



Chapter 2

Exemplos de Regressão Linear em R

[Home Page](#) [Title Page](#) [Contents](#)

[◀◀](#) [▶▶](#) [◀](#) [▶](#)

Page 69 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

2.1. Exemplo: dimensões corporais

O ficheiro **bodyLM.sav** contém dados recolhidos em 507 indivíduos adultos jovens fisicamente activos (praticando várias horas de exercício físico por semana) de vários estados americanos. A amostra consta de 247 homens e de 260 mulheres e, de entre as variáveis consideradas, constaram 12 medidas corporais de circunferência, a idade, o peso, a altura e o sexo. Um dos objectivos do estudo consistiu da modelação do peso através das várias medidas de circunferência recolhidas e da altura.

O ficheiro contém as seguintes variáveis:

- **ShoulderG**: shoulder girth over deltoid muscles (cm)
- **ChestG**: chest girth, nipple line in males and just above breast tissue in females, mid-expiration (cm)
- **WaistG**: waist girth, narrowest part of torso below the rib cage, average of contracted and relaxed position (cm)
- **NavelG**: navel (or "abdominal") girth at umbilicus and iliac crest, iliac crest as a landmark (cm)
- **HipG**: hip girth at level of bitrochanteric diameter (cm)
- **ThighG**: thigh girth below gluteal fold, average of right and left girths (cm)
- **BicepG**: bicep girth, flexed, average of right and left girths (cm)
- **ForearmG**: forearm girth, extended, palm up, average of right and left girths (cm)

- **KneeG**: knee girth over patella, slightly flexed position, average of right and left girths (cm)

- **CalfMaxG**: calf maximum girth, average of right and left girths (cm)

[Home Page](#)

[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Page 70 of 104](#)

[Go Back](#)

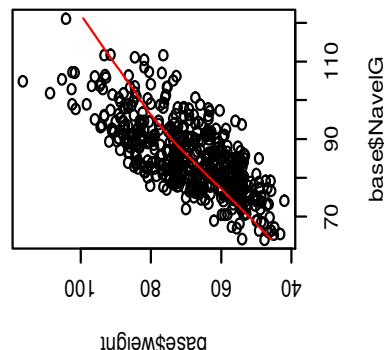
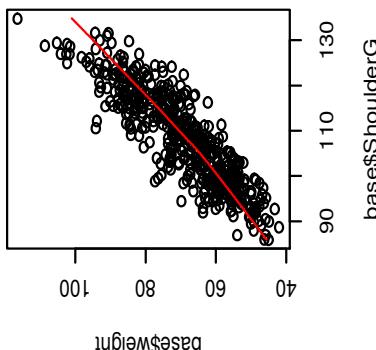
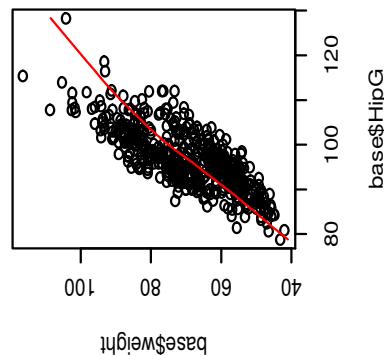
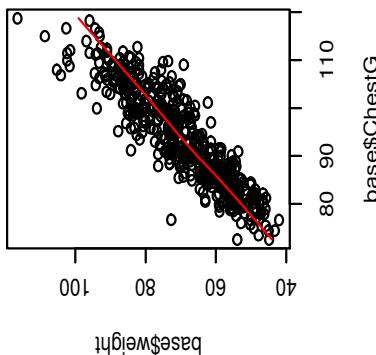
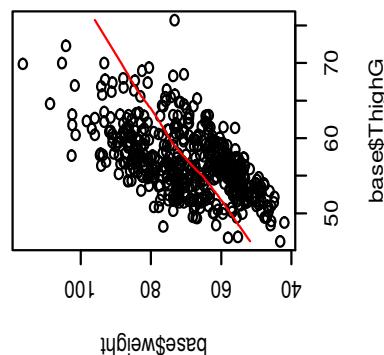
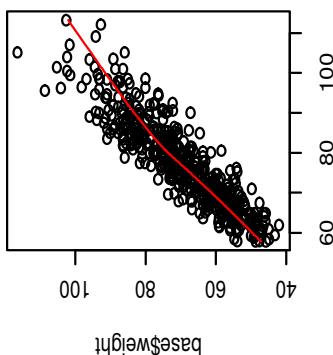
[Full Screen](#)

[Close](#)

[Quit](#)

Considerando o objectivo do estudo e não tendo nenhuma indicação extra sobre as variáveis em causa, podemos começar por considerar o modelo completo, com todas as variáveis explicativas.

$$\begin{aligned} \text{weight} \sim & \text{ShoulderG} + \text{ChestG} + \text{WaistG} + \text{NavelG} \\ & + \text{HipG} + \text{ThighG} + \text{BicepG} + \text{ForearmG} \\ & + \text{KneeG} + \text{CalfMaxG} + \text{AnkleMinG} \\ & + \text{WristMinG} + \text{height} \end{aligned}$$

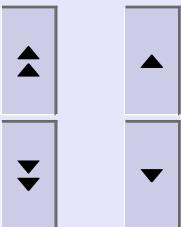




[Home Page](#)

[Title Page](#)

[Contents](#)



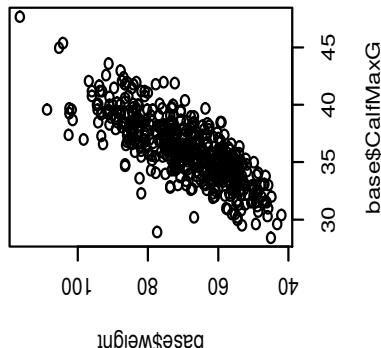
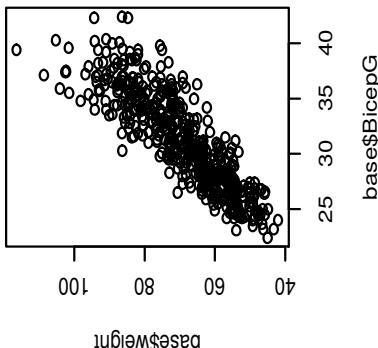
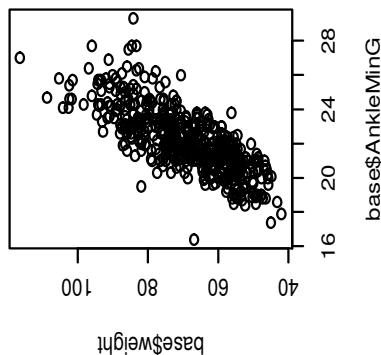
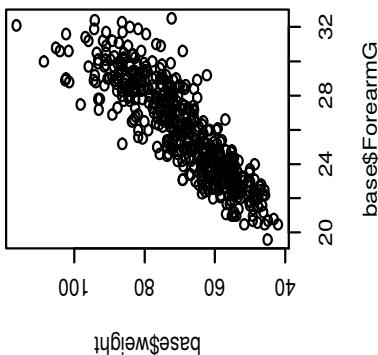
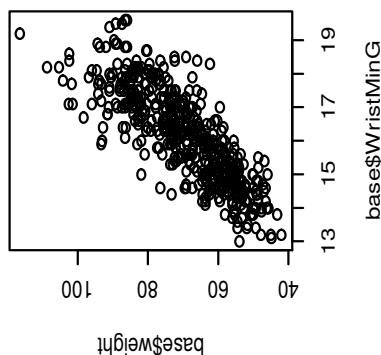
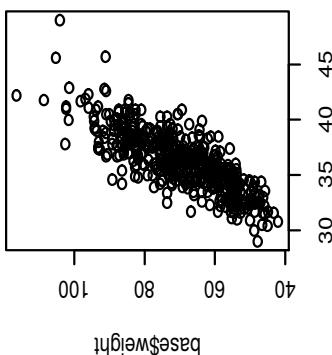
[Page 72 of 104](#)

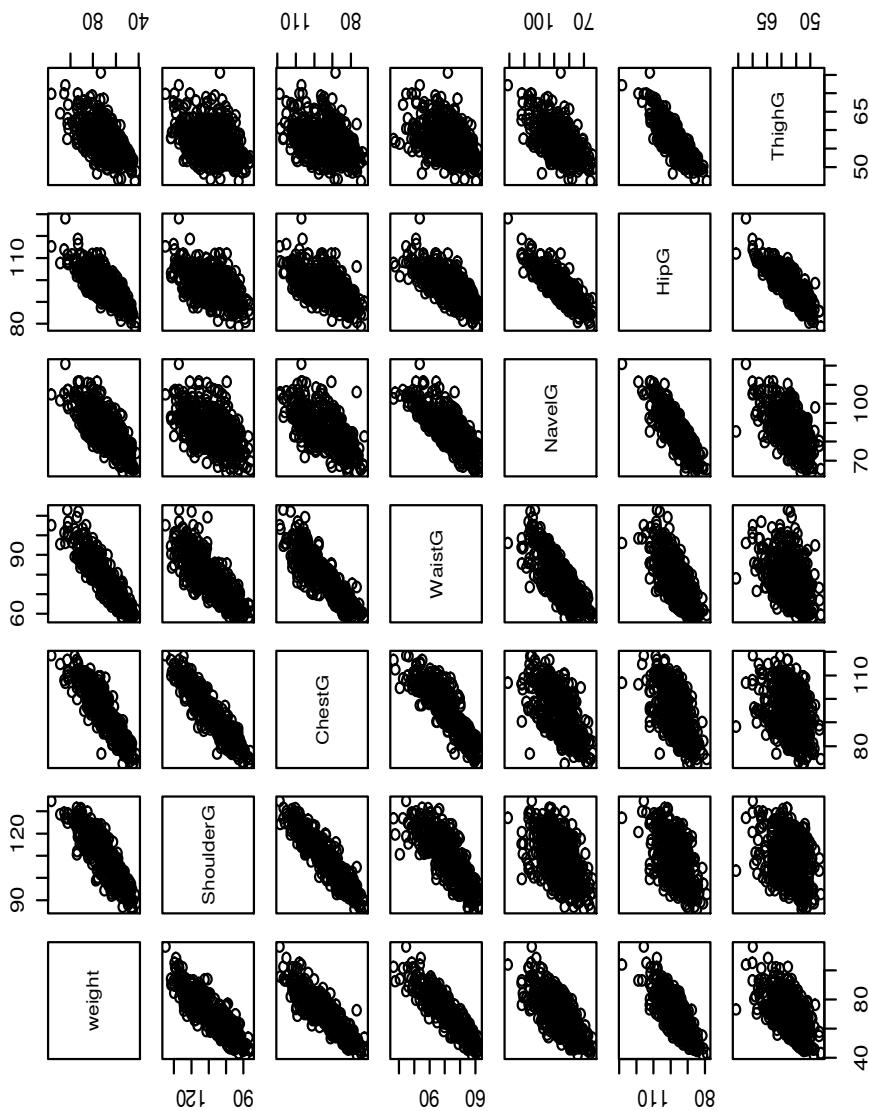
[Go Back](#)

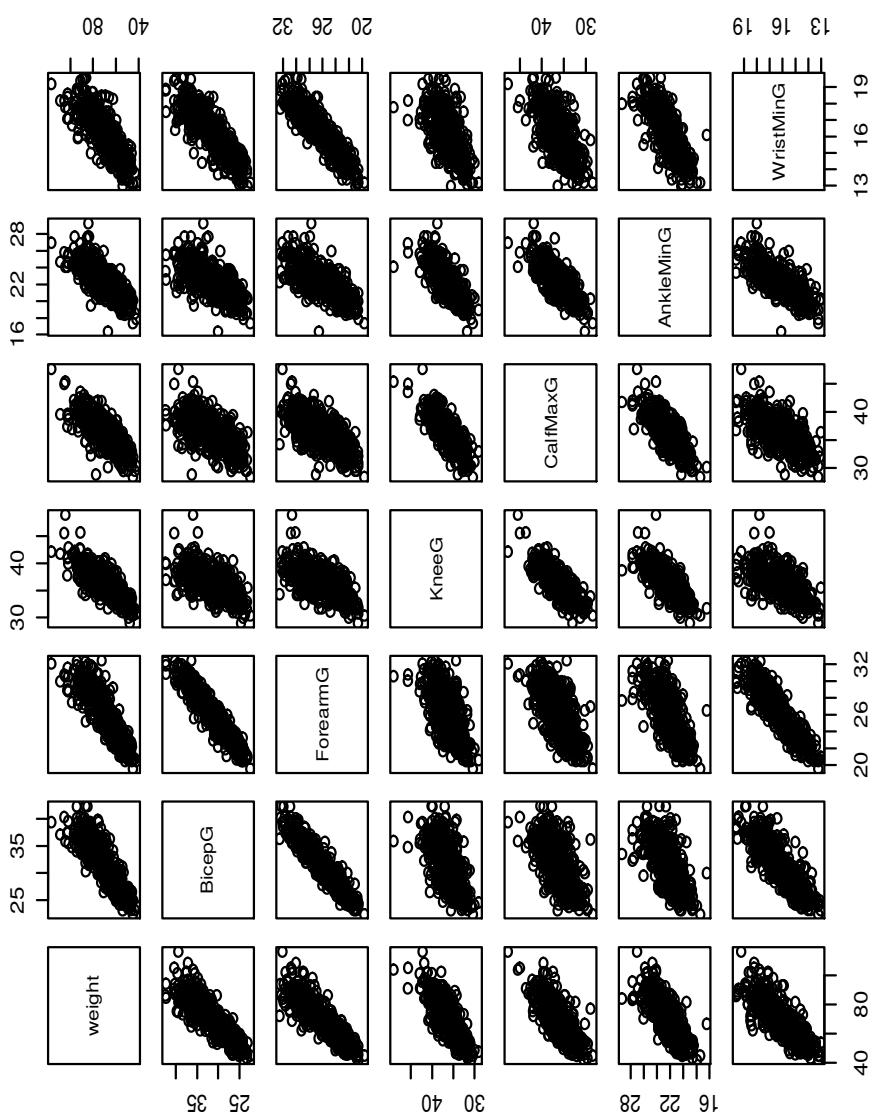
[Full Screen](#)

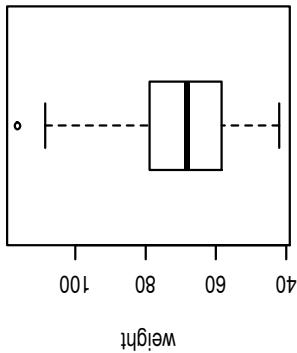
[Close](#)

[Quit](#)

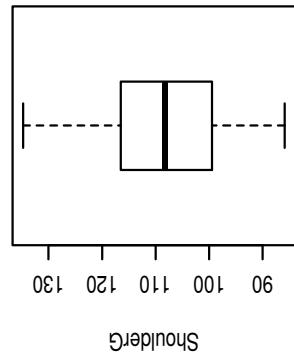




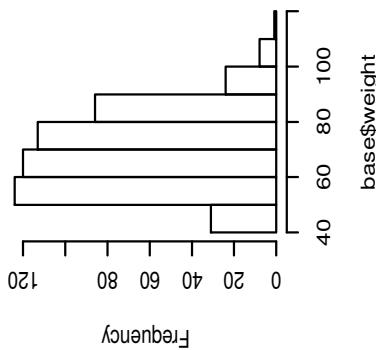




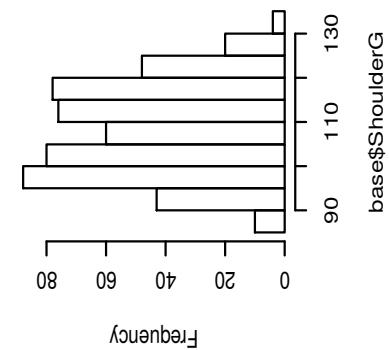
weight



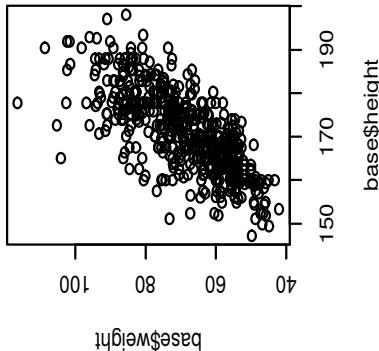
ShoulderG



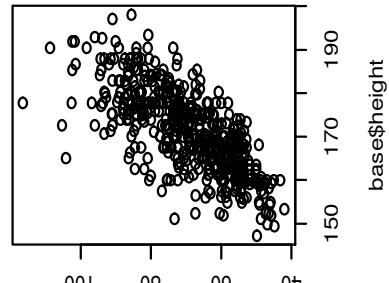
Frequency



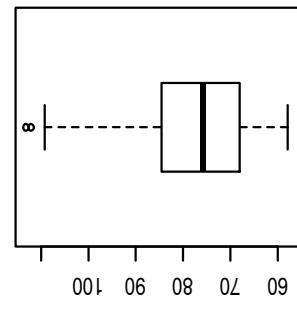
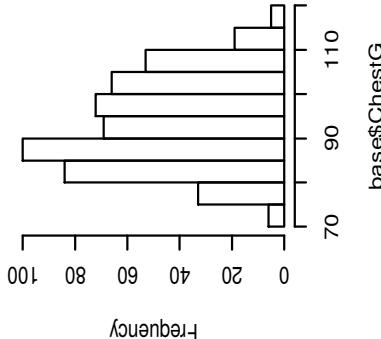
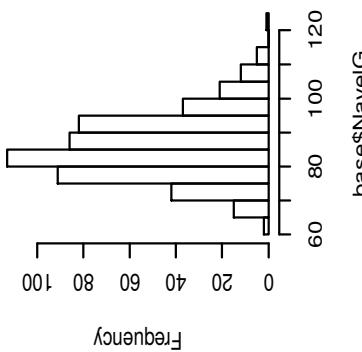
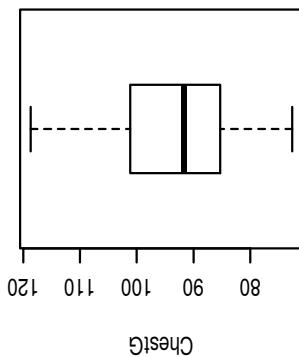
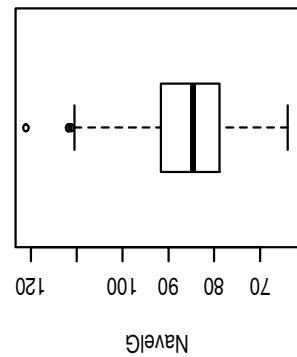
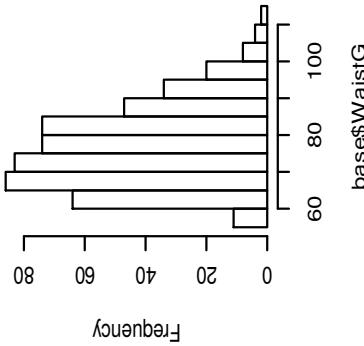
Frequency



base\$weight



base\$weight



```

> base <- read.spss("BodyLM.sav", to.data.frame=TRUE, use.value.labels=FALSE)
Warning message:
In read.spss("BodyLM.sav", to.data.frame = TRUE, use.value.labels = FALSE) :
  BodyLM.sav: Unrecognized record type 7, subtype 18 encountered in system file
>
> names(base)
 [1] "ShoulderG"  "ChestG"      "WaistG"      "NavelG"      "HipG"       "ThighG"
 [7] "BicepG"      "ForearmG"    "KneeG"       "CalfMaxG"   "AnkleMinG"  "WristMinG"
[13] "age"         "weight"      "height"     "sex"
> dim(base)
[1] 507 16

> par(mfrow=c(2,3))
> plot(base$ShoulderG, base$weight)
lines(lowess(base$ShoulderG, base$weight), col="red")

plot(base$ChestG, base$weight)
lines(lowess(base$ChestG, base$weight), col="red")

plot(base$WaistG, base$weight)
lines(lowess(base$WaistG, base$weight), col="red")

plot(base$NavelG, base$weight)
lines(lowess(base$NavelG, base$weight), col="red")

plot(base$HipG, base$weight)
lines(lowess(base$HipG, base$weight), col="red")

plot(base$ThighG, base$weight)
lines(lowess(base$ThighG, base$weight), col="red")

```

```
> cor(base[,c(14, 1:12)])
```

	weight	ShoulderG	ChestG	WaistG	NavelG	HipG	ThighG
weight	1.0000000	0.8788342	0.8989595	0.9039908	0.7118165	0.7629691	0.5585626
ShoulderG	0.8788342	1.0000000	0.9271923	0.8234546	0.5154661	0.5336717	0.3234272
ChestG	0.8989595	0.9271923	1.0000000	0.8837994	0.6229823	0.5834991	0.3630508
WaistG	0.9039908	0.8234546	0.8837994	1.0000000	0.7547704	0.6923506	0.4210849
NavelG	0.7118165	0.5154661	0.6229823	0.7547704	1.0000000	0.8258924	0.6026428
HipG	0.7629691	0.5336717	0.5834991	0.6923506	0.8258924	1.0000000	0.8289411
ThighG	0.5585626	0.3234272	0.3630508	0.4210849	0.6026428	0.8289411	1.0000000
BicepG	0.8666722	0.8951884	0.9081845	0.8047044	0.5578071	0.5598848	0.4114580
ForearmG	0.8695531	0.8949838	0.8875909	0.7807924	0.4862181	0.5143585	0.3452848
KneeG	0.7955518	0.6247826	0.6140547	0.6582072	0.6120932	0.7349017	0.6384400
CalfMaxG	0.7692826	0.6270538	0.6088643	0.6313445	0.5247789	0.6745805	0.6288901
AnkleMinG	0.7619985	0.6797568	0.6691396	0.6558891	0.5194785	0.5770429	0.4217687
WristMinG	0.8164884	0.8407085	0.8246754	0.7289813	0.4354197	0.4588567	0.2416102

	BicepG	ForearmG	KneeG	CalfMaxG	AnkleMinG	WristMinG
weight	0.8666722	0.8695531	0.7955518	0.7692826	0.7619985	0.8164884
ShoulderG	0.8951884	0.8949838	0.6247826	0.6270538	0.6797568	0.8407085
ChestG	0.9081845	0.8875909	0.6140547	0.6088643	0.6691396	0.8246754
WaistG	0.8047044	0.7807924	0.6582072	0.6313445	0.6558891	0.7289813
NavelG	0.55778071	0.4862181	0.6120932	0.5247789	0.5194785	0.4354197
HipG	0.5598848	0.5143585	0.7349017	0.6745805	0.5770429	0.4588567
ThighG	0.4114580	0.3452848	0.6384400	0.6288901	0.4217687	0.2416102
BicepG	1.0000000	0.9423755	0.6207299	0.6374041	0.6693240	0.8479443
ForearmG	0.9423755	1.0000000	0.6575450	0.6701918	0.7125539	0.9047086
KneeG	0.6207299	0.6575450	1.0000000	0.7958277	0.7377154	0.6409596
CalfMaxG	0.6374041	0.6701918	0.7958277	1.0000000	0.7622219	0.6476269
AnkleMinG	0.6693240	0.7125539	0.7377154	0.7622219	1.0000000	0.7536365
WristMinG	0.8479443	0.9047086	0.6409596	0.6476269	0.7536365	1.0000000

```
> pairs(base[,c(14, 1:6)])
> pairs(base[,c(14, 7:12)])
```

[Close](#)

[Home Page](#) [Title Page](#) [Contents](#) [Full Screen](#) [Quit](#)

```

> mod1 <- lm(weight ~ ShoulderG+ChestG+WaistG+NavelG+HipG+ThighG+BicepG+
+ ForearmG+KneeG+CalfMaxG+AnkleMinG+WristMinG+height , data=base)
> summary(mod1)

```

Call:

```

lm(formula = weight ~ ShoulderG + ChestG + WaistG + NavelG +
    HipG + ThighG + BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG +
    WristMinG + height, data = base)

```

Residuals:

	Min	1Q	Median	3Q	Max
	-7.7332	-1.3677	0.0069	1.2171	10.3976

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.202e+02	2.489e+00	-48.306	< 2e-16 ***
ShoulderG	7.813e-02	2.979e-02	2.622	0.009001 **
ChestG	1.979e-01	3.569e-02	5.544	4.83e-08 ***
WaistG	3.404e-01	2.438e-02	13.960	< 2e-16 ***
NavelG	1.172e-03	2.291e-02	0.051	0.959225
HipG	2.404e-01	4.334e-02	5.547	4.76e-08 ***
ThighG	3.141e-01	5.148e-02	6.103	2.11e-09 ***
BicepG	5.468e-02	8.526e-02	0.641	0.521631
ForearmG	5.321e-01	1.371e-01	3.882	0.000118 ***
KneeG	3.013e-01	7.740e-02	3.892	0.000113 ***
CalfMaxG	4.039e-01	7.005e-02	5.765	1.44e-08 ***
AnkleMinG	-9.635e-03	9.992e-02	-0.096	0.923221
WristMinG	-1.180e-01	1.959e-01	-0.602	0.547135
height	3.282e-01	1.560e-02	21.033	< 2e-16 ***

```

Signif. codes: 0 `***' 0.001 `*' 0.01 `.' 0.05 `.' 0.1 `-' 1

```

Residual standard error: 2.204 on 493 degrees of freedom
Multiple R-squared: 0.9734, Adjusted R-squared: 0.9727
F-statistic: 1390 on 13 and 493 DF, p-value: < 2.2e-16

[Close](#)

[Quit](#)

[Title Page](#)

[Contents](#)

[Page 79 of 104](#)

[Go Back](#)

[Full Screen](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)



FACULDADE DE CIÉNCIAS
UNIVERSIDADE DO PORTO

```
> anova(mod1)
Analysis of Variance Table
```

Response: weight

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ShoulderG	1	69607	69607	14332.992	< 2.2e-16 ***
ChestG	1	4544	4544	935.680	< 2.2e-16 ***
WaistG	1	4822	4822	992.990	< 2.2e-16 ***
NavelG	1	1131	1131	232.923	< 2.2e-16 ***
HipG	1	3089	3089	636.027	< 2.2e-16 ***
ThighG	1	226	226	46.615	2.549e-11 ***
BicepG	1	201	201	41.442	2.884e-10 ***
ForearmG	1	997	997	205.275	< 2.2e-16 ***
KneeG	1	755	755	155.542	< 2.2e-16 ***
CalfMaxG	1	151	151	30.991	4.263e-08 ***
AnkleMinG	1	24	24	4.851	0.028092 *
WristMinG	1	34	34	6.929	0.008747 **
height	1	2148	2148	442.391	< 2.2e-16 ***
Residuals	493	2394	5	---	---

```
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `-' 1
```

```
# coeficientes estandardizados
```

```
> library(QuantPsyc)
> lm.beta(mod1)
```

	ShoulderG	ChestG	WaistG	NavelG	HipG	ThighG
0.060735509	0.148661084	0.280903890	0.000827652	0.120334352	0.104980810	
BicepG	ForearmG	KneeG	CalfMaxG	AnkleMinG	WristMinG	
0.017400311	0.112861443	0.059086773	0.086176095	-0.001344473	-0.012212489	
height						
0.231311627						

Page 80 of 104

Go Back

Full Screen

Close

Quit

```

> regressionSS <- sum(anova(mod1)[1:13,2] )

> regressionSS
[1] 87729.14

> residualSS <- anova(mod1)[14,2]
> residualSS
[1] 2394.205

> totalSS <- residuals + regressionSS
> totalSS
[1] 90123.34

# resíduos estandardizados
> hist(rstandard(mod1))
> boxplot(rstandard(mod1))
> rqqnorm(rstandard(mod1))
> qqline(rstandard(mod1))
> hist(rstandard(mod1), breaks=20)

> which(rstandard(mod1)>3.3)
# 4 observações; poder-se-ia explorar a relevância e o significado físico
# dessas observações no contexto dos dados

> library(car)
> qqPlot(residuals(mod1))

# resíduos studentizados; só funciona para modelos lm
> qqPlot(mod1)

> plot(fitted.values(mod1), rstandard(mod1))
> abline(0,0, lty="dashed")
> plot(fitted.values(mod1), base$weight)
> abline(0,1, lwd=2, col="red")
> plot(base$ShoulderG, rstandard(mod1))
> plot(base$height, rstandard(mod1))

```





A tabela da ANOVA mostra que o modelo é estatisticamente significativo ($F=1389.588$, $p=0.000$) sendo que o efeito de quatro variáveis explicativas (NavelG, BicepG, AnkleMinG, WristMinG) sobre o peso não é significativo. Os coeficientes estandardizados indicam que as variáveis WaistG e height são as que mais influenciam o peso dos indivíduos.

O coeficiente de determinação é significativo e bastante elevado $R^2 = 0.973$ ($p=0.000$), $\bar{R}^2 = 0.973$. Antes de retirarmos mais conclusões sobre o modelo verificamos se os seus pressupostos são satisfeitos (gráficos na página seguinte):

- o histograma, o boxplot e o qq-plot dos resíduos estandardizados sugerem ligeiros desvios da normalidade, que parecem contudo toleráveis dado o elevado tamanho amostral (estamos a usar o teorema do limite central); identificam-se ainda algumas observações com resíduos superiores a 3.3 unidades.

- o gráfico dos resíduos estandardizados contra os valores ajustados indica pequenos problemas com a hipótese de homocedasticidade e evidencia 3 pontos com resíduos grandes; os gráficos dos resíduos estandardizados contra cada uma das variáveis regressoras e dos valores ajustados contra a resposta indicam conclusões análogas. A hipótese de homocedasticidade não parece estar a ser grandemente violada.

- o gráfico das leverages contra as distâncias de Cook (nas páginas seguintes) não revela a existência de pontos influentes.

O ponto de corte tradicional para as leverages é de $2(p+1)/n = 2(13+1)/507 = 0.055$ e para as distâncias de Cook é 1.0. Há alguns pontos com leverages superior a 0.055 mas não parecem ser problemáticos porque as suas distâncias de Cook são inferiores a 1.

Acontece porém que

- uma análise à multicolinearidade do modelo revela altos coeficientes de correlação entre várias das variáveis explicativas (gráfico de dispersão das variáveis explicativas duas a duas e tabela dos coeficientes de correlação)
- alguns dos coeficientes de regressão não são significativos, o que sugere que pelo menos algumas das variáveis explicativas que lhes estão associadas possam ser excluídas do modelo

- a interpretação do modelo seria mais simples caso este contivesse menos variáveis.

Uma forma de reduzir o número de variáveis explicativas poderia ser por eliminação sucessiva daquelas que apresentam um maior valor- p , portanto as menos significativas. Em cada eliminação ter-se-ia de interpretar as várias estimativas obtidas, incluindo estimativas relacionadas com o ajustamento do modelo, e averiguar a satisfação das hipóteses das equações de regressão através de análises gráficas adequadas.



FACULDADE DE CIÉNCIAS
UNIVERSIDADE DO PORTO

[Home Page](#)

[Title Page](#)

[Contents](#)



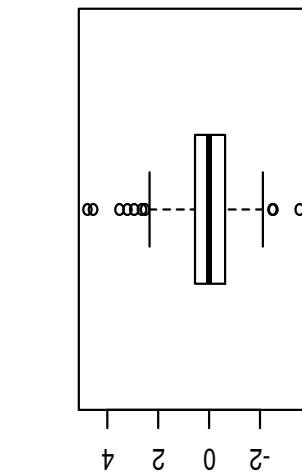
[Page 83 of 104](#)

[Go Back](#)

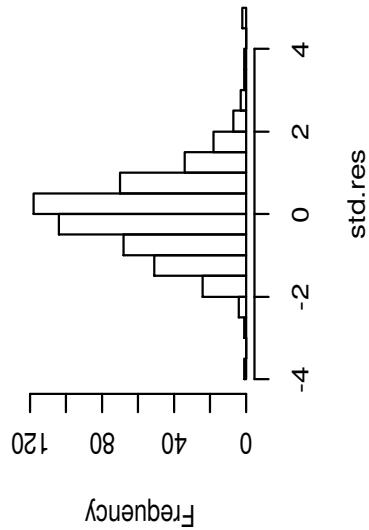
[Full Screen](#)

[Close](#)

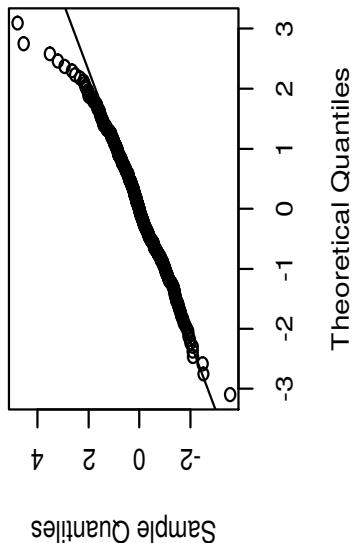
[Quit](#)

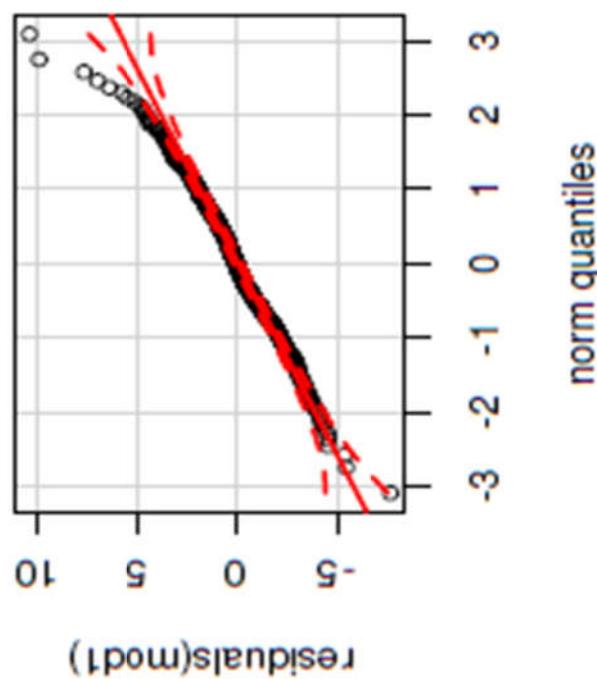
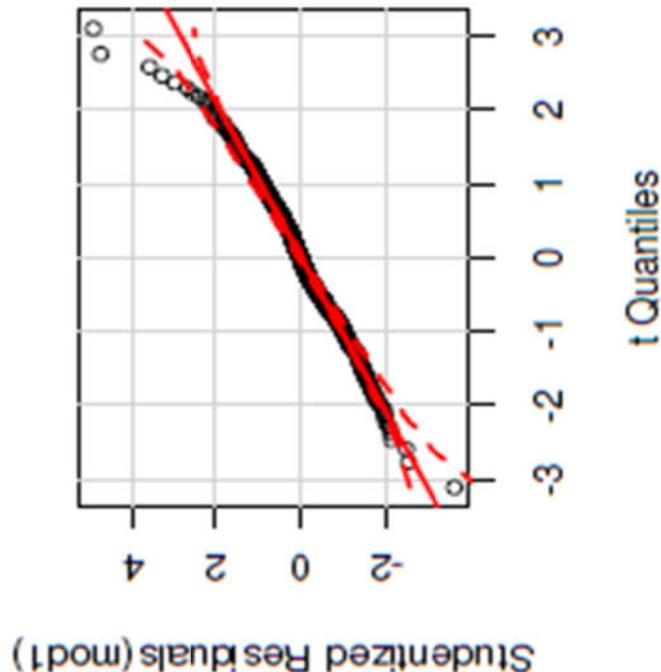


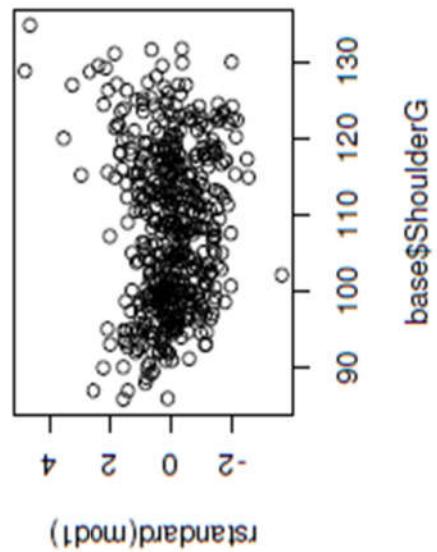
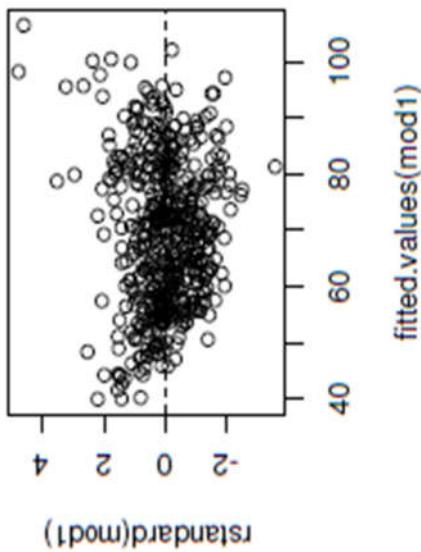
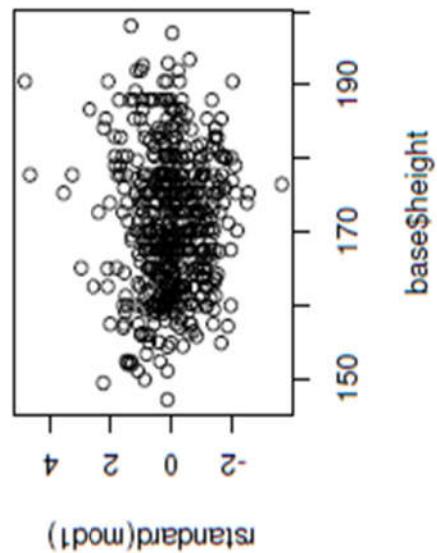
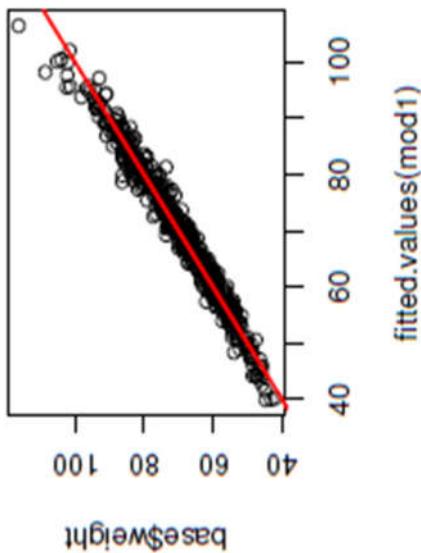
Histogram of std.res



Normal Q-Q Plot



[Home Page](#)[Title Page](#)[Contents](#)[Page 84 of 104](#)[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)



Há 4 observações com resíduos superiores a 3.3 desvios-padrão. Poder-se-ia explorar a relevância e significado físico dessas observações no conjunto de dados.



[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

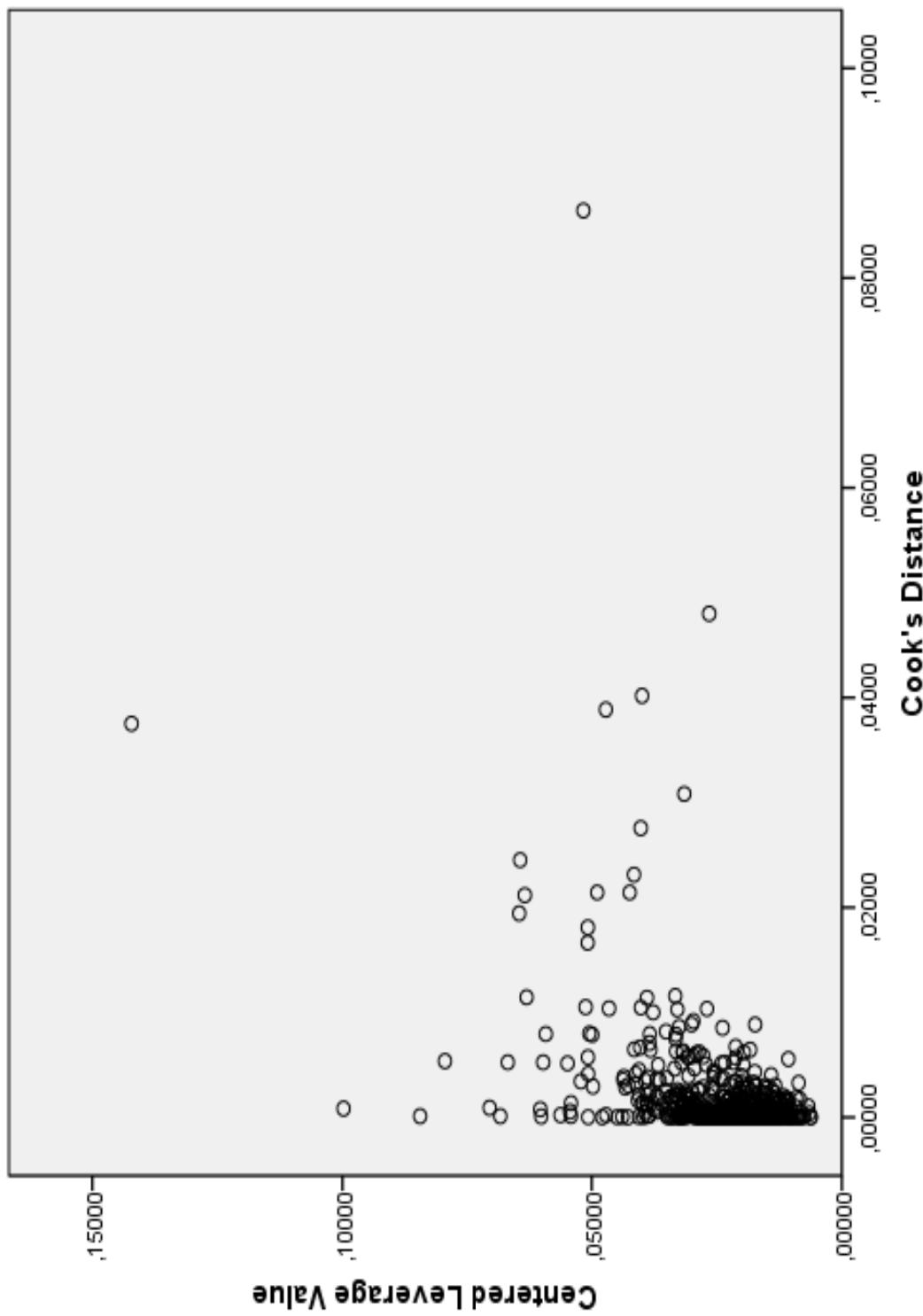
Page 86 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



weight ~ ShoulderG + ChestG + WaistG
+ HipG + ThighG + ForearmG
+ KneeG + CalfMaxG + height

Aproveitamos este exemplo para aplicar alguns métodos de selecção de variáveis. Seguiremos o método stepwise, em ambas as direcções (instruções e resultados mais pormenorizados na página seguinte). O modelo escolhido é

- De entre ShoulderG, ChestG e WaistG, todas muito correlacionados, escolhemos ficar apenas com a variável WaistG por apresentar o maior coeficiente estandardizado e o valor mais significativo da estatística t e portanto nos estar a dar a indicação de que será essa variável que mais contribui para explicar o peso.

Apesar do problema de multicolinearidade persistir, há diagnósticos que se podem fazer que levam a crer que isso não esteja a afectar o ajustamento do modelo:

- os erros padrão dos coeficientes não são exageradamente grandes
- não há incongruências entre os resultados das estatísticas F e t

- retiraram-se algumas variáveis do modelo e os coeficientes das outras variáveis não sofreram alterações substanciais.

Analises gráficas standard (gráficos na página seguinte) não revelam violação das hipóteses de normalidade, homocedasticidade, linearidade e independência. Receamos que os coeficientes estejam a ser demasiado significativos devido à multicolinearidade elevada pelo que, de seguida, tentaremos reduzir essa multicolinearidade e ajustaremos um terceiro modelo.

- De entre HipG e ThighG escolhemos ThighG por apresentar o maior valor para a estatística t
- De entre KneeG e CalfMaxG escolhemos CalfMaxG por apresentar o maior valor para a estatística t
- De entre WaistG e ForearmG não eliminamos nenhuma porque ambas as variáveis parecem estar a ser muito significativas para o modelo e o coeficiente de correlação não é exageradamente elevado.

Os restantes coeficientes de correlação já parecem ser razoáveis pelo que prosseguimos então com o modelo

weight ~ WaistG + ThighG + ForearmG
+ CalfMaxG + height

[Home Page](#) [Title Page](#) [Contents](#)

[Page 87 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

[Home Page](#)[Title Page](#)[Contents](#)[Page 88 of 104](#)[Go Back](#)[Full Screen](#)[Close](#)[Quit](#)

Em relação a este último modelo (exercício):

- $R^2 = 0.966$ ($p = 0.000$), $R^2 = 0.966$
- $F = 2836.678$ ($p = 0.000$)

- não existe multicolinearidade elevada entre as variáveis (coeficientes de correlação razoáveis e VIF < 10 para todas as variáveis explicativas; o índice de condição é elevado mas as proporções de variância não são elevadas em mais de duas variáveis portanto não é de valorizar)
- todos os pressupostos do modelo parecem ser satisfeitos
- o gráfico das leverages contra as distâncias de Cook revela a existência de um ponto influente. O efeito da remoção desse ponto sobre as estimativas dos coeficientes deverá ser analisada.

> mod2 <- step(mod1, direction="both")

Start: AIC=815.01
 weight ~ ShoulderG + ChestG + WaistG + NavelG + HipG + ThighG +
 BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG +
 height

	Df	Sum of Sq	RSS	AIC
- NavelG	1	0.01	2394.2	813.02
- AnkleMinG	1	0.05	2394.3	813.02
- WristMinG	1	1.76	2396.0	813.39
- BicepG	1	2.00	2396.2	813.44
<none>		2394.2	815.01	
- ShoulderG	1	33.40	2427.6	820.04
- ForearmG	1	73.17	2467.4	828.28
- KneeG	1	73.56	2467.8	828.36
- ChestG	1	149.26	2543.5	843.68
- HipG	1	149.40	2543.6	843.70
- CalfMaxG	1	161.43	2555.6	846.10
- ThighG	1	180.86	2575.1	849.94
- WaistG	1	946.48	3340.7	981.91
- height	1	2148.43	4542.6	1137.72

Step: AIC=813.02

weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + BicepG +
 ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG + height

	Df	Sum of Sq	RSS	AIC
- AnkleMinG	1	0.04	2394.3	811.03
- WristMinG	1	1.80	2396.0	811.40
- BicepG	1	2.09	2396.3	811.46
<none>		2394.2	813.02	
+ NavelG	1	0.01	2394.2	815.01
- ShoulderG	1	35.06	2429.3	818.39
- KneeG	1	74.28	2468.5	826.51
- ForearmG	1	74.34	2468.6	826.52
- ChestG	1	153.89	2548.1	842.60

- CalfMaxG 1 163.46 2557.7 844.50
- ThighG 1 183.09 2577.3 848.38
- HipG 1 200.26 2594.5 851.74
- WaistG 1 1080.33 3474.5 999.83
- height 1 2153.30 4547.5 1136.27

Step: AIC=811.03
 weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + BicepG +
 ForearmG + KneeG + CalfMaxG + WristMinG + height

.

.

Step: AIC=807.87
 weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG +
 KneeG + CalfMaxG + height

	Df	Sum of Sq	RSS	AIC
<none>		2398.2	807.87	
+ WristMinG	1	1.88	2396.4	809.47
+ BicepG	1	1.87	2396.4	809.47
+ AnkleMinG	1	0.36	2397.9	809.79
+ NavelG	1	0.13	2398.1	809.84
- ShoulderG	1	38.00	2436.2	813.84
- KneeG	1	72.48	2470.7	820.96
- ChestG	1	169.16	2567.4	840.42
- ForearmG	1	177.77	2576.0	842.12
- CalfMaxG	1	180.18	2578.4	842.60
- HipG	1	196.99	2595.2	845.89
- ThighG	1	240.80	2639.0	854.38
- WaistG	1	1103.10	3501.3	997.72
- height	1	2219.38	4617.6	1138.03

<summary(mod2)

```

Call:
lm(formula = weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG +
    ForearmG + KneeG + CalfMaxG + height, data = base)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7663 -1.3694  0.0286  1.2146 10.3239

```

Coefficients:

	Estimate	Std. Error	t value	F(1,16)	Significance
(Intercept)	-120.83845	2.37022	-50.982	< 2e-16	***
ShoulderG	0.08007	0.02853	2.806	0.005211	**
ChestG	0.20152	0.03404	5.921	5.98e-09	***
WaistG	0.34283	0.02267	15.120	< 2e-16	***
HipG	0.23724	0.03713	6.389	3.84e-10	***
ThighG	0.33178	0.04697	7.064	5.48e-12	***
ForearmG	0.54867	0.09040	6.070	2.55e-09	***
KneeG	0.28703	0.07406	3.876	0.000121	***
CalfMaxG	0.38924	0.06370	6.111	2.01e-09	***
height	0.32519	0.01516	21.446	< 2e-16	***
---	---	---	---	---	---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.197 on 497 degrees of freedom
 Multiple R-squared: 0.9734, Adjusted R-squared: 0.9729
 F-statistic: 2020 on 9 and 497 DF, p-value: < 2.2e-16

```
> anova(mod2, mod1) # F test for nested models
```

Model 1: weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG + KneeG + CalfMaxG + height
 Model 2: weight ~ ShoulderG + ChestG + WaistG + NavelG + HipG + ThighG + BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG + height

```
Res.Df   RSS Df Sum of Sq F Pr(>F)
1     497 2398.2
2     493 2394.2  4  4.0352 0.2077 0.9341
```

```
> lrtest(mod2, mod1)
Likelihood ratio test
```

```
Model 1: weight ~ ShoulderG + ChestG + WaistG + HipG + ThighG + ForearmG +
KneeG + CalfMaxG + height
Model 2: weight ~ ShoulderG + ChestG + WaistG + NavelG + NavelG + HipG + ThighG +
BicepG + ForearmG + KneeG + CalfMaxG + AnkleMinG + WristMinG +
height
#Df LogLik Df Chisq Pr(>Chisq)
1 11 -1113.3
2 15 -1112.9  4  0.8538  0.9311
```

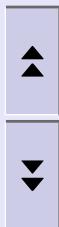


FACULDADE DE CIÉNCIAS
UNIVERSIDADE DO PORTO

[Home Page](#)

[Title Page](#)

[Contents](#)



Page 92 of 104

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)



FACULDADE DE CIÉNCIAS
UNIVERSIDADE DO PORTO

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ | ▶▶

◀ | ▶

[Page 93 of 104](#)

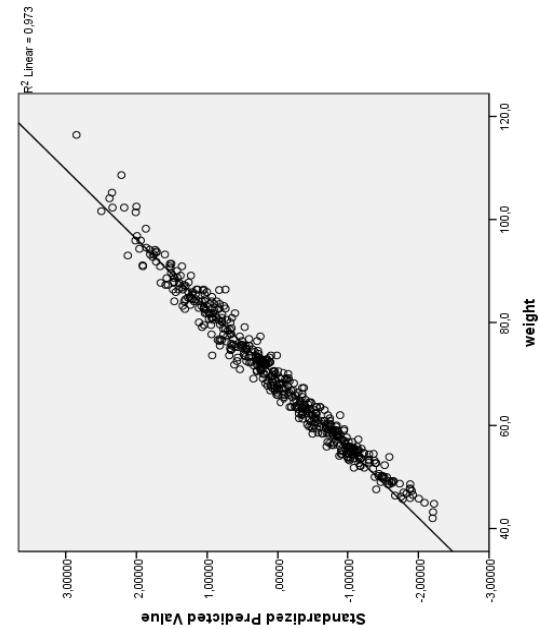
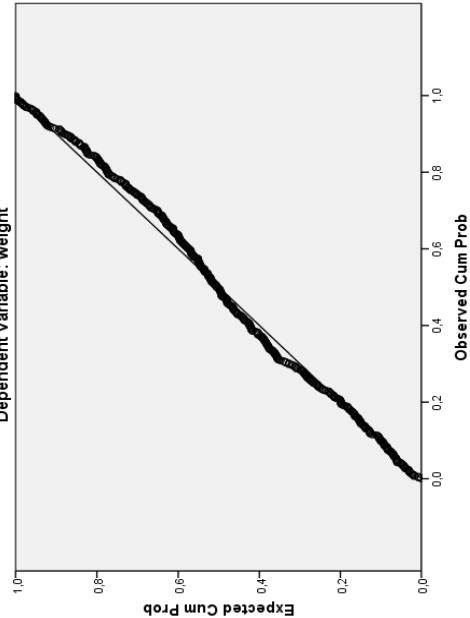
[Go Back](#)

[Full Screen](#)

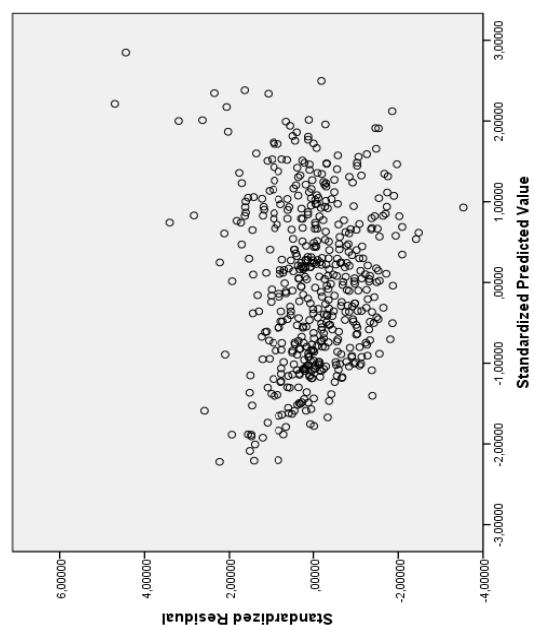
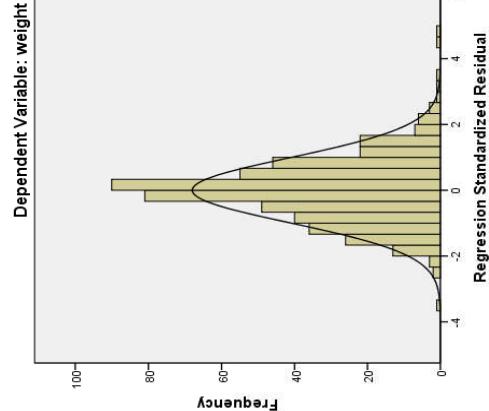
[Close](#)

[Quit](#)

Normal P-P Plot of Regression Standardized Residual



Histogram



2.2. Exemplo: funções respiratórias e tabaco

O ficheiro FEV.sav ^a contém dados recolhidos entre 1975 e 1980 respeitantes a 654 crianças e adolescentes da área de Boston Oriental. O objectivo do estudo consistiu na avaliação das funções pulmonares na presença ou ausência de exposição a fumo de cigarros (quer por exposição activa do próprio por fumar, quer por exposição passiva por contacto com um progenitor fumador).

O ficheiro contém as seguintes variáveis:

- age: idade (anos)
- fev: volume expiratório forçado (litros)
- ht: altura (polegadas)
- sex: sexo (0-rapariga; 1-rapaz)
- smoke: fuma cigarros de forma regular? resposta auto-declarada: 0- não; 1-sim. ^b

Tendo em conta o objectivo da análise, começamos por descrever numericamente e graficamente o volume expiratório forçado observado no grupo dos fumadores e dos não fumadores.

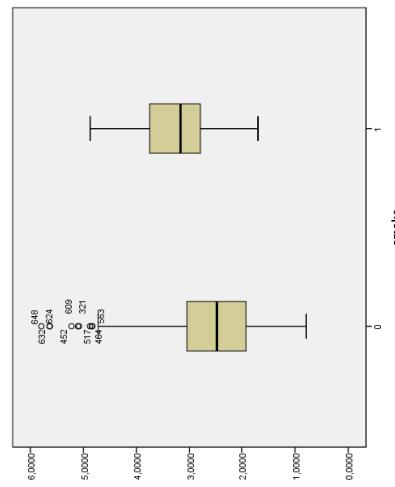
O comando

Analyze → Descript Stats → Explore

com a separação de fev por sexo evidencia uma distribuição assimétrica para os não fumadores, rejeitando a hipótese de normalidade, e uma distribuição relativamente simétrica para os fumadores.

^a Rosner, B. (1999), Fundamentals of Biostatistics, 5th Ed., Pacific Grove, CA: Duxbury
^besta variável diz apenas respeito à exposição activa, apesar de outros dados terem sido também levantados no estudo.

Var.	Smoke	n	min	max
fev	0	589	0.7910	5.7930
fev	1	65	1.6940	4.8720



[Title Page](#)

[Contents](#)

[◀◀](#) [▶▶](#)

[◀](#) [▶](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

Os resultados sugerem que, de uma forma geral, os fumadores apresentam valores superiores de fev (i.e., melhores funções pulmonares) do que os não fumadores! Mas:

- a classificação em fumador vs não-fumador é automaticamente declarada...
- é sabido que os valores de fev dependem da estrutura corporal e a análise anterior não teve isso em consideração...

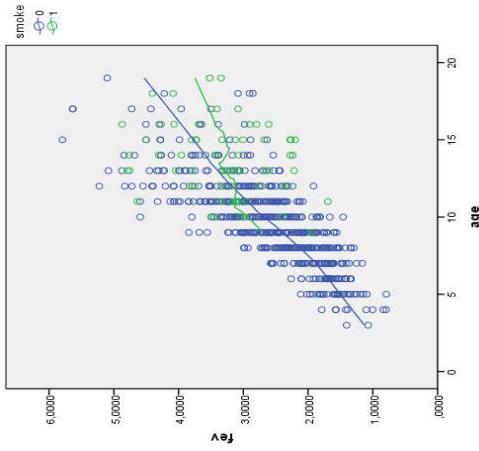
[Page 94 of 104](#)

Tentamos então obter gráficos mais elucidativos sobre a questão, ajustados para variáveis que são julgadas de interesse.

O gráfico abaixo obteve-se das instruções usuais de

Graphs → Chart Builder → Scatter/Dot

com separação das observações, por cor, de acordo com a classe de smoke. As curvas representadas correspondem a um modelo de regressão não linear local. Obtém-se clicando sobre a figura, no output, escolhendo depois o ícone que representa duas retas de regressão diferentes ajustadas às observações e considerando a opção loess.



Repare-se que agora há conclusões diferentes a retirar, sendo que algumas já parecem ir ao encontro daquilo que se esperava.

- a função pulmonar cresce com a idade de uma forma aproximadamente linear; para os jovens a partir dos 14 anos, aproximadamente, parece existir uma distinção entre os valores de fev entre fumadores e não-fumadores, sendo que agora a relação de ordem parece ser a correcta.

Esta questão ultrapassa a análise aqui apresentada mas **os resultados estão a sugerir a existência de uma interacção age*smoke.**

Nota: Diz-se que existe **interacção** $X_1 * X_2$ entre duas variáveis explicativas X_1 e X_2 quando o efeito de uma delas sobre a resposta depende do valor da outra variável.

Havendo interacção, os efeitos principais de cada uma das variáveis não devem ser retirados do modelo de regressão, i.e., devemos considerar

$$Y \sim X_1 + X_2 + X_1 * X_2 + \dots$$

- a função pulmonar cresce com a altura de uma forma que parece quadrática mas não há indicação de que essa relação esteja a ser influenciada pelo facto de o indivíduo ser ou não fumador.

Outra questão que se pode levantar: terá interesse considerar crianças com menos de, por exemplo, 6 anos?! Se por um lado essa crianças são num certo sentido irrelevantes para qualquer avaliação do efeito de fumar, também é verdade que os seus dados podem contribuir para uma melhor explicação do fev em função das outras medidas recolhidas...



Apesar de já se ter visto que a regressão

$$\text{fev} = \beta_0 + \beta_1 \text{smoke} + u, \quad u \sim N(0, \sigma^2 \text{Id})$$

não vai conduzir a resultados concordantes com o expectável, a análise dessa equação vai ser aqui apresentada de forma muito rápida, por motivos pedagógicos.

A variável smoke é um factor com duas categorias (portanto coincide com a dummy que lhe está associada), sendo a classe dos não fumadores a classe de referência.

Os resultados para o ajustamento do modelo aos dados são os seguintes

Modelo	coef	s.e. coef	t	valor-p
constante	2.566	0.035	74.037	0.000
smoke	0.711	0.110	6.464	0.000

$$F = 41.789, p = 0.000; R^2 = 0.060, \bar{R}^2 = 0.059$$

(Por o modelo não ter interesse, não avançamos sequer para as análises gráficas e numéricas de verificação dos pressupostos do modelo.)

Supondo que as condições do modelo eram satisfeitas, a tabela permitiria concluir:

- a existência de diferenças significativas para o fev de acordo com a classe de fumador (observar que a estatística de teste da tabela coincide com a estatística de teste de um teste de comparação de médias em amostras independentes assumindo variâncias iguais - exercício)

- o valor médio de fev para um indivíduo fumador é de $2.566 + 0.711$ enquanto que o correspondente valor médio para um indivíduo não fumador é de 2.566. Em média, um fumador tem um volume expiratório superior em 0.711 litros a um não fumador. Um intervalo a 95% de confiança para esta diferença é aproximadamente $0.7 \pm 2(0.1)$, isto é, (0.5, 0.9).

- caso a variável smoker tivesse mais de duas categorias, o que esta análise de regressão permitiria analisar era a comparação entre os vários volumes expiratórios médios para as diferentes classes de smoker (ANOVA - análise da variância).

Os comandos em **SPSS** correspondentes à tabela apresentada acima são os seguintes:

Analyze → Regression → Linear
Dependent: fev
Independent(s): smoke

deixando todas as opções por defeito que o software tem incluídas.

De acordo com os últimos gráficos considerados, analisamos agora a regressão do volume expiratório contra smoke, ajustada para a idade e a altura.

fev ~ smoke + age + height + height².

^a Um dos gráficos anteriores sugere uma dependência quadrática do volume expiratório em relação à altura. Nestas situações, é usual considerarem-se ambas as variáveis "altura" e "altura"².

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.890 ^a	.791	.780	.3873925	,791	615,303	4	.649	,000

a. Predictors: (Constant), ht_sq, smoke, age, ht

b. Dependent Variable: fev

ANOVA^b

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression 388,481	4	97,120	615,303	,000 ^a
	Residual 102,439		64,9		
	Total 490,920		65,3		

a. Predictors: (Constant), ht_sq, smoke, age, ht

b. Dependent Variable: fev

Coefficients^a

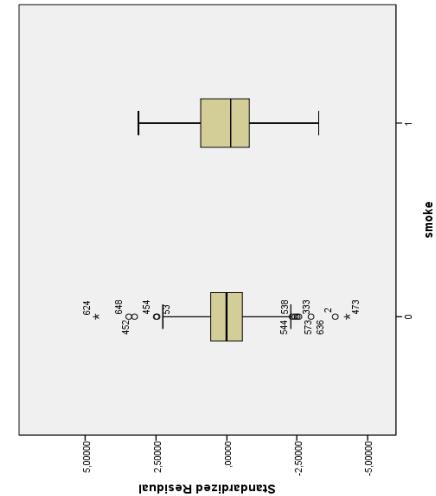
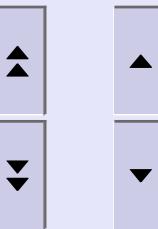
Model	Unstandardized Coefficients	Standardized Coefficients	Beta	t	Sig.	Collinearity Statistics	
						VIF	
1	(Constant) 7,868	1,469		,052	,5,356	,000	
	smoke ,150	,057		-,227	-2,635	,009	,827
	age ,067	,009		,2019	7,313	,000	,334
	ht ,307	,049		-,6,310	,000	,003	,2992
	ht_sq ,003	,000		2,726	8,589	,000	,003

a. Dependent Variable: fev

PASW Statistics Processor is ready

As variáveis explicativas consideradas são todas significativas para o modelo. Pela observação da magnitude dos coeficientes estandardizados observamos que a altura parece ser a variável que mais influencia o volume expiratório. Em relação ao factor smoke, observamos que, para indivíduos da mesma idade e com a mesma altura, o volume expiratório dos fumadores é, em média, 0,15 litros mais baixo do que o dos não fumadores. Um intervalo a 95% de confiança para esta diferença entre os volumes médios, condicionada pela idade e pela altura, é aproximadamente

$$(0.150 - 2(0.057), 0.150 + 2(0.057)) = (-0.264, -0.036).$$



Acresce que o modelo na generalidade é significativo ($F=615.303$, $p=0.000$), com um coeficiente de determinação $R^2 = 0.791$ que é significativamente diferente de zero e um coeficiente de determinação ajustado de $\overline{R}^2 = 0.790$.

Nas instruções efectuadas em SPSS pedimos que fossem guardados os objectos: resíduos estandardizados, valores ajustados estandardizados, leverages e distâncias de Cook. Selecçãoámos ainda a identificação das observações com resíduos estandardizados superiores a 3.3, fazendo

Statistics → Residuals → Casewise Diagnostics

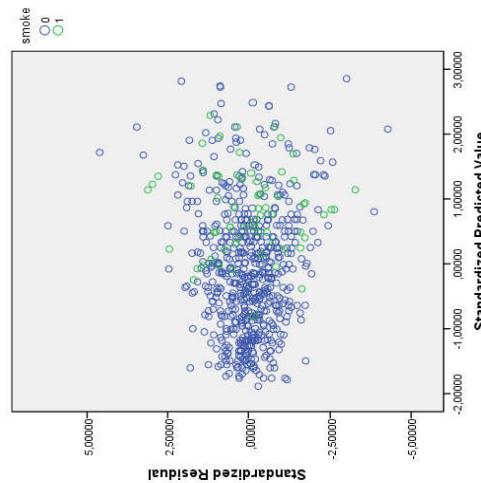
e incluindo o valor de referência de 3.3 desvios padrão. O output respeitante aos resíduos é apresentado na página seguinte.

As conclusões anteriores só são válidas uma vez satisfeitos os pressupostos do modelo. No que se segue, analisamos graficamente a satisfação desse pressuposto [para cada uma das classes de smoke](#).

Poderíamos começar por considerar o gráfico de dispersão dos resíduos estandardizados contra a ordem das observações, quer para detecção de linearidade e homocedasticidade, quer para a eventual detecção de outliers. Contudo, esse gráfico exige a criação de um variável com a indexação dos indivíduos, que não existe no ficheiro, e portanto não vai ser aqui considerado. De qualquer forma podemos observar o que se passa com o boxplot dos resíduos estandardizados para cada uma das categorias de smoke.

Existem 4 indivíduos com resíduos superiores a 3.3 sendo que dois deles têm resíduos grandes negativos e outros têm resíduos grandes positivos. A remoção destas observações do modelo depende de uma análise crítica dos valores de todas as variáveis recolhidas para essas observações.

O gráfico a seguir é o gráfico dos resíduos estandardizados (ZRESID) contra os valores ajustados estandardizados (ZPRED), por categoria de smoke.



[Home Page](#)

[Title Page](#)

[Contents](#)



[Page 99 of 104](#)

[Go Back](#)

[Full Screen](#)

[Close](#)

[Quit](#)

*Output1 [Document1] - PASW Statistics Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help

Descriptives Tests of Normality Standardized Residuals Title Histograms smoke Detrended Fit smoke Detrended Fit Boxplot Log GGraph Title Notes Active Dataset Graph Log Regression Title Notes Active Dataset Variables Entered/Removed Model Summary ANOVA Coefficients Casewise Diagnostics Residuals Statistics

Casewise Diagnostics^a

Case Number	Std. Residual	fey	Predicted Value	Residual
2	-3,856	1,7240	3,256009	-,1,5320099
473	-4,279	2,5380	4,238029	-,7000294
624	4,610	5,7930	3,981472	1,8315293
648	3,461	5,6380	4,262780	1,3752196

a. Dependent Variable: fev

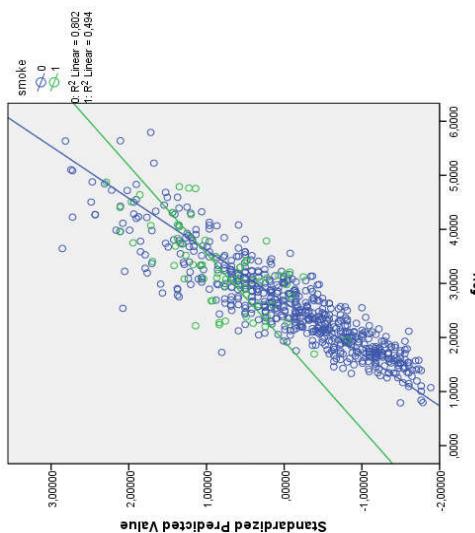
Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,181591	4,830557	2,636780	,7713086	654
Std. Predicted Value	-1,887	2,855	,000	1,000	654
Standard Error of Predicted Value	,021	,083	,032	,013	654
Adjusted Predicted Value	1,186635	4,877048	2,636866	,7712849	654
Residual	-1,7000294	1,8315283	,0000000	,3900738	654
Std. Residual	-4,279	4,610	,000	,997	654
Stud. Residual	-4,309	4,633	,000	1,002	654
Deleted Residual	-1,7236316	1,8500280	-,0000866	,4000384	654
Stud. Deleted Residual	-4,368	4,708	,000	1,005	654
Mahal. Distance	.750	27,739	3,994	4,652	654
Cook's Distance	,000	,060	,002	,006	654
Centered Leverage Value	,001	,042	,006	,007	654

a. Dependent Variable: fev

Parce existir uma tendência para um aumento da variância dos resíduos com um aumento dos valores ajustados, mas nos não fumadores do que nos fumadores, o que poderá violar a hipótese de homocedasticidade dos resíduos. Eventualmente poder-se-ia transformar a resposta e ver se isso resolvia esta questão.

O gráfico dos valores ajustados contra a resposta apresentam uma relação linear para cada uma das categorias de smoke e mais uma vez sugerem uma pequena violação do pressuposto de homocedasticidade.



Observa-se a existência de vários outliers para os resíduos no grupo dos não fumadores o que inevitavelmente acaba por comprometer a normalidade dos resíduos neste grupo mas notamos simultaneamente que o tamanho amostral é grande e a distribuição não é marcadamente assimétrica pelo que, na verdade, não parecem existir problemas com este aspecto. Quanto aos resíduos no grupo dos fumadores, não são identificadas violações de normalidade.

Os gráficos usuais de averiguação de normalidade são obtidos do comando

Analyze → Descriptive Stats → Explore

escolhendo os resíduos estandardizados com variável dependente e smoke como factor (página seguinte)



FACULDADE DE CIÉNCIAS
UNIVERSIDADE DO PORTO

[Home Page](#)

[Title Page](#)

[Contents](#)

◀◀ ▶▶

◀ ▶

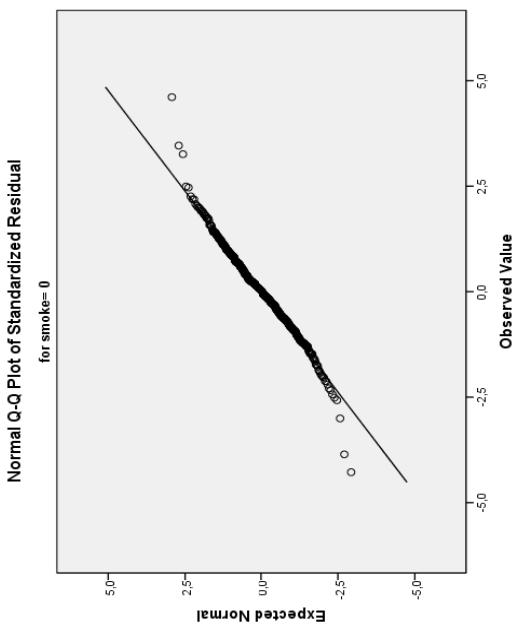
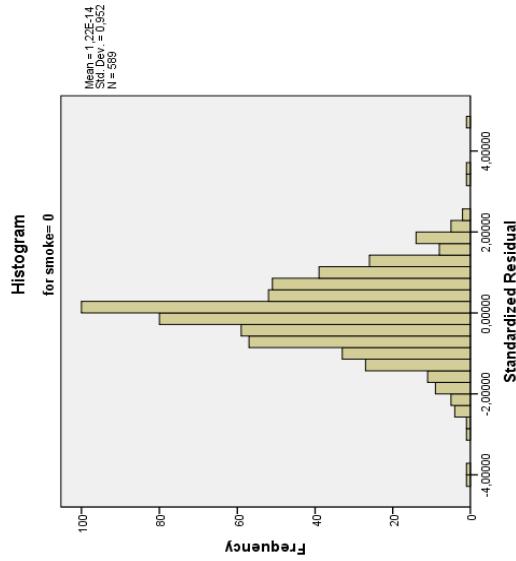
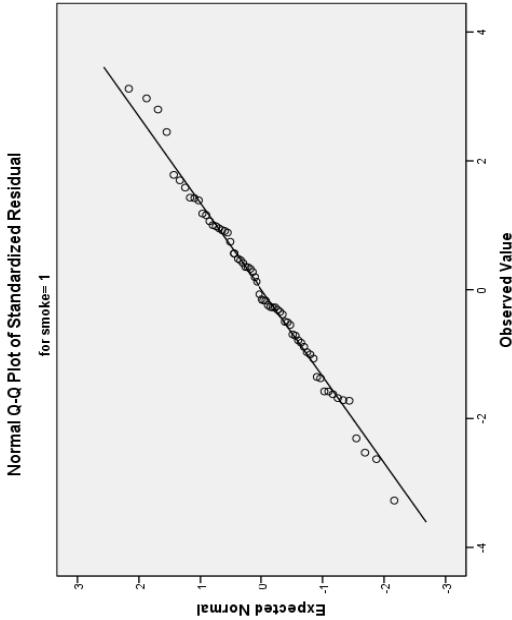
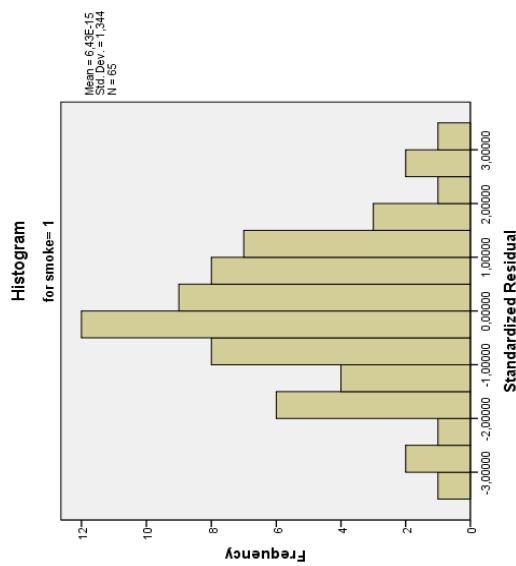
[Page 101 of 104](#)

[Go Back](#)

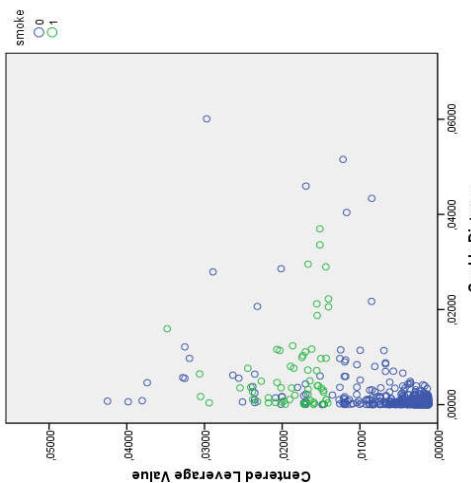
[Full Screen](#)

[Close](#)

[Quit](#)



Finalmente fazemos a detecção de pontos influentes considerando o gráfico das leverages contra as distâncias de Cook (dispondo de uma variável índice para os indivíduos, consideraríamos também os gráficos de dispersão de cada uma destas medidas contra o índice).



Outro

modelos a considerar poderia ser considerando o gráfico das leverages contra as distâncias de Cook (dispondo de uma variável índice para os indivíduos, consideraríamos também os gráficos de dispersão de cada uma destas medidas contra o índice).

Outro modelo a considerar poderia ser

$\text{fev} \sim \text{smoke} + \text{age} + \text{height} + \text{height}^2 + \text{sex}$

fazendo uso de todas as variáveis explicativas disponibilizadas.

Os resultados para o ajustamento do modelo aos dados são os seguintes

Modelo	coef	s.e. coef	t	valor-p
constante	6.895	1.499	4.60	0.000
age	0.069	0.009	7.63	0.000
sex	0.095	0.033	2.88	0.004
ht	-0.274	0.050	-5.52	0.000
ht.sq	0.003	0.000	7.65	0.000
smoke	-0.133	0.057	-2.33	0.020

$$F = 499.416, p = 0.000; R^2 = 0.794 \quad (p = 0.000),$$

$$\bar{R}^2 = 0.792$$

O modelo é muito semelhante ao anterior em termos de interpretação. Quando em presença também do sexo, o efeito do tabaco sobre o volume expiratório é essencialmente o mesmo que no caso anterior, quer em termos de magnitude quer em termos de variância. O mesmo se passa em relação à percentagem da variância total explicada pela regressão.

As análises gráficas desta equação de regressão são deixadas como exercício, sendo que os 4 outliers identificados na situação anterior mantêm-se também neste modelo.

Entre os dois modelos o anterior parece ser preferível, pela simplicidade da apresentação.

Não há portanto observações especiais a considerar.

Outros modelos com interacções de variáveis poderiam também ser considerados.

2.3. Exemplo: resultados eleitorais na Georgia (EUA) nas eleições presidenciais de 2000

- (a) Leia a biblioteca *faraway* no R. Se não a tiver instalada, terá de o fazer previamente. Esta biblioteca contém vários ficheiros de dados, a maior parte usados no livro *Extending the Linear Model with R*, de J.J. Faraway.

(b) Considere as instruções

```
> data(gavote)  
> help(gavote)
```

A primeira lê um ficheiro designado por *gavote*; a segunda fornece elementos descritivos das variáveis que constam no ficheiro.

A coluna *votes* representa a totalidade de votos úteis, e a coluna *ballots* representa a totalidade de boletins de voto preenchidos (dos quais nem todos resultaram num voto útil). Mais precisamente, um eleitor dirige-se à sua secção de voto, onde se identifica a sua legitimidade para a votação. Estando recenseado, é emitido um boletim de voto. Contudo, votos em branco e votos nulos não são considerados. Por vezes, o equipamento que lê e guarda os resultados da votação também tem falhas. A diferença

O objectivo deste exercício consiste da determinação dos factores que afectam a sub-contagem.

- (c) Efectue uma análise estatística descritiva dos dados, usando gráficos e medidas estatísticas numéricas adequados. Observe que o comando

```
> plot(density(gavote$votes), main="Votes")  
> rug(gavote$votes)
```

produz um gráfico que pode ser visto como uma suavização do histograma, e portanto como um bom complemento deste último. O "tapete" que aparece no fundo do gráfico indica a localização das várias observações recolhidas. Consegue identificar alguns problemas com os dados que possam vir a dificultar a análise posterior? Defina também a variável resposta.

- (d) Encontre um modelo para a sub-contagem que lhe pareça adequado ao problema, e estude a sua validade e qualidade do ajustamento. Interprete também os coeficientes de regressão obtidos no modelo anterior.

ballots - votes
é designada por *undercount* (digamos, sub-contagem).

