FCTUC FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Grupo 13 – Henrique Tavares (2017242743)
Diogo Oliveira (2017244910)

Trabalho final

Análise de sintomas de dor lombar

Resumo

Com este trabalho pretendemos estabelecer e analisar um algoritmo de classificação dado em aula ao longo do semestre com base num problema de classificação binário, em função do qual foi selecionado um dataset (conjunto de dados) ao qual serão aplicados diversos testes.

Foi escolhido um conjunto de dados onde podemos encontrar diversos atributos que podem ser a causa de dores lombares em diferentes pacientes (encontram-se 310 casos no dataset). Procurar-se-á, portanto, tirar conclusões relativamente aos sintomas de dores lombares das pessoas no que toca à anormalidade da sua condição aplicando métodos de seleção de variáveis.

Introdução

Em algum momento das suas vidas, quase todas as pessoas sofrem de dor lombar, dor ligeira ou intensa de duração variável que vem interferir de modo significativo nas atividades diárias de quem a carrega.

Esta dor pode ser causada por uma variedade de problemas relativamente à complexa e interconectada rede de músculos espinhais, nervos, ossos, discos ou tendões da espinha lombar. Causas típicas de dores lombares passam por danos em ossos, ligamentos ou articulações, irritação em nervos grandes que façam a conexão entre a zona lombar e as pernas ou em nervos mais pequenos presentes também na zona lombar, alguma deformação nos músculos das costas ou, ainda, degeneração de um disco intervertebral.

Neste conjunto de dados obtido no site recomendado, Kaggle, existem diversos atributos para analisar como a inclinação sacral, cervical, pélvica e torácica, espondilolistese e até escoliose. Conforme sejam aplicados diversos testes (cuja descrição mais detalhada estará presente nos Métodos), tirar-se-ão conclusões relativamente á importância destas variáveis como causa de dores lombares tendo em conta os resultados obtidos.

Métodos

O dataset que se escolheu apresenta treze colunas, sendo uma delas o classificador, que define se um dos casos apresenta uma condição anormal ou normal. Temos como atributos: incidência pélvica *(pelvic_incidence)*; inclinação irregular da pélvis *(pelvic tilt)*; ângulo de lordose lombar *(lumbar_lordosis_angle)*; inclinação sacral *(sacral_slope)*; raio pélvico *(pelvic_radius)*; grau de espondilolistese *(degree_spondylolisthesis)*; inclinação pélvica *(pelvic_slope)*; inclinação direta das costas *(Direct_tilt)*; inclinação torácica *(thoracic_slope)*; inclinação irregular cervical *(cervical_tilt)*; ângulo sacro *(sacrum_angle)* e inclinação associada à escoliose *(scoliosis_slope)*.

Primeiramente, observaram-se os dados através de boxplots de modo a entender a possibilidade de descartar alguma das variáveis com base apenas na observação (dados da variável *versus* classificação).

```
boxplot(pelvic_incidence ~ results,xlab="Classification", ylab="Pelvic Incidence",main='Pelvic I
ncidence Vs. Classification', col=c('limegreen', 'red2'))
```

Figura 1 – Código aplicado na formação de boxplots para cada variável, neste caso, a incidência pélvica

De seguida, aplicou-se o teste de Shapiro-Wilk a cada uma das variáveis, com o objetivo de verificar se os dados apresentam uma distribuição normal, dado o facto de se estar a trabalhar com variáveis quantitativas. A hipótese nula deste teste diz que os dados estão distribuídos normalmente. Tomando 0.05 como nível de significância no que toca ao *p-value* temos que: caso uma variável apresente um *p-value* menor que 0.05, será possível rejeitar a

hipótese nula. Caso apresente um *p-value* superior a 0.05, não será possível rejeitar a hipótese nula. Para variáveis cuja hipótese nula tenha sido rejeitada, não poderemos aplicar testes paramétricos, mas sim testes não-paramétricos.

```
#pelvic incidence
npelvic_incidence <- dataset[results == 'Normal',]$pelvic_incidence #normal pelvic incidences values
apelvic_incidence <- dataset[results == 'Abnormal',]$pelvic_incidence #abnormal pelvic incidence values
shapiro.test(npelvic_incidence)
shapiro.test(apelvic_incidence)
```

Figura 2 – Código aplicado para a realização do teste de Shapiro-Wilk, neste caso à incidência pélvica (para caso normal e para caso anormal) *(pelvic_incidence)*

Avaliada a necessidade de aplicar testes paramétricos ou não-paramétricos, seguiu-se a aplicação do teste de Mann-Whitney (Wilcoxon Rank Sim Test). Um teste não-paramétrico aplicado a duas amostras independentes, o teste de Wilcoxon apresenta como hipótese nula o facto de não existirem diferenças entre os valores das medianas das variáveis. Tomando 0.05 como nível de significância, rejeitar-se-á a hipótese nula para as variáveis cujo *p-value* seja menor que 0.05. As variáveis com p-value superior a 0.05, ou seja, sem grandes diferenças no que toca ao valor das suas medianas, serão descartadas.

```
options(warn=-1)
wilcox.test(pelvic_incidence ~ results, paired = FALSE)
wilcox.test(pelvic_tilt ~ results, paired = FALSE)
```

Figura 3 – Código aplicado para a realização do teste de Wilcoxon, neste caso à incidência pélvica *(pelvic_incidence)* e à inclinação irregular da pélvis *(pelvic_tilt)*

Seguiu-se a aplicação de um modelo linear generalizado (MLG) às variáveis que passaram os testes aplicados anteriormente. De modo a formar um classificador para o MLG, dividiram-se os dados em dois grupos: o grupo de treino (para se ajustar o modelo estatístico) e um grupo de teste (para se testar e validar o modelo) (divididos numa proporção 70%/30%, respetivamente). Com base nos resultados obtidos, é possível avaliar a significância das variáveis colocadas à prova.

```
indice <- sample(2,nrow(dataset),replace = TRUE,
prob=c(0.7,0.3))
train <- dataset[indice==1,]
test1 <- dataset[indice==2,]
```

Figura 4 – Formação dos grupos de treino e de teste

```
GLM <- glm(formula = results ~ pelvic_incidence + p
elvic_tilt + lumbar_lordosis_angle + sacral_slope +
pelvic_radius + degree_spondylolisthesis, data = tr
ain, family = binomial)
summary(GLM)~
GLM <- glm(formula = results ~ pelvic_radius + degr
ee_spondylolisthesis, data = train, family = binomi
al)
summary(GLM)
confint(GLM)
```

Figura 5 – Código aplicado para a aplicação do MLG ao grupo de treino

```
coluna_1 <- coef(summary(GLM))[1,1]
coluna_2 <- coef(summary(GLM))[2,1]
coluna_3 <- coef(summary(GLM))[3,1]
plot(coluna_2*train$pelvic_radius+coluna_3*train
$degree_spondylolisthesis,GLM$fitted.values,
 col=c('#66CCFF','red'),
 xlab = 'Combined Model',
 ylab = 'Corrected Values',
 main = 'Logistic Model',
 pch=19)
legend('bottomright',c('Normal','Abnormal'),cex=
1,col = c('#66CCFF','red'),merge=FALSE,pch = 19)
```

Figura 6 – Formação do plot onde será possível observar o modelo logístico obtido a partir do MLG

Após a aplicação do modelo linear generalizado, traçou-se o gráfico correspondente à regressão logística, obtendo, assim, a curva que melhor representava os dados.

Realizaram-se alguns testes de modo a verificar a precisão do nosso classificador. Aplicou-se o teste de Wald, teste paramétrico estatístico cuja função é comparar o parâmetro estimado com o parâmetro modelado nulo, avaliando assim a significância estatística das variáveis, e o teste de Hosmer-Lemeshow, teste estatístico que avalia a qualidade do ajuste no modelo de regressão logística tendo em conta a distância entre as probabilidades ajustadas e as probabilidades observadas.

```
library(aod)

wald_test <- wald.test(b=coef(GLM), Sigma = vcov(GLM), Terms=1:2)

print(wald_test)

library(ResourceSelection)

hosmer_lemeshow <- hoslem.test(GLM$y, fitted(GLM),g=10)

print(hosmer_lemeshow)
```

Figura 7 – Código aplicado para a realização dos testes de Wald e Hosmer-Lemeshow

Como hipótese nula para o teste de Wald, temos que ambos os parâmetros são iguais, ou seja, o parâmetro estimado não apresenta significado estatístico. Para um valor *p-value* inferior a 0.05, pode ser rejeitada a hipótese nula. Por outro lado, para o teste de Hoslem-Lemeshow, a hipótese nula diz que o modelo se ajusta corretamente. Pretende-se obter um *p-value* superior a 0.05 para este último teste, de modo a não ser possível rejeitar a hipótese nula.

Aplicaram-se os testes de Cox-Snell e Nagelkerke R^2 para, em conjunto com a matriz de confusão, indicar a qualidade do ajuste do nosso modelo, procedendo de seguida à determinação da precisão, sensibilidade e especificidade do modelo com base em 500 ciclos.

```
library(knitr)

prob = predict(GLM,type = c('response'),train)

confusion <- table(prob>0.5,train$results)

kable(confusion)
```

Figura 8 – Código aplicado para a determinação da matriz de confusão

```
GLM_null <- glm(results ~ 1, data=dataset, family=binomial)

LL_null <- logLik(GLM_null)

LL_k <- logLik(GLM)

R_Cox <- 1 - (exp(LL_null[1])/exp(LL_k[1]))^(2/length(dataset$results))

R_Nag <- R_Cox/(1-(exp(LL_null[1]))^(2/length(dataset$results)))

R2 <- R_Cox/R_Nag

print(sprintf('R2 Cox = %s',R_Cox))

print(sprintf('R2 Naguelkerke = %s',R_Nag))

print(sprintf('R2 = %s',R2))
```

Figura 9 – Código aplicado para a realização dos testes de Cox-Snell e Nagelkerke R^2

Fez-se ainda o SVM, Support Vector Machine, para ver se existia um modelo de classificação que proporcionasse um melhor ajuste dos dados. O SVM tem como vantagem o facto de permitir criar um hiperplano que maximiza a separação entre as variáveis.

Determinou-se a precisão, a sensibilidade e a especificidade para os resultados obtidos a partir da aplicação deste método.

```
library(e1071)

modeloSVM<-svm(results ~ pelvic_radius + degree_spondylolisthesis,data=train,
               cost=100,
               gamma=0.75)

plot(train$pelvic_radius +train$degree_spondylolisthesis,modeloSVM$fitted.values,
     col=c('#66CCFF','red'),
     pch=19,
     xlab = 'Combination of variables',
     ylab = 'Adjustment of values',
     main = 'SVM model')
```

Figura 10 – Aplicação do método da máquina de vetores de suporte (SVM, Support Vector Machine)

Discussão de resultados

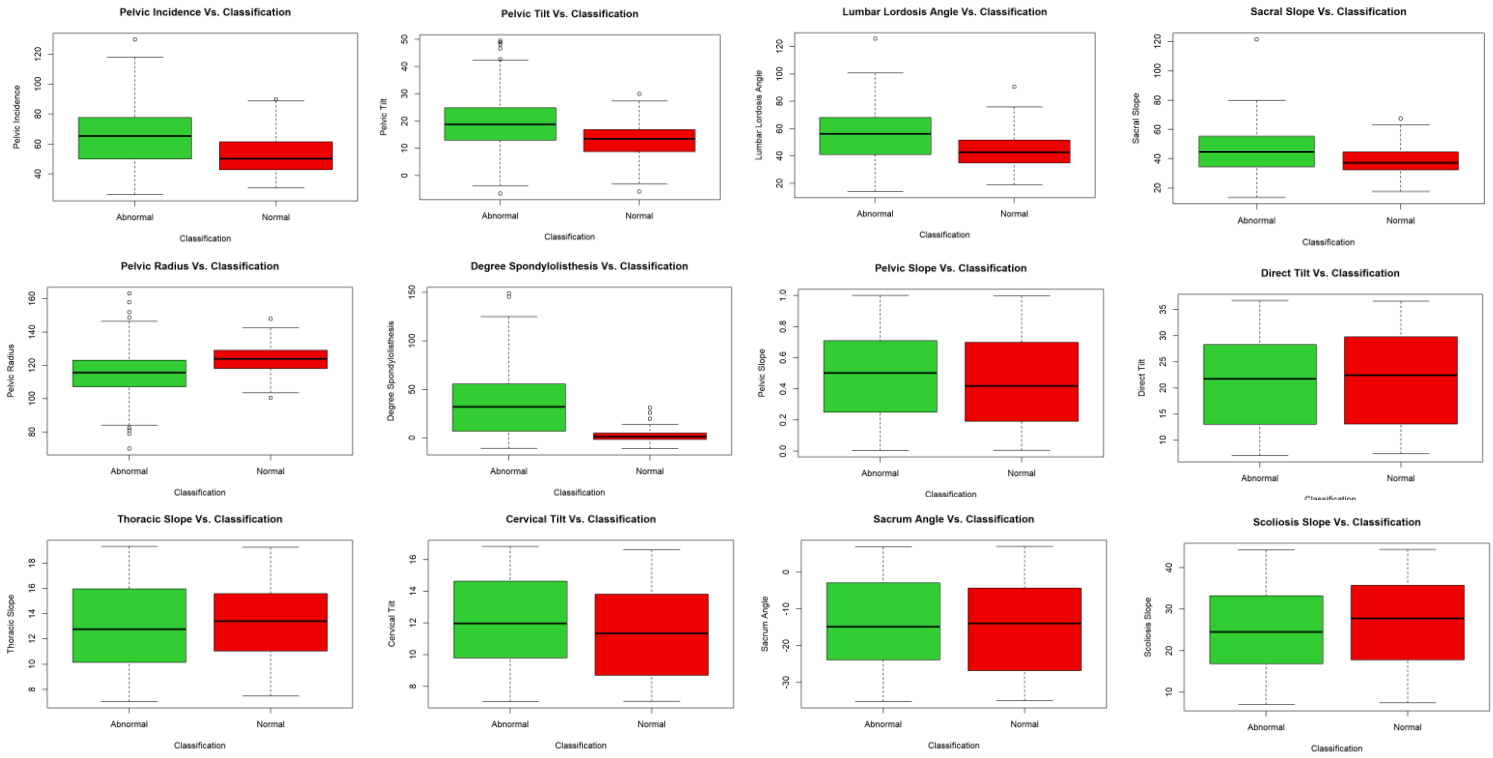

Figura 11 – Boxplots obtidos para cada variável em comparação com a sua classificação

Assim como descrito na seção onde foram descritos os métodos usados para a análise e seleção dos dados, o primeiro método adotado foi a observação dos boxplots obtidos a partir da comparação entre cada uma das variáveis e a sua classificação. Analisado o facto de não existir uma quantidade elevada de outliers e a quantidade de dados, optou-se por não se excluir qualquer variável.

Estudada a existência de outliers, aplicou-se o teste de Shapiro-Wilk, para verificar se os dados apresentam uma distribuição normal. Sendo o nível de significância 0.05, é possível observar pelos resultados apresentados na Tabela 1 que se rejeita a hipótese nula para todas as variáveis uma vez que o p-value é inferior ao nível de significância em pelo menos um dos casos para cada variável. Concluiu-se, portanto, que os dados não apresentam uma distribuição normal, pelo que os testes a aplicar de seguida deverão ser não-paramétricos.

Variáveis	*p-value* para casos normais	*p-value* para casos anormais
Pelvic incidence	2.608×10^{-3}	3.461×10^{-3}
Pelvic tilt	0.792	5.483×10^{-4}
Lumbar lordosis angle	0.017	0.039
Sacral slope	0.255	2.947×10^{-5}
Pelvic radius	0.936	0.028
Degree of spondylolisthesis	2.652×10^{-7}	1.104×10^{-7}
Pelvic slope	1.449×10^{-3}	3.005×10^{-6}
Direct tilt	1.595×10^{-4}	1.801×10^{-6}
Thoracic slope	0.021	7.080×10^{-6}
Cervical tilt	1.929×10^{-4}	5.124×10^{-6}
Sacrum angle	4.974×10^{-4}	2.344×10^{-6}
Scoliosis slope	7.621×10^{-4}	2.897×10^{-5}

Tabela 1 – Resultados do teste de Shapiro-Wilk

Seguiu-se a aplicação do teste de Wilcoxon para testar a existência de diferenças significativas entre os valores das medianas. É possível observar pelos resultados apresentados na Tabela 2 que as variáveis *pelvic slope*, *direct tilt*, *thoracic slope*, *cervical tilt*, *sacrum angle* e *scoliosis slope* apresentam p-value superior a 0.05, ou seja, não se rejeita a hipótese nula. Para estas variáveis, não se verificam diferenças significativas entre as medianas de cada classe, pelo que estas variáveis serão excluídas nos próximos passos.

Variáveis	*p-value*
Pelvic incidence	2.252×10^{-10}
Pelvic tilt	1.467×10^{-8}
Lumbar lordosis angle	3.181×10^{-8}
Sacral slope	1.144×10^{-4}
Pelvic radius	2.930×10^{-10}
Degree of spondylolisthesis	2.200×10^{-16}
Pelvic slope	0.331
Direct tilt	0.476
Thoracic slope	0.378
Cervical tilt	0.086
Sacrum angle	0.630
Scoliosis slope	0.236

Tabela 2 – Resultados do teste de Wilcoxon

Para as variáveis que apresentam p-value inferior a 0.05 rejeita-se a hipótese nula, o que contribui significativamente para a discriminação da classificação, uma vez que apresenta diferenças consideráveis entre as medianas das duas classes.

Construiu-se depois um modelo linear generalizado com as variáveis que tinham passado os restantes testes até ao momento, a partir do qual se concluiu que as variáveis *pelvic radius* e *degree of spondylolisthesis* seriam as variáveis mais significativas (resultados presentes na Figura 12), após o qual se fez um novo MLG só com estas duas variáveis que se provaram ser as mais significativas.

No que toca aos testes aplicados para testar a precisão do modelo, obteve-se um *p-value* de 8.800×10^{-5} para o teste de Wald e um *p-value* de 0.714 para o teste de Hosmer and Lemeshow. Com estes resultados concluiu-se que os dados têm significado estatístico (Wald) e que o modelo se ajusta corretamente (Hosmer and Lemeshow).

Obteve-se um valor de 71,5% para os testes de Cox-Snell e Nagelkerke R^2, ou seja, para a qualidade do ajuste (o que é positivo). Obteve-se ainda, para o grupo de treino, 79,4% de exatidão, 57,5% de especificidade e 89,7% de sensibilidade. Por outro lado, para o grupo de teste, obteve-se 78,6%, 57,3% e 89% respetivamente.

Por sua vez, o modelo SVM para o mesmo conjunto de variáveis obteve:

	Anormal	Normal
Anormal	122	18
Normal	26	51

Tabela 4 – Matriz de confusão para o SVM

Obteve-se com a aplicação do SVM, para o grupo de treino, 79,4% de exatidão, 57,7% de especificidade e 89,7% de sensibilidade. Por outro lado, para o grupo de teste, obteve-se 78,5%, 57,1% e 89,9% respetivamente.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.237e+01  3.248e+00  -3.810 0.000139 ***
pelvic_incidence         5.170e+05  4.468e+07   0.012 0.990767
pelvic_tilt             -5.170e+05  4.468e+07  -0.012 0.990767
lumbar_lordosis_angle    2.920e-02  2.287e-02   1.277 0.201711
sacral_slope            -5.170e+05  4.468e+07  -0.012 0.990767
pelvic_radius            8.257e-02  2.278e-02   3.625 0.000289 ***
degree_spondylolisthesis -1.602e-01  2.529e-02  -6.335 2.37e-10 ***
```

Figura 12 – Coeficientes calculados a partir da aplicação do MLG

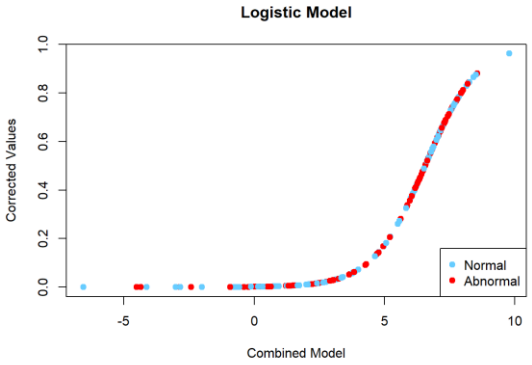

Figura 13 – Modelo logístico obtido para o MLG

	Anormal	Normal
Anormal	127	14
Normal	21	55

Tabela 3 – Matriz de confusão para o MLG

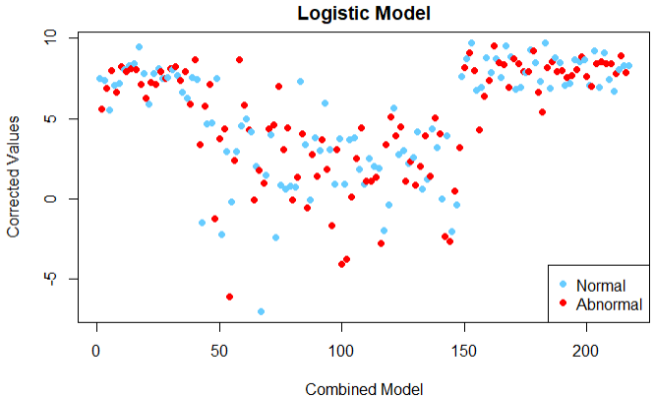

Figura 14 – Modelo logístico obtido para o SVM

Conclusão

Atendendo aos resultados obtidos por ambos os modelos estatísticos, podemos confirmar que o problema em questão apresenta viabilidade no que toca à previsão estatística do mesmo através de processos computacionais, apesar dos resultados obtidos não serem perfeitos constata-se que ainda há margem de manobra, possivelmente através de um maior número de medições e consequentemente mais variáveis e ainda desenvolvimento de técnicas mais exatas de medir determinados ângulos.

De um ponto de vista teórico, é possível concluir através do tratamento de dados que as variáveis com maior poder discriminativo para a classificação foram o grau de espondilolistese (uma condição que consiste no deslocamento de uma vértebra em relação a outra vértebra na sua proximidade) e o raio pélvico, uma vez que são as variáveis que melhor discriminam se o indivíduo apresenta condição normal ou anormal.

Referências

kaggle.com/sammy123/lower-back-pain-symptoms-dataset (28/12/2019)
rdocumentation.org/
ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Low-Back-Pain-Fact-Sheet