5.2.3. Avaliação da qualidade do modelo.

Depois que um modelo de regressão foi ajustado aos dados da amostra (i.e., depois que definimos os coeficientes β_0 e β_1 que fazem a reta melhor se ajustar aos dados), precisamos avaliar a *qualidade* deste modelo, e fazer alguns testes para verificar se os resultados obtidos são *significativos*, no sentido estatístico – isto é, se realmente refletem o que ocorre na população, e não são apenas algo que ocorreu por acaso na amostra. Os testes podem ser *testes t*, feitos sobre os coeficientes, ou uma *ANOVA*. A qualidade do modelo pode ser avaliada também a partir de seu *poder explicativo*. Por fim, temos também que verificar se os pressupostos sobre os quais o modelo foi construído são válidos. Veremos a seguir como são feitos estes testes e verificações.

5.2.3.1. Testes dos coeficientes.

Os coeficientes do modelo foram encontrados com base na informação contida na amostra; não há garantia de que o modelo também serviria para o resto da população. Para verificar se é realmente possível fazermos inferências a partir do modelo, precisamos aplicar alguns testes a seus coeficientes.

Num modelo de regressão em (1), o teste mais importante é o da declividade β_1.

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (1)$$

A hipótese nula é a de que

$$H_0 \rightarrow \beta_1 = 0$$

Se H_0 for verdadeira, o modelo será inútil, já que consistirá apenas numa reta horizontal (neste caso, o valor médio de Y não dependeria do valor de X). Isto acontece no gráfico da Fig. 2, que mostra os erros de previsão de carga elétrica Y como função da temperatura X, obtidos por um modelo de previsão. O gráfico mostra que o valor médio do erro é sempre constante, igual a zero, qualquer que seja o valor da temperatura, o que indica que um modelo de regressão ligando estas duas variáveis não teria nenhuma utilidade.

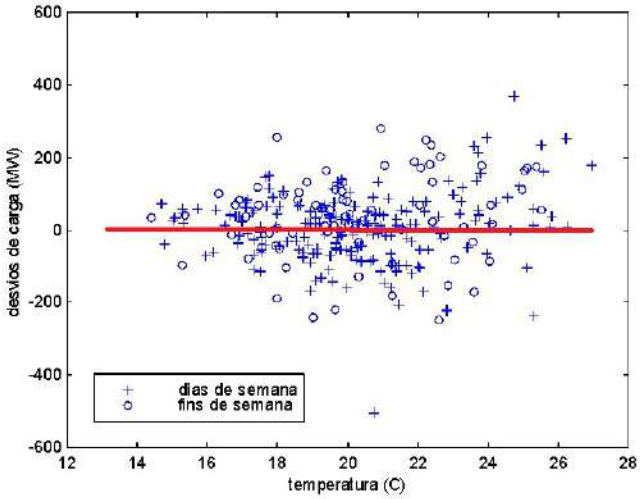


Fig. 6. Reta de regressão, se $\beta_1 = 0$

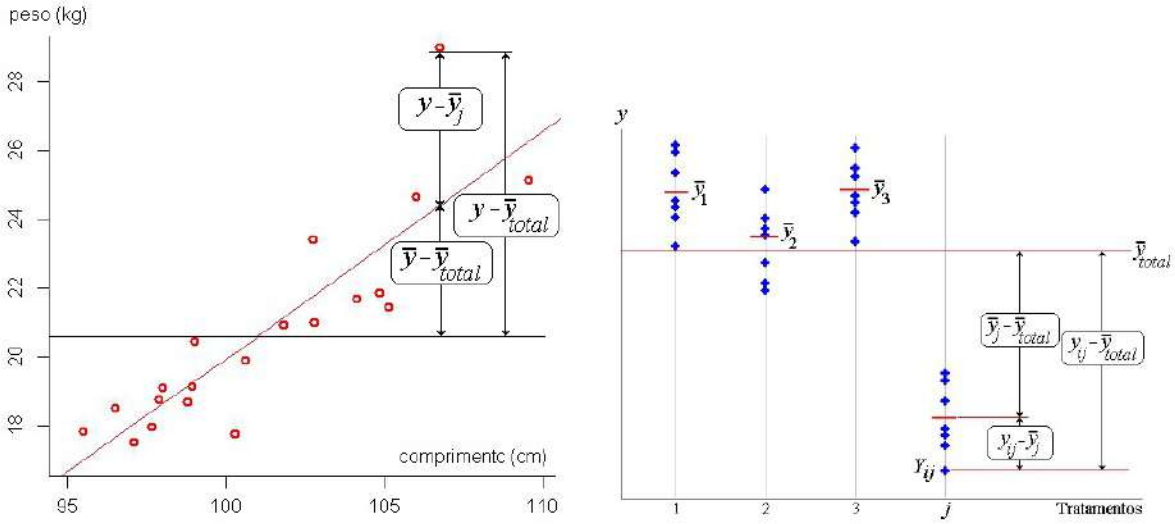
No exemplo sobre peso × comprimento de cães (Fig. **5**), encontramos $t = 5{,}13$ o que equivale a um valor-$p = 0{,}000$, e leva à rejeição da hipótese nula. A reta portanto *não* é horizontal (sua declividade é significativa), e o modelo de regressão faz sentido.

5.2.3.2. Análise de Variância do modelo de regressão

A *ANOVA* (seção **5.1**) também pode ser aplicada um modelo de regressão. A Fig. **7A** ilustra a lógica desta análise, usando os dados sobre comprimento e peso de cães: o desvio entre um ponto qualquer y e a média total $\bar{y}_{total}$ de todos os dados pode ser decomposto em duas parcelas:

$$(y - \bar{y}_{total}) = (y - \bar{y}) + (\bar{y} - \bar{y}_{total})$$

As somas dos quadrados destas duas parcelas servem de base para medidas da variação entre os pontos e a reta ($y - \bar{y}$), e entre a reta e a horizontal ($\bar{y} - \bar{y}_{total}$). Na Fig. **7B**, é mostrada a decomposição feita anteriormente (seção **5.1**), numa ANOVA comparando as médias de quatro tratamentos. Há duas diferenças entre este tipo de ANOVA, usado para comparar as médias de vários tratamentos, e o que usaremos agora para analisar um modelo de regressão. Primeiro, a variável X no eixo horizontal não será mais um conjunto de algumas variáveis qualitativas (os quatro 'tratamentos' da **7B**), e sim uma variável que pode ser contínua e ter infinitos valores. Segundo, a média total $\bar{y}_{total}$ não será mais uma constante, e sim uma função linear da variável X. A hipótese nula é que a média de Y é constante para todo valor de X (i.e., que Y independe de X, e a reta de regressão é horizontal).



A. ANOVA aplicada a um modelo de regressão B. ANOVA aplicada a um conjunto de tratamentos

Figura 7. Decomposição do desvio em relação à média

No R, os resultados desta ANOVA são mostrados como no Quadro **1**, que é praticamente idêntico aos quadros já vistos na seção **5.1**; as únicas diferenças estão na ordenação das colunas. No Quadro **1**, o nome que o R usa para cada coluna está em negrito, na prime-

ira linha; na segunda linha estão os nomes e as siglas equivalentes em português, como usadas na seção **5.1**.

Quadro 1 – ANOVA do modelo de regressão, no R

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	graus de liberdade	SQ	MQ	F	valor-p
regressão	1	SQ_{regr}	$MQ_{regr}=SQ_{regr}/1$	MQ_{regr}/MQ_{resid}	
resíduo	n-2	SQ_{resid}	$MQ_{resid}=SQ_{resid}/(n-2)$		
total	n-1	SQ_{total}	$MQ_{total}=SQ_{total}/(n-1)$		

O Quadro **2** mostra como exemplo os resultados da ANOVA aplicada ao modelo calculado pelo R para os dados do problema *comprimento* × *peso* de cães (Figs. **5** e **7B**).

Quadro 2 – ANOVA no R : comprimento × peso de cães

```
            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 ' ' 1
```

É importante lembrar que a ANOVA exige que os desvios entre os valores observados da variável *Y* e os valores previstos tenham distribuição normal, e variância constante, para qualquer valor de *X*. Estas verificações serão vistas a seguir (seção **5.3.3.4**).

5.2.3.3. Poder explicativo do modelo linear

Outra maneira de avaliar a qualidade de um modelo é calcular o seu *poder explicativo*. Este coeficiente dá a porcentagem da variação de *Y* que pode ser "explicada" pelo modelo. Porque o animal de peso *y* na Fig. **7A** tem peso acima da média $\bar{y}$? Em parte, porque também tem comprimento acima da média. Ou seja, o fato de ele ter um grande peso pode em parte ser explicado pelo fato de ele ter também uma grande comprimento. Esta explicação, contudo, é parcial; o comprimento não explica totalmente o peso (há animais menores que têm muito peso, e animais maiores que têm pouco peso). A razão entre a parte da variação que pode ser "explicada" e a variação total, expressa em porcentagem, é chamada de *poder explicativo* do modelo.

O poder explicativo, representado pelo coeficiente r^2, é calculado pela razão entre as duas somas de quadrados vistas acima:

$$r^2 = \frac{SQ_{\text{regr}}}{SQ_{\text{total}}} = \frac{SQ_{\text{explicada}}}{SQ_{\text{total}}}$$

Note que este coeficiente nada mais é do que o quadrado do coeficiente *r* de correlação linear de Pearson. No modelo mostrado na Fig. **7A**, $r^2 = 0{,}75$. Isto quer dizer que o modelo pode explica 75 % da variação do *peso* dos cães da amostra, usando o *comprimento* como variável explicativa. No entanto, ainda restam 25% de variação que o modelo não consegue explicar. Esta é a chamada *variação não-explicada*. Quanto menor esta variação

não-explicada, melhor será o modelo; se um modelo é *determinístico*, a variação não-explicada será nula, porque o modelo conseguirá explicar toda a variação de *Y* (sabendo-se *X*, será possível determinar exatamente qual será o valor de *Y*).

5.2.3.4. Verificação : os pressupostos do modelo foram atendidos?

O modelos de regressão linear são construídos com base em diversos pressupostos. Alguns destes pressupostos podem ser verificados antes de calcularmos o modelo:

(i) As observações de cada uma das variáveis são independentes entre si.
Nos estudos experimentais, como o estudo da relação entre o peso e o comprimento de cães mencionado acima, as observações de um variável são em geral independentes entre si: o peso e o comprimento de cada cão são independentes do peso e do comprimento de qualquer outro cão. Não é porém possível verificar esta independência a partir dos dados; é preciso saber *como* os dados foram obtidos (como o experimento foi planejado).
Um tipo comum de variável no qual as observações são dependentes entre si são as *séries temporais*, sequências de observações de uma variável feitas ao longo do tempo. Um exemplo destas séries é a temperatura do ar: a temperatura registrada a cada hora do dia depende da temperatura registrada nas horas anteriores. O modelo de regressão visto acima *não* pode ser usado em séries temporais.
(ii) Existe de fato uma relação linear entre as variáveis X e Y.
Isto é fácil de se verificar: basta fazer um diagrama de dispersão entre estas variáveis, e calcular o coeficiente de correlação linear entre elas. Se o diagrama de dispersão não parece mostrar nenhuma relação linear entre as variáveis, e a correlação entre elas é nula, não faz sentido tentar relacioná-las por um modelo de regressão.

Outros pressupostos se referem à distribuição dos *resíduos* do modelo (os desvios entre os valores observados da variável dependente *Y* e os valores previstos pelo modelo):

(i) Os resíduos têm média nula.
(ii) Os resíduos têm distribuição normal.
Isto pode ser verificado a partir de gráficos como o diagrama de quantis (no R, `qqnorm`), ou por meio de testes, como o de Shapiro ou o de Kolmogorov-Smirnov. (ver seção **4.7.1.2.i**).
(iii) A variância dos resíduos é constante, para qualquer valor de *X*.
Esta propriedade é chamada de *homoscedasticidade*. Para verificar se ela existe ou não num modelo, é melhor fazer um diagrama de dispersão dos resíduos pelos valores de X. Se este diagrama indicar que a dispersão dos resíduos é maior para alguns valores de X do que para outros valores, o modelo de regressão não serve.
(iv) Os resíduos são independentes dos valores de *X*.
Isto pode ser verificado se fizermos o diagrama de dispersão entre os resíduos e *X*, e calculamos o coeficiente de correlação linear entre estas variáveis.

No exemplo visto na seção a seguir (**5.2.4**), ilustraremos o uso de algumas destas técnicas.