5.2.4. Exemplo: peso x comprimento de cães

A Tabela **1** mostra os comprimentos e pesos de 20 cães de uma amostra. Iremos ajustar um modelo de regressão linear a estes dados, usando o R.

Tabela 1. Comprimento e peso de cães

comprimento (cm)	peso (kg)	comprimento (cm)	peso (kg)
105.9	24.7	98.0	19.1
102.8	21.0	104.1	21.7
101.8	20.9	97.6	18.0
96.5	18.5	97.1	17.5
100.6	19.9	100.2	17.7
97.9	18.8	98.7	18.7
98.9	19.1	106.7	29.0
102.7	23.4	95.5	17.8
109.6	25.1	99.0	20.4
104.8	21.9	105.1	21.5

O diagrama de dispersão dos dados está mostrado na Fig. **1**. O diagrama mostra também a reta de regressão ajustada pelo R (a linha preta), e o intervalo que têm 0,95 de probabilidade de conter os valores do peso, dado o comprimento (linhas azuis). O modelo usado foi:

$$peso_i = \beta_0 + \beta_1 \times comprimento_i + e_i$$

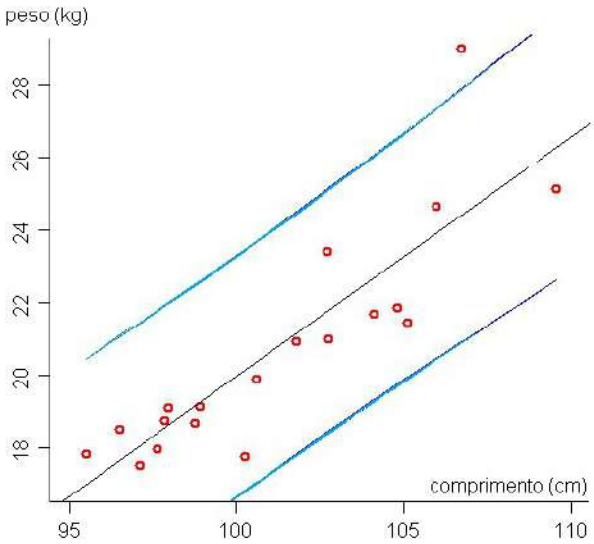


Figura 1. Comprimento x altura de 20 cães

Note que a reta de regressão encontrada parece se ajustar bem aos dados (embora exista um valor discrepante nesta amostra, fora do intervalo de previsão). Os resultados do ajuste deste modelo no R são mostrados no Quadro **1.** Os valores mais importantes (destacados em vermelho no quadro) são:

poder explicativo: $r^2 = 0{,}7488$

coeficiente de correlação: $r = \sqrt{r^2} = 0{,}86$

valor-p dos coeficientes: $\beta_0 \rightarrow p = 0{,}000$ $\beta_1 \rightarrow p = 0{,}000$

Quadro 1. Saída do R, para modelo de regressão

```
Residuals:
    Min      1Q  Median      3Q     Max
-2.3799 -0.8261 -0.3200  0.7926  4.6031

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.05853    9.12619  -5.047 8.39e-05 ***
cmp           0.66018    0.09014   7.324 8.41e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.53 on 18 degrees of freedom
Multiple R-squared:  0.7488,    Adjusted R-squared:  0.7348
F-statistic: 53.64 on 1 and 18 DF,  p-value: 8.414e-07
```

Estes valores indicam que existe uma correlação positiva muito forte entre as duas variáveis, e que o modelo de regressão parece bem ajustado a estes dados.

Uma forma de avaliar a qualidade de um modelo de regressão é fazendo a ANOVA de seus resultados (ver seção **5.2.3.2**). O Quadro **2** mostra a ANOVA do modelo de regressão do *peso* no *comprimento* de cães. Note que o resultado desta ANOVA já está mostrado, resumidamente, na última linha do Quadro **1**, que dá o valor da estatística F, o número de graus de liberdade, e o valor-p calculado.

Quadro 2 – ANOVA do modelo de regressão : comprimento × peso

```
            Df Sum Sq Mean Sq F value   Pr(>F)
cmp          1 125.51  125.51   53.64 8.41e-07 ***
Residuals   18  42.12    2.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Para uma análise mais completa, porém, será preciso ainda verificar se os pressupostos do modelo de regressão foram obedecidos, através da análise de seus resíduos.

Primeiro, estes resíduos devem ter média nula e distribuição normal. A média encontrada foi muito próxima de zero, igual -2.78×10^{-17}. A normalidade deve agora ser verificada, por meio de gráficos ou de testes estatísticos.

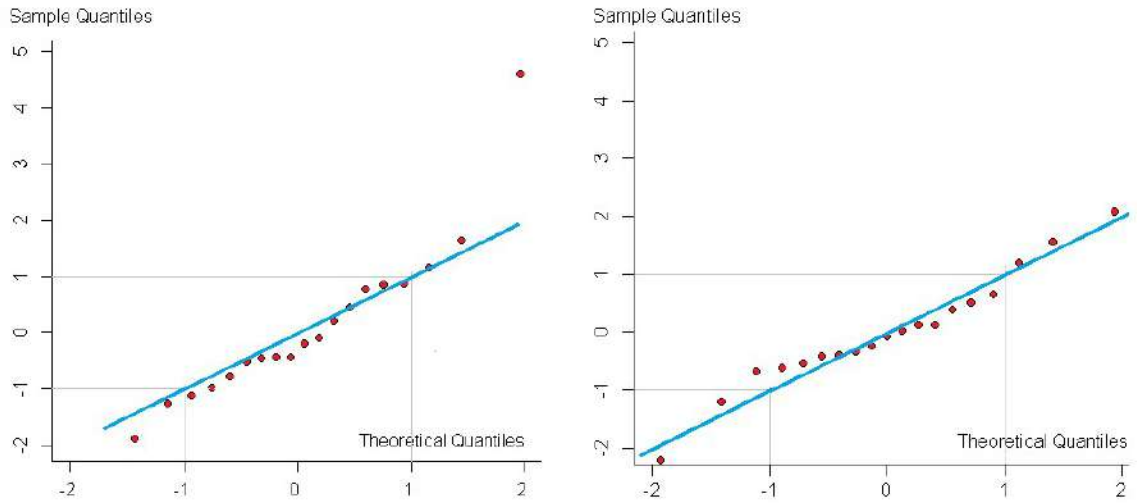
A Fig. **2A** mostra o gráfico de quantis (*qqplot*) dos resíduos, feito no R. Podemos notar que quase todos os pontos estão alinhados ao longo de uma reta (isto é, que distribuição dos resíduos se é aproxima da normal), mas há um ponto muito discrepante no extremo à direita do gráfico. O teste de Shapiro para a normalidade da distribuição dos resíduos dá o resultado no Quadro **3**.

Quadro 3. Teste da normalidade dos resíduos da regressão

```
shapiro.test(caes.lm$resid)
        Shapiro-Wilk normality test
data:  caes.lm$resid
W = 0.89941, p-value = 0.0402
```

O teste teve resultado significativo, o que indica que a distribuição dos resíduos *não* pode ser considerada normal. O valor-p deste resultado contudo é relativamente alto; a hi-

pótese da normalidade é rejeitada se considerarmos um α=0,05, mas não se considerarmos um α=0,01. Note que este resultado é coerente com o que observamos no gráfico de quantis (Fig. **2A**); provavelmente, esta rejeição foi causada pela presença de um valor discrepante na amostra.



A. Amostra original B. Amostra com o ponto discrepante removido

Figura 2. Gráfico dos quantis

Outro pressuposto importante é a *homocedasticidade* – isto é, a variância dos resíduos deve ser constante, para qualquer valor de X. O gráfico da Fig. **3** mostra os resíduos pelos valores de X. A dispersão dos resíduos não parece variar de acordo com o valor de X; no entanto, continua havendo um ponto discrepante, no extremo superior direito, que quebra o padrão do gráfico.

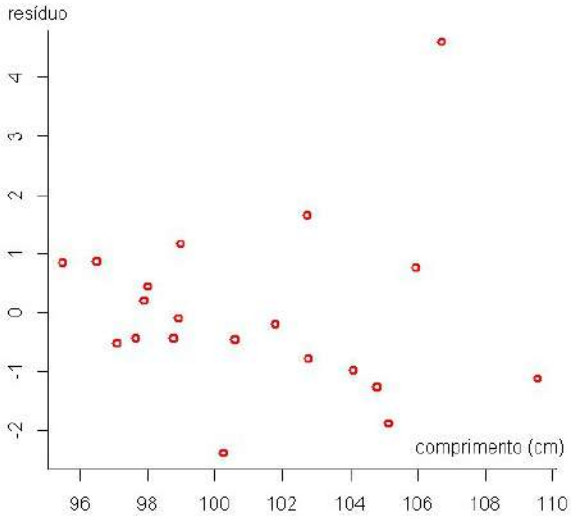


Figura 2. Gráfico resíduos × comprimento

Por fim, é também preciso que os resíduos sejam independentes dos valores previstos $\hat{y}$. O valor coeficiente de correlação entre estas variáveis, calculado pelo R, foi de $r = 1.5\times10^{-16}$, o que corrobora este pressuposto de independência.

Em resumo, alguns dos testes e gráficos e feitos acima fornecem evidências de que os pressupostos do modelo foram todos atendidos; outros, porém sugerem o contrário. Que fazer, num caso destes?

Neste exemplo, a quebra dos pressupostos ocorreu sempre por causa de uma observação – um animal cujo peso era muito maior do que o que seria esperado. Uma possibilidade é voltar aos dados, e procurar descobrir por quê aquele animal (cujos valores estão destacados em vermelho na Tabela **1**) tem peso tão discrepante. Se considerarmos que houve erro na amostragem (por exemplo, que aquele animal era de uma raça diferente dos outros, e portanto não deveria ter sido incluído no estudo), podemos talvez excluir aqueles valores da amostra, e refazer a análise com a nova amostra. Neste caso, o gráfico de quantis ficará como o da Fig. **2B**, e o teste de Shapiro terá o resultado não-significativo (Quadro **4**), não rejeitando a hipótese de normalidade dos resíduos:

Quadro 4. Teste da normalidade dos resíduos da regressão

```
shapiro.test(caes.lm$resid)
      Shapiro-Wilk normality test
data:  caes.lm$resid
W = 0.96654, p-value = 0.7058
```

Este tipo de manipulação dos dados, porém, deve ser feito com cuidado, pois é muito fácil confundi-lo com a falsificação deliberada dos resultados. É preciso que haja justificativas muito convincentes para qualquer alteração dos dados da amostra, e o que foi feito seja explicado claramente no relatório ou artigo que descreva a estudo; caso contrário, os autores poderão vir a ser acusados de *cherry-picking* - isto é, de manipular as amostras, mantendo os dados que apóiam a hipótese que se pretende testar e descartando os demais, o que é considerado uma forma de fraude (veja seção **2.2.4.3**).