PSALevel - Regressão linear

Wesley Nunes Marques Torres

April 01, 2016

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Loading required package: reshape2

## Loading required package: ggplot2

## Loading required package: caret

## Loading required package: lattice
```

Objetivo da análise

Prever o nível do PSA em pacientes através de uma regressão linear.

Um pouco sobre os dados

Temos para análise um dataset com 97 observações com 8 variáveis preditoras, 1 variável para especificar se é treino ou não e 1 variável resposta. Abaixo podemos verificar um pouco sobre a descrição de cada variável e seu tipo:

```
dim(dados)
```

```
## [1] 97 10
```

```
str(dados)
```

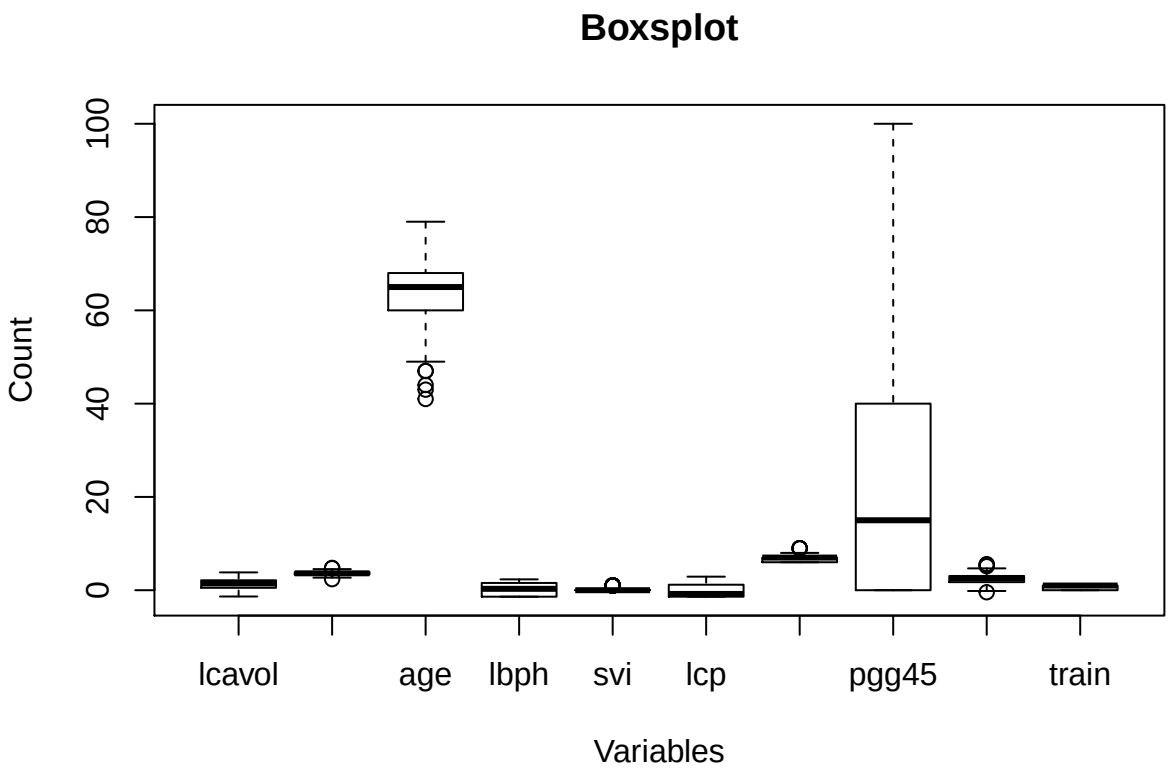
```
## 'data.frame':    97 obs. of  10 variables:
##  $ lcavol : num  -0.58 -0.994 -0.511 -1.204 0.751 ...
##  $ lweight: num  2.77 3.32 2.69 3.28 3.43 ...
##  $ age    : int  50 58 74 58 62 50 64 58 47 63 ...
##  $ lbph   : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ svi    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lcp    : num  -1.39 -1.39 -1.39 -1.39 -1.39 ...
##  $ gleason: int  6 6 7 6 6 6 6 6 6 6 ...
##  $ pgg45  : int  0 0 20 0 0 0 0 0 0 0 ...
##  $ lpsa   : num  -0.431 -0.163 -0.163 -0.163 0.372 ...
##  $ train  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
```

Ao sumarizar os dados, podemos verificar que há uma boa consistência nos dados e que não se tem alguma discrepância, mas com um boxsplot podemos identificar outliers, mas que neste caso, não irá nos interessar;

```
summary(dados)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.629   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.876   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp              gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa            train        
##  Min.   :-0.4308   Mode :logical  
##  1st Qu.: 1.7317   FALSE:30       
##  Median : 2.5915   TRUE :67       
##  Mean   : 2.4784   NA's :0        
##  3rd Qu.: 3.0564                  
##  Max.   : 5.5829
```

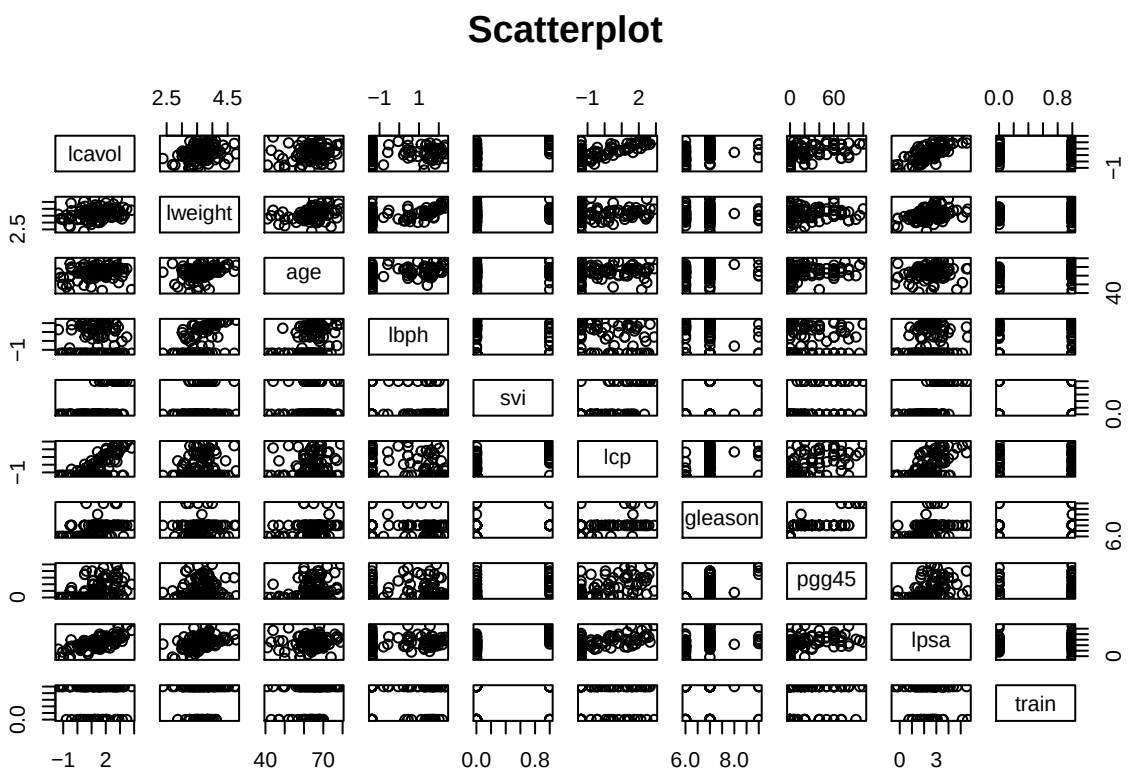
```
boxplot(dados, main="Boxsplot", xlab="Variables", ylab="Count")
```



Scatter plot dos dados

Podemos verificar o relacionamento entre as variáveis com a variável resposta através de um scatter plot. Abaixo, o que chama atenção é poder avaliar a dispersão entre as variáveis e mais precisamente a da variável resposta(lpsa)

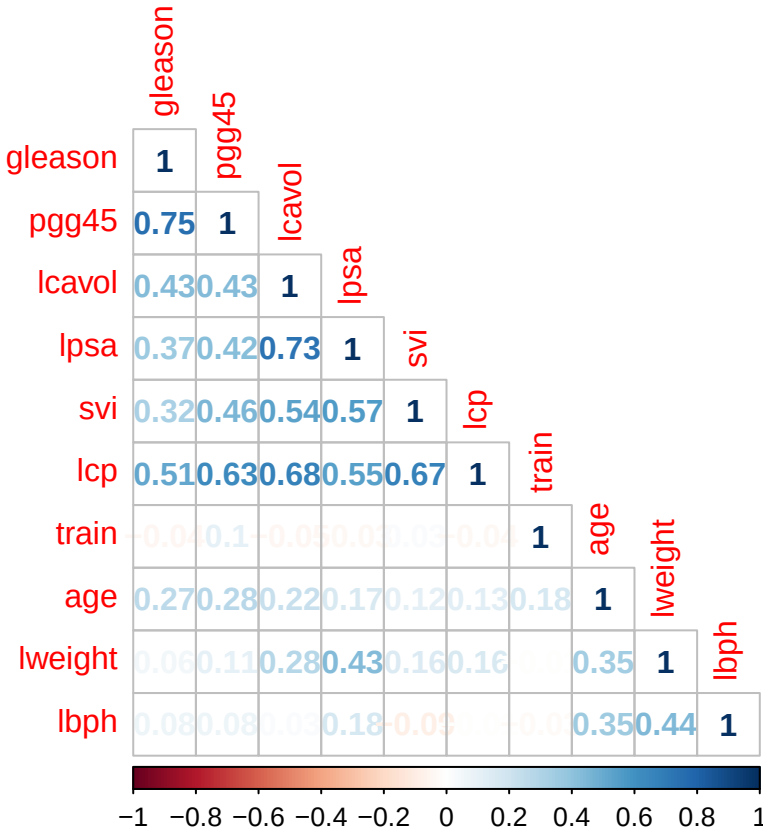
```
plot(dados, main="Scatterplot", pch=1)
```



Análise de correlação

Como se pode analisar no gráfico abaixo, a correlação com sí próprio é redundante e logicamente, tem um valor alto. A correlação será de suma importância para escola dos preditores da variável resposta. Com o plot abaixo podemos perceber que temos ótimos candidatos para preditores.

```
correlationMatrix <- cor(dados)
corrplot(correlationMatrix, method="number", type="lower", order="hclust")
```



Separando dados para treino

Como foi dado, existe um campo para nos dizer se uma amostra é do tipo treino ou teste. Abaixo, vamos separar esses dados em diferentes dataframes para realizar o treino do modelo de forma correta.

```
train <- filter(dados,train)
test <- filter(dados,!train)
```

Treinando meu modelo

Para realizar o treino, precisamos verificar quais preditores são melhores para se ter um modelo com a melhor acurácia possível. Para isso, a princípio, uma das melhores formas de se realizar o treino é escolhendo preditores com um alto valor de correlação com a variável resposta. Acima, temos os valores das correlações entre as variáveis, e com base nisso, foram escolhidas as variáveis: lcavol, svi e lcp.

```
lm <- lm(lpsa ~ lcavol+svi+lcp,data = train)
```

Temos agora um modelo e basta verificar a acurácia desse modelo

```
lm
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + svi + lcp, data = train)
```

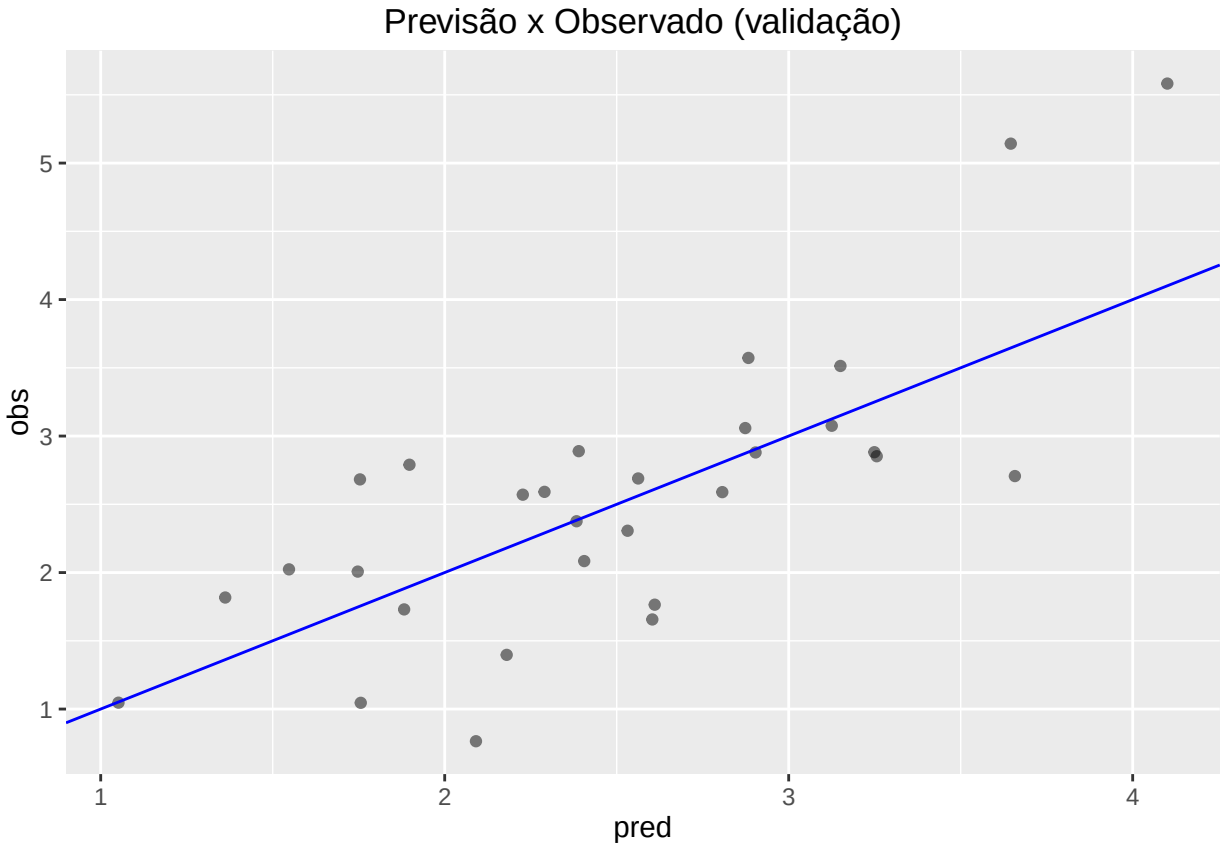
```
##
## Coefficients:
## (Intercept)       lcavol          svi          lcp
##      1.3722       0.6754       0.7290      -0.1395
```

```
summary(lm)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + svi + lcp, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7015 -0.5090  0.1243  0.5160  1.6985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3722     0.1952   7.030 1.77e-09 ***
## lcavol        0.6754     0.1144   5.903 1.55e-07 ***
## svi           0.7290     0.3296   2.212   0.0306 *
## lcp          -0.1395     0.1103  -1.265   0.2104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8093 on 63 degrees of freedom
## Multiple R-squared:  0.5714, Adjusted R-squared:  0.551
## F-statistic:    28 on 3 and 63 DF,  p-value: 1.256e-11
```

Realizando a previsão

```
prediction <- predict(lm, select(test,lcavol, svi, lcp))
lm_prediction <- data.frame(pred = prediction, obs = test$lpsa)
ggplot(lm_prediction, aes(x = pred, y = obs)) + geom_point(alpha = 0.5, position = position_jitter(widt
```



Com os dados expostos, podemos responder as seguintes perguntas:

Há evidência de relação entre os preditores e a variável alvo?

Podemos identificar evidências de relação entre preditores e variável alvo analisando a correlação entre elas. Com isso, temos que as variáveis lcavol(Volume do câncer), svi(invasão das vesículas seminais) e cp(penetração capsular) são fortes candidatos para preditores que me darão uma boa acurácia no meu modelo.

Havendo relação, quão forte é essa relação?

Como já foi mostrado no gráfico acima, mas temos os seguintes valores para representar um relação entre os preditores escolhidos e a variável resposta. Com valores de 0-1, indicando nível de correlação, de mais fraco para mais forte, temos:

- lcavol X lpsa = 0.73
- svi X lpsa = 0.57
- lcp X lpsa = 0.55

Que variável parece contribuir mais?

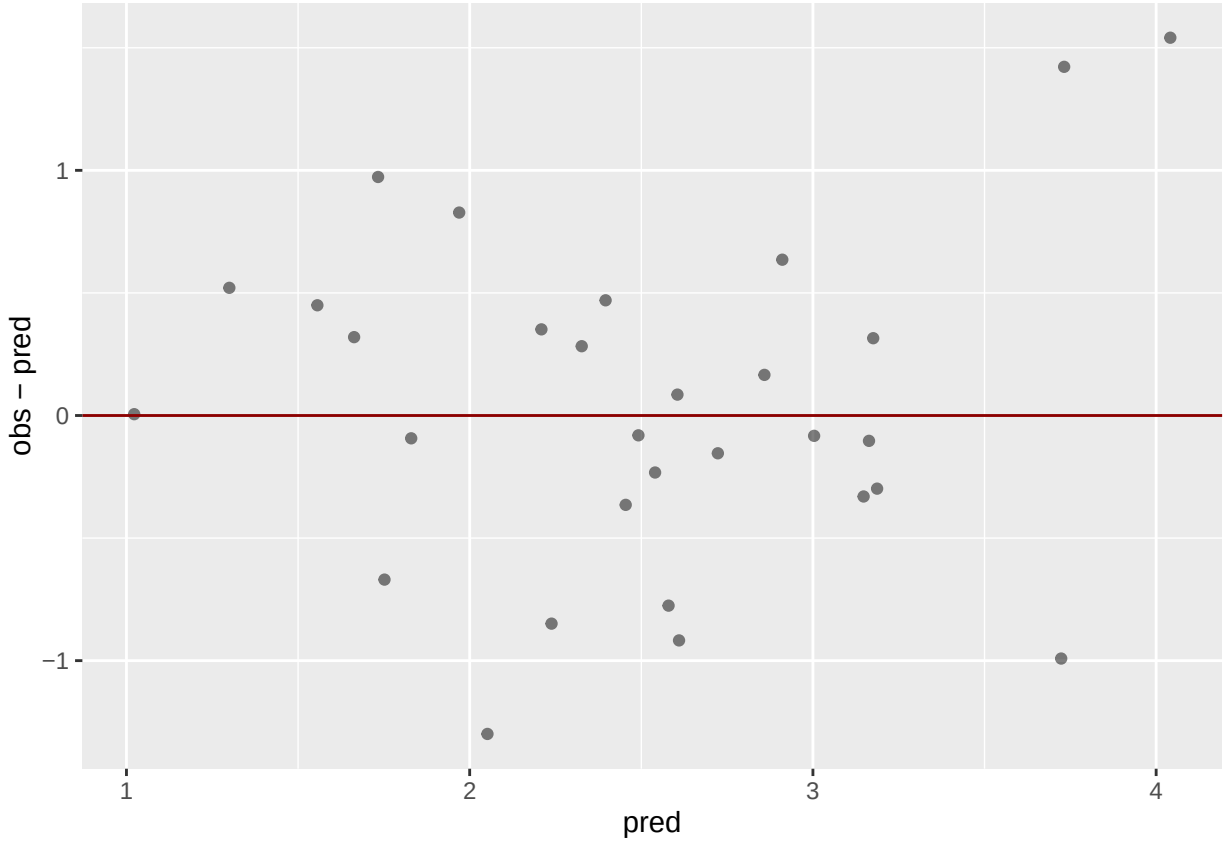
Tendo como base o nível de correlação, a variável **lcavol** aparentemente possui uma maior contribuição para o modelo.

A relação sugere um modelo de regressão linear?

```r
summary(lm)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + svi + lcp, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7015 -0.5090  0.1243  0.5160  1.6985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3722     0.1952   7.030 1.77e-09 ***
## lcavol        0.6754     0.1144   5.903 1.55e-07 ***
## svi           0.7290     0.3296   2.212   0.0306 *
## lcp          -0.1395     0.1103  -1.265   0.2104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8093 on 63 degrees of freedom
## Multiple R-squared:  0.5714, Adjusted R-squared:  0.551
## F-statistic:    28 on 3 and 63 DF,  p-value: 1.256e-11
```

```r
ggplot(lm_prediction, aes(y = obs - pred, x = pred)) +
  geom_point(alpha = 0.5, position = position_jitter(width=0.1)) +
  geom_abline(slope = 0, intercept = 0, colour = "darkred")
```



Pelo gráfico de resíduo e uma sumarização dos dados, vemos que temos um grande indício para se utilizar a regressão linear. Um dos fatores é o da Mediana dos resíduos se encontrar próximo de zero e módulo do 1Q e do 3Q serem aproximados.

Por fim, temos os valores de RMSE = 0.665, que é muito bom, pois nos diz a o erro dos valores esperados para os observados e como foi baixo, já nos diz que o uso de Regressão é uma boa opção para esse caso. Outra informação importante é o Rsquared = 58,5%, que nos diz o quando minhas variáveis explicam meu modelo. Com isso, temos os valores finais abaixo para a predição do nível do PSA em pacientes:

```
round(defaultSummary(lm_prediction), digits = 3)
```

```
##     RMSE Rsquared
##    0.665    0.584
```