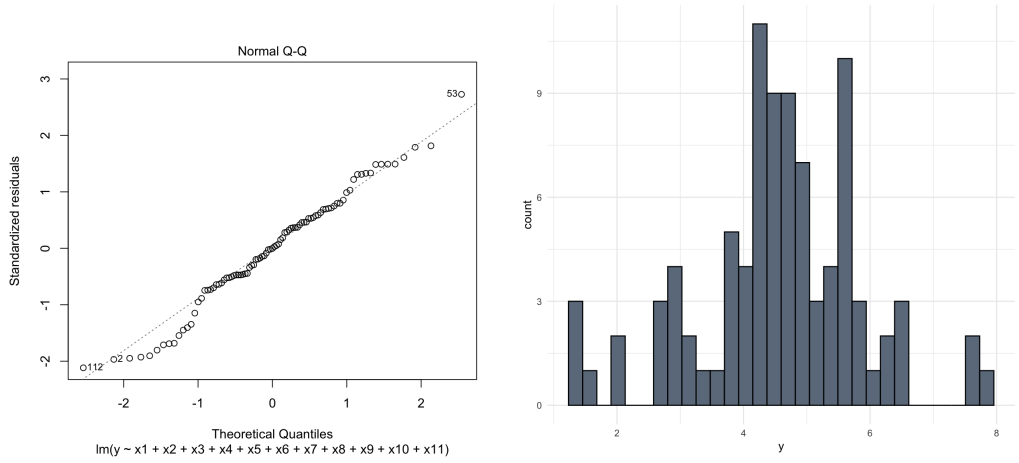

Figura 1: *QQplot* e Histograma de Y

O histograma obtido não aparenta normalidade, e a interpretação do gráfico *qqnorm* leva-nos a concluir também que a variável resposta não cumpre a suposição em questão: num gráfico *qqnorm*, queremos que os pontos (observações) estejam mais perto da linha a traçejado quanto possível, o que não acontece - nota-se a presença de várias observações bastante afastadas desta linha, o que indica que não coincidem com o esperado numa distribuição normal.

Deste modo, temos reunida informação suficiente para concluir que Y não segue, de facto, distribuição normal. O próximo passo será proceder a transformações desta variável, até ser obtida a distribuição pretendida.

Realizaram-se algumas transformações, nomeadamente $log(Y)$, $\sqrt{Y}$, $1/Y$ e $sin(Y)$, na tentativa de cumprir a suposição de normalidade. Abaixo apresentamos os resultados obtidos para a transformação $sin(Y)$.

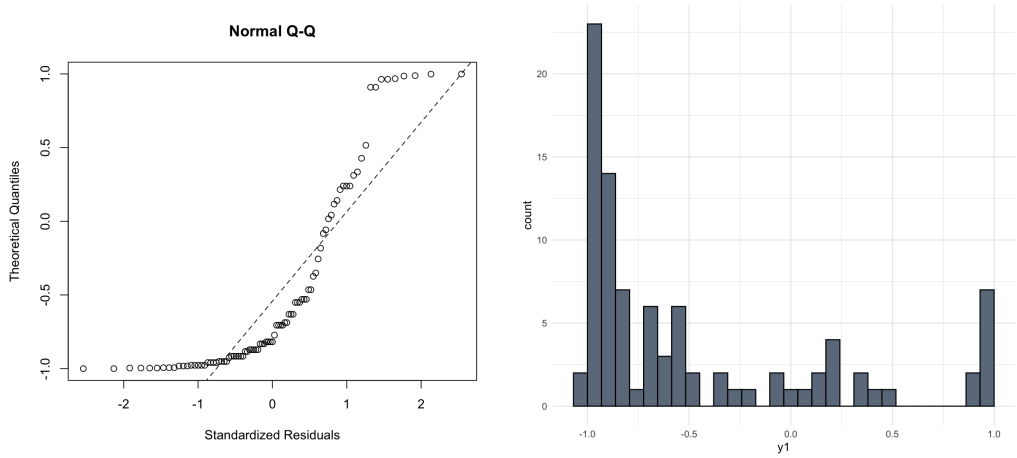

Figura 2: *QQplot* e Histograma de $sin(Y)$

Nenhuma das transformações anteriores mostrou ser uma melhor opção em relação aos dados originais, pelo que se procedeu à transformação de *Box-Cox*. O valor obtido para λ foi 1.235436 isto é, $Y^{1.235436}$ representa a melhor transformação da variável resposta no sentido de cumprir a suposição de normalidade.

Realizou-se, novamente, o teste de Shapiro-Wilk, cujo resultado pode ser consultado em 3, e traçou-se o gráfico *qqplot*:

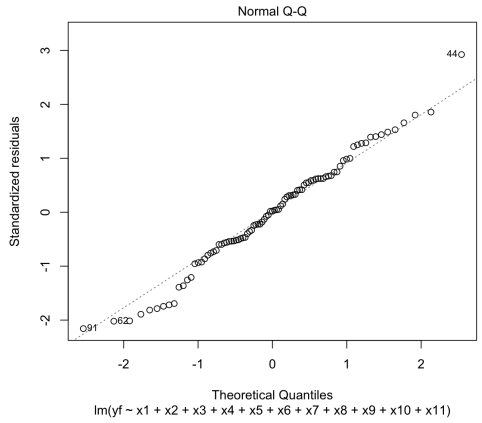

Figura 3: *QQplot* de $Y^{1.235436}$

Começando pelo teste de hipóteses, verificamos que o *valor-p* permite aceitar H_0 aos níveis de significância usuais de 1% e 5%. O gráfico *qqplot* também evidencia melhorias, uma vez que as observações estão mais próximas da linha a traçejado. Podemos então considerar a suposição de normalidade da variável resposta como cumprida.

2.3 Análise Preliminar do Modelo Completo

Implementou-se, utilizando os dados de teste, o modelo linear completo, através do comando *lm* do R. Obteve-se os seguintes indicadores:

Modelo	R^2	R^2_{aj}	AIC	BIC
Completo	0.6119	0.5464	90.53488	125.6869

Idealmente, tanto para R^2 como para R^2_{aj}, espera-se um valor tão perto de 1 quanto possível. Atendendo a este facto, evidentemente os valores obtidos são insatisfatórios, o que indica que o modelo não estará a explicar a variável resposta da melhor maneira.

Temos agora de realizar algoritmos que visam melhorar os valores dos indicadores acima, que por sua vez traduzem a qualidade do modelo. Nos próximos capítulos, será desenvolvido o estudo do modelo de regressão, e sua consequente alteração, seja pela exclusão de covariáveis, introdução de interações de segunda ordem entre as variáveis, ou tratamento de *outliers* e pontos influentes.

3 Melhor modelo de Regressão Linear

3.1 Métodos de Eliminação de Covariáveis

Uma possível justificação para o desempenho insuficiente dos valores obtidos anteriormente é o facto de existirem covariáveis no modelo que não explicam cabalmente a variável de resposta.
Por exemplo, no contexto do problema, é intuitivo supor que a variável *Identification Number* não influencia a variável resposta *Infection Risk*, o risco de infeção de um paciente; deste modo, prevê-se que essa variável venha a ser retirada do modelo em estudo.

Foram, então, aplicados 3 métodos de eliminação de covariáveis: *Stepwise*, *forward selection*, e *backward elimination*.

Os 3 métodos devolveram o mesmo resultado. Assim, as variáveis que contribuem significativamente no modelo são as seguintes:

- *Length of stay*
- *Age*
- *Routine culturing ratio*
- *Routine chest X-ray ratio*
- *Region*
- *Available facilities and services*

Foram, portanto, retiradas as variáveis *Identification Number* - como já tinhamos previsto, *Number of beds*, *Medical school affiliation*, *Average daily census* e *Number of nurses*.

Realizou-se um teste de hipóteses F parcial, de modo a confirmar que as covariáveis retiradas têm, efetivamente, contribuição nula no modelo.
Analisando o *valor-p* em 4, temos que os níveis de significância usuais 1%, 5% e 10% se encontram fora da Região de Rejeição, pelo que H_0 é aceite, isto é, as variáveis retiradas do modelo não explicam cabalmente a variável resposta Y, logo, é pertinente retirá-las do modelo.

Analisa-se novamente algumas medidas indicadoras da qualidade da regressão:

Modelo	R^2	R^2_{aj}	AIC	BIC
Completo	0.6119	0.5464	90.53488	125.6869
Reduzido	0.6079	0.5697	81.46747	104.0652

Mais uma vez, note-se que R^2 é sensível ao número de covariáveis, pelo que a diminuição do seu valor não é surpreendente. Tendo em conta uma medida mais adequada, R^2_{aj}, observa-se uma melhoria no valor, como desejado. Os critérios AIC e BIC também apresentam um valor mais baixo, indicando que a exclusão de covariáveis foi pertinente.

Em suma, temos evidência de que o modelo reduzido tem qualidade superior ao modelo completo com que se inicializou este estudo.

3.2 Covariáveis de Segunda Ordem

Esta etapa consiste em repetir o passo anterior, utilizando modelos de seleção, desta vez para tentar encontrar interações de primeira ordem entre as covariáveis selecionadas para o modelo novo.
Foram aplicados os mesmos 3 métodos de seleção: *Stepwise*, *forward selection*, e *backward elimination*. O resultado obtido foi igual para os três métodos e corresponde a interações entre:

- Age *e Routine chest X-ray ratio*, $x_3 : x_5$
- *Routine culturing ratio* e *Available facilities and services*, $x_4 : x_{11}$
- *Region* e *Available facilities and services*, $x_8 : x_{11}$

3.3 Modelo Final (provisório)

Por fim, é obtido o modelo final provisório, isto é, o melhor modelo obtido com o conjunto de dados de teste.

$$Y \sim x2 + x3 + x4 + x5 + x8 + x11 + x3 : x5 + x4 : x11 + x8 : x11$$

Podemos comparar todos os modelos até agora estudados:

Modelo	R^2	R^2_{aj}	AIC	BIC
Completo	0.6119	0.5464	90.53488	125.6869
Reduzido	0.6079	0.5697	81.46747	104.0652
Provisório	0.6739	0.6189	74.70345	109.8555

Quanto a R^2_{aj}, o crescimento é evidente, o que nos indica que a proporção de Y que é explicada pelo modelo aumentou, isto é, a contribuição do modelo com interações de segunda ordem na variável resposta é maior que a do modelo completo.
Os critérios AIC e BIC diminuíram em relação aos seus valores no modelo completo, sendo que AIC apresenta o seu valor mais baixo no modelo com interações de segunda ordem.

Podemos então concluir, com base nos indicadores, que o modelo com interações de segunda ordem é o que melhor se ajusta ao problema em estudo.

3.4 Análise de Outliers e Tratamento de Pontos Influentes

Às observações "aberrantes", ou discordantes das restantes, chamamos de *outliers*. É necessário verificar se estes existem , e se estão potencialmente a prejudicar a qualidade/capacidade preditora do modelo de regressão.

Os pontos que têm elevada influência nos parâmetros do modelo são designados *pontos influentes*. Assim, procura-se pontos influentes que sejam discordantes do resto das observações, de modo a suscitar uma alteração positiva na qualidade do modelo em estudo, quando estes forem retirados.

Os *outliers* e pontos influentes podem ser detetados analiticamente, mas por vezes é útil, para fornecer alguma intuição, procurá-los graficamente. Para tal, traçámos um gráfico *boxplot* e um $influencePlot$:

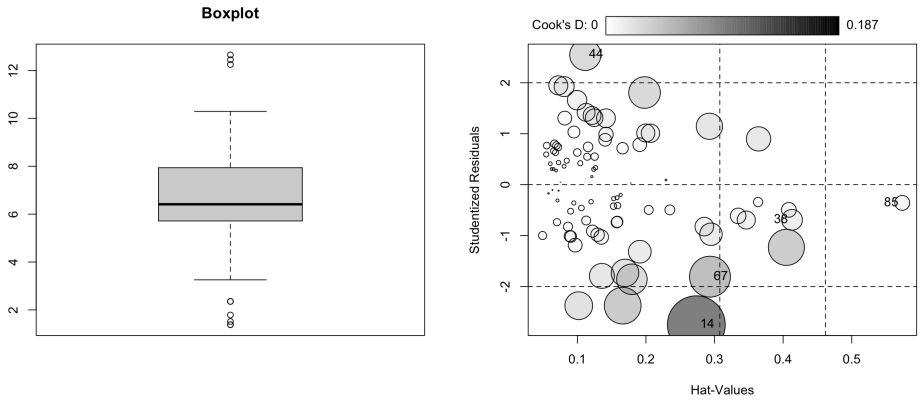

Figura 4: *Boxplot* e $InfluencePlot$, respetivamente

Os pequenos círculos mais acima e mais abaixo do gráfico *boxplot* representam potenciais outliers. No entanto, vale a pena estudá-los com mais cuidado, antes de os retirar do conjunto de dados.

No gráfico $influencePlot$, os círculos maiores correspondem a observações mais discordantes. Quanto mais escuro um círculo está colorido, maior a distância de Cook da observação correspondente. Por outro lado, a distância à reta $y = 0$ traduz um maior valor de resíduo estudantizado, o que pode indicar que a observação é discrepante. Analisando o gráfico, podemos aferir que provavelmente a observação 14 será considerada ponto influente, uma vez que o círculo correspondente está pintado de cinzento escuro, e que também é potencialmente *outlier*, pela maior distância à reta $y = 0$.

Procedemos a análise algébrica para detetar *outliers* e pontos influentes de forma mais rigorosa. Através do comando $influence.measures$, que combina várias medidas de influência, como a distância de Cook, encontrámos pontos influentes. De seguida, recorremos a análise dos resíduos padronizados, a comparação dos resíduos eliminação e internamente estudantizados, e à medida Hampel Filter, para deteção de *outliers*.

	Nº da Observação
Pontos Influentes	$1, 7, 14, 15, 32, 38, 39, 44, 45, 85, 89$
Padronizados	$14, 29, 44$
Estudantizados	$14, 29, 44, 67, 70, 77, 82, 84$
Hampel Filter	45
Escolhidos	$14, 29, 44, 45, 84$

Procedeu-se então à exclusão dos pontos escolhidos, correspondentes à última linha da tabela. Mantendo a estrutura do modelo obtida no final do capítulo anterior, obtiveram-se as seguintes medidas:

Modelo	R^2	R^2_{aj}	AIC	BIC
Completo	0.6119	0.5464	90.53488	125.6869
Reduzido	0.6079	0.5697	81.46747	104.0652
Provisório (com outliers)	0.6739	0.6189	74.70345	109.8555
Provisório (sem outliers)	0.7151	0.6636	47.12117	88.431

Ainda que os resultados sejam os melhores de entre todos os anteriores, foram repetidos os passos do capítulo 3, isto é, foi construído de raiz um novo modelo, aplicado ao novo conjunto de dados sem outliers. As variáveis explicativas obtidas foram as mesmas (apesar de os seus coeficientes terem alterado o seu valor), no entanto surgiu mais uma interação de segunda ordem entre as covariáveis *Routine chest X-ray ratio* e *Routine culturing ratio*.

O modelo final sem *outliers* é então:

$$Y \sim x2 + x3 + x4 + x5 + x8 + x11 + x3 : x5 + x4 : x5 + x4 : x11 + x8 : x11$$

As medidas indicadoras obtidas apresentam-se em baixo:

Modelo	R^2	R^2_{aj}	AIC	BIC
Completo	0.6119	0.5464	90.53488	125.6869
Reduzido	0.6079	0.5697	81.46747	104.0652
Provisório (com outliers)	0.6739	0.6189	74.70345	109.8555
Provisório (sem outliers)	0.7151	0.6636	53.90828	88.431
Final	0.729	0.6755	52.93256	92.38709

Conclui-se, finalmente, que o modelo que melhor se adequa ao problema em estudo é o último obtido - com remoção de outliers, redução de covariáveis e interações de segunda ordem. Predemos à interpretação dos coeficientes e do modelo em geral.

Os coeficientes obtidos foram:

β_0	β_2	β_3	β_4	β_5	β_{82}	β_{83}	β_{84}
−2.9558362	0.2553293	−0.0003573	−0.0515711	0.0344879	0.1058643	−2.6129135	2.0585336

β_{11}	$\beta_{3:5}$	$\beta_{4:5}$	$\beta_{4:11}$	$\beta_{82:11}$	$\beta_{83:12}$	$\beta_{84:11}$
0.0423357	0.0075310	−0.0015291	−0.0020314	0.0069239	0.0733227	−0.0115665

O coeficiente β_0 representa o risco de infeção de um paciente quando a variável x_8 tem valor 1, e as restantes têm valor nulo. No entanto, atendendo ao facto de que o seu valor é negativo, esta interpretação não tem sentido no contexto do problema - não existe significado para percentagem de risco de infeção negativa.

Os restantes coeficientes (excepto β_{82}, β_{83} e β_{84}) representam a variação do risco de infeção do paciente, quando a variável respetiva aumenta em 1 unidade, mantendo as restantes constantes - no contexto do problema, se considerarmos dois indivíduos cujos valores em todos os parâmetros sejam iguais, excepto na variável x_i em que o valor medido difere de 1 unidade, a diferença no seu risco de infeção será de β_i.

β_{82}, β_{83} e β_{84} correspondem, respetivamente, à variação do risco de infeção do paciente, quando este se encontra na região NC, S e W, e mantendo as restantes covariáveis constantes.

Algumas conclusões relevantes são, por exemplo, que a idade de um indivíduo (*Age*, x_2) está relacionada com um maior risco de infeção nosocomial.

4 Diagonóstico

Após a escolha de modelo e o tratamento de *outliers*, pode-se, então, proceder a uma análise do modelo final, de modo a averiguar a sua qualidade e tirar possíveis conclusões no contexto do problema.

4.1 Validação do Modelo Obtido

Começamos por notar, novamente através do comando *gvlma*, que as suposições apresentadas são todas verificadas para o modelo final. 5

De modo a aprofundar o estudo de validação, serão aplicadas medidas de diagnóstico e realizados testes de hipóteses adicionais.

Em primeiro lugar, pretende-se testar a multicolinearidade das covariáveis, e, para tal, recorre-se ao Fator de Inflação na Variância (VIF). Por ter sido obtido o valor $\overline{VIF} = 1.596164$, que não é muito superior a 1, pode-se concluir que não existem problemas de multicolinearidade que afem negativamente o modelo.

De seguida, recorrendo ao teste de hipóteses Breusch-Pagan, testa-se a suposição de homocedasticidade, utilizando o comando *ncvTest* do *R*. Resultou um *valor-p* de 0.43705, então aceita-se H_0 para todos os níveis de significância habituais, ou seja, os dados apresentam variância constante, resultado corroborado pelo output mencionado anteriormente.

Por último, através do teste de hipóteses Durbin-Watson, pode-se detetar a presença de autocorrelação nos resíduos das variáveis preditoras, com auxílio do comando *durbinWatsonTest*. Obteve-se um *valor-p* de 0.798, logo aceita-se H_0 para todos os níveis de significância habituais, ou seja, os resíduos não se encontram autocorrelacionados.

Assim, como as suposições do modelo são verificadas, pode-se concluir que o modelo final é um modelo de regressão linear adequado aos dados fornecidos.

4.2 Capacidade de Previsão

Com o objetivo de testar a capacidade de previsão do modelo, procede-se à determinação do erro quadrático médio de previsão (MSPE), valor que pode ser interpretado como a média do quadrado da distância entre os valores observados e os valores previstos pelo modelo.

Note-se que anteriormente à obtenção do modelo, foi realizada a trasnformaçaõ *Box-Cox* de modo a normalizar a variável resposta. Assim, é necessário reverter este processo, caso contrário estar-se-ão a tirar conclusões baseadas em valores enviesados.

Dado o alcance $[1.3, 7.8]$ de valores das observações, como a raíz de MSPE é da mesma ordem de grandeza de Y, conclui-se que o modelo apresenta uma baixa capacidade de previsão: a distância média entre uma observação e o seu valor previsto pelo modelo é da mesma ordem que a própria observação, podendo, em alguns casos - nomeadamente no mínimo do alcance, representar um erro de quase 100%.

Por fim, foi construído um gráfico de previsão, de modo a verificar a conclusão acima visualmente.

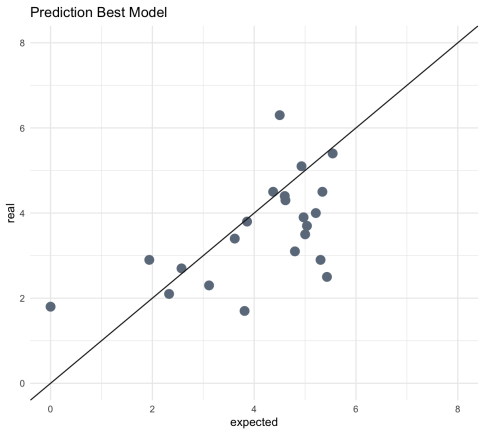

Figura 5: Gráfico de Previsão do Modelo Final

Caso a capacidade do modelo fosse ideal, os dados cairiam sobre a bissetriz dos quadrantes ímpares, ou seja, os valores das observações e das previsões seriam iguais. Como podemos observar, de maneira geral, apesar de ser observado um afastamento pequeno, os pontos apresentam-se afastados da reta, novamente comprovando a conclusão.

Este resultado pode ser explicado pelo número reduzido de observações no conjunto de teste.

5 Conclusões

Neste estudo, foram realizados vários algoritmos e transformações, visando alcançar o melhor modelo de regressão linear, que explicasse cabalmente a variável *Infection Risk*.
A utilidade deste modelo seria prever o risco de infeção nosocomial de um paciente, tendo acesso aos dados das variáveis explicativas: *Length of stay*, *Age*, *Routine culturing ratio*, *Routine chest X-ray ratio*, *Region* e *Available facilities and services*.

É de salientar que, para o objetivo do estudo, a informação de que se dispõe apresenta algumas limitações, nomeadamente a dimensão da amostra, que pode levar a conclusões menos exatas nas várias etapas do processo, que seriam evitadas caso contrário. Para além disso, deve-se notar que o número elevado de covariáveis aumenta a complexidade do modelo, o que dificulta a sua estrutura, e, consequentemente, o seu estudo e prestação.

Na fase de validação, confirmou-se que o modelo obtido verifica todas as suposições do modelo de regressão linear, e, no entanto, a sua capacidade de previsão não é considerada satisfatória. Assim, pode-se concluir que o método de seleção do modelo foi adequado e produz resultados aceitáveis, mas que a amostra reduzida leva a erros elevados nos valores previstos pelo modelo, quando avaliado com conjuntos de dados distintos.

Em suma, considera-se que o resultado não constitui o modelo mais adequado para prever o risco de infeção de um indivíduo num dado hospital, ainda que explique satisfatoriamente os dados com os quais foi construído. Em alternativa, propõe-se a recolha de dados adicionais, de forma a desenvolver um novo modelo, ou a pesquisa de um modelo com características diferentes das utilizadas neste estudo.

Referências

[1] S. M. Ross (2009), *Introduction to Probability and Statistics for Engineers and Scientists,* Elsevier, Academic Press, 4th ed

[2] J. E. Gentle, (2003), *Random Number Generation and Monte Carlo Methods,* Springer, 2nd ed

Apêndice

Código

Tratamento Preliminar dos Dados

```
install.packages("gvlma")
install.packages("MASS")
install.packages("car")
install.packages("ggplot2")
install.packages("carData")
install.packages("VGAM")

library(MASS)
library(gvlma)
library(ggplot2)
library(car)
library(VGAM)


#2 TRATAMENTO DE DADOS


#2.1 Tratamento Preliminar dos Dados

d <- read.table("Trabalho2_Dados.txt")
d[,8] <- as.factor(d[,8])
d[,9] <- as.factor(d[,9])
y <- d$V4
dados <- data.frame(y, x1=d$V1, x2=d$V2, x3=d$V3, x4=d$V5, x5=d$V6, x6=d$V7, x7=d$V8, x8=d$V9,
    x9=d$V10, x10=d$V11, x11=d$V12)

# Conjunto de Treino e Conjunto de Teste
s <- sample(113,22)
S = c(6, 87, 3, 35, 25, 97, 49, 59, 108, 70, 99, 76, 17, 61, 106,79, 9, 96, 14, 85, 113, 8)
teste.dados <- dados[S,]
treino.dados <- dados[-S,]
y <- treino.dados$y

# Modelo Completo
n<-dim(treino.dados)[1]
modc=lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11, treino.dados)
gvlma(modc)
plot(modc)
```

Normalidade da Variável Resposta

```
#2.2 Verificar Normalidade da Variável Resposta

# Histograma
ggplot(treino.dados, aes(x=y)) + geom_histogram(color = "black", fill = "slategray4") + theme_bw()

# Teste Shapiro-Wilk - antes de transformação
shapiro.test(treino.dados$y)

# sin(y)
y1 <- sin(y)
qqnorm(y1, main = "Normal Q-Q Plot",
       xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
qqline(y1, datax = FALSE, distribution = qnorm,probs = c(0.25, 0.75), qtype = 7)
ggplot(treino.dados, aes(x=y4)) + geom_histogram(color = "black", fill = "slategray4") +
    theme_bw()

# Transformação Box-Cox
pt <-powerTransform(treino.dados$y, family = "bcPower")
lambda <- pt$lambda
yf <- treino.dados$y^lambda

# Teste Shapiro-Wilk - depois de transformação
shapiro.test(yf)
```

Análise Preliminar do Modelo Completo

```
#2.3 Análise Preliminar do Modelo Completo

dadosf <- data.frame(yf, x1=treino.dados$x1, x2=treino.dados$x2, x3=treino.dados$x3,
    x4=treino.dados$x4, x5=treino.dados$x5, x6=treino.dados$x6, x7=treino.dados$x7,
    x8=treino.dados$x8, x9=treino.dados$x9, x10=treino.dados$x10, x11=treino.dados$x11)
modcf=lm(yf ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11, dadosf)
summary(modcf)
anova(modcf)
extractAIC(modcf)
extractAIC(modcf, k = log(n))
plot(modcf)
gvlma(modcf)
```

Métodos de Eliminação de Covariáveis

```r
#3 MELHOR MODELO DE REGRESSÃO LINEAR


#3.1 Métodos de Eliminação de Covariáveis

# Forward
fit <- lm(yf ~ 1, dadosf) #modelo base
modfow <- stepAIC(fit, direction ="forward", scope=list(upper=modcf, lower=fit))
summary(modfow)

# Backward
modback <- stepAIC(modcf, direction = "backward")
summary(modback)

# Stepwise
modstep <- step(modcf)
summary(modstep)

# Escolhido
modr <- modfow

# Teste de Hipóteses: F parcial
anova(modr,modcf)

# info modelo completo
Xc <- model.matrix(modcf)
Hc <- Xc%*%solve(t(Xc)%*%Xc)%*%t(Xc)
anova(modcf)
ssec <- sum((fitted(modcf)- yf)^2)
ssrc <- sum((fitted(modcf) - mean(yf))^2)

# info modelo reduzido
Xr <- model.matrix(modr)
Hr <- Xr%*%solve(t(Xr)%*%Xr)%*%t(Xr)
anova(modr)
sser <- sum((fitted(modr)- yf)^2)
ssrr <- sum((fitted(modr) - mean(yf))^2)
extractAIC(modr)
extractAIC(modr, k = log(n))
```

Covariáveis de Segunda ordem

```
#3.2 Variáveis de Segunda Ordem

dados2 <- data.frame(yf, x2=treino.dados$x2, x3=treino.dados$x3, x4=treino.dados$x4,
    x5=treino.dados$x5, x8=treino.dados$x8, x11=treino.dados$x11)
modc2 <- lm(yf ~(.)^2, dados2)

# Métodos
modfow2 <- stepAIC(modr, direction ="forward", scope=list(upper=modc2, lower=modr)) ##ESCOLHIDO
summary(modfow2)
extractAIC(modfow2)
extractAIC(modfow2, k = log(n))

modback2 <- stepAIC(modc2, direction ="backward")
summary(modback2)

modstep2 <- step(modc2)
summary(modstep2)


# Modelo Final (provisório)
modr2 <- modfow2
```

Análise de *Outliers* e Pontos Influentes

```
#3.4 Outliers e Pontos Influentes

# info modelo final
tabela <- anova(modr2)
ssr <- sum((fitted(modr2) - mean(yf))^2)
sse <- sum((fitted(modr2) - yf)^2)
sst = ssr +sse
mse= sse/tabela$Df[length(tabela$Df)]
msr = ssr/sum(tabela$Df[-length(tabela$Df)])

# Visualizar possíveis Outliers
boxplot(yf, main="Boxplot")

# Pontos Influentes
influencePlot(modr2, fill.col=carPalette()[1])
which(apply(influence.measures(modr2)$is.inf, 1, any))

# Alguns Resíduos

# habituais
X <- model.matrix(modr2)
H <- X%*%solve(t(X)%*%X)%*%t(X)
e <- yf - H%*%yf

# matriz de covariância dos resíduos
rcovmatrix <- mse * (diag(n) - H)
s2 <- diag(rcovmatrix)
```

```r
# standerdized residual
stanr <- e/sqrt(mse)
which(sapply(stanr, function(x) x^2 > 3*sqrt(mse)))

# internally studentized residual
SR <- sqrt(s2)
istudr <- e/SR
denom <- diag(diag(n)-H)
d <- e/denom
which(sapply(istudr-d, function(x) abs(x) > 1))

# hampel filter
mediana <- median(yf)
MAD <- mad(yf, center=median(yf))
upper.bound <- mediana + 3*MAD
lower.bound <- mediana - 3*MAD
upper.bound
lower.bound
which(sapply(yf, function(x) x>upper.bound || x< lower.bound))


# Retirar Outliers
dadosout <- dadosf[-c(14,29,44,45,84),]
n2 <- dim(dadosout)[1]

# Modelo Provisório (sem outliers)
modout <- lm(yf ~ x4 + x2 + x11 + x8 + x5 + x3 + x4:x11 + x11:x8 + x5:x3, data = dadosout)
anova(modout)
ssrout <- sum((fitted(modout) - mean(yf))^2)
summary(modout)
plot(modout)
extractAIC(modout)
extractAIC(modout, k = log(n2))


# Construção do Modelo Final
novo.modout <- lm(yf ~. , dadosout)

# Eliminação de Covariáveis

# primeira ordem
fitout <- lm(yf ~ 1, dadosout)
modfowout <- stepAIC(fitout, direction ="forward", scope=list(upper=novo.modout, lower=fitout))
summary(modfowout)
modrout <- modfowout

# segunda ordem
dadosout2 <- data.frame(yout=dadosout$yf, x2=dadosout$x2, x3=dadosout$x3, x4=dadosout$x4,
    x5=dadosout$x5, x8=dadosout$x8, x11=dadosout$x11)
modout2 <- lm(yout~(.)^2, dadosout2)
modfowout2 <- stepAIC(modrout, direction ="forward", scope=list(upper=modout2, lower=modrout))
    ##ESCOLHIDO
summary(modfowout2)
extractAIC(modfowout2)
extractAIC(modfowout2, k = log(n2))
final.mod <- modfowout2
```

Validação do Modelo Obtido

```
#4 TÉCNICAS DE DIAGNÓSTICO

#4.1 Validação do Modelo

gvlma(final.mod)
anova(final.mod)
summary(final.mod)

# Variance Inflation Factor - teste de multicolinearidade
v <-vif(final.mod, type="predictor")
mean(v$`GVIF^(1/(2*Df))`)

# Homocedasticidade dos Resíduos - teste Breusch-Pagan
ncvTest(final.mod)

# Autocorrelação dos Resíduos - teste Durbin-Watson
durbinWatsonTest(final.mod)
```

Capacidade de Previsão

```
#4.2 Capacidade de Previsão

# RMSPE
expected <-predict(final.mod, newdata= teste.dados)
expected[7] <- 0
expected2 = expected^(1/lambda)
yteste <- teste.dados$y
SSPE <- (expected2- yteste)^2
MSPE <- sum(SSPE)/length(yteste)
RMSPE <-sqrt(MSPE)

# Range de Observações Originais
range(dados$y)

# Gráfico de Previsão
teste <- data.frame(real=yteste, expected=expected2)
ggplot(teste, aes(expected, real, color=real),scale) +
  geom_point(color="slategray4" ,shape = 16, size = 4,show.legend = FALSE) +
  coord_cartesian(x=c(0,8),y=c(0,8))+
  theme_minimal() + labs(title = "Prediction Best Model")+
  geom_abline(slope=1, intercept=0, color="grey10")
```

Outputs do R

Listing 1: Output de *gvlma*

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11, data = treino.dados)

Coefficients:
(Intercept)           x1           x2           x3
x4           x5           x6
 -2.3324130   -0.0008616    0.1815398    0.0259427    0.0724330
0.0100715   -0.0012068
        x72          x82          x83          x84
x9          x10          x11
  0.3311792    0.3761192    0.4701634    1.1073406    0.0022867
-0.0003188    0.0229252


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = modc)

                     Value  p-value                   Decision
Global Stat        10.30136 0.035646 Assumptions NOT satisfied!
Skewness            0.01145 0.914778    Assumptions acceptable.
Kurtosis            0.60678 0.436002    Assumptions acceptable.
Link Function       9.31200 0.002277 Assumptions NOT satisfied!
Heteroscedasticity  0.37113 0.542388    Assumptions acceptable.
```

Listing 2: Primeiro teste de Shapiro

```
        Shapiro-Wilk normality test

data:  treino.dados$y
W = 0.97068, p-value = 0.03769
```

Listing 3: Segundo teste de Shapiro

```
        Shapiro-Wilk normality test

data:  yf
W = 0.97523, p-value = 0.07989
```

Listing 4: Teste F parcial

```
Analysis of Variance Table

Model 1: yf ~ x4 + x2 + x11 + x8 + x5 + x3
Model 2: yf ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     82 182.78
2     77 180.92  5    1.8637 0.1586 0.9768
```

Listing 5: Output de *gvlma* para Modelo Final

```
Call:
lm(formula = yf ~ x4 + x2 + x11 + x8 + x5 + x3 + x11:x8 + x4:x5 +
    x8:x3, data = dadosout)

Coefficients:
(Intercept)           x4           x2          x11
x82          x83          x84           x5
 -11.130033     0.256810     0.254469    -0.018621     7.457716
5.405815     2.706688     0.038660
         x3      x11:x82      x11:x83      x11:x84        x4:x5
x82:x3       x83:x3       x84:x3
   0.195833     0.058043     0.106828     0.035047    -0.001486
-0.184492    -0.178787    -0.051823


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = final.mod)

                     Value p-value                Decision
Global Stat        1.0123257  0.9079 Assumptions acceptable.
Skewness           0.0006819  0.9792 Assumptions acceptable.
Kurtosis           0.3094010  0.5780 Assumptions acceptable.
Link Function      0.5337222  0.4650 Assumptions acceptable.
Heteroscedasticity 0.1685206  0.6814 Assumptions acceptable.
```