

## Estatística Aplicada II

### Segunda Lista de Exercícios

#### SAMUEL KOJICOVSKI

Utilizar a base de dados “prostate” que encontra-se na aba arquivos do TEAMS, explicações do que se trata esta base de dados pode ser encontrado em: <https://rafalab.github.io/pages/649/prostate.html>

Com a base de dados “prostate” obter os seguintes resultados com o auxílio do “R”

a) Elaborar a regressão linear preliminar com as seguintes variáveis:

Variável dependente: lcavol;

Variáveis explicativas: lweight, age, lbph, svi, lcp, gleason, pgg45, lpsa

```
Call:
lm(formula = lcavol ~ lweight + age + lbph + svi + lcp + gleason +
    pgg45 + lpsa, data = prostate)

Residuals:
    Min       1Q   Median       3Q      Max
-1.88603 -0.47346 -0.03987  0.55719  1.86870

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.260101   1.259683  -1.794   0.0762 .
lweight      -0.073166   0.174450  -0.419   0.6759
age           0.022736   0.010964   2.074   0.0410 *
lbph          -0.087449   0.058084  -1.506   0.1358
svi           -0.153591   0.253932  -0.605   0.5468
lcp           0.367300   0.081689   4.496 0.00002098941 ***
gleason       0.190759   0.154283   1.236   0.2196
pgg45        -0.007158   0.004326  -1.654   0.1016
lpsa          0.572797   0.085790   6.677 0.00000000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6998 on 88 degrees of freedom
Multiple R-squared:  0.6769,    Adjusted R-squared:  0.6475
F-statistic: 23.04 on 8 and 88 DF,  p-value: < 2.2e-16
```

b) Testar outliers pelas técnicas de Grubbs, qui-quadrado, Bonferroni e da Distância de Cook e eliminar observações se necessário.

#### Teste de Grubbs

Como p-value é maior que 0.05, não há necessidade de remoção de outlier.

```
Grubbs test for one outlier
data: prostate$residuos
G.69 = 2.81502, U = 0.91659, p-value = 0.3996
alternative hypothesis: lowest value -1.88603397103562 is an outlier
```

#### Teste qui-quadrado

Como p-value é menor que 0.05, existe outlier e o seu valor é -1.88603397103562

```
chi-squared test for outlier
data: prostate$residuos
X-squared.69 = 7.9244, p-value = 0.004877
alternative hypothesis: lowest value -1.88603397103562 is an outlier
```

Nova rodada com qui-quadrado, com p-value agora menor que 0,05.

```
chi-squared test for outlier
data: prostate$residuos
X-squared.69 = 7.9244, p-value = 0.004877
alternative hypothesis: lowest value -1.88603397103562 is an outlier
```

#### Teste Bonferroni

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
69 -2.976271      0.0037776      0.36643
```

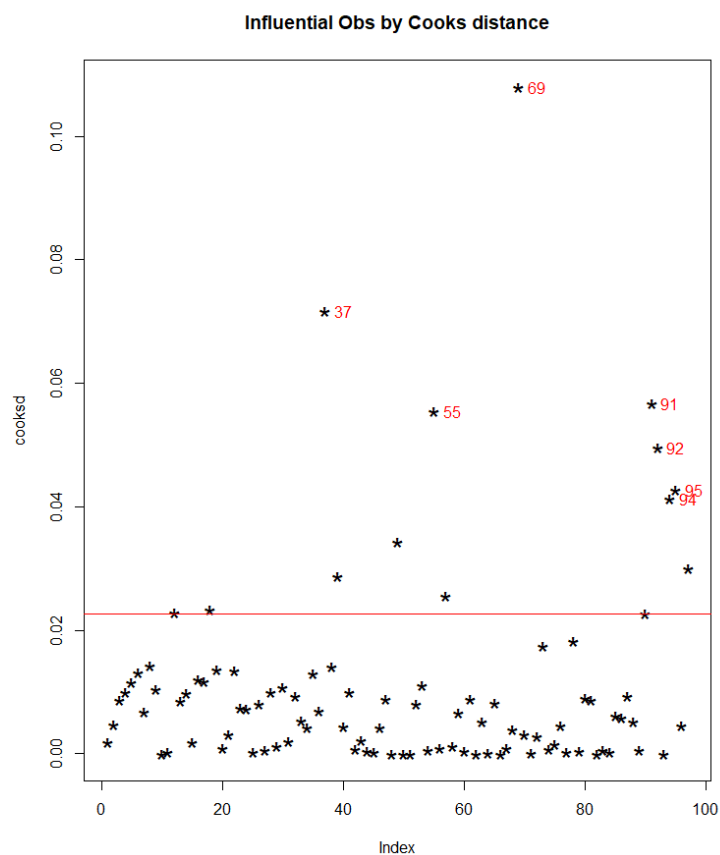
Nova rodada com teste Bonferroni, agora com p-value maior que 0,05, portanto, não necessário remoção de outlier.

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferroni p
55 2.875127      0.0050893      0.48858
```

## Distância de cook

Outliers encontrados no teste

```
# A tibble: 7 x 13
  lccavol lweight age lbph svi lcp gleason pgg45 lpsa ObsNumber residuos
  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>   <dbl>
1  1.42    3.66  73 -0.580  0  1.66    8  15  2.16    37 -1.02
2  3.15    3.52  59 -1.39  0 -1.39    7   5  2.70    55  1.87
3 -0.446   4.41  69 -1.39  0 -1.39    6   0  2.96    69 -1.89
4  3.25    4.10  68 -1.39  0 -1.39    6   0  4.03    91  1.20
5  2.53    3.68  61  1.35  1 -1.39    7  15  4.13    92  0.863
6  3.82    3.90  44 -1.39  1  2.17    7  40  4.68    94  0.869
7  2.91    3.40  52 -1.39  1  2.46    7  10  5.14    95 -0.849
# ... with 2 more variables: cooks_d <dbl>, outlier <chr>
```



c) Testar normalidade por Kolmogorov-Smirnov e Shapiro-Wilk.

### Kolmogorov-Smirnov

Para  $n=96 \rightarrow D = 1/\sqrt{96} = 0,102$ , como  $D$  calculado é menor que tabelado, aceita-se  $H_0$ .  
Portanto, como  $p$ -value é maior que 0.05, a distribuição é normal, aceita-se a normalidade.

```
Lilliefors (Kolmogorov-Smirnov) normality test
data:  residuos
D = 0.066595, p-value = 0.3689
```

## Shapiro-Wilk

Valor de “W” calculado menor que “W” tabelado, portanto, aceita-se H0. Com um p-value maior que 0.05, distribuição normal.

```
Shapiro-Wilk normality test
data:  residuos
W = 0.98198, p-value = 0.2106
```

d) Testar autocorrelação e corrigir com HAC se for o caso.

## Dwtest

Não há autocorrelação, dado que p-value é maior que 0,05.

```
Durbin-Watson test
data:  lcavol ~ lweight + age + lbph + svi + lcp + gleason + pgg45 + lpsa
DW = 2.3129, p-value = 0.9318
alternative hypothesis: true autocorrelation is greater than 0
```

e) Testar heterocedasticidade e corrigir se for o caso, com regressão robusta ou HC1.

## Bptest

```
Breusch-Pagan test
data:  lcavol ~ lweight + age + lbph + svi + lcp + gleason + pgg45 + lpsa
BP = 14.854, df = 8, p-value = 0.06205
```

```
> chisup <- qchisq(0.95, df=8)
> chisup
[1] 15.50731
```

Como o resultado do teste BP (14.854) é menor que o tabelado (15.50731), aceita-se a hipótese de homoscedasticidade.

f) Fazer regressão stepwise e obter a melhor regressão, apresentar o resultado do modelo.

Melhor modelo stepwise → result\_stepwise

```
Call:
lm(formula = lcavol ~ lcp + lpsa, data = prostate)

Residuals:
    Min       1Q   Median       3Q      Max
-1.65744 -0.54398 -0.05502  0.57163  2.07959

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09135    0.20527   0.445   0.657
lcp          0.32837    0.06193   5.303 7.54e-07 ***
lpsa         0.53162    0.07501   7.087 2.49e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7092 on 94 degrees of freedom
Multiple R-squared:  0.6455,    Adjusted R-squared:  0.6379
F-statistic: 85.57 on 2 and 94 DF,  p-value: < 2.2e-16
```

**g) Obter o AIC, BIC e AICc do melhor modelo selecionado.**

O modelo selecionado para os testes AIC, BIC e AICc é o modelo utilizando todas as variáveis, pois, o melhor modelo stepwise teve um rendimento menor que o modelo preliminar.

```
> AIC(result)
[1] 216.5729
> BIC(result)
[1] 242.32
> library(AICcmodavg)
Warning message:
package 'AICcmodavg' was built under R version 4.0.5
> # AICc used for small samples
> AICc(result)
[1] 219.131
```

**h) Estimar os Intervalos de confiança dos parâmetros.**

Teste de performance e intervalos de confiança, comparando melhor modelo stepwise com modelo preliminar utilizando todas as variáveis.

```
> model_performance(result)
# Indices of model performance
```

AIC	BIC	R2	R2 (adj.)	RMSE	Sigma
216.573	242.320	0.677	0.647	0.667	0.700

```
> model_performance(result_stepwise)
# Indices of model performance
```

AIC	BIC	R2	R2 (adj.)	RMSE	Sigma
213.564	223.863	0.645	0.638	0.698	0.709

```
> confint(result)
```

	2.5 %	97.5 %
(Intercept)	-4.7634552328	0.243253740
lweight	-0.4198499330	0.273517018
age	0.0009476298	0.044524471
lbph	-0.2028782827	0.027979871
svi	-0.6582274921	0.351044923
lcp	0.2049612566	0.529638732
gleason	-0.1158467106	0.497363732
pgg45	-0.0157554674	0.001440317
lpsa	0.4023082197	0.743286502

```
> confint(result_stepwise)
```

	2.5 %	97.5 %
(Intercept)	-0.3162228	0.4989147
lcp	0.2054207	0.4513289
lpsa	0.3826856	0.6805565