Notas sobre Regressão Linear com apoio do R

Alessandro de Castro Corrêa*

2 de maio de 2020

1 Apresentação e procedimentos

A regressão é utilizada para explicar e modelar a relação entre uma variável Y, denominada resposta ou variável dependente e uma ou mais variáveis independentes, explicativas ou fatores, $X_1, \ldots, X_k$ (FARAWAY, 2002). Quando só há uma variável explicativa, $k = 1$, chama-se REGRESSÃO SIMPLES, mas quando há mais de uma variável explicativa, $k > 1$, chama-se REGRESSÃO MÚLTIPLA.

$$y = f(X_1, X_2, \ldots, X_k) \tag{1}$$

A variável dependente deve ser contínua, mas a(s) variável(is) independentes de podem ser contínuas, discretas ou categóricas.

A regressão podem ser utilizadas para três finalidades básicas:

a) Previsão: estimar valores da variável dependente com base nos valores das variáveis independentes;

b) Analisar os efeitos de variações nas variáveis independentes ou fatores sobre a variável dependente ou resposta;

c) Descrição geral da estrutura de dados.

O valor esperado de cada y para cada valor de x pode ser descrito pelo seguinte modelo:

$$E(y|x) = \beta_0 + \beta_i X \tag{2}$$

*Grupo de Estudos Avançados em Gestão (GEAG/IFPA-Campus Belém), Programa de Pós-Graduação em Engenharia de Materiais (PPGEMat/IFPA), Programa de Pós-Graduação em Engenharia Industrial (PPGEI/UFPA), alessandro.correa4@gmail.com

em que o intercepto β_0 e ao coeficiente de inclinação β_i são constantes desconhecidas. Os valores de y podem ser descrito pelo modelo

$$y = \beta_0 + \beta_i X + \epsilon \tag{3}$$

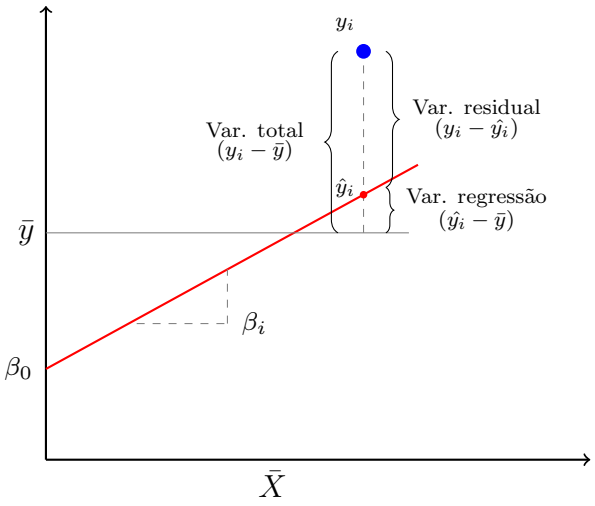
em que ϵ representa os resíduos do modelo e que são variáveis aleatórias com média zero e variância σ^2.

2 Interpretando o modelo

Para entender os principais indicadores da regressão é preciso compreender como um modelo deve ser avaliado. O modelo procura explicar o comportamento de uma variável. Caso não haja informações adicionais, a melhor estimativa a disposição de seu valor esperado é pode ser seu valor médio.

$$E(y_i) = \bar{y} \tag{4}$$

Figura 1: Decomposição da variância na regressão linear.



Como se pode observar, na Figura 1. O valor médio não passa exatamente sobre o valor observado y_i. A diferença entre a observação e o valor esperado $(y_i - \bar{y})$ é a variação total que reflete, no momento, a parcela não explicada do comportamento da variável e que servirá de referência para futuros modelos, que somente valerão a pena se forem capazes de se aproximarem mais dos valores observados.

$$E(y_i|x) = \beta_0 + \beta_i x \tag{5}$$

Agora introduzimos uma variável x que supomos que possa ajudar a explicar o comportamento de y, segundo um modelo de regressão linear. Os valores esperados ou ajustados da variável $\hat{y}$ serão o valor de uma constante mais o produto do coeficiente angular com x. Como se trata de uma estimativa, sabe-se que há um erro e_i associado a esses valores esperados, logo o modelo estimado terá a seguinte forma:

$$\hat{y}_i = b_0 + b_1 x + e_i \tag{6}$$

A linha vermelha representa os valores estimados pelo modelo, b_0, denominado intercepto, representa o valor estimado de y quando $x = 0$ e a b_1 indica a inclinação da reta. Observa-se que o valor estimado da variável resposta $\hat{y}_i$, indicado pelo ponto vermelho, não coincide com o valor observado na realidade, todavia, o modelo se aproximou do valor observado, agregando mais informação e melhorando a capacidade explicativa do modelo.

A diferença entre o valor estimado e o valor médio $(\hat{y}_i - \bar{y})$ indica a parte da variação total explicada pela regressão, indicado o poder explicativo do modelo.

A diferença entre o valor observado e o estimado $(y_i - \hat{y}_i)$ é a variação residual que ficou sem explicação, também denominados resíduo ou erro. Quanto menores forem os resíduos melhor a capacidade preditiva do modelo.

Note que com a introdução de x, os resíduos passaram a ser a parcela não explicada do comportamento de y_i, substituindo a variação total do modelo que continha somente a média, pois parte da variação total passou a ser explicada pelo modelo de regressão. Esse será o foco central das avaliações dos modelos a capacidade explicativa ou, o outro lado da moeda, os resíduos que reflete as variações que o modelo não conseguiu captar.

Como se verá adiante, as medidas de variação são expressas em desvios elevados ao quadrado, uma vez que as somatórias dessas diferenças resultarão em zero, pois se tratam afastamentos em relação a médias.

Tabela 1: Fontes de variação.

Fonte	Total	Explicada	Não explicada (Erro)
Português	SQT	SQR	SQE (Resíduos)
Inglês	SST	SSR	SSE (Residuals)
Matemática	$\sum(y_i - \bar{y})^2$	$\sum(\hat{y}_i - \bar{y})^2$	$\sum(y_i - \hat{y}_i)^2$

A Tabela 1 exibe as denominações de fontes de erro bem como suas expressões em forma de quadrados dos desvios. A variação total é a Soma

dos Quadrados Totais (SQT), ou *Sum of Squares Total - SST*. A variação da regressão é a Soma dos Quadrados da Regressão (SSR), *Sum of Squares Regression (SSR)*. A variação residual é a Soma dos Quadrados do Erros (SQE), ou *Sum of Squares Regression (SSE)*

Observemos os dados de um experimento, apresentado por Neto, Scarminio e Bruns (2010), no qual foram coletadas cinco observações (y_i) do rendimento percentual de uma reação.

Tabela 2: Variação total com base na média de y

i	y_i	$\bar{y}$	$y_i - \bar{x}$	$(y_i - \bar{x})^2$
1	60	76.8	-16.80	282.24
2	70	76.8	-6.80	46.24
3	77	76.8	0.20	0.04
4	86	76.8	9.20	84.64
5	91	76.8	14.20	201.64
Σ			0.00	614.80

A média das observações é 76,8%, sendo esse o valor esperado de uma reação estimada $\hat{y} = \bar{y}$, uma vez que não há nenhuma outra variável que possa explicar o comportamento da reação. Os desvios das observações em relação à média estão na quarta coluna e a sua soma é zero uma vez que os valores positivos anulam os negativos, provocando a perda de informação útil sobre a precisão do modelo. A soma dos quadrados resolve esse problema pois elimina os sinais negativos. Agora a precisão do modelo baseado somente na média amostral é medido pela SQT que é 614,80, esse valor corresponde na totalidade, nesse momento, à variação não explicada pela média simples e qualquer modelo alternativo deve exibir diminuição na variação não explicada. Como será demonstrado adiante, a parcela não explicada serão os resíduos.

A decomposição do variação total de um modelo linear é apresentado e testada na tabela de Análise de Variância (Anova), que serão exibidas nas estimações no decorrer do texto.

3 Estimação dos parâmetros

A regressão linear simples utiliza somente um fator como variável independente para explicar o comportamento de uma variável resposta ou dependente.

$$\hat{y} = b_0 + b_1 x_1 + e_i \tag{7}$$

na qual y representa a variável dependente, ou resposta, x a variável independente ou fator, b_0 e b_1 são o coeficiente linear e angular respectivamente, e e são os resíduos associados à observação i. As estimativas dos coeficientes do modelos são calculados conforme as equações a seguir:

$$b_0 = \bar{y} - b_1\bar{x} \tag{8}$$

$$b_1 = \frac{\sum(x_i - \bar{y})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)} \tag{9}$$

onde $\bar{y}$ e $\bar{x}$ são médias das variáveis y e x, y_i e x_i representam cada observação das variáveis e b_0 e b_1 são os estimadores dos parâmetros β_0 e β_1.

4 Inferências

As principais inferências sobre um modelo de regressão são se o modelo é capaz de explicar a variável dependente, se é capaz de fazer previsões e quais as variáveis ajudam nessa explicação. A seguir são explicadas brevemente as medidas que permitem essas avaliações.

A estatística F Anova (*F-statistic*) mede o efeito conjunto das variáveis independentes (fatores) sobre a variável dependente (resposta). Testa a hipótese nula de que nenhuma das variáveis independentes é capaz de explicar a variável dependente. Para que a regressão seja significativa, a hipótese nula deve ser rejeitada, ou seja (Valor-P $< \alpha = 0,05$).

O Coeficiente de determinação (*Multiple R-squared*) mede a proporção de variação explicada pela regressão no caso de uma regressão simples, isto é, com uma única variável explicativa, sendo calculado pela equação:

$$R^2 = \frac{SQR}{SQT} \tag{10}$$

onde a soma dos quadrados da regressão (SQR) é uma proporção da soma dos quadrados total (SQT), seu valor varia de 0 a 1 e quanto maior esse valor, maior poder explicativo do modelo.

Uma hipótese que pode ser testada é se $H_0 : R^2 = 0$ contra $H_0 : R^2 > 0$. Essa hipótese é testada pela significância da estatística F.

No caso de haver mais de uma variável independente, a medida a ser avaliada é o Coeficiente de determinação ajustado (*Adjusted R-squared*) que penaliza o acréscimo de novas variáveis independentes ao modelo, uma vez que modelos parcimoniosos são preferíveis por sua simplicidade, logo, o acréscimo de uma variável e da complexidade devem ser justificados por um aumento razoável no poder explicativo.

O Erro Padrão da Estimativa, ou *Residual Standard Error*, também denominado Raiz do Erro Médio Quadrático, ou *Root Mean Squared Error (RMSE)*, mede o grau de dispersão em torno da linha de regressão, sendo calculado como a raiz da variância dos resíduos, ou como o desvio padrão dos desvios centrado nos valores previstos.

$$RMSE = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k}} \tag{11}$$

onde RMSE é expressa na mesma unidade de y, $\sum(y_i - \hat{y}_i)^2$ é SSE e n é o número de observações e k, o número de parâmetros a serem estimado, o demoninador indica o número de graus de liberdade.

O RMSE mede como os valores previstos pelo modelo se aproximam dos valores reais, ou seja, o grau de ajuste, podendo auxiliar na comparação de precisão, já que quanto menor for o RMSE, maior a precisão do modelo, pois a dispersão resultante é menor.

5 Aplicação do Modelo de Regressão

5.1 Modelo 1 - Regressão Simples

Retomando o exemplo de Neto, Scarminio e Bruns (2010), no qual um engenheiro químico interessado no efeito da temperatura sobre o rendimento percentual de um processo coletou os seguintes dados:

Tabela 3: Dados de temperatura em rendimento percentual de uma reação.

Temperatura (^{o}C)	40	45	50	55	60
Rendimento (%)	60	70	77	86	91

Importante recordar que um tentativa de explicação do comportamento do rendimento baseado somente na média amostral foi exibido na Tabela 2, o qual resultou numa SQT de 614,80 que correspondia ao mesmo valor dos resíduos.

5.1.1 Avaliando e interpretando o Modelo 1

Uma regressão linear simples utilizando a Temperatura como variável independente foi rodada para avaliar o efeito desse fator sobre o Rendimento percentual da reação. Esse modelo regressão, denominado de Model1 é descrito pela equação abaixo:

Tabela 4: Output da Anova no R

```
Analysis of Variance Table

Response: Rend
          Df  Sum Sq  Mean Sq F value    Pr(>F)
Temp       1  608.4   608.40  285.19 0.0004522
Residuals  3    6.4     2.13
```

$$\hat{Rend} = b_0 + b_1 Temp + e_i \tag{12}$$

Os *outputs* da Análise de Variância e as estimativas da regressão do Model1 são exibidos nas Tabelas 4 e 5.

A Anova revela agora a variação explicada pela regressão do Model1 (SQR=608,4) corresponde a quase a totalidade da variação total (SQT=614,8[1]), deixando pouquíssima variação não explicada como resíduos (SQE=6.4)

A primeira inferência se refere à significância do modelo como todo pela estatística F presente nas duas tabelas. O valor-P (0.00045) da estatística F com 1 e 3 graus de liberdade ($F = 285.19$) indica que é significante. A maneira usual de reportar esse resultado é ($F_{1,3} = 285, 19, p < .01$).

Tabela 5: Resultado da regressão no R

```
Call:
lm(formula = Rend ~ Temp)

Coefficients:
            Estimate    Std. Error t value Pr(>|t|)
(Intercept) -1.20000     4.66476  -0.257 0.813623
Temp         1.56000     0.09238  16.887 0.000452
---
Residual standard error: 1.461 on 3 degrees of freedom
Multiple R-squared:0.9896, Adjusted R-squared:  0.9861
F-statistic: 285.2 on 1 and 3 DF,  p-value: 0.0004522
```

A segunda inferência complementa a primeira. Trata-se do poder explicativo medido pelo Coeficiente de Determinação, apresentado pelo output da regressão (Multiple R-squared=0.9896), indicando que o modelo é capaz de explicar 98,61% das variações no rendimento percentual da reação. Esse

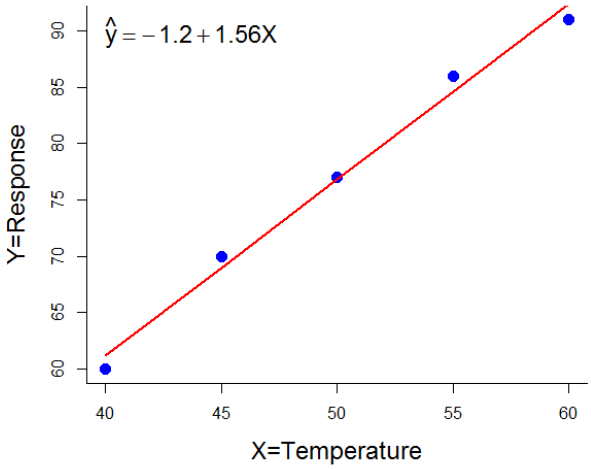
[1]Basta somar os valores dos quadrados na tabela Anova para se chegar a SQT: 608.4+6.4=614,8

indicador pode ser reportado juntamente com o teste F, uma vez que este corresponde ao seu teste de significância ($R^2 = .99, F_{1,3} = 285.2, p < .01$).

O Erro Padrão da Estimativa, ou *Residual Standard Error* do modelo é 1,461% é baixo, todavia será mais útil quando for feita uma comparação de precisão com outro modelo.

As análise gerais indicam que o modelo é muito bem ajustado o que pode ser visualmente constatado pela Figura 2, onde os valores reais estão praticamente sobre a reta de regressão.

Figura 2: Regressão estimada pelo Modelo 1.



Tendo em conta que o modelo é bem ajustado, as inferências se deslocam para as estimativas dos coeficientes.

O coeficiente angular estimado associado à temperatura foi 1,56 e o no seu teste de hipóteses, $H_0 : b_1 = 0$ é rejeitado, porque o efeito da temperatura é altamente significativo, P-value$<$ 0.01, e o valor de t-value$>$ $t_{0.05/2,3} = 3.1824$. Esse resultado deve ser reportado da seguinte forma: ($b_1 = 1.56, p < .001$). Sua interpretação é que, para cada unidade de graus Celsius adicional na temperatura, é esperado um acréscimo de médio de 1.56% no rendimento percentual da reação (y).

Já no caso do teste de hipótese do Intercepto, $H_0 : b_0 = 0$ é aceito, porque P-value$>$ 0.05 e t-value$<$ $t_{0.05/2,3} = 3.1824$, esse resultado deve ser reportado da seguinte maneira ($b_0 = -1.2,$ n.s).

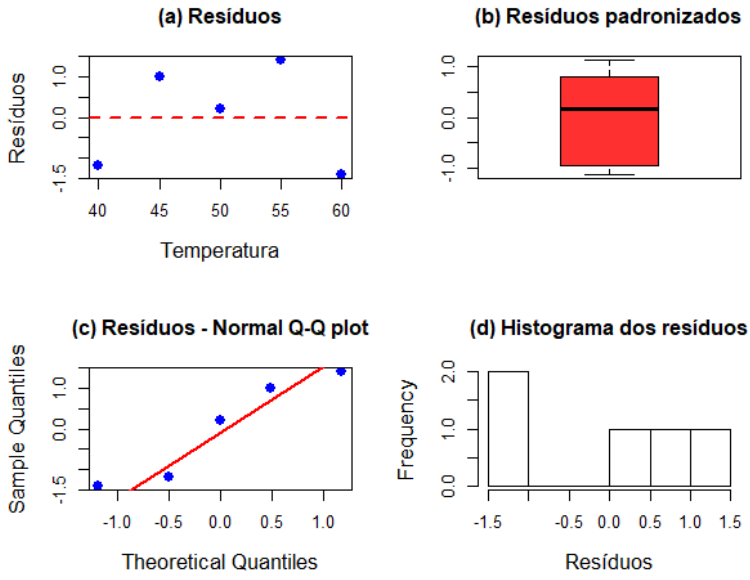
5.1.2 Análise dos resíduos do Modelo 1

Muito embora, estejamos realizados análise dos resíduos no final, o fazemos apenas para que nos concentremos na análise do modelo de regressão,

na prática, o mais recomendado é que se análise o resíduos antes de qualquer conclusão, pois as estimativas e inferências são dependentes dos resíduos se comportarem conforme supõe o método dos mínimos quadrados ordinários.

Os resíduos devem ter média zero e variância constate e não apresentarem sinais de autocorrelação, isso indica que não restaram padrões não captados pelo modelo e que o mesmo é corretamente especificado, não omitindo variáveis explicativas.

Figura 3: Resíduos do Modelo 1



A Figura 3 exibe quatro gráficos dos resíduos do Modelo 1. No Gráfico (a), os resíduos são plotados contra a variável independente, a única do Modelo 1, a Temperatura, no intuído de examinar a presença de padrões. Como são apenas 5 observações, é difícil visualizar padrões sutis. Não há padrões aparentes [2]. A dispersão parece homogênea (Variância homogênea) e ná indícios de tendências ou autocorrelação.

Na Figura 3(b), o *boxplot* dos resíduos padronizados indicam valor central bem próximo a zero, e uma distribuição relativamente simétrica.

O Gráfico Normal *Q-Q plot*, na Figura 3(c), mostra que os resíduos não se afastam muito da reta de normalidade. A avaliação da distribuição pelo histograma, Figura 3(d), ficou prejudicada pela pequena quantidade de observações.

A análise dos resíduos do modelo 1, sugerem que as suposições do modelo

[2]Apesar de minha filha de 4 anos, Maria, insistir de que há uma letra "M"ali.

de regressão estão satisfeitas e que a sua especificação, as suas estimativas e inferências são válidas.

5.2 Modelo 2 - Regressão Simples: faixa maior de dados

Nesta sessão, serão demonstradas como identificar um erro de especificação do modelo e como a generalização de um modelo estimado numa faixa de valores para valores fora da faixa deve ser feito com muito cautela.

No exemplo de Neto, Scarminio e Bruns (2010), o pesquisador realiza nova experimentação na expende-se a faixa avaliada, adicionando-se duas observações com temperaturas abaixo da faixa anterior e duas observações acima, dados exibidos na Tabela 6.

Tabela 6: Dados de temperatura em rendimento percentual ampliados

Temperature (^{o}C)	30	35	40	45	50	55	60	65	70
Yield (%)	24	40	60	70	77	86	91	86	84

Utilizou-se a mesma Equação 12 para se estimar os parâmetros desta segunda estimativa, que foi denominada Modelo 2, ou seja, os modelos 1 e dois possuem a mesma especificação, distinguindo-se apenas na faixa de dados utilizados para a estimação dos parâmetros.

$$\hat{Rend2} = b_0 + b_1 Temp2 + e_i$$

As estatísticas da estimativa do Modelo 2 estão na Tabela 7. A estatística F com 1 e 7 graus de liberdade é significativo e o Coeficiente de Determinação ajustado indica que o modelo é capaz de explicar 81% da variação do Rendimento percentual ($\bar{R}^2 = .81, F_{1,7} = 29.1, p < .01$). O coeficiente angular é significativo ($b_1 = 1.52, p < .001$) e o intercepto não difere de zero ($b_0 = -7.33,$ n.s), similares aos valores do Modelo 1.

Todavia, erro padrão da estimativa (RMSE) do Modelo 2 é 10,9, maior do que o Modelo 1, como pode ser comparado na Tabela 8, indicando que o segundo modelo é menos preciso. De fato, o valor do $\bar{R}^2$ do Modelo 2 também exibe um podem explicativo menor, apesar ser significativo. Por que teria ocorrido a perda de poder explicativo com a ampliação da faixa de dados?

5.2.1 Análise dos resíduos do Modelo 2

A análise dos resíduos do Modelo 2 na Figura 4 revela a presença de padrões que violam os pressupostos do modelo de regressão linear. O gráfico

Tabela 7: Output de regressão do Modelo 2.

```
Call:
lm(formula = Rend2 ~ Temp2)

Residuals:
Min      1Q  Median      3Q     Max
-15.067  -5.867   6.533   8.333   9.733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.3333    14.5397  -0.504  0.62949
Temp2         1.5200     0.2816   5.398  0.00101
---
Residual standard error: 10.9 on 7 degrees of freedom
Multiple R-squared: 0.8063, Adjusted R-squared: 0.7787
F-statistic: 29.14 on 1 and 7 DF,  p-value: 0.00101
```

Tabela 8: Comparação entre as estatísticas dos modelos 1 e 2.

	Model 1	Model 2
Intercepto	-1.20	-7.33
	(4.66)	(14.54)
Temperatura	1.56^{***}	1.52^{***}
	(0.09)	(0.28)
R^2	0.99	0.81
Adj. R^2	0.99	0.78
Num. obs.	5	9
RMSE	1.46	10.90

Notas: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Erros padrão entre parênteses.

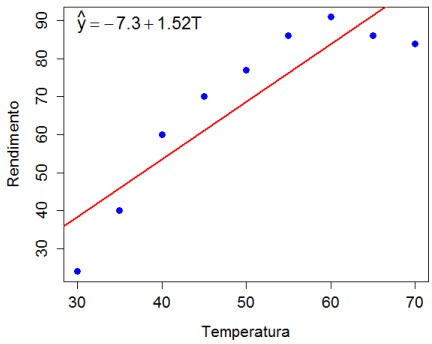
4(a) mostra que as observações de rendimento da reação não está aleatoriamente disposta em torno da reta e o gráfico 4(b) exibe o comportamento não linear nos resíduos.

5.3 Modelo 3 - Regressão Múltipla com elemento não linear

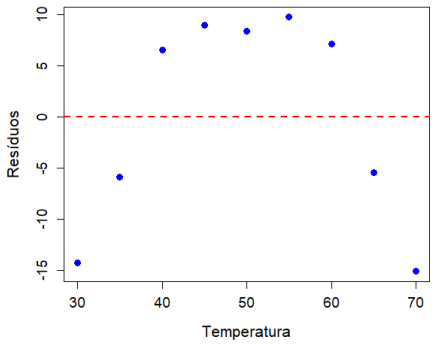
Logo, como há indícios de problemas de especificação do Modelo 2, pois há um componente não linear não captado, é recomendado que se examine um modelo que inclua os valores da temperatura ao quadrado, conforme o modelo de regressão a seguir.

Figura 4: Modelo 2

(a) Regressão



(b) Resíduos vs Temperatura



$$\hat{Rend2} = b_0 + b_1 Temp2 + b_2 Temp2^2 + e_i \tag{13}$$

Tabela 9: Output de regressão do Modelo 3.

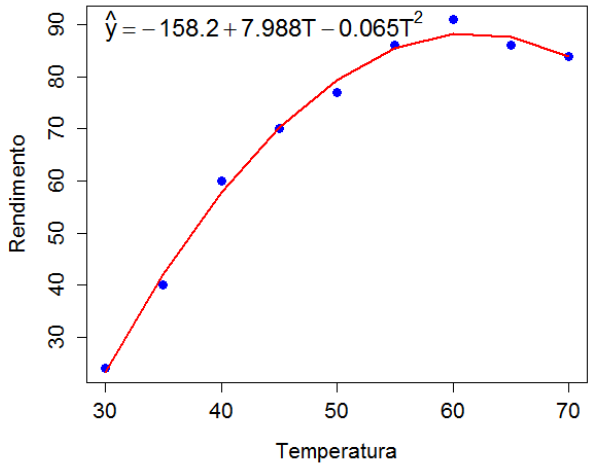
```
Call:
lm(formula = Rend2 ~ Temp2 + I(Temp2^2))

Residuals:
Min        1Q    Median        3Q       Max
-2.44589 -1.69394  0.02424  0.82424  2.82165

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -158.242424   11.672005  -13.56 0.00000999
Temp2          7.987532    0.488315   16.36 0.00000332
I(Temp2^2)    -0.064675    0.004852  -13.33 0.00001103
---
Residual standard error: 2.129 on 6 degrees of freedom
Multiple R-squared: 0.9937, Adjusted R-squared: 0.9916
F-statistic: 471.2 on 2 and 6 DF, p-value: 0.00000025
```

As estatísticas do Modelo 3 estão na Tabela 9 e indicam uma modelo muito bem ajustado. A estatística F com 2 e 6 graus de liberdade é significativa e o Coeficiente de Determinação ajustado indica que o modelo é capaz de explicar 99% da variação do Rendimento percentual($\bar{R}^2 = .99, F_{2,6} = 471.2, p < .01$). Figura 5 oferece uma boa amostra visual do bom ajuste do modelo.

Figura 5: Modelo 3



Os coeficientes angulares são significativos ($b_1 = 7.99, p < .001$ e $b_2 = -0.06, p < .001$), confirmando que a decisão de incluir a temperatura ao quadro foi acertada.

Tabela 10: Comparação entre as estatísticas dos modelos 1, 2 e 3.

	Model 1	Model 2	Model 3
(Intercept)	−1.20	−7.33	−158.24***
	(4.66)	(14.54)	(11.67)
Temp	1.56***		
	(0.09)		
Temp2		1.52***	7.99***
		(0.28)	(0.49)
Temp^2			−0.06***
			(0.00)
R^2	0.99	0.81	0.99
Adj. R^2	0.99	0.78	0.99
Num. obs.	5	9	9
RMSE	1.46	10.90	2.13

Notas: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. Erros padrão entre parênteses.

A comparação entre os três modelos na Tabela 10 mostra a sensível melhora entre do Modelo 3 para o 2 ao incluir a não linearidade na sua especificação, o RMSE do Modelo 3 é sensivelmente menor do que o do Modelo 2, comprovando sua maior precisão preditiva.

Assim, o Modelo 3 é o mais indicado para explicar e prever o comporta-

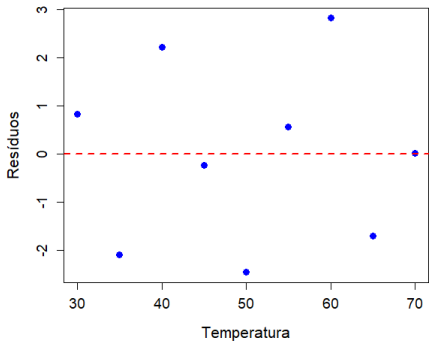
mento do Rendimento percentual na faixa de 30 a 70 graus.

5.3.1 Análise dos resíduos do Modelo 3

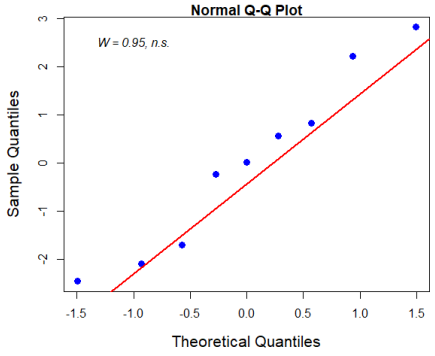
Análise dos resíduos, na Figura 6(a) não acusa a presença de padrões, estado aleatoriamente dispostos em torno do valor zero, exibido homogeneidade na variância e sem autocorrelação.

Figura 6: Resíduos do Modelo 3

(a) Resíduos vs Temperatura



(b) Q-Q Plot dos Resíduos



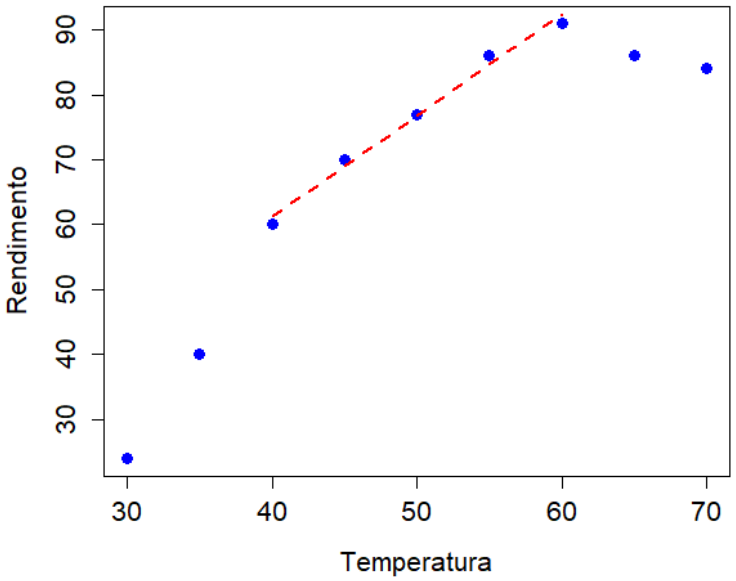
A Figura 6(b) não indica afastamento grave da reta de normalidade. Para uma avaliação ainda mais formal, foi incluída a estatística de Shapiro-Wilk (W), que testa a hipótese nula de que os dados são normalmente distribuídos. A estatística W não é significante de modo que os dados não se afastam severamente da normalidade.

5.3.2 A extrapolação da faixa de experimento

Então porque o Modelo 1 é tão bom sendo linear? Porque os dados são lineares na faixa estudada, como se pode constatar na Figura 7, na qual a linha tracejada é a reta de regressão do Modelo 1, sendo plenamente válida, pois o que se tem é uma estimativa com base numa amostra daqueles valores, mas quando a faixa é ampliada, elementos não lineares passam a se refletirem, exigindo nova especificação.

Como nos adivertem Neto, Scarminio e Bruns (2010) e Hines et al. (2006), deve-se muita cautela em extrapolar para além da região dos dados coletados sobre os quais se estimou um modelo. A medida em que se afasta dessa faixa de dados maior a incerteza e a validade do modelo estimado. O mais recomendável é que, se possível, realizem-se novos experimentos ou se colete novas observações para a amplitude pretendida.

Figura 7: Região aproximadamente linear de uma curva



6 Abordagem Matricial

$$\hat{y} = Xb \tag{14}$$

$$b = (X^t X)^{-1} X^t y \tag{15}$$

$$e = y - \hat{y} \tag{16}$$

$$s^2 = \frac{e^t e}{(n-k)} \tag{17}$$

Referências

FARAWAY, J. J. *Practical regression and ANOVA using R.* [S.l.]: University of Bath, 2002.

HINES, W. et al. *Probabilidade e Estatística na Engenharia, 4a edição, Ed.* Rio de Janeiro-RJ: LTC, 2006.

NETO, B. B.; SCARMINIO, I. S.; BRUNS, R. E. *Como Fazer Experimentos: Pesquisa e Desenvolvimento na Ciência e na Indústria.* [S.l.]: Bookman Editora, 2010.