**Modelos de regressão polinomial
problemas e soluções**

NELIO PEREIRA MACHADO

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO DE MESTRE
EM CIÊNCIAS

Área de Concentração: Estatística
Orientador: Professora Dra. Silvia Nagib Elia

São Paulo, outubro de 2006.

Modelos de regressão polinomial

problemas e soluções

NELIO PEREIRA MACHADO

Este exemplar corresponde à redação final da dissertação devidamente corrigida e defendida por Nélcio Pereira Machado e aprovada pela Comissão Julgadora.

São Paulo, 31 de outubro de 2006.

Banca Examinadora:

Profa. Dra. Silvia Nagib Elia (Orientadora) – IME/USP

Prof. Dr. Rinaldo Artes – IBMEC

Profa. Dra. Linda Lee Ho – POLI/USP

Dedicatória

A meus pais Sebastião Euzébio Machado e Romilda
Pereira Costa Machado pelo amor e apoio ilimitado!

Agradecimentos

Primeiramente, agradeço a Deus. Somente Ele e eu sabemos das alegrias e tristezas que vivi durante a confecção deste trabalho. Só quem já passou pela experiência de escrever uma dissertação de mestrado sabe avaliar o trabalho que tal tarefa demanda e a importância do apoio dos amigos e familiares, sem o qual seria impossível levar adiante o projeto. Felizmente, tive o apoio e incentivo de várias pessoas e, por isso, gostaria de lhes agradecer por terem acreditado em mim e em meus sonhos.

Obrigado a meus queridos pais Sebastião Euzébio Machado e Romilda Pereira Costa Machado pela paciência e estímulo para meu crescimento pessoal e profissional.

Obrigado a meus grandes amigos Joelmir Divino Carlos Feliciano e Ricardo de Souza Freitas. Vocês estiveram comigo em momentos importantes! Merecem todo meu respeito e consideração por toda a vida!

Obrigado à família da Cleuza e Toninho por terem me acolhido em seu lar em momentos difíceis de minha caminhada. Vocês me estenderam a mão num momento de extrema necessidade e sou-lhes eternamente grato por isso!

Obrigado à Professora Dagmar Junqueira pelo apoio financeiro; sem seu valioso patrocínio eu não teria conseguido.

Obrigado à Professora Doutora Silvia Nagib Elia por me aceitar como orientando e dedicar a mim tamanha atenção. Aprendi bastante com a Senhora.

Resumo

Modelos de regressão polinomial, embora possam ser considerados como casos particulares do modelo de regressão linear geral, apresentam características próprias e problemas específicos. Destaca-se especialmente o problema da multicolinearidade, consequência da existência de relações lineares entre as potências da variável independente. No presente trabalho, nos dedicamos inicialmente à discussão do problema da multicolinearidade em modelos de regressão polinomial. Serão analisadas técnicas para identificação e correção do problema, bem como as consequências da presença de tal anomalia. É apresentada a técnica dos polinômios ortogonais como uma possível solução do problema, a metodologia da geração dos coeficientes ortogonais e as vantagens de sua utilização. Posteriormente, trataremos de alguns modelos alternativos ao de regressão polinomial, basicamente, à construção de *Splines* e modelos com termos trigonométricos. Finalizando, alguns dos procedimentos descritos são aplicados a um conjunto de dados reais.

Abstract

Polynomial Regression Models, even so can be considered as a particular case of the general linear regression model, they present proper characteristics and specific problems. The problem of the multicollinearity is distinguished especially, fruit of the existence of linear relations between the powers of the independent variable. In the present work, we dedicate initially to the discussion of the problem of the multicollinearity in Polynomial Regression Models. Techniques for identification and correction of the problem will be analyzed, as well as the fruit of the presence of such anomaly. It is presented the technique of the ortogonais polynomials as a possible solution of the problem, the methodology of the generation of the ortogonais coefficients and the advantages of its use. Later, we will deal with some alternative models to the one of Polynomial Regression, basically, to the construction of Splines and models with Trigonometrical Terms. Finishing, some of the described procedures are applied to a real data set.

Sumário

1	Introdução	1
2	Modelos de Regressão Polinomial	4
2.1	Ajuste de modelos de regressão polinomial	4
2.2	Exemplo	11
3	Multicolinearidade em modelos de regressão polinomial	16
3.1	Introdução	16
3.2	Colinearidade em modelos de regressão polinomial	23
3.3	Escolha do grau do polinômio	29
3.4	Transformações para Redução de Colinearidade entre as Variáveis Explicativas	32
3.5	Correlação entre potências de uma variável aleatória e centralização	41
4	Polinômios ortogonais	51
4.1	Introdução	51
4.2	Polinômios ortogonais	53

4.3	Polinômios ortogonais para espaçamento e frequências desiguais	59
4.4	Transformações Matriciais	67
5	Modelos alternativos aos modelos de regressão polinomial	72
5.1	O uso de <i>splines</i> em regressão polinomial	72
5.2	Polinômios e termos trigonométricos	79
6	Aplicação e conclusões	81
6.1	Aplicação	81
6.2	Recomendações e considerações finais	89
	Referências Bibliográficas	89

Capítulo 1

Introdução

A motivação inicial do presente trabalho foi o Relatório de Análise Estatística sobre o projeto *Dinâmica de espécies arbóreas em uma floresta estacional semidecídua na Estação de Caetetus, Gália - SP, RAE-CEA-04P01*. Neste trabalho, no ajuste dos modelos de regressão polinomial surgiu o problema de multicolinearidade. Isso ocorre quando as variáveis preditoras são altamente correlacionadas entre si. A correlação entre duas (colinearidade) ou mais (multicolinearidade) variáveis independentes é um problema que deve ser evitado. MANSFIELD e HELMES (1982) afirmam que examinar o conjunto de variáveis com o objetivo de detectar multicolinearidade deveria ser o primeiro passo no ajuste dos modelos de regressão múltipla.

O objetivo do estudo era determinar uma relação entre a idade da árvore e o comprimento da circunferência do tronco na altura do peito ou a 1,30m do solo (CAP) para espécies encontradas em uma área florestal da região sudeste do Brasil. Para este estudo, empregou-se modelos de regressão polinomial. Neste ajuste, deparou-se com o problema da multicolinearidade, violando a hipótese básica de que cada variável independente X não pode ser combinação linear das demais. Dentre os principais problemas da multicolinearidade está a imprecisão dos Estimadores de Mínimos Quadrados (EMQ); diante disso, as

estimativas dos parâmetros envolvidos na análise se tornam sensíveis a pequenas mudanças nos dados e os correspondentes estimadores têm variância grande. No referido estudo, a solução para o problema de multicolinearidade foi uma centralização da variável independente. No entanto, muitas vezes esta correção não elimina completamente o problema. Portanto, a presente dissertação trata dos modelos de regressão polinomial, bem como da análise dos problemas e soluções em seu ajuste.

Dedicamos o Capítulo 2 a uma breve introdução dos modelos de regressão polinomial, pertencentes à classe dos modelos de regressão linear. Tais modelos são utilizados em situações onde a função que relaciona a variável resposta y às variáveis explicativas $x_1, x_2, \dots, x_k$ é curvilínea, sendo representada por um polinômio. Estes modelos também são utilizados para estimar funções complexas; ou seja, quando uma função polinomial é uma boa aproximação para a verdadeira função complexa ou desconhecida.

No Capítulo 3, nos dedicamos às consequências, à identificação e, se for o caso, ao tratamento para eliminação das fontes de multicolinearidade entre as variáveis explicativas. Em outras palavras, vamos abordar a presença de multicolinearidade nos modelos de regressão polinomial, problema que tem sido exaustivamente analisado em vários trabalhos de renomados autores, dentre os quais se destacam FARRAR e GLAUBER (1967). Segundo MELA e KOPALLE (2002), uma recente busca pelo assunto utilizando o EconLit revelou que 154 estudos discutiam colinearidade ou multicolinearidade em seus abstracts. Uma busca na revista *Applied Economics* utilizando o Infotrac encontrou 220 artigos desde 1991. Trata-se de um assunto importante, pois, como veremos adiante, de nada valerá nosso modelo se tivermos problemas de multicolinearidade.

No Capítulo 4, trataremos dos polinômios ortogonais, uma possível solução para erradicar os problemas da multicolinearidade. Em geral, pesquisas teóricas e empíricas apontam para a superioridade computacional e acurácia dos polinômios ortogonais em detrimento dos polinômios ordinários. Embora superiores, polinômios ortogonais não têm sido usados nos modelos estatísticos devido à sua complexidade. Nesta parte do trabalho, apresentamos a metodologia de geração dos coeficientes ortogonais e as vantagens de sua utilização.

Reservamos o Capítulo 5, para tratarmos dos modelos alternativos aos modelos de regressão polinomial. Trataremos basicamente de dois métodos muito utilizados hoje na literatura estatística que são Splines e Termos Trigonômétricos. Splines são utilizados nos casos onde é difícil calcular o valor da função f ou quando não conhecemos a expressão da função, mas dispomos de um conjunto de valores que em geral são obtidos através de experimentos. Por sua vez, a combinação de polinômios com termos trigonométricos pode nos levar a bons ajustes de modelos de regressão.

Uma parte do conjunto de dados do Projeto *Dinâmica de espécies arbóreas em uma floresta estacional semidecídua na Estação de Caetetus, Gália - SP* (RAE-CEA-04P01, ELIAN, LIMA e LOPES (2004)) será reanalisada no Capítulo 6, com o uso de algumas técnicas descritas nos capítulos anteriores.

Capítulo 2

Modelos de Regressão Polinomial

2.1 Ajuste de modelos de regressão polinomial

Modelos de regressão polinomial são um caso especial de *modelos de regressão linear múltipla* e são utilizados em situações onde a função que relaciona a variável resposta y às variáveis explicativas $x_1, x_2, \dots, x_k$ é curvilínea, sendo representada por um polinômio. Estes modelos também são utilizados para estimar funções complexas; ou seja, quando uma função polinomial é uma boa aproximação para a verdadeira função complexa ou desconhecida.

Modelos de regressão linear podem ser representados matricialmente por:

$$y = X\beta + \epsilon, \quad (2.1)$$

onde:

1. y é um vetor de dimensão $n \times 1$ de variáveis aleatórias observadas (variável resposta ou variável dependente);

2. $\mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]$ é uma matriz de dimensão $n \times p$, com $p = k + 1$ composta de variáveis regressoras, explicativas ou independentes.
3. ϵ é um vetor de dimensão $n \times 1$ de variáveis aleatórias ϵ_i não observadas e independentes, com

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, 2, \dots, n,$$

e σ^2 é um parâmetro desconhecido.

4. $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ é um vetor de dimensão $k \times 1$ de parâmetros desconhecidos.

Matricialmente, escrevemos cada componente da expressão (2.1) da seguinte forma:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Sem perda de generalidade, podemos padronizar a matriz $\mathbf{X}$, de forma que:

- $\mathbf{1}'\mathbf{X}_j = 0$;
- $\mathbf{X}_j'\mathbf{X}_j = 1$, para $j = 1, 2, \dots, k$.

Padronizar significa subtrair de cada coluna da matriz $\mathbf{X}$ sua média e dividir pelo desvio-padrão.

Conforme comentado anteriormente, modelos de regressão polinomial são um caso particular dos modelos de regressão linear múltipla; basta notar que o modelo polinomial de grau k em uma variável é dado por:

$$y = \sum_{j=0}^k \beta_j x^j + \epsilon. \quad (2.2)$$

Tomando $\mathbf{x}_j = \mathbf{x}^j$ para $j = 0, 1, 2, \dots, k$, a expressão (2.2) se torna um Modelo de regressão linear múltipla com k variáveis regressoras $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$. Daí, podemos escrever

$$\mathbf{X}_i = (1, \mathbf{x}_i, \mathbf{x}_i^2, \dots, \mathbf{x}_i^k), \quad i = 1, 2, \dots, n, n \geq p.$$

Como consequência das suposições, temos:

$$E[\mathbf{y}/\mathbf{x}] = \mathbf{x}\beta.$$

A matriz de Covariância de $\mathbf{y}$ é dada por:

$$\text{Var}[\mathbf{y}/\mathbf{x}] = \sigma^2 \mathbf{I}$$

Estatisticamente escrevemos $\mathbf{y} \sim NM_n(\mathbf{X}\beta; \sigma^2 \mathbf{I})$, onde $NM_n(\mathbf{X}\beta; \sigma^2 \mathbf{I})$ indica a distribuição normal n variada com vetor de médias $\mathbf{X}\beta$ e matriz de covariâncias $\sigma^2 \mathbf{I}$. Este modelo é utilizado para ajustar a relação funcional linear entre $\mathbf{y}$ e $\mathbf{X}$ que por sua vez deve ser linear nos parâmetros presentes em β . Portanto, de fato, os modelos de regressão polinomial são um caso especial dos modelos de regressão linear múltipla, permitindo-nos escrever:

$$y = \sum_{j=0}^k \beta_j x_j + \epsilon, \quad \mathbf{x}_j = \mathbf{x}^j, \quad j = 0, 1, 2, \dots, k. \quad (2.3)$$

Agora, os resultados estatísticos para regressão linear múltipla tais como estimativas dos parâmetros, testes estatísticos, ...,etc poderão ser utilizados em modelos de regressão polinomial de ordem k . Definimos $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ como sendo o vetor de valores ajustados ou valores preditos. O termo ϵ é chamado de erro aleatório ou termo randômico e supomos que ele tenha média zero e variância constante σ^2 . O vetor $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ é chamado de vetor de resíduos e o princípio dos mínimos quadrados é baseado nos resíduos. A análise de resíduos é de primordial importância na avaliação de modelos de regressão linear múltipla, pois através deste recurso, podemos avaliar a adequação do modelo com relação aos pressupostos básicos, bem como avaliar as contribuições individuais das variáveis regressoras. Para mais detalhes sobre análise de resíduos, boas referências são DRAPER e SMITH (1998) e MONTGOMERY, PECK e VINNING (2001).

O estimador $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ de Mínimos Quadrados de $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ é calculado com base no método de Mínimos Quadrados a partir de uma amostra de

tamanho n de valores observados. Portanto, temos $(\mathbf{X}_i, \mathbf{y}_i)$, onde $\mathbf{X}_i = (x_{0i}, x_{1i}, \dots, x_{ki})$, para $i = 1, 2, \dots, n$, $n \geq p$ que é a i -ésima linha da matriz $\mathbf{X}$. Após cálculos, temos que o estimador de Mínimos Quadrados do vetor β é dado por:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}. \quad (2.4)$$

Além disso, $\hat{\beta} \sim NM_{(k+1)}(\beta; \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$. Segue que o modelo ajustado aos dados é dado por:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}. \quad (2.5)$$

O vetor de resíduos é dado por $e = \mathbf{y} - \mathbf{X}\hat{\beta}$ e como visto anteriormente, temos que:

$$\begin{aligned} \mathbf{y} &\sim NM_n(\mathbf{X}\beta; \sigma^2 \mathbf{I}), \\ \hat{\beta} &\sim NM_{(k+1)}\left(\beta; \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right), \\ \hat{\mathbf{y}} &\sim NM_n\left(\mathbf{X}\beta; \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\right). \end{aligned}$$

Podemos calcular a Esperança e a matriz de Covariância do vetor de resíduos:

$$\begin{aligned} E[e] &= E[\mathbf{y} - \mathbf{X}\hat{\beta}] \\ &= E[\mathbf{y}] - E[\mathbf{X}\hat{\beta}] \\ &= \mathbf{X}\beta - \mathbf{X}\beta = \mathbf{0} \\ Var(e) &= Var[\mathbf{y} - \mathbf{X}\hat{\beta}] \\ &= Var[\mathbf{y}] + Var[\mathbf{X}\hat{\beta}] - 2Cov[\mathbf{y}, \mathbf{X}\hat{\beta}] \\ &= \sigma^2 \mathbf{I} + \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' - 2Cov\left[\mathbf{y}, \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y}\right] \\ &= \sigma^2 \mathbf{I} + \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' - 2Cov[\mathbf{y}, \mathbf{y}] \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ &= \sigma^2 \mathbf{I} + \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' - 2\sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ &= \sigma^2 \mathbf{I} - \sigma^2 \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \\ &= \sigma^2 \left[\mathbf{I} - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right] \end{aligned}$$

Agora podemos escrever:

$$e \sim NM_n \left(\mathbf{0}; \sigma^2 \left[\mathbf{I} - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right] \right).$$

De (2.5), temos que $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$, com $\mathbf{H} = \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$.

Uma medida bastante usada em análise de regressão é o *Coefficiente de Determinação* R^2 , interpretado como a proporção de variabilidade presente na variável dependente $\mathbf{y}$ que é explicada pelas variáveis regressoras no modelo ajustado aos dados. O valor de R^2 varia de 0 a 1 e quanto mais próximo de 1, melhor o ajuste do modelo; em contrapartida, quanto mais próximo de 0, pior será o modelo. É prática comum compararmos modelos através de seus coeficientes de determinação, no entanto, seu uso em modelos de regressão polinomial deve ser cuidadoso, pois na presença de multicolinearidade, R^2 estará sempre próximo de 1 devido à alta correlação existente entre as variáveis regressoras. Na verdade, basta que somente um par de variáveis regressoras tenha alta correlação para que o valor de R^2 seja elevado.

O Estimador de Mínimos Quadrados de σ^2 é dado por:

$$\hat{\sigma}^2 = \frac{SQE}{n - p} \quad (2.6)$$

onde $SQE = (\mathbf{y} - \hat{\mathbf{y}})' (\mathbf{y} - \hat{\mathbf{y}})$ é a *Soma de Quadrados dos Resíduos*.

Há vários problemas e recomendações no ajuste de modelos de regressão polinomial. Os principais são:

- (i) O modelo deve possuir a menor ordem possível. Polinômios de ordem elevada ($k > 2$) devem ser evitados, a menos que existam motivos teóricos ou práticos que justifiquem sua adoção. Para modelos com ordem elevada, a interpretação dos parâmetros se torna complicada, não sendo também confiáveis para se fazer previsões. Portanto, devemos utilizar modelos mais simples sempre que possível;
- (ii) Há várias estratégias para escolher o grau de um polinômio. Talvez a mais comum seja ajustar vários modelos aumentando sucessivamente a ordem do polinômio até que a estatística do teste $H_0 : \beta_j = 0$, em que β_j é o coeficiente de $\mathbf{x}^j$, seja não significativa. Reservamos a Seção 3.3 para discutir esse assunto;

- (iii) De acordo com SHACHAM e BRAUNER (1997), extrapolações com modelos polinomiais são complicadas. Portanto, deve-se evitar realizar extrapolações com modelos polinomiais. Em geral, os modelos polinomiais podem ser bem ajustados para os dados originais; mas podem levar-nos a tomar conclusões errôneas ou incorretas em situações de extrapolação. Na Seção 3.3 trataremos desse ponto;
- (iv) Se os valores de $\mathbf{x}$ são limitados a intervalos de amplitude pequena, podemos ter multicolinearidade ou mal-condicionamento na matriz $\mathbf{X}$. Como exemplo, considere que x varie entre 1 e 2; daí x^2 varia entre 1 e 4, podendo haver fortes problemas de multicolinearidade entre x e x^2 ;
- (v) Deve-se sempre utilizar a variável transformada $x_i - \bar{x}$ ao invés da variável regressora x_i . Em modelos de regressão polinomial, a sequência de variáveis regressoras $\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \dots$ geralmente possuem alta correlação. Em face disso, teremos mal-condicionamento da matriz de planejamento $\mathbf{X}$. Como consequência, vários problemas de ordem computacional podem surgir durante os cálculos da inversa da matriz $\mathbf{X}'\mathbf{X}$, essencialmente importante para a estimativa dos coeficientes do modelo. Portanto, centralizando a variável explicativa em torno de sua média estaremos minimizando o problema de mal-condicionamento. É importante enfatizar que na maioria dos casos a centralização elimina completamente o problema da multicolinearidade; em outros casos a centralização minimiza o problema e em casos mais complexos a centralização não produz nenhum efeito para reduzir o problema da multicolinearidade ou mal-condicionamento.

Modelos Polinomiais com uma variável regressora

Considere o modelo dado por:

$$y = \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{X}^2 + \epsilon. \quad (2.7)$$

Esse modelo é chamado de *modelo polinomial de segunda ordem com uma variável explicativa*, ou simplesmente, *modelo quadrático*, já que o valor esperado (*esperança*)

de y que é dado por $E[y] = \beta_0 + \beta_1 x + \beta_2 x^2$, é uma função quadrática. Um típico exemplo do modelo acima é dado em MONTGOMERY, PECK e VINNING (2001). A título de ilustração, os autores usa a função $E[y] = 5 - 2x + \frac{1}{4}x^2$ para elaborar o gráfico. Utilizando o software *MatLab* versão 7.0.1, teremos o gráfico ilustrativo dado pelos autores. Os comandos usados para gerar os resultados são:

```
> y=@(x)(5-2.*x+ 0,25.*x^2)
> fplot(y,[0 10])
```

Esses comandos geram o gráfico da Figura 2.1. No modelo (2.7), o parâmetro β_1 mede o efeito linear de x , enquanto que β_2 mede o efeito quadrático. Já o parâmetro β_0 é a média de y quando $x = 0$, desde que a faixa de variação da variável x inclua o valor 0. Se x não assumir o valor zero, então β_0 não tem interpretação prática.

Figura 2.1: Exemplo de modelo polinomial

Modelos Polinomiais com duas variáveis regressoras

O modelo de regressão polinomial de *segunda ordem com duas variáveis regressoras* (explicativas) pode ser escrito matematicamente como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon. \quad (2.8)$$

Neste modelo, os parâmetros β_1 e β_2 são associados aos efeitos lineares das variáveis x_1 e x_2 , respectivamente e os parâmetros β_{11} e β_{22} estão associados aos efeitos quadráticos de x_1 e x_2 , respectivamente. O parâmetro β_0 é a média de y quando $x_1 = x_2 = 0$, desde que a faixa de variação das variáveis regressoras permita que $x_1 = x_2 = 0$. Da mesma forma que antes, se isso não acontecer, β_0 não tem interpretação prática. Afim de evitar problemas computacionais na estimação dos parâmetros, aqui também a centralização é recomendada. No entanto, como estamos trabalhando com duas variáveis regressoras, então substituímos x_1 por $x_1 - \bar{x}_1$ e x_2 por $x_2 - \bar{x}_2$, com o objetivo de estimar os parâmetros do modelo, voltando em seguida para o modelo com as variáveis originais.

2.2 Exemplo

Os dados deste exemplo foram retirados de MONTGOMERY, PECK e VINCING (2001), e são relativos a uma indústria química que está investigando a influência de duas variáveis denominadas *tempo de reação* (x_1) e *temperatura* (x_2) na produção de um processo. Com esse objetivo, foram coletadas 24 observações aleatórias para a variável Y , rendimento. Vale ressaltar que os valores de x_1 e x_2 são fixos e as informações são apresentadas na Tabela 2.2. Através do software **S-Plus**, ajustamos o modelo:

$$\begin{aligned} \hat{y} = & 50,41708 - 0,7198062.(x_1 - 85,46667) + 0,1052813.(x_2 - 175,7917) - \\ & 0,05965306.(x_1 - 85,46667)^2 - 0,03767571.(x_2 - 175,7917)^2 + \\ & + 0,01257698.(x_1 - 85,46667)(x_2 - 175,7917). \end{aligned}$$

Com as variáveis transformadas, temos as novas variáveis

- $A = x_1 - \bar{x}_1$;
- $B = x_2 - \bar{x}_2$;
- $C = A^2 = (x_1 - \bar{x}_1)^2$;
- $D = B^2 = (x_2 - \bar{x}_2)^2$;
- $E = A.B = (x_1 - \bar{x}_1) \cdot (x_2 - \bar{x}_2)$.

Neste exemplo, temos que $R^2 = 0,6353202$ e, de acordo com os resultados apresentados na Tabela 2.1, todos os termos são estatisticamente significantes. A Figura 2.2 se refere aos gráficos de diagnóstico e, de através dos gráficos, podemos constatar que o modelo não apresenta problemas.

A adequação do modelo de regressão ajustado aos dados foi avaliada tomando como base a análise de resíduos. Através da análise dos gráficos, chegamos à conclusão de que não há qualquer indicação de falta de adequação do modelo e das suposições a ele associados.

Tabela 2.1: Análise de Variância

Termo	gl	Soma de Quadrados	Quadrado Médio	Estatística F	Nível Descritivo
A	1	312,0072	312,0072	772,9997	0,00000000
B	1	21,0809	21,0809	52,228	0,00000101
C	1	49,0279	49,0279	121,467	0,00000000
D	1	31,8626	31,8626	78,9399	0,00000005
E	1	2,3325	2,3325	5,7788	0,02720764
Resíduos	18	7,2650	0,4040		

Tabela 2.2: Rendimento (y), Tempo de Reação (x_1) e Temperatura (x_2)

Obs.	Rendimento (y)	Tempo de Reação (x_1)	Temperatura (x_2)
1	50,95	76,0	170
2	47,35	80,5	165
3	50,99	78,0	182
4	44,96	89,0	185
5	41,89	93,0	180
6	41,44	92,1	172
7	51,79	77,8	170
8	50,78	84,0	180
9	42,48	87,3	165
10	49,80	75,0	172
11	48,74	85,0	185
12	46,20	90,0	176
13	50,49	85,0	178
14	52,78	79,2	174
15	49,71	83,0	168
16	52,75	82,0	179
17	39,41	94,0	181
18	43,63	91,4	184
19	38,19	95,0	173
20	50,92	81,1	169
21	46,55	88,8	183
22	44,28	91,0	178
23	48,72	87,0	175
24	49,13	86,0	175

Figura 2.2: Gráficos de Diagnóstico

Neste capítulo vimos que os modelos de regressão polinomial são um caso especial dos modelos de regressão linear múltipla e que tais modelos são utilizados nos casos em que a função que relaciona a variável resposta y às variáveis explicativas é curvilínea, sendo representada por um polinômio e nos casos em que é necessário estimar funções que são complexas. Também, elencamos os problemas e recomendações dos autores no ajuste de modelos de regressão polinomial. A principal recomendação por parte dos autores é que se deve utilizar a variável transformada $x_i - \bar{x}$ no lugar da variável regressora x_i a fim de se evitar o mal-condicionamento da matriz de planejamento X . Tal recomendação minimiza os problemas de ordem computacional que surgem durante os cálculos da inversa da matriz $X'X$. Por fim, extraímos um exemplo de MONTGOMERY, PECK e VINNING (2001) no qual se utilizou modelos de regressão polinomial com a variável regressora x_i centralizada.

No próximo capítulo trataremos da multicolinearidade em modelos de regressão polinomial assim como dos problemas que ela causa no modelo.

Capítulo 3

Multicolinearidade em modelos de regressão polinomial

3.1 Introdução

Este capítulo tem o propósito de definir, detectar e propor soluções para eliminação das fontes de multicolinearidade entre as variáveis explicativas no modelo de regressão polinomial. Multicolinearidade é definida em Econometria como sendo a presença de relações lineares perfeitas ou quase perfeitas entre as variáveis explicativas em modelos de regressão linear, segundo SILVEY (1968), FARRAR e GLAUBER (1967) e GUNNS (1984). MASON, GUNST e WEBSTER (1975), por sua vez, afirmam que multicolinearidade ou dependência linear entre os vetores de variáveis regressoras em um modelo de regressão linear múltipla pode provocar severos efeitos na estimação dos parâmetros e na seleção de variáveis do modelo. De modo geral, *multicolinearidade* ou *mal-condicionamento* pode surgir em modelos de regressão linear múltipla por vários motivos distintos e seus danos não são ocasionados simplesmente pela sua presença, mas sim pelo grau com que

se manifesta. De acordo com FREITAS (1988), as principais causas da multicolinearidade são:

1. Especificação do modelo;
2. Propriedades inerentes à população;
3. Deficiências na amostragem;
4. Modelos mal definidos;
5. Presença de outliers.

Modelos mal definidos estão relacionados ao problema de existir mais variáveis regressoras do que observações. Esse tipo de problema está relacionado à pesquisas médicas, onde são recolhidas muitas informações para cada indivíduo no estudo. Para solucionar o problema, podemos nos utilizar das seguintes opções relacionadas a seguir:

1. O modelo deve ser redefinido, sendo utilizadas somente algumas variáveis (as mais importantes);
2. Utilizar técnicas de componentes principais para selecionar variáveis regressoras.

Obviamente, a primeira solução requer a experiência do pesquisador para selecionar as variáveis úteis à regressão. No entanto, julgamento do pesquisador falhar, ele pode estar retirando variáveis úteis à análise, comprometendo o ajuste do modelo. A segunda solução, por sua vez, é recomendada por vários autores para selecionar as variáveis explicativas.

Neste trabalho, estamos interessados em estudar o tipo de multicolinearidade que não é inerente ao modelo. Exemplo disso aparecem em análises de dados em que as variáveis estão altamente relacionadas. Por exemplo, custos de produção e valor de remédios. Altos custos de produção implicam em altos custos dos remédios nas drogarias e farmácias. Como outro exemplo, suponha que o consumo seja função da renda e da taxa real de juros. Qualquer pessoa de bom senso seria capaz de nos informar que obviamente o consumo

está altamente relacionado à renda e esta variável. Se estas variáveis forem utilizadas como explicativas no modelo de regressão, a alta correlação entre elas causa o problema da multicolinearidade. Não é difícil ver a forte dependência linear positiva entre as variáveis e essa dependência não é inerente ao modelo. Várias técnicas tem sido propostas na literatura estatística e econômica a fim de detectar a presença de multicolinearidade. O cálculo das correlações entre as variáveis pode não ser suficiente em se tratando de modelos com mais de duas variáveis explicativas.

Este problema tem sido exaustivamente analisado em vários trabalhos de renomados autores. Aqui, discutiremos procedimentos estatísticos para detectar e eliminar fontes de colinearidade entre as variáveis preditoras. Segundo MELA e KOPALLE (2002), uma recente busca pelo tema utilizando o EconLit revelou que 154 estudos discutiam colinearidade ou multicolinearidade em seus abstracts. Uma busca na revista *Applied Economics* utilizando o Infotrac encontrou 220 artigos desde 1991. É um assunto importante, pois, como veremos a seguir, de nada valerá o modelo ajustado na presença de multicolinearidade.

A limitação imposta pela multicolinearidade é facilmente identificada e corrigida através de várias técnicas que serão apresentadas neste capítulo. Levando em consideração que $\mathbf{X} = [1, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k]$, então, de acordo com MONTGOMERY, PECK e VINNING (2001) e MASON, GUNST e WEBSTER (1975), os vetores $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k$ são linearmente dependentes se existem constantes $a_1, a_2, \dots, a_k$ não todas nulas tais que

$$\sum_{j=1}^k a_j \cdot \mathbf{X}_j = 0. \quad (3.1)$$

Os autores concordam em afirmar que existe multicolinearidade nos casos em que

$$\sum_{j=1}^k a_j \cdot \mathbf{X}_j \approx 0, \quad (3.2)$$

para relações lineares quase perfeitas entre as variáveis explicativas.

Independentemente da fonte, aqui nos preocuparemos principalmente com a alta correlação entre as variáveis $\mathbf{X}^l$ e $\mathbf{X}^h$, para $l < h$ e $h = 1, 2, \dots, k$, que consiste na multico-

linearidade em modelos de regressão polinomial. Alguns indícios da existência da multicolinearidade são fáceis de serem inspecionados. Os sinais mais claros são:

1. Ocorrência de valores de coeficientes de correlação linear de Pearson (r_{ij}) entre as variáveis explicativas $\mathbf{x}_i$ e $\mathbf{x}_j$ para $i \neq j$ próximos de 1 ou -1 . O coeficiente de correlação linear de Pearson é definido por

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}},$$

onde

$$S_{ij} = \sum_{l=1}^n (\mathbf{x}_{li} - \bar{x}_i)(\mathbf{x}_{lj} - \bar{x}_j), \quad i, j = 1, 2, \dots, k, i \neq j.$$

Para entender melhor esse item, considere o modelo dado em (2.1) com k variáveis regressoras. Conforme dito no capítulo anterior, sem perda de generalidade, as variáveis $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k$ podem ser padronizadas, de modo que:

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{pmatrix} n & nr_{12} & \dots & nr_{1p} \\ nr_{21} & n & \dots & nr_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ nr_{p1} & nr_{p2} & \dots & n \end{pmatrix}^{-1} \begin{pmatrix} nr_{y1} \\ nr_{y2} \\ \vdots \\ nr_{yP} \end{pmatrix} \\ &= \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_{y1} \\ r_{y2} \\ \vdots \\ r_{yP} \end{pmatrix} = \mathbf{R}_{xx}^{-1} \mathbf{R}_{xy}. \end{aligned} \quad (3.3)$$

Nesta equação, r_{ij} representa o coeficiente de correlação entre $\mathbf{x}_i$ e $\mathbf{x}_j$, para $i, j = 1, 2, \dots, k$; r_{yi} representa o coeficiente de correlação entre $\mathbf{x}_i$ e $\mathbf{y}$ e n é o número de observações. Através dessa representação, fica fácil ver o impacto da correlação na estimativa dos parâmetros da regressão. Por outro lado, baixos valores de r_{ij} não implicam na inexistência de multicolinearidade, já que esse coeficiente mede apenas o grau de relação linear entre pares de variáveis.

2. Grandes alterações nas estimativas dos coeficientes de regressão devido à adição ou retirada de variáveis regressoras ou quando uma observação é alterada ou excluída do conjunto de dados;
3. A hipótese $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ é rejeitada utilizando-se o *teste F*, mas nenhuma das hipóteses $H_0 : \beta_j = 0$ para $j = 1, 2, \dots, k$ é rejeitada por meio da realização dos *testes t* dos coeficientes individuais da regressão;
4. Ocorrência de sinais algébricos das estimativas dos coeficientes diferentes daqueles esperados a partir de conhecimentos teóricos ou experiência sobre o fenômeno que está sob estudo;
5. Obtenção de intervalos de confiança para os coeficientes da regressão com elevadas amplitudes;
6. Variáveis importantes são descartadas com a aceitação da hipótese $H_0 : \beta_j = 0$, devido à alta variância de $\hat{\beta}_j$ provocada pela multicolinearidade.

Sabemos que para calcular o estimador $\hat{\beta}$ via método dos Mínimos Quadrados, precisaremos da inversa da matriz $\mathbf{X}'\mathbf{X}$. As propriedades do estimador $\hat{\beta}$ são bem conhecidas e bem explicadas em DRAPER e SMITH (1981) e MONTGOMERY, PECK e VINNING (2001). De acordo com FARRAR e GLAUBER (1967), SILVEY (1968) e MASON, GUNST e WEBSTER (1975), na presença de multicolinearidade teremos vários problemas com o estimador $\hat{\beta}$, pois a matriz $\mathbf{X}$ será quase singular, implicando na inexistência da inversa da matriz $\mathbf{X}'\mathbf{X}$.

Para que as estimativas sejam confiáveis, é de fundamental importância identificar e quantificar o grau de multicolinearidade entre as variáveis explicativas presentes no modelo. Nos casos de multicolinearidade severa, de nada servirá o modelo ajustado aos dados através de técnicas estatísticas e matemáticas. O modelo ajustado, sob a influência de multicolinearidade, superestima os valores dos coeficientes do modelo, inflacionando, também a variância. Especialistas em Econometria têm apontado que, na presença de multicolinearidade, também teremos altos valores de R^2 em face de baixa significância das estimativas dos parâmetros da regressão.

Os valores dos coeficientes de correlação entre pares de variáveis explicativas permitem ao pesquisador prever as alterações em uma determinada variável em função da outra; portanto, o coeficiente de correlação é apenas uma medida de associação entre as variáveis, consideradas duas a duas. Este coeficiente pode ser baixo e mesmo assim existir forte multicolinearidade entre as variáveis explicativas. De acordo com MONTGOMERY, PECK e VINNING (2001) e BELSLEY, KUH e WELSCH (1980), analisar os autovalores da matriz $\mathbf{X}'\mathbf{X}$ é aconselhável para identificar a natureza de dependência linear existente entre as variáveis, identificando aquelas que contribuem para o aparecimento da multicolinearidade, pois em sua presença, a matriz $\mathbf{X}'\mathbf{X}$ é mal condicionada (singular, $\|\mathbf{X}'\mathbf{X}\| = 0$), fazendo com que as estimativas dos coeficientes de mínimos quadrados não sejam confiáveis. Quase todos os autores são unânimes em dizer que a análise dos autovalores e autovetores ortogonalizados da matriz $\mathbf{X}'\mathbf{X}$ fornecem importantes informações para analisarmos a multicolinearidade presente na estrutura da matriz $\mathbf{X}'\mathbf{X}$. Trata-se de um sério problema, pois o pressuposto do modelo é que as variáveis explicativas são linearmente independentes e se elas interagem entre si, então as previsões efetuadas com a equação estimada são inaceitáveis. Além disso, é desnecessário utilizar no modelo variáveis independentes que, sendo altamente correlacionadas, mediriam a mesma informação.

Matematicamente, o problema se resume no fato de que a matriz $\mathbf{X}$ não tem posto completo e, neste caso, o determinante da matriz $\mathbf{X}'\mathbf{X}$ é zero, tendo como consequência o fato de que essa matriz não poderá ser invertida. Assim sendo, o determinante da matriz $\mathbf{X}'\mathbf{X}$ pode ser utilizado para detectar multicolinearidade. Se a matriz $\mathbf{X}$ for padronizada, então $0 \leq \|\mathbf{X}'\mathbf{X}\| \leq 1$. Quando $\|\mathbf{X}'\mathbf{X}\| = 0$, então existe relação de perfeita dependência entre as colunas (ou linhas) da matriz $\mathbf{X}$. Por outro lado, se tivermos $\|\mathbf{X}'\mathbf{X}\| = 1$, então as colunas (ou linhas) de $\mathbf{X}$ são mutuamente ortogonais.

Conforme comentado anteriormente, a multicolinearidade pode ser expressa por:

$$\sum_{j=1}^k a_j \cdot \mathbf{X}_j = 0, \quad \mathbf{X} = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_k] \quad (3.4)$$

e $a_1, a_2, \dots, a_k$ são constantes não todas nulas. Podemos resolver (3.4) para uma das

variáveis, digamos,

$$\mathbf{X}_r = -a_r^{-1} \sum_{j=1, j \neq r}^k a_j \cdot \mathbf{X}_j, \quad r = 1, 2, \dots, k. \quad (3.5)$$

A expressão (3.4) indica que a variável $\mathbf{X}_r$ é perfeitamente predita pelas variáveis remanescentes.

Uma medida utilizada para detecção da multicolinearidade é $R_{(j)}^2$, o coeficiente de determinação entre a variável resposta $\mathbf{y}$ com todas as outras variáveis regressoras, exceto $\mathbf{X}_j$. Quando a multicolinearidade está presente nos dados, então a diferença entre R^2 e o maior valor dos $R_{(j)}^2$, para algum j , será pequena. O problema da utilização desse critério é que a diferença entre R^2 e o maior valor de $R_{(j)}^2$, para algum j , pode ser pequena, mesmo que os dados não sejam colineares.

Outra possível medida é o valor de R_j^2 , o coeficiente de determinação, obtido realizando a regressão da variável $\mathbf{X}_j$ com as $k - 1$ variáveis remanescentes. Sem perda de generalidade, quando a matriz $\mathbf{X}$ é padronizada, então R_j pode ser calculado da seguinte forma:

$$R_j^2 = 1 - C_{jj}^{-1}, \quad j = 1, 2, \dots, k,$$

onde C_{jj} é o j -ésimo elemento da diagonal de $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ e, de acordo com MASON, GUNST e WEBSTER (1975), valores de R_j^2 próximos de 1 são indícios de multicolinearidade severa.

Valores da estatística t que testam a hipótese $H_0 : \beta_j = 0$ são diretamente afetados pelos valores de R_j^2 , pois

$$t_j = \frac{\sqrt{1 - R_j^2}}{S} \cdot \hat{\beta}_j, \quad S = \sqrt{QME},$$

sendo QME o *Quadrado Médio dos Resíduos*.

Desta forma, qualquer variável envolvida em multicolinearidade terá pequeno valor para a estatística t , uma vez que $\sqrt{1 - R_j^2}$ é pequeno. Portanto, os valores de R_j^2 são importantes para determinar quais variáveis podem oferecer problemas.

3.2 Colinearidade em modelos de regressão polinomial

Os efeitos da multicolinearidade na estimativa dos coeficientes da regressão no caso de duas variáveis regressoras no modelo são facilmente vistos, bastando, para isso, notar que

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}, \quad (3.6)$$

onde r_{12} é o coeficiente de Correlação de Pearson entre $\mathbf{X}_1$ e $\mathbf{X}_2$. Uma forte dependência linear entre $\mathbf{X}_1$ e $\mathbf{X}_2$ resulta em alta variância e covariância para os estimadores dos coeficientes da regressão.

Para o caso geral de k variáveis regressoras é mais difícil detectarmos a presença e os efeitos da multicolinearidade nas estimativas dos coeficientes da regressão. Suponha que a relação dada em (3.2) se verifique entre $p \leq k$ variáveis regressoras. Os elementos da diagonal principal da matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ são dados por

$$C_{jj} = (1 - R_j^2)^{-1}, \quad j = 1, 2, \dots, k. \quad (3.7)$$

sendo que R_j^2 o coeficiente de determinação múltipla, resultante da regressão de x_j nos outros $k-1$ regressores. Claramente, quanto mais forte for a dependência linear de x_j nos regressores restantes, e por conseguinte mais forte a colinearidade, maior será o valor de R_j^2 e o de C_{jj} , implicando em maior variância de $\hat{\beta}_j$.

Considere um conjunto de n observações de uma variável dependente y_i associada a uma variável independente $\mathbf{x}_i$. O modelo de regressão polinomial de grau k tem a forma:

$$y_i = \beta_0 + \beta_1 \mathbf{x}_i + \beta_2 \mathbf{x}_i^2 + \dots + \beta_k \mathbf{x}_i^k + \varepsilon_i = \sum_{j=0}^k \beta_j \mathbf{x}_i^j + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (3.8)$$

onde $\beta_0, \beta_1, \dots, \beta_k$ são parâmetros desconhecidos e ε_i é o erro aleatório em y_i . O vetor de estimativas dos parâmetros $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ é usualmente calculado através do procedimento de Mínimos Quadrados através da resolução das equações normais:

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \quad (3.9)$$

e as linhas da matriz $\mathbf{X}$ são dadas por:

$$\mathbf{x}_i = [1, \mathbf{x}_i, \mathbf{x}_i^2, \dots, \mathbf{x}_i^k], \quad i = 1, 2, \dots, n,$$

e a matriz quadrada $\mathbf{X}'\mathbf{X}$ é chamada de matriz normal. Uma das principais suposições para utilização do procedimento de Mínimos Quadrados é que não há erro nas variáveis independentes e isso na prática é raramente satisfeito. BRAUNER e SCHACHAM (1998) e SCHACHAM e BRAUNER (1997) discutem vários problemas associados ao ajuste dos modelos de regressão polinomial. Analisam brevemente a situação em que a variável explicativa $\mathbf{X}$ apresenta erro de medida, a detecção e correção do problema de multicolinearidade e a escolha do grau do polinômio. Geralmente, os valores de uma variável independente podem ser representados por:

$$\mathbf{x}_i = \tilde{x}_i + \mu_i, \quad i = 1, 2, \dots, n,$$

onde $\tilde{x}_i$ é o valor esperado de $\mathbf{x}_i$ e μ_i é o erro (ou incerteza) associada a esse valor. Colinearidade entre as diferentes variáveis podem limitar severamente a precisão e a confiabilidade do ajuste do modelo de regressão. Conforme comentado anteriormente, uma típica consequência da multicolinearidade é a grande mudança nas estimativas dos parâmetros mediante a adição ou retirada de um simples ponto, ou seja, uma simples perturbação pode causar grande mudança nas estimativas dos parâmetros do modelo. Esse efeito é geralmente chamado de mal-condicionamento da matriz $\mathbf{X}$. De acordo com GUNNS (1984), se $\mathbf{X} = [1, \mathbf{x}, \mathbf{x}^2, \dots, \mathbf{x}^k]$, onde

$$\mathbf{x}^j = \begin{bmatrix} \mathbf{x}_1^j \\ \mathbf{x}_2^j \\ \vdots \\ \mathbf{x}_n^j \end{bmatrix} \quad e \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

dizemos que existe colinearidade entre as colunas (ou linhas) de $\mathbf{X}$ se para um valor de $\eta > 0$, pré-determinado e satisfatório, existirem constantes $c_0, c_1, c_2, \dots, c_k$ não todas nulas tais que

$$\sum_{j=0}^k c_j \mathbf{x}^j = \Delta; \quad (3.10)$$

sendo Δ um vetor n dimensional tal que $\|\Delta\| < \eta \|\mathbf{c}\|$ e $\mathbf{c} = (c_0, c_1, \dots, c_k)$, onde $\|\cdot\|$ é qualquer norma matricial válida. A definição dada em (3.10) não pode ser usada diretamente para diagnosticar problemas de multicolinearidade devido ao fato de não sabermos o quão pequeno deve ser η a fim de que os efeitos de multicolinearidade sejam notados. A equação (3.10) pode ser dividida por c_j ($c_j \neq 0$) para dar origem à expressão

$$\sum_{i=0}^k c_{ij} \mathbf{x}^i = \Delta_j, \quad j = 0, 1, 2, \dots, k, \quad (3.11)$$

onde $c_{ij} = \frac{c_i}{c_j}$, $i = 0, 1, 2, \dots, k$ e $c_{jj} = 1$. Para um conjunto finito de pontos, os valores do vetor $\mathbf{c}$ podem ser obtidos através da regressão de $\mathbf{x}^j$ contra as demais potências da variável $\mathbf{x}$. O vetor Δ_j dessa representação pode ser encarado como o resíduo e é denominado *erro de truncagem*, pois representa a informação perdida ou truncada quando $\mathbf{x}^j$ é explicada pelas demais potências de $\mathbf{x}$. Sabemos que se os valores da variável regressora estão sujeitos à erros, então os valores Δ_j também estão sujeitos à erros, que são denotados por δ_j . SHACHAM e BRAUNER (1997) sugerem que para que $\mathbf{x}^j$ ser incluído no modelo é necessário que a norma de Δ_j seja maior que a norma de δ_j . O valor δ_j é chamado de erro propagado e pode ser calculado, dentre outras formas, através de perturbação numérica. Para se utilizar essa técnica, faz-se necessário a utilização de dois conjuntos de dados, a saber, o conjunto de dados originais $\mathbf{x}$ e o conjunto de dados perturbados $\mathbf{x}_e = \mathbf{x} + \eta \delta_j$, de forma que os valores aleatórios do vetor η sejam os desvios-padrão dos erros. Assim, tomando $\eta_i \in [-\frac{5}{3}, \frac{5}{3}]$, a norma da diferença entre Δ_j calculado usando os dados originais e os dados perturbados são uma ótima estimativa para $\|\delta_j\|$.

O valor do erro de truncagem (TNR) pode ser definido como

$$TNR = \frac{\|\Delta_j\|}{\|\delta_j\|}. \quad (3.12)$$

O grau de multicolinearidade pode ser avaliado com base no número condição, neste capítulo simbolizado por $\kappa(\mathbf{X}'\mathbf{X})$ e definido em MONTGOMERY, PECK e VINNING (2001). De forma simples, o número condição é a razão entre o maior e o menor autovalor da matriz $\mathbf{X}'\mathbf{X}$, sendo uma medida de estabilidade ou sensibilidade de uma matriz. BELSLEY (1991) sugere o uso da norma que leva em consideração o maior e o menor autovalor da matriz $\mathbf{X}'\mathbf{X}$, que representaremos por λ_{max} e λ_{min} , respectivamente. Essa norma geralmente é utilizada em Análise Numérica e na literatura estatística em geral, consistindo na razão entre λ_{max} e λ_{min} . Matematicamente, temos que o número condição pode ser dado por:

$$\kappa(\mathbf{X}'\mathbf{X}) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}. \quad (3.13)$$

De modo geral, os autores classificam a multicolinearidade da seguinte forma:

1. **Fraca**, se $\kappa(\mathbf{X}'\mathbf{X}) < 100$;
2. **Moderada**, se $100 \leq \kappa(\mathbf{X}'\mathbf{X}) \leq 1000$;
3. **Forte ou severa**, se $\kappa(\mathbf{X}'\mathbf{X}) > 1000$.

Segundo CARVALHO (1995), não teremos problemas com o modelo (e consequentemente com a análise) nos casos em que a multicolinearidade é fraca. Nos demais casos, tanto moderada quanto forte, deve-se aplicar métodos de atenuação dos problemas.

Considere o sistema de equações dado por

$$\mathbf{A}.\mathbf{z} = \mathbf{b}, \quad (3.14)$$

onde $\mathbf{A}$ é uma matriz quadrada. Analizando a estabilidade do sistema, estamos interessados em saber o quanto muda sua solução $\mathbf{z}$ quando uma pequena mudança ou erro é introduzido na matriz $\mathbf{A}$ ou no vetor $\mathbf{b}$. Diremos que o sistema é estável se pequenas mudanças ou erros nos dados nos levam, também, a pequenas mudanças na solução.

De acordo com FREITAS (1988) e GOLUB e LOAN(1996), na ausência total de multicolinearidade, a matriz $\mathbf{X}'\mathbf{X}$ é ortogonal, de modo que e como consequência, os

autovalores de $\mathbf{X}'\mathbf{X}$ são iguais entre si, levando a $\kappa(\mathbf{X}'\mathbf{X}) = 1$. Neste caso, dizemos que a matriz $\mathbf{X}$ é bem condicionada. A matriz $\mathbf{P}_{n \times n}$ é ortogonal se $\mathbf{P}_i'\mathbf{P}_i = 1$ e $\mathbf{P}_i'\mathbf{P}_j = 0$ para $i, j = 1, 2, \dots, n, i \neq j$, sendo $\mathbf{P}_i$ o vetor de dimensão $n \times 1$, a i -ésima coluna de $\mathbf{P}$.

Esta identificação e seleção das variáveis explicativas (na verdade, potências da variável explicativa $\mathbf{X}$), mas com baixa multicolinearidade pode ser realizada, por exemplo, através do procedimento *stepwise* dos pacotes computacionais *SAS* e *S-Plus*, utilizados neste trabalho. O procedimento *stepwise* implementado na maioria dos pacotes estatísticos, possui uma opção que define a tolerância para a inclusão ou exclusão de variáveis no modelo. De modo geral, pode-se especificar um valor T de tolerância que varia entre 0,01 e 1,0. A variável $\mathbf{X}_j$ não entrará no modelo, mesmo apresentando contribuição significativa se $R_j^2 > 1 - T$, ou seja, estiver altamente correlacionada com as demais variáveis independentes presentes no modelo ou se sua entrada fizer com que $R_k^2 > 1 - T$ onde $\mathbf{X}_k$ é qualquer outra variável presente no modelo.

Conforme comentado anteriormente, na presença de multicolinearidade, ocorre inflação dos elementos da diagonal da matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, aqui representados por C_{jj} . Esses elementos são usados para construir intervalos de confiança para os parâmetros do modelo, da forma:

$$\hat{\beta}_i - t_{(\nu, \alpha)} s \sqrt{C_{jj}} \leq \beta_i \leq \hat{\beta}_i + t_{(\nu, \alpha)} s \sqrt{C_{jj}}, \quad (3.15)$$

onde C_{jj} é o i -ésimo elemento da diagonal da matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$; $t_{(\nu, \alpha)}$ é o valor crítico da distribuição *t-Student* com ν graus de liberdade para $\alpha = 1 - \gamma$, γ coeficiente de confiança e s é o desvio padrão estimado, geralmente igual à raiz quadrada do quadrado médio do resíduo.

A variância amostral é um valioso indicador numérico que frequentemente é utilizado para mensurar e testar a qualidade do ajuste, dada por

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\nu}, \quad (3.16)$$

que é o quadrado médio do resíduo e $\nu = n - k - 1$. Uma pequena variância amostral indica bom ajuste do modelo aos dados. O desvio padrão da estimativa, s , é a medida de variabilidade em torno da função polinomial e foi utilizado na expressão (3.15). Grandes

amplitudes nos intervalos de confiança podem ser consequência de grandes variâncias dos estimadores ou indicativo de que pequenas mudanças nos dados causariam grandes mudanças nas estimativas dos parâmetros. De particular importância é o caso em que o valor $\beta_i = 0$ é incluído no intervalo definido pela equação (3.15). Isto acontece quando $|\hat{\beta}_i| \leq t_{(\nu, \alpha)} s \sqrt{C_{jj}}$ e em tais casos, não há justificativa estatística para incluir a variável associada na equação de regressão. O intervalo de confiança geralmente é utilizado para determinar se um termo parâmetro β_j deve ou não ser mantido no modelo.

Ambos o número condição e o intervalo de confiança possuem desvantagens para identificar os efeitos da multicolinearidade. Para intervalos de confiança, é difícil saber se o intervalo é amplo devido à multicolinearidade ou se pela imprecisão do ajuste.

Uma medida alternativa muito utilizada para detectar multicolinearidade é o **FIV** (Fator de Inflação de Variância), sugerido por BELSLEY (1991) e definido por:

$$FIV_j = \frac{1}{1 - R_j^2}, \quad (3.17)$$

onde R_j^2 é o coeficiente de correlação múltipla da regressão de $\mathbf{x}^j$ em função das demais $j - 1$ variáveis $\mathbf{x}^i$. Na presença de multicolinearidade, teremos $R_j^2 \simeq 1$, levando-nos a $FIV_j \rightarrow \infty$. Os valores de FIV_j são calculados para $j = 1, \dots, k$. O valor $FIV_{max} = \max(FIV_j)$, para $j = 1, \dots, k$ usualmente é utilizado para diagnosticar multicolinearidade. A literatura indica que há problemas de multicolinearidade para $FIV_j > 5$. Para dados não centralizados, o valor de R_j^2 é dado por

$$R_j^2 = 1 - \frac{\sum_{i=1}^n (\hat{x}_i^j - x_i^j)^2}{\sum_{i=1}^n (x_i^j - \bar{x}_j^j)^2},$$

onde $\hat{x}_i^j$ é o valor previsto de $\mathbf{x}_i^j$ quando é feita a regressão de $\mathbf{x}_i^j$ com as demais potências de $\mathbf{x}_i$.

3.3 Escolha do grau do polinômio

Em Análise de regressão polinomial nem sempre sabemos o grau do polinômio a ser ajustado. Entretanto, a literatura tem mostrado que polinômios com grau maior que 5 não tem utilidade prática. Apesar de muitos autores recomendarem que termos polinomiais de alto grau sejam ajustados até que ocorra o primeiro resultado não significativo, GRAYBILL (1961), página 182, recomenda que seja realizado ajustes de ordens superiores até que dois resultados não significativos sejam encontrados. A justificativa para isso, de acordo com MONTGOMERY, PECK e VINNING (2001), é que se o polinômio de ordem p é apropriado para ajustar o modelo, então o $(p-1)$ -ésimo termo pode ser não-significante. Dessa forma, o modelo adequado deverá ser aquele após se obter o primeiro resultado não-significativo. Por exemplo, utilizando a centralização e termo quadrático significativo, o termo linear pode não ser significativo devido ao fato de que o termo quadrático centralizado é função do termo linear. Quando variáveis explicativas não centralizadas são utilizadas, é possível que todos os termos polinomiais sejam não-significativos devido à multicolinearidade.

Para compreender melhor o problema relacionado ao grau do polinômio, considere os dados contidos na Tabela 3.1. Na Tabela 3.2 temos o comportamento do coeficiente de determinação R^2 em função do grau do polinômio. Aqui utilizamos a centralização da variável independente que, conforme será discutido nas próximas seções, pode evitar problemas computacionais. Os gráficos presentes na Figura 3.1 mostram o comportamento dos valores ajustados pelo modelo de regressão polinomial em função do grau do polinômio. Nestes gráficos podemos observar os perigos contidos na extrapolação para a regressão polinomial. Observa-se que os valores extrapolados são bem distantes da tendência observada para o conjunto de dados.

Tabela 3.1: Custos de seguro de automóveis em New Brunswick (1971-1980)

Anos	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980
Custo	45,13	51,71	60,17	64,83	65,24	65,17	67,65	79,80	96,13	115,19

Tabela 3.2: Coeficientes de determinação para vários graus do polinômio

Grau	2	3	4	5	6	7	8	9
R^2	0,9213	0,9922	0,9922	0,9996	0,9996	0,9998	0,9999	1

Figura 3.1: Perigos da extrapolação ao se utilizar regressão polinomial

De acordo com SHACHAM e BRAUNER (1997), o j -ésimo grau de $\mathbf{x}_i$ pode ser representado por um polinômio de menor grau:

$$\mathbf{x}_i^j = \sum_{l=0}^{(j-1)} \alpha_l \mathbf{x}_i^l + E_{ij}, \quad i = 1, 2, \dots, n, \quad (3.18)$$

onde $\alpha_0, \alpha_1, \dots, \alpha_{j-1}$ são constantes e E_{ij} é um indicador de multicolinearidade que chamaremos de erro de truncagem. Esta medida, introduzida por SHACHAM e BRAUNER (1997) é usada para avaliar a colinearidade entre as várias potências da variável independente. Em outras palavras, ela dimensiona o erro de truncagem na representação de $\mathbf{x}_i^j$ por uma combinação linear de $\mathbf{x}_i, \mathbf{x}_i^2, \dots, \mathbf{x}_i^{j-1}$. Suponha que a medida dada por (3.11) seja utilizada para detectar multicolinearidade. Quando os coeficientes de α são estimados através do uso de regressão da variável $\mathbf{x}_i^j$ em função das outras $j - 1$ variáveis independentes, seja Δ_j o resíduo desta representação. Os autores utilizam $\|\Delta_j\|$ como medida da multicolinearidade. Grandes valores de $\|\Delta_j\|$ são indicativos de níveis aceitáveis de multicolinearidade. Consequentemente, a variável $\mathbf{x}_i^j$ pode ser incluída no modelo.

SEBER (1977) sugere ainda o uso do procedimento *Backward*. Nesse caso, o grau máximo do polinômio deve ter sido fixado a priori e os termos de mais alto grau são eliminados um a um usando o teste F parcial. O processo pára no primeiro valor de F significativo e alguns autores sugerem que se utilize nível de significância de 0,10 nos testes parciais. A adequação do modelo selecionado deveria ser julgada através da análise de resíduos.

3.4 Transformações para Redução de Colinearidade entre as Variáveis Explicativas

Nesta seção discutiremos a implementação de algumas das transformações mais utilizadas na literatura estatística a fim de se eliminar ou reduzir os problemas ocasionados pela multicolinearidade. De acordo com SHACHAM e BRAUNER (1997), as melhores transformações são aquelas que alteram as variáveis independentes de modo que seus valores estejam no intervalo simétrico $[-1, 1]$ e esse tipo de transformação é a que efetivamente se usa para redução de multicolinearidade. Veremos posteriormente, através de um exemplo, o impacto das transformações na matriz $\mathbf{X}'\mathbf{X}$.

Dentre as várias possibilidades, talvez a mais usada seja a divisão dos valores de $\mathbf{x}_i$ por x_{max} , onde x_{max} é o maior valor absoluto de $\mathbf{x}$, ou seja, utiliza-se a transformação V dada por

$$v_i = \frac{\mathbf{x}_i}{x_{max}}, \quad i = 1, 2, \dots, n.$$

Se todos os $\mathbf{x}_i$ possuem o mesmo sinal (digamos, $\mathbf{x}_i > 0 \quad \forall i$), tem-se que $0 < v_{min} \leq v_i \leq 1$.

Temos também a transformação W, dada por

$$w_i = \frac{(\mathbf{x}_i - x_{min})}{(x_{max} - x_{min})},$$

garante que

1. $0 \leq w_i \leq 1$,
2. $w_{max} = 1$.

A transformação Z é definida como

$$z_i = \frac{2\mathbf{x}_i - x_{max} - x_{min}}{(x_{max} - x_{min})},$$

de modo que $-1 \leq z_i \leq 1$.

Assumindo que $\mathbf{x}_i$ seja distribuído uniformemente e o tamanho da amostra n seja grande o bastante, é possível obtermos uma aproximação para os elementos da matriz normal. FORSYTHE (1957) derivou a expressão do rs -ésimo elemento da matriz $\mathbf{X}'\mathbf{X}$ neste caso. Verifica-se que o rs -ésimo elemento é dado por:

$$\left(\mathbf{X}'\mathbf{X}\right)_{rs} \approx \frac{n}{r+s-1} \quad (3.19)$$

onde r e s representam, respectivamente, os índices de linha e coluna. Podemos ver que os elementos da matriz $\mathbf{X}'\mathbf{X}$ dados por (3.19) são n vezes os elementos da matriz de Hilbert de mesma dimensão de $\mathbf{X}'\mathbf{X}$ e, como sabemos, tal matriz é mal-condicionada para k grande.

Em particular, verifica-se que o número de condição para a matriz de Hilbert de ordem $k+1$ é $e^{3,5(k+1)}$.

Para a transformação V, assumindo que v_i seja distribuído uniformemente em $[v_{min}; 1]$, onde $0 \leq v_{min} \leq 1$, temos:

$$\left(\mathbf{X}'\mathbf{X}\right)_{rs} \approx n \frac{\left[1 - v_{min}^{(r+s-1)}\right]}{(1 - v_{min})(r+s-1)} \quad (3.20)$$

Para a transformação Z, temos:

$$\left(\mathbf{X}'\mathbf{X}\right)_{rs} \simeq \begin{cases} 0, & \text{se } r+s-1 \text{ for par;} \\ \frac{n}{r+s-1}, & \text{se } r+s-1 \text{ for ímpar.} \end{cases} \quad (3.21)$$

Para esta transformação, os elementos da matriz $\mathbf{X}'\mathbf{X}$ são similares aos elementos da matriz de Hilbert exceto para os valores nulos. Verifica-se, portanto, que esta transformação oferece uma melhor redução do Número Condição, como veremos futuramente através dos exemplos.

A Figura 3.2 exhibe o valor do número condição para diferentes graus do modelo de regressão polinomial e para as transformações propostas. Observa-se que a reta de maior inclinação está associada à transformação V[0, 9; 1]; a segunda corresponde à transformação V[0, 5; 1]; a terceira reta se refere a transformação W[0; 1] e a reta de menor inclinação é

relativa à transformação $Z[-1; 1]$. Desta forma, percebemos que a transformação $Z[-1; 1]$ oferece a maior redução nos valores do Número Condição. O Número Condição sozinho é insuficiente para estimar o grau do polinômio. Entretanto, ele pode ser utilizado para determinar o máximo grau do polinômio com base em informações numéricas. Através das *Matrizes de Hilbert* de várias ordens, verifica-se que para $\kappa(\mathbf{X}'\mathbf{X}) \sim 10^{18}$, a maioria dos elementos da matriz inversa de $\mathbf{X}'\mathbf{X}$ não possuem sequer um dígito significativo (usando dupla precisão, as variáveis contêm aproximadamente 15 dígitos significantes). Este limite parece ser muito grande se o dados pertencerem ao intervalo $[-1; 1]$. Mas, se o intervalo original é tal que a transformação fornece valores no intervalo $[0, 9; 1]$, então até mesmo um polinômio de 5º grau pode alcançar este limite numérico.

Para ilustrar o uso das transformações aqui discutidas, vamos utilizar os dados extraídos de GIAUQUE e WIEBE (1928). Considere as variáveis $\mathbf{C}(\text{Cal/mol.K})$, contendo informações relativas à capacidade dos sólidos e a variável $\mathbf{T}(K)$, com informações relativas à temperatura. As medidas associadas às 18 observações dessas variáveis encontram-se na Tabela 3.18. Construímos as variáveis $\mathbf{T}_2 = \mathbf{T}^2, \mathbf{T}_3 = \mathbf{T}^3, \mathbf{T}_4 = \mathbf{T}^4$ e $\mathbf{T}_5 = \mathbf{T}^5$. O modelo, neste caso, é dado por (2.1), com $\mathbf{y} = \mathbf{C}$, $\mathbf{X} = [1, \mathbf{T}, \mathbf{T}_2, \mathbf{T}_3, \mathbf{T}_4, \mathbf{T}_5]$ e $\beta = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]$. Apresentamos na Tabela 3.6 a matriz de correlação entre as variáveis presentes em $\mathbf{X}$ e podemos notar através dela a alta correlação (como era de se esperar) existente entre as potências da variável $\mathbf{T}$. Os autovalores da matriz $\mathbf{X}'\mathbf{X}$ são dados na Tabela 3.3. É importante notar que a matriz de correlação quando os dados são centralizados não difere muito da matriz de correlação apresentada na Tabela 3.6. Isso nos leva à conclusão de que a centralização não foi suficiente para reduzir a correlação entre as potências da variável $\mathbf{T}$. No entanto, temos uma considerável redução dos autovalores, que são apresentados na Tabela 3.4.

Conforme pode ser notado na Tabela 3.5, a maior redução dos autovalores, se dá pelo uso da padronização aplicada aos valores de $\mathbf{T}$. Numa etapa seguinte, ajustamos polinômios de grau 3 e 4 utilizando o pacote estatístico *S-Plus*. A Tabela 3.7 apresenta os valores de $\kappa(\mathbf{X}'\mathbf{X})$ para o ajuste no caso em que não é feita transformação, ou seja, utiliza-se a variável original. As Tabelas 3.8, 3.9 e 3.10 contêm essas informações quando realizadas as transformações V, W e Z, respectivamente. Através da Tabela 3.17, vemos que o aumento da ordem do polinômio afeta a amplitude dos intervalos de confiança.

Removendo o ponto $T_{12} = 153,45$ do conjunto de dados, teremos que o polinômio de 4ª ordem terá estimativas dos parâmetros dados por

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) = (-3,96594; 22,694; -39,4994; 30,0849; -8,31485).$$

Note que esses valores são completamente diferentes daqueles apresentados na Tabela 3.17, evidenciando que pequenas mudanças no conjunto de dados acarretam grandes mudanças nas estimativas dos parâmetros. Com relação às transformações, vemos que a transformação Z é a que mais reduz os efeitos da multicolinearidade. Para as transformações V, W e Z, temos que:

- $\hat{\sigma}^2 = 0,002677$;
- Estatística $F = 820,3$;
- Significância (p-valor) = $1,776.10^{-15}$;
- $R^2 = 0,9961 = 99,61$.

Através da simples inspeção da matriz de correlação das transformações V, W e Z, vemos o quanto a transformação Z é superior às outras duas.

Os resultados dos ajustes utilizando as transformações são apresentados nas Tabelas 3.12, 3.13 e 3.14. Transformações similares têm sido altamente recomendadas por outros autores a fim de evitar problemas gerados por mal-condicionamento. Em particular destacamos SEBER (1977).

Tabela 3.3: Autovalores da matriz $\mathbf{X}'\mathbf{X}$

$1,617103.e^{23}$	$5,951890.e^{16}$	$3,749281.e^{10}$	$-1,486584.e^3$	$-3,470117.e^6$

Tabela 3.4: Autovalores da matriz $\mathbf{X}'\mathbf{X}$ para os dados centralizados

$5,299353.e^{22}$	$3,515018.e^{15}$	$5,271870.e^8$	$-8,835891.e^1$	$-6,485081.e^6$

Tabela 3.5: Autovalores da matriz $\mathbf{X}'\mathbf{X}$ para os dados padronizados

4,97285	0,02709	0,00006	0,00000	0,00000

Tabela 3.6: Matriz de correlação das potências da Variável $\mathbf{T}$

	$\mathbf{T}$	$\mathbf{T}_2$	$\mathbf{T}_3$	$\mathbf{T}_4$	$\mathbf{T}_5$
$\mathbf{T}$	1	0,9985344	0,9942605	0,9874369	0,9783961
$\mathbf{T}_2$	0,9985344	1	0,9985892	0,9945165	0,9880785
$\mathbf{T}_3$	0,9942605	0,9985892	1	0,9986626	0,9948358
$\mathbf{T}_4$	0,9874369	0,9945165	0,9986626	1	0,9987492
$\mathbf{T}_5$	0,9783961	0,9880785	0,9948358	0,9987492	1

Figura 3.2: Número Condição versus ordem do polinômio para várias transformações

Tabela 3.7: Valores de $\kappa(\mathbf{X}'\mathbf{X})$ para várias ordens - Variável original

Ordem	$\kappa(\mathbf{X}'\mathbf{X})$
3	$4,1844.10^{18}$
4	$3,15.10^{23}$

Tabela 3.8: Valores de $\kappa(\mathbf{X}'\mathbf{X})$ para várias ordens - transformação V

Ordem	$\kappa(\mathbf{X}'\mathbf{X})$
3	$3,1443.10^7$
4	$1,4096.10^{10}$

Tabela 3.9: Valores de $\kappa(\mathbf{X}'\mathbf{X})$ para várias ordens - transformação W

Ordem	$\kappa(\mathbf{X}'\mathbf{X})$
3	$9,538.10^3$
4	$3,113.10^5$

Tabela 3.10: Valores de $\kappa(\mathbf{X}'\mathbf{X})$ para várias ordens - transformação Z

Ordem	$\kappa(\mathbf{X}'\mathbf{X})$
3	$6,0623.10$
4	$3,7726.10^2$

Tabela 3.11: Matriz de correlação para a transformação V

	Intercepto	v	v_2	v_3
v	-0,9997			
v_2	0,9986	-0,9997		
v_3	-0,9969	0,9986	-0,9997	
v_4	0,9946	-0,9970	0,9987	-0,9997

Tabela 3.12: Resultados do ajuste utilizando a transformação V

	Valor	Erro Padrão	t valor	Nível Descritivo
intercepto	-6,3430	3,5543	-1,7846	0,0977
v	34,4208	17,5979	1,9560	0,0723
v_2	-60,9899	32,4549	-1,8792	0,0828
v_3	47,4234	26,4239	1,7947	0,0960
v_4	-13,5127	8,0142	-1,6861	0,1156

Tabela 3.13: Resultados do ajuste utilizando a transformação W

	Valor	Erro Padrão	t valor	Nível Descritivo
intercepto	0,8752	0,0020	438,5446	0,0000
w	0,1319	0,0322	4,0915	0,0013
w_2	-0,3173	0,1455	-2,1799	0,0482
w_3	0,5036	0,2292	2,1970	0,0468
w_4	-0,1949	0,1156	-1,6861	0,1156

Tabela 3.14: Resultados do ajuste utilizando a transformação Z

	Valor	Erro Padrão	t valor	Nível Descritivo
intercepto	0,9126	0,0014	650,2692	0,0000
z	0,0475	0,0026	17,9576	0,0000
z_2	0,0365	0,0074	4,9364	0,0003
z_3	0,0142	0,0036	4,0101	0,0015
z_4	-0,0122	0,0072	-1,6861	0,1156

Tabela 3.15: Matriz de correlação para a transformação W

	intercepto	w	w_2	w_3
w	-0,7910			
w_2	0,6382	-0,9588		
w_3	-0,5496	0,9008	-0,9852	
w_4	0,4920	-0,8494	0,9578	-0,9924

Tabela 3.16: Matriz de correlação para a transformação Z

	intercepto	z	z_2	z_3
z	0,0207			
z_2	-0,7902	0,1312		
z_3	-0,0360	-0,9246	-0,0850	
z_4	0,6572	-0,1580	-0,9645	0,1278

Tabela 3.17: Estimativas dos Parâmetros e Intervalos de Confiança em função da ordem do polinômio

Estimativas	1ª ordem	2ª ordem	3ª ordem	4ª ordem
$\hat{\beta}_0$	$0,66 \pm 0,03$	$1,17 \pm 0,13$	$-0,38 \pm 0,83$	$-6,34 \pm 7,67$
$\hat{\beta}_1$	$0,31 \pm 0,04$	$-0,95 \pm 0,33$	$4,83 \pm 3,11$	$34,42 \pm 38,01$
$\hat{\beta}_2$	0	$0,77 \pm 0,20$	$-6,34 \pm 3,81$	$-60,98 \pm 70,1$
$\hat{\beta}_3$	0	0	$2,88 \pm 1,54$	$47,42 \pm 57,08$
$\hat{\beta}_4$	0	0	0	$-13,51 \pm 17,31$
s^2	$8,009 \cdot 10^{-5}$	$1,624 \cdot 10^{-5}$	$8,11 \cdot 10^{-6}$	$7,166 \cdot 10^{-6}$

Tabela 3.18: Transformações para temperatura de sólidos

<i>Obs</i>	<i>C</i>	<i>T</i>	$V = \frac{T}{T_{max}}$	$W = \frac{T-T_{min}}{T_{max}-T_{min}}$	$Z = \frac{2T-T_{max}-T_{min}}{T_{max}-T_{min}}$
1	10,79	118,99	0,653468065	0	-1
2	10,8	120,76	0,663188533	0,028050713	-0,943898574
3	10,86	122,71	0,673897523	0,058954041	-0,882091918
4	10,93	125,48	0,689109781	0,102852615	-0,79429477
5	10,99	127,31	0,699159756	0,1318542	-0,736291601
6	10,96	130,06	0,714262178	0,175435816	-0,649128368
7	10,98	132,41	0,727167884	0,212678288	-0,574643423
8	11,03	135,89	0,746279312	0,267828843	-0,464342314
9	11,08	139,02	0,763468614	0,317432647	-0,365134707
10	11,10	140,25	0,770223516	0,336925515	-0,32614897
11	11,19	145,61	0,799659509	0,421870048	-0,156259905
12	11,25	153,45	0,842715141	0,546117274	0,092234548
13	11,40	158,03	0,867867538	0,618700475	0,237400951
14	11,61	162,72	0,893624032	0,693026941	0,386053883
15	11,69	167,67	0,920808391	0,771473851	0,542947702
16	11,91	172,86	0,94931078	0,853724247	0,707448494
17	12,07	177,52	0,974902521	0,927575277	0,855150555
18	12,32	182,09	1,000000000	1,000000000	1,000000000

3.5 Correlação entre potências de uma variável aleatória e centralização

Nesta seção, apresentaremos o estudo desenvolvido por BRADLEY e SRIVASTAVA (1979), relativo ao nível de correlação entre as variáveis explicativas presente no modelo de regressão polinomial e sobre o efeito da centralização na tentativa de eliminar esse problema. Centralizar ou padronizar as variáveis regressoras significa que estamos realizando uma translação.

Para medir a correlação entre as variáveis $\mathbf{X}^l$ e $\mathbf{X}^h$, os autores utilizam o coeficiente de correlação amostral r_{lh} que matematicamente pode ser escrito como:

$$r_{lh} = \frac{c_{lh}}{\sqrt{S_l S_h}}, \quad (3.22)$$

onde $c_{lh} = \sum_{i=1}^n (\mathbf{x}_i^l - \bar{\mathbf{x}}^l) x_i^h$, $S_\delta = \sum_{i=1}^n (\mathbf{x}_i^\delta - \bar{\mathbf{x}}^\delta)^2$ e $\bar{\mathbf{x}}^\delta = \frac{\sum_{i=1}^n \mathbf{x}_i^\delta}{n}$, para $\delta = l, h$, sendo $l, h = 1, 2, \dots, l \neq h$. Para esse coeficiente de correlação, valem os resultados a seguir:

Teorema 3.1 Para c_{lh} , S_l e S_h definidos em (3.22),

$$c_{lh} = \sum_{j=0}^{l+h-2} a_j c^j \quad (3.23)$$

e

$$S_l S_h = \sum_{j=0}^{2(l+h)-4} b_j c^j, \quad (3.24)$$

com a_j e b_j funções dos momentos centrais da amostra, dados por:

$$m_s = \frac{1}{n} \sum_{j=0}^n (x_i - \bar{x})^s, \quad s = 1, 2, \dots \quad (3.25)$$

e

$$c = \frac{\bar{x}}{\sqrt{m_2}}.$$

Prova. Sejam c_{lh} , S_l e S_h definidos em (3.22). Nessas condições, a expressão c_{lh} pode ser reescrita como:

$$c_{lh} = \sum_{i=1}^n \mathbf{x}_i^{l+h} - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^l \sum_{i=1}^n \mathbf{x}_i^h, \quad (3.26)$$

e utilizando o fato de que $(\mathbf{x}_i - \bar{x}) + \bar{x} = \mathbf{x}_i$, obtemos:

$$[(x_i - \bar{x}) + \bar{x}]^{l+h} = \sum_{j=0}^{l+h} \binom{l+h}{j} (x_i - \bar{x})^j \bar{x}^{l+h-j}.$$

Como

$$\begin{aligned} [(x_i - \bar{x}) + \bar{x}]^l &= \sum_{j=0}^l \binom{l}{j} (x_i - \bar{x})^j \bar{x}^{l-j} \\ [(x_i - \bar{x}) + \bar{x}]^h &= \sum_{j=0}^h \binom{h}{j} (x_i - \bar{x})^j \bar{x}^{h-j}, \end{aligned}$$

então

$$\begin{aligned} c_{lh} &= \sum_{i=1}^n \sum_{j=0}^{l+h} \binom{l+h}{j} (x_i - \bar{x})^j \bar{x}^{l+h-j} - \\ &\quad - \frac{1}{n} \sum_{i=1}^n \sum_{j=0}^l \binom{l}{j} (x_i - \bar{x})^j \bar{x}^{l-j} \sum_{i=1}^n \sum_{j=0}^h \binom{h}{j} (x_i - \bar{x})^j \bar{x}^{h-j} \\ &= n \left[\sum_{j=0}^{l+h} \binom{l+h}{j} m_j \bar{x}^{l+h-j} - \sum_{j=0}^l \binom{l}{j} m_j \bar{x}^{l-j} \sum_{j=0}^h \binom{h}{j} m_j \bar{x}^{h-j} \right] \\ &= n \sum_{j=0}^{l+h} \binom{l+h}{j} m_j m_2^{\frac{l+h-j}{2}} c^{l+h-j} - \\ &\quad - n \sum_{j=0}^l \binom{l}{j} m_j m_2^{\frac{l-j}{2}} c^{l-j} \cdot \sum_{j=0}^h \binom{h}{j} m_j m_2^{\frac{h-j}{2}} c^{h-j}. \end{aligned} \quad (3.27)$$

As últimas duas expressões de (3.27) resultam da substituição de $\bar{x}$, advindo de $c = \frac{\bar{x}}{\sqrt{m_2}}$ e de m_j , já que $m_j = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{x})^j$, $j = 0, 1, 2, \dots$. De forma similar demonstra-se que

$$S_\delta = n \left\{ \sum_{j=0}^{2\delta} \binom{2\delta}{j} m_j m_2^{\frac{2\delta-j}{2}} c^{2\delta-j} - \left[\sum_{j=0}^{\delta} \binom{\delta}{j} m_j m_2^{\frac{\delta-j}{2}} c^{\delta-j} \right]^2 \right\}. \quad (3.28)$$

Analisando-se a expressão (3.27) para $j = 0$, temos que

$$\begin{aligned} m_j &= m_0 = 1 \\ \binom{l+h}{j} &= 1 \\ \binom{l}{j} &= \binom{h}{j} = 1 \\ c^{l-j} \cdot c^{h-j} &= c^{l+h}. \end{aligned}$$

Como consequência, o coeficiente de c^{l+h} é igual a zero. Analogamente, para $j = 1$, $m_j = m_1 = 0$, de modo que

$$\binom{l+h}{1} m_1 m_2^{\frac{l+h-1}{2}} c^{l+h-1} = 0.$$

Na expressão

$$\sum_{j=0}^l \binom{l}{j} m_j m_2^{\frac{l-j}{2}} c^{l-j} \cdot \sum_{j=0}^h \binom{h}{j} m_j m_2^{\frac{h-j}{2}} c^{h-j}, \quad (3.29)$$

para que tenhamos $c^{l-j} \cdot c^{h-j} = c^{l+h-1}$, deveríamos ter um dos valores de j igual a 1 (um) e outro igual a 0 (zero). Tal fato implicaria em $m_j = 0$, anulando o correspondente coeficiente de c^{l-j} ou de c^{h-j} e portanto o coeficiente de c^{l+h-1} . Assim, a expressão (3.27) é um polinômio de grau $l + h - 2$ em c , podendo ser escrito na forma (3.23) para a_j convenientemente escolhido.

De maneira similar, verifica-se que os coeficientes de $c^{2\delta}$ e $c^{2\delta+1}$ em (3.28) são nulos e com isso a expressão $S_l.S_h$ reduz-se a (3.24) para escolhas convenientes de b_j . $\square$

Teorema 3.2 *O Coeficiente de correlação entre $\mathbf{X}^l$ e $\mathbf{X}^h$, r_{lh} , pode ser escrito como uma função de c , na forma*

$$r_{lh}(c) = \frac{c^{l+h-2}}{|c|^{l+h-2}} + f(c),$$

com $f(c)$ tal que

$$\lim_{|c| \rightarrow \infty} f(c) = 0.$$

Prova. Da expressão (3.27), segue que o coeficiente de c^{l+h-2} em (3.23) é

$$n \left\{ \binom{l+h}{2} - \binom{l}{2} - \binom{h}{2} \right\} m_2^{\frac{l+h}{2}},$$

sendo que os coeficientes

$$\binom{l}{2} m_2^{\frac{l+h}{2}}$$

$$\binom{h}{2} m_2^{\frac{l+h}{2}}$$

são obtidos quando $j = 0$ e $j = 2$ ou $j = 2$ e $j = 0$, respectivamente em

$$\sum_{j=0}^l \binom{l}{j} m_j m_2^{\frac{l-j}{2}} c^{l-j}$$

e

$$\sum_{j=0}^h \binom{h}{j} m_j m_2^{\frac{h-j}{2}} c^{h-j}.$$

Após cálculos, temos que

$$n \left\{ \binom{l+h}{2} - \binom{l}{2} - \binom{h}{2} \right\} . m_2^{\frac{l+h}{2}} = n.l.h.m_2^{\frac{l+h}{2}}.$$

De maneira análoga, verifica-se que o coeficiente de $c^{2\delta-2}$ em (3.28) é

$$n \left\{ \binom{2\delta}{2} - 2 \binom{\delta}{2} \right\} . m_2^\delta = n.\delta^2.m_2^\delta.$$

Como consequência, c_{lh} , $S_l^{\frac{1}{2}}$ e $S_h^{\frac{1}{2}}$ podem ser escritos como

$$c_{lh} = n.l.h.m_2^{\frac{l+h}{2}} c^{l+h-2} + n.l.h.m_2^{\frac{l+h}{2}} c^{l+h-2}.g_1(c) = n.l.h.m_2^{\frac{l+h}{2}} c^{l+h-2} [1 + g_1(c)]$$

e

$$S_l^{\frac{1}{2}}.S_h^{\frac{1}{2}} = n.l.h.m_2^{\frac{l+h}{2}} |c|^{l+h-2} [1 + g_2(c)]$$

com $g_1(c)$ e $g_2(c)$ funções de c tais que $g_1(c) \rightarrow 0$ e $g_2(c) \rightarrow 0$ para $|c| \rightarrow \infty$. Substituindo-se em (3.22), e denotando r_{lh} como $r_{lh}(c)$, temos que

$$r_{lh}(c) = \frac{c^{l+h-2}}{|c|^{l+h-2}} \cdot \frac{1 + g_1(c)}{1 + g_2(c)},$$

com

$$\lim_{|c| \rightarrow \infty} \frac{1 + g_1(c)}{1 + g_2(c)} = 1.$$

Esta expressão pode ainda ser escrita como

$$r_{lh}(c) = \frac{c^{l+h-2}}{|c|^{l+h-2}} + f(c),$$

com $f(c)$ tal que $\lim_{|c| \rightarrow \infty} f(c) = 0$. □

Segue que $r_{lh}(c)$ pode ser próximo de 1 ou -1 se c é grande, sendo o sinal de $r_{lh}(c)$ o mesmo de c (ou $\bar{x}$), se $l+h-2$ for ímpar e positivo caso $l-h+2$ for par. Concluimos então que a correlação entre X^l e X^h , $l < h$, $l, h = 1, 2, \dots, p$ será numericamente próxima de

1 à medida que o valor de c aumenta. Desta forma, quando o valor de $\bar{x}$ for relativamente maior que o valor de $\sqrt{m_2}$, teremos problemas computacionais na inversão da matriz $\mathbf{X}'\mathbf{X}$.

Para ilustrar o fato, trabalharemos com $l = 1$ e $h = 2$, ou seja, vamos analisar a correlação entre as variáveis $\mathbf{X}$ e $\mathbf{X}^2$. Os resultados obtidos podem ser facilmente estendidos ao caso geral. De (3.27) e (3.28), segue que

$$\begin{aligned} c_{12} &= n \left(2m_2^{\frac{3}{2}}c + m_3 \right) \\ S_1 &= nm_2 \\ S_2 &= n \left(m_4 - m_2^2 + 4m_2^{\frac{1}{2}}m_3c + 4m_2^2c^2 \right). \end{aligned}$$

Utilizando (3.22), temos:

$$r_{12}(c) = \frac{2m_2^{\frac{3}{2}}c + m_3}{\left[m_2(m_4 - m_2^2 + 4m_2^{\frac{1}{2}}m_3c + 4m_2^2c^2) \right]^{\frac{1}{2}}} = \frac{2c + a_3}{(a_4 - 1 + 4a_3c + 4c^2)^{\frac{1}{2}}}, \quad (3.30)$$

onde a_3 e a_4 são respectivamente os 3º e 4º momentos amostrais padronizados de $\mathbf{x}$, dados por:

$$a_3 = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{m_2^{\frac{1}{2}}} \right)^3 = \frac{m_3}{m_2^{\frac{3}{2}}} \quad (3.31)$$

$$a_4 = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{m_2^{\frac{1}{2}}} \right)^4 = \frac{m_4}{m_2^2}. \quad (3.32)$$

De (3.30), temos que para $c = \frac{-a_3}{2}$, $r_{12} = 0$. Lembrando que $c = \frac{\bar{x}}{m_2^{\frac{1}{2}}}$, esta condição implica em $\bar{x} = \frac{-m_3}{2m_2}$. Desta forma, se os dados forem planejados de modo que $\bar{x} = \frac{-m_3}{2m_2}$, as variáveis $\mathbf{X}$ e $\mathbf{X}^2$ serão não correlacionadas. Este planejamento é equivalente à centralização quando os valores de $\mathbf{x}$ são distribuídos simetricamente em torno de $\bar{x}$. Neste caso, o 3º momento da variável centrada, m_3 e a média amostral da variável independente são nulos, valendo portanto a igualdade $\bar{x} = \frac{-m_3}{2m_2}$. No entanto, quando $c = 0$ mas $m_3 \neq 0$, temos que $r_{12} = \frac{a_3}{\sqrt{a_4 - 1}}$, e esse valor da correlação pode ser muito distante de zero. Dessa forma, a centralização merece um estudo mais aprofundado.

No ajuste de modelos de regressão polinomial de grau k , valores das variáveis independentes $\mathbf{x}_i, \mathbf{x}_i^2, \dots, \mathbf{x}_i^k$ podem ser escolhidos aleatoriamente de várias formas distintas.

Para cada seleção, teremos diferentes valores dos momentos amostrais; conseqüentemente, também teremos valores diferentes para $r_{12}(c)$ para um dado c . Suponha que os valores de $\mathbf{x}$ sejam aleatoriamente selecionados em um intervalo de acordo com uma distribuição Uniforme. Assim, utilizando a expressão (3.31) teremos $a_3 = 0$ e $a_4 = 1,8$, que são os momentos padronizados da distribuição Uniforme.

Assim, temos que

$$r_{12}^2(c) = \frac{1}{1 + \frac{1}{5c^2}}.$$

A Tabela 3.19 exibe os valores de $r_{12}^2(c)$ para alguns valores de c . Notamos que $|r_{12}(c)|$ se aproxima da unidade quando a média padronizada de $\mathbf{x}$, igual a c , difere de zero por dois ou mais desvios-padrão. O mesmo efeito pode ser notado para o caso em que os valores de $\mathbf{x}$ são escolhidos igualmente espaçados em qualquer intervalo. Por dados igualmente espaçados, entendemos que

$$x_{i+1} - x_i = \delta, \quad i = 1, 2, \dots, n-1, \quad (3.33)$$

para algum δ .

Tabela 3.19: Valores de $r_{12}^2(c)$ para $\mathbf{X}$ com distribuição Uniforme

c	1	2	3
$r_{12}^2(c)$	0,83	0,95	0,98

Como segundo exemplo, suponha que os valores de $\mathbf{x}$ sejam escolhidos aleatoriamente de uma distribuição Normal. Daí, $a_3 = 0$, $a_4 = 3$ e $\bar{x} = c\sqrt{m_2}$. Segue de (3.30) que

$$r_{12}^2(c) = \frac{1}{1 + \frac{1}{2c^2}}.$$

É importante notar que a Tabela 3.20 apresenta resultados similares aos que foram apresentados na Tabela 3.19. Através do gráfico apresentado na Figura 3.3, temos idéia do comportamento de $r_{12}^2(c)$ em função dos valores de c .

Tabela 3.20: Valores de $r_{12}^2(c)$ para $\mathbf{X}$ com distribuição Normal

c	1	2	3
$r_{12}^2(c)$	0,67	0,94	0,97

Figura 3.3: Comportamento de $r_{12}^2(c)$ em função dos valores de c .

Ao centralizarmos, a variável independente $\mathbf{X}$ do modelo (2.1) terá média zero e como consequência, $c = 0$ e das expressões (3.22), (3.27) e (3.28), temos que:

$$r_{lh}(0) = \frac{m_{l+h} - m_l m_h}{\sqrt{(m_{2l} - m_l^2)(m_{2h} - m_h^2)}}. \quad (3.34)$$

Analisando-se essa expressão, verifica-se que os momentos amostrais de $\mathbf{x}$ podem ser tais que nenhum valor de $r_{lh}(0)$ seja zero para $l < h$, $l, h = 1, 2, \dots, p$. Se além disso os valores de $\mathbf{x}$ estão simetricamente distribuídos em torno de $\bar{x}$, então $r_{lh}(0) = 0$ quando $l + h$ for ímpar, pois neste caso, ou l ou h será ímpar, implicando em $m_l m_h = 0$. Para o caso em que $l + h$ é par e l e h são ambos pares, a expressão dada por (3.34) pode ser muito diferente de zero. Por outro lado, quando $l + h$ é par e ambos l e h são ímpares, então valores de $\mathbf{x}$ simetricamente distribuídos em torno de $\bar{x}$ levam à seguinte expressão:

$$r_{lh}(0) = \frac{m_{l+h}}{\sqrt{m_{2l} m_{2h}}}. \quad (3.35)$$

Dando continuidade ao exemplo em que os valores de $\mathbf{x}$ são aleatoriamente selecionados de uma distribuição Uniforme, caso em que ocorre simetria, verifica-se que

$$\frac{m_{2l}}{m_2^l} = \frac{3^l}{2a + 1}, \quad a = 1, 2, \dots$$

Utilizando a expressão (3.35), encontraremos $r_{13}(0) = 0,92$, $r_{15}(0) = 0,82$ e $r_{35}(0) = 0,97$ e com a expressão (3.34), temos $r_{24}(0) = 0,96$ e $r_{26}(0) = 0,9$.

De forma semelhante, se os momentos amostrais forem padronizados e selecionados de uma distribuição Normal, teremos:

$$\frac{m_{2l}}{m_2^l} = \frac{2l!}{l!2^l}$$

e como consequência, $r_{13}(0) = 0,77$, $r_{15}(0) = 0,49$, $r_{35}(0) = 0,88$, $r_{24}(0) = 0,87$ e $r_{26}(0) = 0,63$.

Em ambos os exemplos, se $h = l + 2\lambda$, sendo que $\lambda \in \mathbf{Z}_+$, então

$$\lim_{l \rightarrow \infty} r_{l, l+2\lambda}(0) = 1.$$

Portanto, nos casos em que $\mathbf{x}$ for escolhido de forma aproximadamente Uniforme ou igualmente espaçados (conforme definido em (3.33)), ter-se-à problemas de multicolinearidade, mesmo quando feita a centralização.

De forma geral, os autores recomendam:

- Selecionar os valores da variável independente de modo que sua média seja nula;
- Sempre que possível, selecionar os valores da variável independente simetricamente em torno de sua média;
- Selecionar os valores da variável independente equi-espaçados, pois este procedimento teria a vantagem da futura utilização de polinômios ortogonais. Esta recomendação é muito importante porque uma alta correlação entre $\mathbf{X}^l$ e $\mathbf{X}^h$ pode ocorrer quando $l + h$ é par, mesmo que satisfeitas as duas recomendações anteriores.

Neste capítulo vimos a definição, como detectar e propomos solução para os problemas impostos pela multicolinearidade em modelos de regressão polinomial e elencamos as principais causas da multicolinearidade. Nos propusemos a estudar o problema relacionado ao tipo de multicolinearidade que não é inerente ao modelo, caso em que há alta correlação natural entre as variáveis. Para detectar o problema da multicolinearidade, lançamos mão de ferramentas tais como $\|\mathbf{X}\|$, $R_{(j)}^2$, R_j^2 , TNR , κ e FIV . Tratamos, também, da difícil tarefa da escolha do grau do polinômio. Vimos que o grau do polinômio deve ser o menor possível, pois um modelo de grau alto pode levar a problemas quando se quer fazer previsões com o modelo (extrapolações). A seguir, falamos das principais transformações para redução de multicolinearidade e da centralização entre as potências da variável regressora $\mathbf{x}$. No próximo capítulo trataremos dos polinômios ortogonais, outra alternativa para a correção do problema de mal-condicionamento da matriz $\mathbf{X}$ e eliminação dos problemas de multicolinearidade.

Capítulo 4

Polinômios ortogonais

4.1 Introdução

Este capítulo tem por objetivo descrever a utilização de polinômios ortogonais, altamente indicados para solução dos problemas relacionados com multicolinearidade e fazer comparações entre os *Modelos Polinomiais Ordinários* e os *Modelos Ortogonais*. Para a implementação dos polinômios ortogonais, dois casos precisam ser considerados. São eles:

1. Valores das variáveis independentes igualmente espaçados, com frequências iguais, considerado por FORSYTHE (1957), GRANDAGE (1958) e ROBSON (1959);
2. Valores das variáveis independentes desigualmente espaçados, com frequências também desiguais, problema considerado por GUEST (1950) e WISHART e METAKIDES (1953).

Geralmente não se sabe o grau do polinômio que melhor descreverá os dados de

sua pesquisa de forma adequada. Portanto, além de estimar os parâmetros do modelo, o pesquisador tem a preocupação adicional de determinar o grau do polinômio do modelo. Quanto maior o grau do polinômio, melhor se torna a curva ajustada aos dados; em contrapartida, conforme já comentado, o modelo encontrado pode ser pouco confiável, não servindo para realizar predições ou extrapolações devido à falta de precisão.

NARULA (1979) descreve os problemas de se ajustar modelos de regressão polinomial e traz como solução a implementação dos polinômios ortogonais. O artigo faz uma série de comparações entre os dois modelos, mostrando as desvantagens dos modelos polinomiais ordinários e as vantagens dos modelos ortogonais. Verifica-se que com relação aos testes de significância dos parâmetros, polinômios ordinários não são independentes, ao passo que os polinômios ortogonais são. Além disso, no ajuste dos modelos Polinomiais ordinários, as estimativas dos parâmetros dependem do grau do polinômio, ao passo que nos modelos ortogonais não. Por último, o tempo computacional gasto pelos modelos polinomiais ordinários é duas vezes maior que os modelos ortogonais. Em geral, pesquisas teóricas e empíricas apontam para a superioridade computacional e acurácia dos polinômios ortogonais em detrimento dos polinômios ordinários. Embora superiores, polinômios ortogonais não tem sido usados no ajuste de modelos estatísticos devido à sua complexidade. Descrevemos a seguir o ajuste de polinômios ortogonais na sua forma mais comum, que admite valores de X igualmente espaçados. Posteriormente, apresentaremos a formulação geral, proposta por NARULA (1979).

4.2 Polinômios ortogonais

Consideremos o modelo de regressão polinomial dado em (2.2). Conforme discutido anteriormente, pode haver correlação entre as variáveis X^l e X^h , $l \neq h$, $l, h \leq k$. Além disso, a introdução de um novo termo $\beta_{k+1}X^{k+1}$ exigiria um novo reajuste do modelo e novas estimativas $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ seriam obtidas. Agora suponha o modelo

$$Y_i = \alpha_0 + \alpha_1 P_1(x_i) + \alpha_2 P_2(x_i) + \dots + \alpha_k P_k(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (4.1)$$

de forma que

$$P_r(x_i) = a_{r,r}x_i^r + a_{r,r-1}x_i^{r-1} + \dots + a_{r,0},$$

seja um polinômio ortogonal de r -ésima ordem, isto é, tal que

$$\sum_{i=1}^n P_r(x_i)P_s(x_i) = 0, \quad P_0(x_i) = 1, \quad r \neq s \quad r = s = 0, 1, 2, \dots, k. \quad (4.2)$$

Assim, o modelo na forma matricial é dado por $\mathbf{y} = \mathbf{X}\beta + \epsilon$ onde

$$\beta = \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix} \quad e \quad \mathbf{X} = \begin{pmatrix} P_0(x_1) & P_1(x_1) & \cdots & P_k(x_1) \\ P_0(x_2) & P_1(x_2) & \cdots & P_k(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ P_0(x_n) & P_1(x_n) & \cdots & P_k(x_n) \end{pmatrix}$$

tem colunas mutuamente ortogonais. Segue que a matriz $\mathbf{X}'\mathbf{X}$ será dada por:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum_{i=1}^n P_0^2(x_i) & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n P_1^2(x_i) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sum_{i=1}^n P_k^2(x_i) \end{pmatrix}.$$

Afim de simplificar a notação, sejam $A_{jj} = \sum_{i=1}^n P_j^2(x_i)$, permitindo-nos reescrever a expressão para $\mathbf{X}'\mathbf{X}$ da seguinte forma:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} A_{00} & 0 & \cdots & 0 \\ 0 & A_{11} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{kk} \end{pmatrix}.$$

Essa estrutura ortogonal de $\mathbf{X}$ implica que a estimativa de mínimos quadrados de α_j , para $j = 1, 2, \dots, k$, é independente do grau k do polinômio. Os estimadores de mínimos quadrados de β são encontrados através da solução da equação $(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$, ou seja,

$$\hat{\alpha}_j = \frac{\sum_{i=1}^n P_j(x_i) y_i}{\sum_{i=1}^n P_j^2(x_i)} = \frac{A_{yj}}{A_{jj}}, \quad j = 0, 1, 2, \dots, k. \quad (4.3)$$

Visto que $\text{Var}(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-1} \sigma^2$, então $\text{Var}(\hat{\alpha}_j) = \frac{\sigma^2}{A_{jj}}$. Como $P_0(x_i) = 1$, então $\hat{\beta}_0 = \bar{y}$. Também teremos as somas de quadrados do resíduo e da regressão dadas por:

$$SQE(k) = S_{yy} - \sum_{j=1}^k \hat{\alpha}_j \sum_{i=1}^n P_j(x_i) y_i \quad (4.4)$$

$$SQR(\alpha_j) = \hat{\alpha}_j \sum_{i=1}^n P_j(x_i) y_i \quad (4.5)$$

$$F_0 = \frac{SQR(\alpha_j)}{\frac{SQE(k)}{n-k-1}} = \frac{\hat{\alpha}_j \sum_{i=1}^n P_j(x_i) y_i}{\frac{SQE(k)}{n-k-1}}. \quad (4.6)$$

onde F_0 é a estatística F para testar a hipótese $H_0 : \beta_j = 0$. Conforme dito anteriormente, uma propriedade interessante do ajuste de modelos de regressão via polinômios ortogonais é que os coeficientes $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ não dependem do grau do polinômio. Com isso, se desejarmos ajustar modelos de ordem $k + l$, apenas os l novos coeficientes devem ser estimados.

Para $k = 5$, os polinômios ortogonais são facilmente computados através das

seguintes equações:

$$\begin{aligned}
P_0(x_i) &= 1 \\
P_1(x_i) &= \lambda_1 \left(\frac{x_i - \bar{x}}{d} \right) \\
P_2(x_i) &= \lambda_2 \left[\left(\frac{x_i - \bar{x}}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right] \\
P_3(x_i) &= \lambda_3 \left[\left(\frac{x_i - \bar{x}}{d} \right)^3 - \left(\frac{x_i - \bar{x}}{d} \right) \left(\frac{3n^2 - 7}{20} \right) \right] \\
P_4(x_i) &= \lambda_4 \left[\left(\frac{x_i - \bar{x}}{d} \right)^4 - \left(\frac{x_i - \bar{x}}{d} \right)^2 \left(\frac{3n^2 - 13}{14} \right) + \frac{3(n^2 - 1)(n^2 - 9)}{560} \right]
\end{aligned}$$

de forma que d seja a distância entre os níveis de $\mathbf{x}$.

Exemplo

Este exemplo foi extraído de MONTGOMERY, PECK e VINNING (2001), página 228 e os dados estão contidos na Tabela 4.1. Trata-se do problema de investigar os custos de inventário $\mathbf{Y}$ em função da variável $\mathbf{X}$, quantidade de operações executadas. Para isso, o analista desenvolveu um simulador. Sabe-se, entretanto, que o custo anual do inventário é uma função convexa da quantidade de operações. Decidiu-se ajustar um modelo polinomial de segunda ordem através de polinômios ortogonais,

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \alpha_2 P_2(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, 10.$$

Aqui, $d = 25$, pois os valores de $\mathbf{x}$ são igualmente espaçados. Para o cálculo do coeficiente

do polinômio ortogonal P_1 , temos:

$$\begin{aligned}
P_1(x_1) = P_1(50) &= \lambda_1 \cdot \left(\frac{50 - 162,5}{25} \right) = \lambda_1 \cdot \frac{-9}{2}; \\
P_1(x_2) = P_1(75) &= \lambda_1 \cdot \left(\frac{75 - 162,5}{25} \right) = \lambda_1 \cdot \frac{-7}{2}; \\
P_1(x_3) = P_1(100) &= \lambda_1 \cdot \left(\frac{100 - 162,5}{25} \right) = \lambda_1 \cdot \frac{-5}{2}; \\
P_1(x_4) = P_1(125) &= \lambda_1 \cdot \left(\frac{125 - 162,5}{25} \right) = \lambda_1 \cdot \frac{-3}{2}; \\
P_1(x_5) = P_1(150) &= \lambda_1 \cdot \left(\frac{150 - 162,5}{25} \right) = \lambda_1 \cdot \frac{-1}{2}; \\
P_1(x_6) = P_1(175) &= \lambda_1 \cdot \left(\frac{175 - 162,5}{25} \right) = \lambda_1 \cdot \frac{1}{2}; \\
P_1(x_7) = P_1(200) &= \lambda_1 \cdot \left(\frac{200 - 162,5}{25} \right) = \lambda_1 \cdot \frac{3}{2}; \\
P_1(x_8) = P_1(225) &= \lambda_1 \cdot \left(\frac{225 - 162,5}{25} \right) = \lambda_1 \cdot \frac{5}{2}; \\
P_1(x_9) = P_1(250) &= \lambda_1 \cdot \left(\frac{250 - 162,5}{25} \right) = \lambda_1 \cdot \frac{7}{2}; \\
P_1(x_{10}) = P_1(275) &= \lambda_1 \cdot \left(\frac{275 - 162,5}{25} \right) = \lambda_1 \cdot \frac{9}{2}.
\end{aligned}$$

Obviamente, tomando $\lambda_1 = 2$, teremos

$$\begin{aligned}
P_1(x_1) &= -9; \\
P_1(x_2) &= -7; \\
P_1(x_3) &= -5; \\
P_1(x_4) &= -3; \\
P_1(x_5) &= -1; \\
P_1(x_6) &= 1; \\
P_1(x_7) &= 3; \\
P_1(x_8) &= 5; \\
P_1(x_9) &= 7; \\
P_1(x_{10}) &= 9.
\end{aligned}$$

De modo análogo encontramos os valores para $P_2(x_i)$ para $i = 1, 2, \dots, 10$. Na Tabela 4.2 encontramos os valores dos polinômios ortogonais.

O modelo ajustado é $\hat{y} = 324,30 + 0,7424.P_1(x) + 2,7955.P_2(x)$, todos os coeficientes do modelo ajustado são altamente significantes e o coeficiente de explicação obtido foi $R^2 = 0,9785$. Os resultados da Análise de Variância estão na Tabela 4.3, onde podemos concluir que o modelo quadrático ajustado é satisfatório. Para voltar ao modelo original, basta proceder conforme descreveremos a seguir:

$$\hat{y} = 324,30 + 0,7424.P_1(x) + 2,7955.P_2(x),$$

que se reduz a

$$\begin{aligned}\hat{y} &= 324,30 + 0,7424.(2). \left(\frac{x - 162,5}{25} \right) + 2,7955.\frac{1}{2}. \left[\left(\frac{x - 162,5}{25} \right)^2 - \left(\frac{10^2 - 1}{12} \right) \right] \\ \hat{y} &= 312,7686 + 0,0594(x - 162,5) + 0,0022(x - 162,5)^2.\end{aligned}$$

Tabela 4.1: Simulação de Inventário Anual

Quantidade Reordenada	Custo Médio Anual (y)
50	335
75	326
100	316
125	313
150	311
175	314
200	318
225	328
250	337
275	345

Tabela 4.2: Coeficientes dos polinômios ortogonais

i	$P_0(x_i)$	$P_1(x_i)$	$P_2(x_i)$
1	1	-9	6
2	1	-7	2
3	1	-5	-1
4	1	-3	-3
5	1	-1	-4
6	1	1	-4
7	1	3	-3
8	1	5	-1
9	1	7	2
10	1	9	6
$\sum_{i=1}^{10} P_0^2(x_i) = 10$ $\sum_{i=1}^{10} P_1^2(x_i) = 330$ $\sum_{i=1}^{10} P_2^2(x_i) = 132$			
		$\lambda_1 = 2$	$\lambda_2 = \frac{1}{2}$

Tabela 4.3: Análise de Variância

Fonte	Soma de Quadrados	gl	Quadrado Médio	Estatística F
Regressão	1.213,43	2	606,72	159,24
Termo Linear	181,89	1	181,89	47,74
Termo Quadrático	1.031,52	1	1.031,52	270,75
Resíduo	26,68	7	3,81	
Total	1.240,10	9		

4.3 Polinômios ortogonais para espaçamento e frequências desiguais

NARULA (1979) trata do problema da regressão polinomial ortogonal para o caso em que os valores da variável independente não são igualmente espaçados e são observados com frequências desiguais.

Considere $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_l$ representando os l valores de uma variável independente e $q_1, q_2, \dots, q_l$ suas respectivas frequências. A idéia básica do procedimento é utilizar como variável resposta

$$Y_i = \sum_{j=1}^{q_i} Y_{ij},$$

onde Y_{ij} é o valor da variável resposta $\mathbf{Y}$ associada ao i -ésimo valor de $\mathbf{X}$, $\mathbf{X}_i$, $i = 1, 2, \dots, l$. Assim, sejam $Y_1, Y_2, \dots, Y_l$ os novos valores da variável resposta. Então a equação de regressão polinomial é dada por:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_m X_i^m + \varepsilon_i, \quad i = 1, 2, \dots, l > m, \quad (4.7)$$

sendo que m é o grau do polinômio proposto. Podemos escrever as equações em (4.7) da seguinte forma:

$$Y_i = \beta_0 + \beta_1 \xi_{1i} + \beta_2 \xi_{2i} + \dots + \beta_m \xi_{mi} + \varepsilon_i, \quad i = 1, 2, \dots, l > m, \quad (4.8)$$

com

$$\xi_{ri} = a_{r,r} + a_{r,r-1} X_i + \dots + a_{r,1} X_i^{r-1} + X_i^r, \quad r = 1, 2, \dots, m < l, \quad i = 1, 2, \dots, l, \quad (4.9)$$

sendo um polinômio de grau r em X_i . Os termos ξ_{ri} , $r = 1, 2, \dots, m$ representam os coeficientes do efeito de r -ésima ordem. Também, ξ_r para $r = 1, 2, \dots, m$ representa os coeficientes ortogonais. Portanto, devem satisfazer às relações a seguir:

$$\sum_i \xi_{ri} q_i = 0 \quad r = 1, 2, \dots, m \quad (4.10)$$

e

$$\sum_i \xi_{ri} \xi_{si} q_i = 0 \quad r \neq s = 1, 2, \dots, m. \quad (4.11)$$

O objetivo é determinar os valores de ξ_r para $r = 1, 2, \dots, m$ quando os valores de $\mathbf{X}_i$ não são equi-espaçados e não ocorrem com frequência constante. O procedimento de obtenção desses valores foi apresentado com base em um exemplo.

Exemplo

Seja $\mathbf{Y}$ a variável aleatória que representa o custo de manutenção de tratores e $\mathbf{X}'$ a idade do trator. É razoável supor que quanto mais velho o trator, maior será seu custo de manutenção. Para este exemplo, foram coletados 10 observações e os valores de $\mathbf{X}'_i$ estão contidos na Tabela 4.4. Temos 4 valores para a variável independente $\mathbf{X}'_i$ e foi feita a transformação

$$X_i = \frac{X'_i - 4}{\frac{1}{2}}.$$

A Tabela 4.4 apresenta os valores de $\mathbf{X}'_i$ e sua respectiva frequência. Matematicamente, escrevemos:

- $q_1 = q_2 = 3$;
- $q_3 = q_4 = 2$.

Para obter os valores de ξ_r , inicialmente devemos resolver as equações para $a_{r,j}$ em (4.9) recursivamente usando as propriedades fundamentais (4.10) e (4.11) dos polinômios ortogonais. Substituindo-se os valores encontrados de $a_{r,j}$ em (4.9), obtem-se ξ_{ri} .

Tabela 4.4: Valores referentes à idade de tratores

X_i'	4,0	4,5	5,0	6,0
q_i	3	3	2	2

Tabela 4.5: Valores observados da idade de tratores transformadas

X_i	0	1	2	4
q_i	3	3	2	2

Coeficientes dos polinômios ortogonais para a regressão linear

Sejam $\xi_{1i} = a_{11} + X_i$ os coeficientes para a regressão linear. A Tabela 4.6 exibe os cálculos necessários para encontrar os valores de ξ_{1i} . As etapas são as seguintes:

- Escrever os valores de X_i na coluna (1) e os valores correspondentes de q_i na coluna (2);
- Substituir os valores de X_i sucessivamente em ξ_{1i} e registrar os resultados na coluna (3);
- Multiplicar os valores de q_i da coluna (2) por seu correspondente ξ_{1i} na coluna (3) e registrar os resultados na coluna (4).

Uma das propriedades fundamentais de ξ_1 é que $\sum_i \xi_{1i} q_i = 0$, que corresponde à soma dos elementos da coluna (4). Isso leva a $a_{11} = \frac{-3}{2}$. Substituindo esse valor de a_{11} na coluna (3), obtem-se os valores ξ_{1i} da coluna (5). Em seguida, aplicando-se simplificações adequadas, chega-se aos resultados contidos na coluna (6). No caso, dividiu-se os elementos da coluna (5) por $\frac{1}{2}$.

Portanto, observamos que trata-se de um método simples e fácil para se encontrar os coeficientes ortogonais da regressão polinomial.

Tabela 4.6: Procedimento para determinar os coeficientes ortogonais da regressão linear

X_i	q_i	$\xi_{1i} = a_{11} + X_i$	$\xi_{1i}q_i$	ξ_{1i}	ξ'_{1i}
(1)	(2)	(3)	(4)	(5)	(6)
0	3	a_{11}	$3a_{11}$	$-\frac{3}{2}$	-3
1	3	$a_{11} + 1$	$3(a_{11} + 1)$	$-\frac{1}{2}$	-1
2	2	$a_{11} + 2$	$2(a_{11} + 2)$	$\frac{1}{2}$	1
4	2	$a_{11} + 4$	$2(a_{11} + 4)$	$\frac{5}{2}$	5

Coeficientes dos polinômios ortogonais para regressão quadrática

Sejam $\xi_{2i} = a_{22} + a_{21}X_i + X_i^2$ os coeficientes para a regressão quadrática. Na Tabela 4.7 encontram-se os cálculos necessários para determinar os valores de ξ_{2i} . As etapas são as seguintes:

1. Substituir os valores de X_i em ξ_{2i} sucessivamente e registrar os valores de ξ_{2i} na coluna (3) da Tabela 4.7
2. A coluna (4) é obtida multiplicando os valores de q_i da coluna (2) por seu correspondente ξ_{2i} na coluna (3)
3. Escrever os valores de ξ_{1i} obtidos anteriormente na coluna (5).

Valem as expressões $\sum_i \xi_{2i}q_i = 0$, $\sum_i \xi_{1i}\xi_{2i}q_i = 0$, ou $\sum_i \xi'_{1i}\xi_{2i}q_i = 0$. Dessa última igualdade, encontramos $a_{21} = \frac{-165}{41}$. Substituindo esse valor na coluna (4) e utilizando-se a primeira igualdade, encontra-se $a_{22} = \frac{356}{205}$. Usando a_{22} e a_{21} em ξ_{2i} da coluna (3), obteremos os resultados da coluna (7). Simplificando os resultados da coluna (7), teremos os valores de ξ'_{2i} da coluna (8).

Tabela 4.7: Procedimento para determinar os coeficientes ortogonais da regressão quadrática

X_i	q_i	ξ_{2i}	$\xi_{2i}q_i$	ξ'_{1i}	$\xi'_{1i}\xi_{2i}q_i$	ξ_{2i}	ξ'_{2i}
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	3	a_{22}	$3a_{22}$	-3	$-9a_{22}$	$\frac{356}{205}$	178
1	3	$a_{22} + a_{21} + 1$	$3(a_{22} + a_{21} + 1)$	-1	$-3(a_{22} + a_{21} + 1)$	$-\frac{264}{205}$	-132
2	2	$a_{22} + 2a_{21} + 4$	$2(a_{22} + 2a_{21} + 4)$	1	$2(a_{22} + 2a_{21} + 4)$	$-\frac{474}{205}$	-237
4	2	$a_{22} + 4a_{21} + 16$	$2(a_{22} + 4a_{21} + 16)$	5	$10(a_{22} + 4a_{21} + 16)$	$\frac{336}{205}$	168

Coeficientes dos polinômios ortogonais para a regressão cúbica

Os coeficientes ortogonais $\xi_{3i} = a_{33} + a_{32}X_i + a_{31}X_i^2 + X_i^3$ da regressão cúbica são encontrados de forma similar à dos coeficientes para a regressão linear e quadrática discutidos anteriormente. As relações a seguir devem ser verificadas:

$$\begin{aligned}\sum_i \xi_{3i}q_i &= 0 \\ \sum_i \xi_{1i}\xi_{3i}q_i &= 0 \\ \sum_i \xi_{2i}\xi_{3i}q_i &= 0.\end{aligned}$$

Estas relações se reduzem ao sistema de equações:

$$\begin{cases} 147 + 43a_{31} + 15a_{32} + 10a_{33} = 0 \\ 653 + 165a_{31} + 41a_{32} = 0 \\ 17316 + 3084a_{31} = 0 \end{cases}$$

Ao resolvermos este sistema, obtemos

$$\begin{aligned} a_{31} &= \frac{-1.443}{257} \\ a_{32} &= \frac{70.274}{10.537} \\ a_{33} &= \frac{-5.904}{10.537}. \end{aligned}$$

Após substituírmos esses valores em ξ_{3i} , teremos

$$\begin{aligned} \xi_{31} &= \frac{-5.904}{10.537} \\ \xi_{32} &= \frac{15.744}{10.537} \\ \xi_{33} &= \frac{-17.712}{10.537} \\ \xi_{34} &= \frac{2.952}{10.537}. \end{aligned}$$

Multiplicando-se esses valores por $\frac{10537}{984}$, teremos, finalmente,

$$\begin{aligned} \xi_{31} &= -6 \\ \xi_{32} &= 16 \\ \xi_{33} &= -18 \\ \xi_{34} &= 3. \end{aligned}$$

Finalizando o trabalho, NARULA (1979) comenta que apesar de sua simplicidade, o método pode gerar cálculos bastante extensos. Para evitar o problema, pode-se utilizar o procedimento que descreveremos a seguir que possibilita o cálculo de ξ_{rj} sem a necessidade de se calcular os termos $a_{r,j}$ através da expressão abaixo, levando em consideração que os valores de $\mathbf{X}$ são igualmente espaçados:

$$\xi_{rj} = X_j^r - \sum_{p=0}^{r-1} \xi_{pj} \frac{\sum_i \xi_{pi} X_i^r q_i}{\sum_i \xi_{pi} X_i^p q_i}, \quad (4.12)$$

para $r = 1, 2, \dots, m$, $j = 1, 2, \dots, l$ e $\xi_{0j} = 1$ para $j = 1, 2, \dots, l$.

A seguir, calcularemos para o exemplo os coeficientes ortogonais da regressão

linear, quadrática e cúbica utilizando este procedimento. Para obter os coeficientes ortogonais da regressão linear, tomamos $r = 1$ na equação (4.12), reduzindo-a à equação:

$$\xi_{1j} = X_j - \frac{\sum_i X_i q_i}{\sum_i q_i}, \quad j = 1, 2, 3, 4. \quad (4.13)$$

Para calcular os coeficientes ortogonais da regressão quadrática, tomamos $r = 2$ na equação (4.12), reduzindo-a à equação a seguir:

$$\xi_{2j} = X_j^2 - \xi_{1j} \frac{\sum_i \xi_{1i} X_i^2 q_i}{\sum_i \xi_{1i} X_i q_i} - \frac{\sum_i X_i^2 q_i}{\sum_i q_i}, \quad j = 1, 2, 3, 4. \quad (4.14)$$

Por último, para calcular os coeficientes ortogonais da regressão cúbica, tomamos $r = 3$ na equação (4.12), obtendo:

$$\xi_{3j} = X_j^3 - \xi_{2j} \frac{\sum_i \xi_{2i} X_i^3 q_i}{\sum_i \xi_{2i} X_i^2 q_i} - \xi_{1j} \frac{\sum_i \xi_{1i} X_i^3 q_i}{\sum_i \xi_{1i} X_i q_i} - \frac{\sum_i X_i^3 q_i}{\sum_i q_i}, \quad j = 1, 2, 3, 4.$$

Os resultados dos cálculos para a regressão linear e quadrática estão presentes nas Tabelas 4.8 e 4.9 respectivamente. No caso da regressão cúbica, chegamos aos valores a seguir:

$$\xi_{31} = \frac{-5904}{10537}, \quad \xi_{32} = \frac{15744}{10537}, \quad \xi_{33} = \frac{-17712}{10537}, \quad \xi_{34} = \frac{2952}{10537}.$$

Note que os resultados são exatamente os mesmos encontrados anteriormente.

Tabela 4.8: Cálculo dos Coeficientes Lineares

X_j	q_j	$X_j q_j$	$\xi_{ij} = X_j - \frac{\sum_i X_i q_i}{\sum_i q_i}$
0	3	0	$0 - \frac{15}{10} = -\frac{3}{2}$
1	3	3	$1 - \frac{15}{10} = -\frac{1}{2}$
2	2	4	$2 - \frac{15}{10} = \frac{1}{2}$
4	2	8	$4 - \frac{15}{10} = \frac{5}{2}$
Total	10	15	

Tabela 4.9: Cálculo dos coeficientes Quadráticos

X_j	q_j	$X_j^2 q_j$	$\xi_{1j} X_j^2 q_j$	$\xi_{1j} X_j^2 q_j$	$\xi_{2j} = X_j^2 - \xi_{1j} \frac{\sum_i \xi_{1i} X_i^2 q_i}{\sum_i \xi_{1i} X_i q_i} - \frac{\sum_i X_i^2 q_i}{\sum_i q_i}$
0	3	0	0	0	$0 - \left(\frac{-3}{2}\right) \left(\frac{165}{41}\right) - \frac{43}{10} = \frac{356}{205}$
1	3	3	$\frac{-3}{2}$	$\frac{-3}{2}$	$1 - \left(\frac{-1}{2}\right) \left(\frac{165}{41}\right) - \frac{43}{10} = \frac{-264}{205}$
2	2	8	2	4	$4 - \left(\frac{1}{2}\right) \left(\frac{165}{41}\right) - \frac{43}{10} = \frac{-474}{205}$
4	2	32	20	80	$16 - \left(\frac{5}{2}\right) \left(\frac{165}{41}\right) - \frac{43}{10} = \frac{336}{205}$
Total	10	43	$\frac{41}{2}$	$\frac{165}{2}$	

4.4 Transformações Matriciais

Uma metodologia geralmente indicada para solucionar os problemas provenientes da multicolinearidade é a transformação matricial na matriz $\mathbf{X}$ a fim de se obter colunas ortogonais. Este procedimento é chamado de **Ortogonalização de Gram-Schmidt**. Seja $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m]$, com $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ m vetores n -dimensionais. O método consiste em obter um novo conjunto de vetores $\mathbf{Z}_{1T}, \mathbf{Z}_{2T}, \dots, \mathbf{Z}_{mT}$ tais que $\mathbf{Z}_{1T} = \mathbf{x}_1$ e para $j > 1$, $\mathbf{Z}_{jT}$ é a projeção ortogonal de $\mathbf{x}_j$ no espaço linear gerado por $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{j-1}$. CLAYTON (1971) verifica que este método é equivalente a escrever a matriz $\mathbf{X}$ como $\mathbf{X} = \mathbf{Z} \cdot \mathbf{C}$, onde $\mathbf{C}$ é uma matriz $m \times m$ triangular superior com elementos da diagonal principal iguais a 1; $\mathbf{Z}$ é a matriz $n \times m$ de vetores coluna $\mathbf{Z}_{jT}$, $j = 1, 2, \dots, m$ com $\mathbf{Z}'\mathbf{Z} = \mathbf{D}$, sendo $\mathbf{D}$ matriz diagonal.

No modelo de regressão linear múltipla, tal procedimento pode ser aplicado à matriz $\mathbf{X}$ para detectar se uma dada coluna é dependente de outra. Caso a dependência seja detectada, o algoritmo consiste em encontrar novas colunas através da transformação matricial. Seja

$$\mathbf{Z}_{iT} = \left[\mathbf{I} - \mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}' \right] \mathbf{Z}_i = \mathbf{Z}_i - \mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Z}_i, \quad (4.15)$$

onde $\mathbf{Z}$ é a matriz de vetores colunas transformados até o momento, $\mathbf{Z}_i$ é o próximo vetor coluna da matriz $\mathbf{X}$ que deverá ser transformada e $\mathbf{Z}_{iT}$ é o vetor ortogonal transformado resultante. Observa-se que $\mathbf{Z}_{iT}$ é um vetor de resíduos da regressão de $\mathbf{Z}_i$ nas colunas de $\mathbf{Z}$.

DRAPER e SMITH (1998) ilustra o uso dessa metodologia para $n = 5$ dados e um Modelo de regressão polinomial da forma $\mathbf{Y} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$. A matriz

$\mathbf{X}$ original é dada por

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \end{pmatrix}.$$

A partir desses fatos, o modelo original pode ser escrito como

$$\mathbf{y} = \mathbf{X}.\boldsymbol{\beta} + \varepsilon = \mathbf{Z}.\mathbf{C}.\boldsymbol{\beta} + \varepsilon = \mathbf{Z}.\boldsymbol{\beta}^* + \varepsilon,$$

com $\mathbf{Z}$ ortogonal e o vetor de estimadores de mínimos quadrados de $\boldsymbol{\beta}^*$ é

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{Z}'.\mathbf{Z})^{-1}.\mathbf{Z}'\mathbf{y}.$$

Do fato de que $\boldsymbol{\beta}^* = \mathbf{C}.\boldsymbol{\beta}$ segue que

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{-1}.\hat{\boldsymbol{\beta}}^*. \quad (4.16)$$

Dessa forma, é possível a volta ao modelo original $\mathbf{y} = \mathbf{X}.\boldsymbol{\beta} + \varepsilon$, sendo o modelo ajustado $\hat{\mathbf{y}} = \mathbf{X}.\hat{\boldsymbol{\beta}}$ para $\hat{\boldsymbol{\beta}}$ dado em (4.16).

CLAYTON (1971) apresenta um algoritmo e o correspondente programa computacional para a obtenção das matrizes $\mathbf{Z}$ e $\mathbf{C}$. Para ilustrar a importância das transformações matriciais a fim de tornar as colunas da matriz $\mathbf{X}$ ortogonais, sejam

$$\mathbf{I} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{A} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \mathbf{A}_2 = \begin{pmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \end{pmatrix}, \mathbf{A}_3 = \begin{pmatrix} 1 \\ 8 \\ 27 \\ 64 \\ 125 \end{pmatrix}.$$

A matriz de correlação entre as variáveis presentes na matriz $\mathbf{X} = [\mathbf{I}, \mathbf{A}, \mathbf{A}_2, \mathbf{A}_3]$ é dada na Tabela 4.10 e os autovalores de $\mathbf{X}'\mathbf{X}$ são apresentados na Tabela 4.11. Note que

o software não calculou as correlações entre as colunas **I** e qualquer outra coluna; além do mais, observe a grandeza do maior autovalor. Os valores de FIV_j para $j = 1, 2, 3$ são dados Tabela 4.12. Procedendo a transformação matricial a fim de termos ortogonalidade nas colunas da matriz **X**, obtemos a matriz **Z**, dada por

$$\mathbf{Z} = \begin{pmatrix} 1 & -2 & 2 & -1,2 \\ 1 & -1 & -1 & 2,4 \\ 1 & 0 & -2 & 0 \\ 1 & 1 & -1 & -2,4 \\ 1 & 2 & 2 & 1,2 \end{pmatrix}$$

Novamente, temos que

$$\mathbf{I} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}, \mathbf{B}_2 = \begin{pmatrix} 2 \\ -1 \\ -2 \\ -1 \\ 2 \end{pmatrix}, \mathbf{B}_3 = \begin{pmatrix} -1,2 \\ 2,4 \\ 0 \\ -2,4 \\ 1,2 \end{pmatrix}.$$

A matriz de correlação entre as variáveis presentes na matriz **Z** é dada na Tabela 4.13. Também, os autovalores são dados na Tabela 4.14 e finalmente, os valores de FIV_j , para $j = 1, 2, 3$ são dados na Tabela 4.15. É fácil ver que as colunas da matriz **Z** são ortogonais, eliminando a dependência linear que existia entre as colunas e resolvendo completamente o problema de multicolinearidade.

De acordo com DRAPER e SMITH (1998), este procedimento possui a vantagem de informar as colunas da matriz **X** que dependem linearmente das demais, pois de modo geral, feita a transformação, a correspondente coluna transformada terá muitos zeros.

Tabela 4.10: Matriz de correlação entre as variáveis da matriz $\mathbf{X}$

	A	A_2	A_3
A	1	0,9811049	0,9431175
A_2	0,9811049	1	0,9892158
A_3	0,9431175	0,9892158	1

Tabela 4.11: Autovalores da matriz $\mathbf{X}'\mathbf{X}$

21.520,05489	32,93118	0,99970	0,01423

Tabela 4.12: Valores de FIV_j para as variáveis da matriz $\mathbf{X}$

A	A_2	A_3
413,4921	2130,4643	721,5278

Tabela 4.13: Matriz de Correlação entre as variáveis da matriz $\mathbf{Z}$

	B	B_2	B_3
B	1	0	0
B_2	0	1	0
B_3	0	0	1

Tabela 4.14: Autovalores da matriz $\mathbf{Z}'\mathbf{Z}$

14,4	14	10

Tabela 4.15: Valores de FIV_j para as variáveis da matriz $\mathbf{Z}$

B	B_2	B_3
1	1	1

Neste capítulo descrevemos a utilização de polinômios ortogonais e elencamos as vantagens e desvantagens de sua utilização. Como se pôde observar, polinômios ortogonais podem ser obtidos através de fórmulas ou tabelas. Uma breve tabela de valores numéricos para os polinômios ortogonais podem ser obtidos em MONTGOMERY (2005) e, caso se queira uma tabela mais extensa, consultar DELURY (1960). Caso se queira calcular os polinômios ortogonais através de fórmulas, a referência sugerida é MONTGOMERY, PECK e VINING (2001). No capítulo 6 veremos a aplicação de polinômios ortogonais ao conjunto de dados. Veremos, também, que embora tenhamos descrito técnicas para situações de dados desigualmente espaçados, recursos presentes em pacotes computacionais modernos permitem o ajuste nessas situações de forma muito simples.

Capítulo 5

Modelos alternativos aos modelos de regressão polinomial

5.1 O uso de *splines* em regressão polinomial

A natureza oscilatória dos polinômios de ordem elevada e a instabilidade de estimação devido à multicolinearidade restringem a utilização de modelos de regressão polinomial. Por outro lado, muitas vezes, polinômios de grau baixo não fornecem um bom ajuste. Em outros casos, a função de regressão apresenta diferentes comportamentos em diferentes intervalos do conjunto de valores de $\mathbf{X}$. Um enfoque alternativo é dividir o intervalo em uma série de subintervalos e construir polinômios de graus (geralmente) diferentes para cada subintervalo. A aproximação mais comum faz uso de polinômios cúbicos e é chamada de *interpolação com spline cúbico*. Um polinômio cúbico genérico possui quatro constantes, de modo que o procedimento com *splines* cúbicos tem a flexibilidade suficiente para assegurar que o interpolador não só seja continuamente derivável no intervalo, mas também que ele tenha derivada de segunda ordem contínua.

Interpolação é uma técnica antiga e largamente utilizada em Cálculo Numérico, sendo empregada nos casos em que é realmente difícil calcular o valor da função ou quando não conhecemos a expressão da função, mas dispomos de um conjunto de valores que em geral são obtidos através de experimentos. Dessa forma, sendo fornecida uma série de dados (x_i, y_i) , para $i = 1, 2, \dots, n$, correspondentes aos valores de argumentos e valores de uma função f , definida em $[a, b]$ tal que $y = f(x)$, deseja-se obter os valores de $f(x_m)$, onde $x_m \neq x_i$, utilizando-se os pontos dados.

Definição 5.1 Dada uma função f definida em $[a, b]$ e um conjunto de nós $a = x_0 < x_1 < x_2 < \dots < x_n = b$, um **spline cúbico interpolador** S para f é uma função que satisfaz às seguintes condições:

1. $S(x)$ é um polinômio cúbico, indicado por $S_j(x)$, no subintervalo $[x_j, x_{j+1}]$ para cada $j = 0, 1, 2, \dots, n-1$;
2. $S_j(x_j) = f(x_j)$ para cada $j = 0, 1, 2, \dots, n$;
3. $S_{j+1}(x_{j+1}) = S_j(x_{j+1})$ para cada $j = 0, 1, 2, \dots, n-2$;
4. $S'_{j+1}(x_{j+1}) = S'_j(x_{j+1})$ para cada $j = 0, 1, 2, \dots, n-2$;
5. $S''_{j+1}(x_{j+1}) = S''_j(x_{j+1})$ para cada $j = 0, 1, 2, \dots, n-2$;
6. Uma das seguintes condições de contorno é satisfeita:
  - $S''(x_0) = S''(x_n)$, chamado de **contorno livre ou natural**;
  - $S'(x_0) = f'(x_0)$ e $S'(x_n) = f'(x_n)$, chamado de **contorno restrito**.

Quando as condições de contorno livre ocorrem, o **spline cúbico** é chamado de **spline natural** e seu gráfico se assemelha à forma que uma haste longa e flexível assumiria se fosse forçada a passar pelos pontos $(x_0, f(x_0)), (x_1, f(x_1)), \dots, (x_n, f(x_n))$. Em geral as condições de contorno restrito nos levam a aproximações mais precisas. Para utilizar esse tipo de contorno, faz-se necessário conhecer os valores das derivadas nos pontos extremos ou pelo menos termos uma aproximação acurada desses valores.

De acordo com MONTGOMERY, PECK e VINNING (2001) usualmente, os *splines* cúbicos são adequados na maioria dos problemas práticos. Além disso, sua utilização reduz o problema de mal condicionamento que é comum quando da utilização de outros tipos de *splines* na presença de uma grande quantidade de nós.

Para construir o *spline* cúbico interpolador para uma função f dada, as condições indicadas na definição 5.1 são aplicadas aos polinômios cúbicos

$$S_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3$$

para cada $j = 0, 1, 2, \dots, n - 1$.

O objetivo da interpolação é obter o valor de $f(x_m)$ aproximadamente. Para isso, construímos, a partir dos dados, uma nova função ϕ que interpola f , tal que:

1. $\forall x_i, x_0 \leq x_i \leq x_n \mapsto \phi(x_i) = f(x_i)$
2. $\forall x \in [x_0, x_n] \mapsto \phi(x_i) \cong f(x_i)$

Há vários tipos de função que interpolam a função f . A função ϕ que interpola a função f pertence a uma das seguintes famílias:

1. P: Polinômios: $y = a_0x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n$;
2. F: Fourier: $f_i(x) = a_i \cdot \cos(ix) + b_i \cdot \sin(ix)$;
3. E: Exponencial: $y = a \cdot e^{bx}$;
4. S: *Splines*, que são pedaços de polinômios.

Neste capítulo nos concentraremos, inicialmente, no uso dos *splines*, assunto esse bastante abordado em pesquisas recentes. A vantagem dos *splines* em relação aos outros métodos é que em certas aplicações precisamos de uma função interpolante que possa ser diferenciada a fim de se obter uma aproximação tão suave quanto possível. De forma simples, *splines* são polinômios de ordem k que interpolam um conjunto de dados e têm

$k - 1$ derivadas contínuas em todo o intervalo. Os *splines* devem passar pelos nós e não formar ângulos agudos. Supondo k nós; os denominaremos de $\xi_1, \xi_2, \dots, \xi_k$. Assim, *splines* são polinômios por partes de forma que os diferentes segmentos sejam unidos nos nós a fim de garantir continuidade. Matematicamente, um *spline* de ordem m com nós em $\xi_1, \xi_2, \dots, \xi_k$ é qualquer função da forma:

$$s(t) = \sum_{i=0}^{m-1} \theta_i t^i + \sum_{i=1}^k \delta_i (t - \xi_i)_+^{m-1} \quad (5.1)$$

onde $(x)_+ = \max(0, x)$ e os coeficientes $\theta_0, \dots, \theta_{m-1}, \delta_1, \dots, \delta_k$ são números reais. Da definição, vemos que um *spline* satisfaz as seguintes condições:

1. s é um polinômio por partes de ordem m em qualquer subintervalo $[\xi_i, \xi_{i+1})$;
2. s tem $m - 2$ derivadas contínuas;
3. s tem a $(m - 1)$ -ésima derivada e é uma função escada com saltos nos nós $\xi_1, \xi_2, \dots, \xi_k$.

Observa-se que para um *spline* de ordem m , assim definido, o grau do polinômio associado é no máximo $m - 1$.

Seja $S^m(\xi_1, \dots, \xi_k)$ o conjunto formado por todos os *splines* de ordem m com nós $\xi_1, \xi_2, \dots, \xi_k$ definidos anteriormente. Verifica-se que $S^m(\xi_1, \dots, \xi_k)$ é um espaço vetorial de dimensão $m + k$, desde que as funções

$$1, t, \dots, t^{m-1}, (t - \xi_1)_+^{m-1}, \dots, (t - \xi_k)_+^{m-1}$$

sejam linearmente independentes. Neste caso, os parâmetros

$$\theta_0, \theta_1, \dots, \theta_{m-1}, \delta_1, \delta_2, \dots, \delta_k$$

podem ser estimados por mínimos quadrados, ou seja, minimiza-se a função

$$\sum_{i=1}^n \left[y_i - (\theta_0 + \theta_1 t + \dots + \theta_{m-1} t^{m-1}) + \sum_{i=1}^k \delta_i (t - \xi_i)_+^m \right]^2.$$

Considere o modelo de regressão:

$$y_j = f(t_j) + \varepsilon_j, \quad j = 1, 2, \dots, n, \quad (5.2)$$

de forma que as variáveis aleatórias ε_j tenham média zero, sejam independentes e não correlacionadas, com variância comum dada por σ^2 . As observações consistem nos pares (t_j, y_j) sendo que $f(t_j)$ são valores de uma função f desconhecida calculada nos pontos $t_1, t_2, \dots, t_n$. A função f é chamada de função de regressão ou curva de regressão. Admite-se ainda que $f \in W_2^m[a, b]$, onde $W_2^m[a, b] = \{f : f' \text{ é absolutamente contínua e } \int_a^b [f^{(m)}]^2 < \infty \text{ com } f^{(m)} \text{ a } m\text{-ésima derivada da função } f\}$.

Nosso objetivo é encontrar um estimador para a curva de regressão f que se ajuste bem aos dados e que ao mesmo tempo possua um bom grau de suavidade. Segundo WAHBA (1987), uma possível estimativa da função f é o valor de f que minimiza o funcional

$$g(f) = \frac{1}{n} \sum_{i=1}^n [y_i - f(t_i)]^2 + \lambda \int_a^b [f^{(m)}(u)]^2 du. \quad (5.3)$$

DIAS (2002) define a medida da qualidade do ajuste que é dada por:

$$\frac{\sum_{i=1}^n [y_i - f(t_i)]^2}{n},$$

e uma medida de suavidade para f dada por $\int_a^b [f^{(m)}]^2$; combinando essas duas medidas, define o funcional $A_\alpha(f)$ com $\alpha \in [0, 1]$. Matematicamente, tem-se

$$A_\alpha(f) = (1 - \alpha) \sum_{i=1}^n [y_i - f(t_i)]^2 + \alpha \int_a^b [f^{(m)}]^2. \quad (5.4)$$

O objetivo é minimizar $A_\alpha(f)$. No entanto, tomando-se $\lambda = \frac{\alpha}{1-\alpha}$ teremos um problema equivalente de otimização, de forma a obtermos uma solução f_α que minimiza a função

$$A_\lambda(f) = \sum_{i=1}^n [y_i - f(t_i)]^2 + \lambda \int_a^b [f^{(m)}]^2, \quad \lambda > 0.$$

De acordo com o autor, a solução f_λ é o estimador de suavização por *splines* da função de regressão e λ é o parâmetro de suavização. Se λ for grande, maior importância estaremos

dando para o critério de suavização, ao passo que se λ for pequeno (próximo de zero), estaremos dando mais importância para a medida de qualidade do ajuste.

O modelo de regressão polinomial está relacionado com *splines* porque $f \in W_2^m[a, b]$ pode ser escrita como

$$f(t) = \sum_{i=0}^{m-1} \theta_i t^i + resto(t),$$

onde

$$resto(t) = \frac{1}{(m-1)!} \int_a^b f^{(m)}(x) (t-x)_+^{m-1} dx,$$

sendo que $\theta_0, \theta_1, \dots, \theta_{m-1}$ são constantes. Agora, o modelo (5.2) é equivalente a

$$y_j = \sum_{i=0}^{m-1} \theta_i t^i + resto(t) + \varepsilon_j, \quad j = 1, 2, \dots, n.$$

De acordo com DIAS (2002), se $resto(t_j)$ for próximo de zero, então é razoável estimar f através de um estimador de regressão polinomial de ordem m . Caso contrário, há dúvidas quanto a aplicabilidade deste modelo polinomial.

O autor mostra matematicamente a relação entre os modelos de regressão polinomial e os *splines*. Seja o funcional dado por

$$J_m(f) = \int_a^b [f^{(m)}]^2 dt.$$

Aplicando a Desigualdade de Cauchy-Schwartz na expressão $resto(t)$, temos que

$$resto(t)^2 \leq \frac{(b-a)^{2m-1}}{[(2m-1)(m-1)!]^2} J_m(f).$$

A partir desta expressão, notamos que existe uma constante C que depende de m e não de f , tal que

$$\sup_{t \in [a, b]} |resto(t)| \leq \sqrt{C J_m(f)}.$$

Essa expressão nos diz da relação de $J_m(f)$ com $\text{resto}(t)$, pois se $J_m(f)$ é pequeno, então temos grande credibilidade no modelo polinomial.

Para construção dos *splines*, precisamos escolher uma base adequada. Seja $\phi_1, \phi_2, \dots, \phi_n$ tal base de modo que $f_\lambda = \sum_{j=1}^n \theta_{\lambda_j} \phi_j(t)$. De acordo com DIAS (2002), a escolha conveniente para a base $\phi_1, \phi_2, \dots, \phi_n$ é a sequência formada pelos B-*splines* dados a seguir

$$N_{km}(x) = \sum_{j=0}^m \alpha_{jm}^{[k]} (t_{k+m} - x)_+^{m-1},$$

onde N_{km} são os B-*splines* de BOOR e os valores de α são coeficientes definidos em De BOOR (1978, página 108). Portanto, o problema de otimização anterior, ou seja, o de minimizar $A_\lambda(f)$, se reduz à minimizar

$$\|Y - X\theta\|^2 + \frac{\lambda}{2} \theta' \Omega \theta \quad (5.5)$$

onde $\Omega_{ij} = \int_a^b N_{i2}''(t) N_{j2}''(t) dt$, $Y = (y_1, y_2, \dots, y_n)'$, $\theta_\lambda = (\theta_{\lambda_1}, \theta_{\lambda_2}, \dots, \theta_{\lambda_n})'$ e $x_{ij} = N_{j2}(t_i)$ é o elemento de ordem ij de $\mathbf{X}$. A minimização de (5.5) tem uma única solução dada por

$$\theta_\lambda = [\mathbf{X}'\mathbf{X} + \lambda\Omega]^{-1} \mathbf{X}'\mathbf{Y}.$$

No entanto, esta expressão depende do parâmetro λ desconhecido. Segundo WAHBA (1987), uma possível forma de se estimar o parâmetro de suavização é através do método de validação cruzada. Neste caso, determina-se λ que minimiza

$$CV(\lambda) = \frac{1}{n} \sum_{j=1}^n [y_j - f_{\lambda(j)}(t_j)]^2,$$

com $f_{\lambda(j)}(t_j)$ sendo o preditor da observação y_j obtido através do procedimento aplicado à amostra, excluída a j -ésima observação, para cada valor de λ fixado. Existem várias outras formas para escolher o valor do parâmetro de suavização. As mais utilizadas são a regra prática de SILVERMAN (1986) e a validação cruzada de máxima verossimilhança.

5.2 Polinômios e termos trigonométricos

Podemos combinar polinômios com termos não lineares a fim de se obter modelos parcimoniosos. Aqui, veremos como combinar polinômios e termos trigonométricos. Suponha que os gráficos de resíduos nos mostrem alguma evidência de periodicidade. A combinação de polinômios e termos trigonométricos pode produzir um bom modelo de regressão. Os benefícios do uso concomitante de polinômios e termos trigonométricos podem ser vistos em GRAYBILL (1976). O modelo é escrito como a seguir:

$$Y = \beta_0 + \sum_{i=1}^d \beta_i X^i + \sum_{j=1}^{\lambda} [\gamma_j \cos(jX) + \delta_j \sin(jX)] + \varepsilon_j \quad (5.6)$$

onde os valores de d e γ devem ser determinados. EUBANK e SPECKMAN (1990) indicam sua preferência por $d = 2$. De acordo com os autores, tomar $d = 2$ é melhor do que determinar o valor de d através dos dados. São também apresentados duas formas de se encontrar o valor desconhecido γ .

O valor de λ é aquele que minimiza a função de risco apresentada a seguir:

$$R(\lambda) = \frac{SQE(\lambda) + 2\hat{\sigma}^2 (2\lambda + d + 1)}{n},$$

com $\hat{\sigma}^2$ sendo o estimador da variância do erro ε proposto por GASSER, SROKA e JENNEN-STEINMETZ (1986) e $SQE(\lambda)$ é a soma de quadrados do resíduo obtida a partir do método de validação cruzada, análoga à $CV(\lambda)$ da seção 5.1.

Se os valores de $\mathbf{X}_i$ são igualmente espaçados (definido em (3.33)), então $\sin(\mathbf{X})$ e $\cos(\mathbf{X})$ são ortogonais, da mesma forma como são $\sin(j\mathbf{X})$ e $\cos(j\mathbf{X})$, para $j = 2, 3, \dots$. Nos exemplos dados por EUBANK e SPECKMAN (1990) os valores de $\mathbf{X}_i$ são igualmente espaçados. Para o caso em que os dados não são igualmente espaçados, não teremos ortogonalidade entre $\sin(\mathbf{X})$ e $\cos(\mathbf{X})$ como também não teremos ortogonalidade entre $\sin(j\mathbf{X})$ e $\cos(j\mathbf{X})$, para $j = 2, 3, \dots$. No entanto, de acordo com MONTGOMERY, PECK e VINNING (2001), as correlações entre esses termos serão menores do que no caso em que não usássemos termos trigonométricos.

Exemplo

Para ilustrar o uso de polinômios e termos trigonométricos, EUBANK e SPECKMAN (1990) utiliza o exemplo da voltagem dado em MONTGOMERY, PECK e VINNING (2001), página 213. Esses dados foram escolhidos porque os gráficos dos dados sugerem uma função de regressão periódica. A variável independente é o tempo, em intervalos igualmente espaçados, conforme definido em (3.33). MONTGOMERY, PECK e VINNING (2001) ajustou um *spline* cúbico com 5 termos. EUBANK e SPECKMAN (1990) ajustou um modelo com 4 termos utilizando 2 termos polinomiais e 2 termos trigonométricos. Os autores obtiveram um valor de $R^2 = 0,9895$ e $\hat{\sigma}^2 = 0,0767$ utilizando as funções trigonométricas seno e cosseno. MONTGOMERY, PECK e VINNING (2001) chegaram a um valor de $R^2 = 0,9904$ e $\hat{\sigma}^2 = 0,0717$. Verificamos uma pequena diferença entre os valores de R^2 encontrados pelos dois autores. Assim sendo, prefere-se o modelo com menor quantidade de parâmetros.

Vimos, neste capítulo, que *splines* é uma técnica de interpolação muito utilizada em cálculo numérico e seu uso é uma alternativa nos casos em que a multicolinearidade restringe o uso de modelos de regressão polinomial. Vimos, também, que regressão polinomial está relacionado com *splines* através da função f , pois $f \in W_2^m[a, b]$. A função f é chamada de função de regressão ou curva de regressão e o objetivo é encontrar um estimador para tal função que se ajuste bem aos dados e que ao mesmo tempo possua bom grau de suavidade. O tipo de *spline* mais utilizado é o *spline* cúbico e MONTGOMERY, PECK e VINNING (2001) afirmam que esse tipo de *spline* é utilizado na maioria dos problemas práticos. Em seguida, vimos o emprego de termos trigonométricos e polinômios a fim de se encontrar modelos parcimoniosos. Tais modelos podem ser utilizados quando se verifica que o gráfico dos resíduos apresenta alguma evidência de periodicidade. Pesquisas mostram que se os dados forem igualmente espaçados, então a combinação de polinômios e termos trigonométricos geram coeficientes ortogonais. Caso contrário, ou seja, se os dados não são igualmente espaçados, as correlações entre os termos da regressão serão menores do que no caso em que não se usasse a combinação de termos trigonométricos e polinômios.

Capítulo 6

Aplicação e conclusões

6.1 Aplicação

O objetivo desse capítulo é o da aplicação de algumas técnicas de ajuste de modelos de regressão polinomial descritas ao conjunto de dados do Projeto ***Dinâmica de espécies arbóreas em uma floresta estacional semidecídua na Estação de Caetetus, Gália - SP*** (RAE-CEA-04P01, ELIAN, LIMA e LOPES (2004)). Para tanto, selecionamos três espécies de árvores, a saber:

1. Ocotea Odorífera;
2. Metrodorea Nigra;
3. Trichilia Catigua.

A literatura associada ao problema relaciona a taxa de crescimento de uma árvore em determinado período de tempo, G , com o logaritmo do comprimento da circunferência

na altura do peito a 1,3 m do solo, L , através de um modelo de regressão polinomial. O ajuste inicial de tais modelos apresentou problemas de heterocedasticidade e multicolinearidade. Devido à similaridade dos resultados, será apresentada somente a análise da espécie *Ocotea Odorífera*.

Para essa espécie, ajustamos o modelo

$$G_i = \alpha + \gamma L_i + \beta L_i^2 + \varepsilon_i,$$

onde, por suposição, devemos ter $\varepsilon_i \sim Normal(0, \sigma^2)$, $i = 1, 2, \dots, n$, independentes, com σ^2 constante real positiva. Dispunha-se de uma amostra de tamanho $n = 331$ árvores e após o ajuste do modelo proposto, encontramos os resultados presentes nas Tabelas 6.1 a 6.3.

Tabela 6.1: Análise de Variância

Fonte	gl	SQ	QM	Estatística F	p-valor	Significância
L	1	0,00948	0,00948	8,0786	0,00476	**
L²	1	0,00014	0,00014	1,1227	0,72637	
Resíduos	328	0,38506	0,00117			

Tabela 6.2: Ajuste do modelo de regressão quadrático

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,111025	0,089965	1,234	0,218	
L	-0,027285	0,049582	-0,550	0,582	200, 1185
L²	0,002354	0,006720	0,350	0,726	200, 1185

Tabela 6.3: Matriz de Correlação entre **L** e **L²**

	L	L²
L	1,0000	0,9975
L²	0,9975	1,0000

Observa-se um alto coeficiente de correlação entre as variáveis L e L^2 , indicando a necessidade de soluções para o problema de multicolinearidade existente. Como consequência, tem-se um alto valor do Fator de Inflação de Variância ($FIV = 200,1185$). A análise de resíduos realizada apontou a existência de uma possível heterocedasticidade dos erros. Dado que nosso interesse é avaliar as soluções propostas para a multicolinearidade, partimos para a aplicação das transformações discutidas no Capítulo 3.

Centralização

Devido aos problemas de multicolinearidade visíveis entre as variáveis $\mathbf{L}$ e $\mathbf{L}^2$, procedemos à centralização das variáveis. Utilizamos a variável independente $\mathbf{M} = \mathbf{L} - \bar{\mathbf{L}}$, onde

$$\bar{L} = \frac{\sum_{i=1}^n L_i}{n}.$$

Dessa forma, o modelo pode ser escrito como:

$$G_i = \alpha (L_i - \bar{L})^2 + \beta (L_i - \bar{L}) + \gamma + \varepsilon_i,$$

com as suposições usuais dos modelos de regressão. Feito o ajuste, as Tabelas 6.4 e 6.5 resumem os resultados obtidos.

Tabela 6.4: Ajuste do modelo de regressão quadrático centralizado

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,042995	0,002704	15,901	$< 2e^{-16} (***)$	
$\mathbf{M}$	-0,010197	0,003568	-2,858	0,00454 (**)	1,036504
$\mathbf{M}^2$	0,002354	0,006720	0,350	0,726637	1,036504

Tabela 6.5: Correlação entre $\mathbf{M}$ e $\mathbf{M}^2$

	$\mathbf{M}$	$\mathbf{M}^2$
$\mathbf{M}$	1,0000	0,1877
$\mathbf{M}^2$	0,1877	1,0000

Como podemos observar, a centralização de variáveis preditoras corrigiu o problema de multicolinearidade, fornecendo um novo modelo adequado para relacionar as variáveis nesta espécie. Observa-se ainda que o nível descritivo associado ao termo quadrático levaria à exclusão da variável $\mathbf{M}^2$ do modelo. Essa exclusão, no entanto, seria feita sem o risco de ter sido provocada por um possível aumento da variância do estimador causado pela multicolinearidade.

Uso da Transformação V

Nesta etapa, ajustamos um modelo utilizando a transformação V , dada por:

$$v_i = \frac{x_i}{x_{max}}, \quad i = 1, 2, \dots, n.$$

Os resultados do ajuste encontram-se na Tabela 6.6.

Tabela 6.6: Ajuste do modelo de regressão utilizando-se a transformação V

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,11102	0,08997	1,234	0,218	
V	-0,13448	0,24437	-0,550	0,582	200,1185
V²	0,05718	0,16325	0,350	0,726	200,1185

A matriz de correlação da variável V é equivalente à da variável L , ou seja, a correlação entre V e V^2 é 0,9975, muito alta. Observa-se, dessa forma, que esta transformação foi ineficiente, mantendo a multicolinearidade nos mesmos níveis, conforme podemos conferir pelo alto valor de $FIV = 200,1185$. Tal fato era esperado pois se $r(L, L^2)$ é o coeficiente de correlação de Pearson entre L e L^2 , então

$$r(V, V^2) = r\left(\frac{L}{l_{max}}, \frac{L^2}{l_{max}^2}\right) = r(L, L^2)$$

pois $r(aX, bY) = r(X, Y)$ para $a, b > 0$.

Uso da Transformação W

A transformação W é dada por

$$w_i = \frac{(x_i - x_{\min})}{(x_{\max} - x_{\min})}.$$

Como podemos observar nas Tabelas 6.7 e 6.8, houve uma apenas uma ligeira redução na multicolinearidade existente, apresentando $FIV = 13,61352$. A alta correlação entre as variáveis regressoras W e W^2 ainda é persistente.

Tabela 6.7: Ajuste do modelo de regressão quadrático com transformação W

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,053739	0,005242	10,252	$< 2e^{-16}(* * *)$	
W	-0,031151	0,028127	-1,107	0,269	13,61352
W²	0,011135	0,031791	0,350	0,726	13,61352

Tabela 6.8: Correlação entre W e W^2

	W	W²
W	1,0000	0,9626
W²	0,9626	1,0000

Assim como a transformação V , a transformação W também foi ineficiente para corrigir a multicolinearidade.

Uso da Transformação Z

As Tabelas 6.9 e 6.10 exibem os resultados associados ao ajuste do modelo de regressão polinomial para a variável Z , de forma que

$$z_i = \frac{2x_i - x_{max} - x_{min}}{(x_{max} - x_{min})}.$$

Observa-se uma considerável redução na multicolinearidade com valores de FIV aceitáveis para esse conjunto de dados, corroborando a asserção de que a transformação Z é efetivamente melhor que as transformações V e W . Mesmo assim, a centralização para esse conjunto de dados mostrou-se mais eficiente.

Tabela 6.9: Ajuste do modelo de regressão quadrático com transformação Z

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,040947	0,002696	15,187	$< 2e^{-16}(* * *)$	
Z	-0,010008	0,004482	-2,233	0,0262 (*)	1,382828
Z²	0,002784	0,007948	0,350	0,7264	1,382828

Tabela 6.10: Correlação entre **Z** e **Z²**

	Z	Z²
Z	1,0000	-0,5262
Z²	-0,5262	1,0000

Uso de polinômios ortogonais

CHAMBERS e HASTIE (1992) e KENNEDY e GENTLE (1980) implementaram no pacote computacional *R* rotinas que retornam polinômios ortogonais. Os comandos para execução do ajuste por polinômios ortogonais de grau 2 são *poly* e *polym* e, através dessas rotinas, ajustamos o modelo polinomial e os resultados podem ser observados nas Tabelas 6.11 e 6.12.

Tabela 6.11: Ajuste do modelo de regressão ortogonal

	Estimativa	Desvio-Padrão	t-valor	Significância	FIV
<i>intercepto</i>	0,043675	0,001883	23,191	$< 2e^{-16}(* * *)$	
P	-0,097386	0,034263	-2,842	0,00476 (**)	1
P²	0,012001	0,034263	0,350	0,72637	1

Tabela 6.12: Correlação entre **P** e **P²**

	P	P²
P	1,0000	0,0000
P²	0,0000	1,0000

Novamente aqui, observamos a não significância do termo quadrático, contrariando a suposição de que

$$G_i = \alpha + \gamma L_i + \beta L_i^2 + \varepsilon_i.$$

Caso, no entanto o termo L_i^2 apresentasse contribuição significativa, dada a alta correlação entre **L** e **L²**, sugeriríamos o uso de polinômios ortogonais, com a volta ao modelo original, conforme apresentado no exemplo da Seção 4.2.

Como podemos observar pelos resultados apresentados, efetivamente, os polinômios ortogonais são a melhor solução para eliminação dos problemas da multicolinearidade.

6.2 Recomendações e considerações finais

Esta seção se ocupará de descrever um breve roteiro para solucionar os problemas de multicolinearidade em modelos de regressão polinomial. Para tanto, sugerimos a realização dos seguintes passos:

- Passo 1: Definir o grau do polinômio, pois, como vimos na Seção 2.1., o modelo deve possuir a menor ordem possível;
- Passo 2: Sempre utilizar a variável transformada $x_i - \bar{x}$ ao invés da variável regressora x_i . Como vimos na Seção 2.1., a sequência de variáveis regressoras $\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \dots$ geralmente possui alta correlação resultando no mal-condicionamento da matriz $\mathbf{X}$ que, como consequência, traz problemas de ordem computacional para o cálculo da inversa da matriz $\mathbf{X}'\mathbf{X}$;
- Passo 3: Sempre que possível, selecionar os valores da variável independente de modo que sua média seja nula ou simetricamente em torno de sua média;
- Passo 4: Para detectar a multicolinearidade, calcular $\|\mathbf{X}'\mathbf{X}\|$, $R_{(j)}^2$, R_j^2 , $\kappa(\mathbf{X}'\mathbf{X})$ e FIV ;
- Passo 5: Uma vez detectada a multicolinearidade através das ferramentas descritas no item anterior, usar polinômios ortogonais. Como vimos neste trabalho, polinômios ortogonais foi a estratégia que efetivamente eliminou os problemas da multicolinearidade. Além do mais, recursos presentes em pacotes estatísticos modernos permitem ajustar tais modelos sem dificuldade.

Finalizando, gostaríamos ainda de destacar alguns artigos, referentes a planejamentos de experimentos ótimos em modelos de regressão polinomial, assunto não abordado neste trabalho. Consistem basicamente na utilização dos chamados planejamentos D-ótimos, nos quais o determinante de $\mathbf{X}'\mathbf{X}$ é maximizado para uma particular região de valores das variáveis explicativas. Nesta linha encontram-se os trabalhos de ATKINSON (1982), HAINES (1987) e GAFFKE e KRAFT (1982).

Referências Bibliográficas

1. ATKINSON, A. C. (1982), *Developments in the design of experiments*. International Statistical Review, 50, 161-177.
2. BELSLEY, D. A. (1991), *Condition Diagnostics*. New York: John Wiley & Sons.
3. BELSLEY, D. A.; KUH, E. e WELSCH, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Nova York: John Wiley & Sons, Capítulo 3.
4. BRAUNER, N.; SHACHAM, M. (1998), *The role of range and precision of the independent variable in regression of data*. AIChE J., 44 (3), 603-611
5. BRADLEY, R. A.; SRIVASTAVA, S. S. (1979), *Correlation in Polynomial Regression*. American Statistician, vol. 33, p. 11-14.
6. CARVALHO, S. P. (1992), *Métodos alternativos de estimação de coeficientes de trilha e índices de seleção, sob multicolinearidade*. Viçosa: UFV, 1995. 163p.
7. CHAMBERS. J. M.; HASTIE, T. J. *Statistical Models in S*. Wadsworth & Brooks/Cole.
8. CLAYTON, D. G. (1971), *Gram-Schmidt Orthogonalization*. Applied Statistics, p. 335-338.
9. DELURY, D. B. (1960), *Values and Integrals of the orthogonal polynomials up to $n=26$* . Univ of Toronto Press, 33 pp.

10. DIAS, R. (2001), *O uso de Splines em Regressão não Paramétrica*. Notas de aula. IMECC/UNICAMP.
11. de BOOR, C. (1978), *A practical guide to Splines*. Springer Verlag, New York.
12. DRAPER, N.R.; SMITH, H. (1998), *Applied Regression Analysis*. New York: John Wiley & Sons, 3. ed., Probability and Mathematical Statistics.
13. ELIAN, S. N. ; LIMA, P. H.; LOPES, S. M. (2004), *Dinâmica de espécies arbóreas em uma floresta estacional semidecídua na Estação de Caetetus, Gália - SP*. Relatório de Análise Estatística 04P01. São Paulo: Instituto de Matemática e Estatística da Universidade de São Paulo - USP.
14. EUBANK, R. L.; SPECKMAN, P. (1990), *Curve fitting by polynomial-trigonometric regression*. Biometrika, vol. 77, p. 1-9.
15. FARRAR, D.E.; GLAUBER, R.R. (1967), *Multicollinearity in regression analysis: the problem revisited*. Review of Economics and Statistics, vol. 49, p. 92-107.
16. FORSYTHE, G. E. (1957), *Generation and Use of Orthogonal Polynomials for Data-Fitting with Digital*. J. Soc. Ind. Appl. Math., vol. 5, p. 74-87. Computer.
17. FREITAS, M. C. (1988), *Estimação de Beta no modelo de Regressão Linear na presença de Multicolinearidade*. São Paulo: Instituto de Matemática e Estatística da Universidade de São Paulo - IME/USP - Dissertação de Mestrado.
18. FRISCH, R. (1934), *Statistical Confluence Analysis by Means of Complete regression Systems*. Institute of Economics, Universidade de Oslo, Publicação 5.
19. GAFFKE, N.; KRAFFT, O. (1982), *Exact D-optimum designs for quadratic regression*. Journal of the Royal Statistical Society, Series B, 44, 394-397.
20. GASSER, T.; SROKA, L.; JENNEN-STEINMETZ, C. (1986), *Residual variance and residual pattern in nonlinear regression*. Biometrika, vol. 73, p. 625-633.

21. GIAUQUE and WIEBE (1928), *The Heat Capacity of Nickel Nitrate Hexammoniate from 54 to 300° K. The transition at 243° K*, 50, 101.
22. GOLUB; V. L.; LOAN, C. F. (1996), *Matrix Computations*. Johns Hopkins University Press, 3rd edition.
23. GRANDAGE, A. (1958), *Orthogonal coefficients for unequal intervals*. Biometrics, vol. 14, p. 287-289.
24. GRAYBILL, F. A. (1961), *An Introduction to Linear Statistical Models*. New York: McGraw-Hill, vol. 1.
25. GRAYBILL, F. A. (1976), *Theory and Application of the Linear Models*. North Scituate, MA: Duxbury.
26. GUEST, P.G. (1950), *Orthogonal Polynomials in the least squares fitting of observations*. Philosophical Magazine, vol. 41, p. 124-134.
27. GUNNS, R. F. (1984), *Towards a balanced assessment of Collinearity diagnostics*. American Statistician, vol. 38 (2), p. 79-82.
28. GUNST, R. F. e MASON, R. I. (1977), *Advantages of Examining Multicollinearity in regression Analysis*. Biometrics, vol. 33, p. 249-260.
29. HAINES, L. M. (1987), *The application of the annealing algorithm to the construction of exact optimal designs for linear-regression models*. Technometrics, 29, 439-447.
30. KENNEDY, W. J. Jr.; GENTLE, J. E. (1980), *Statistical Computing*. Marcel Dekker.
31. MANSFIELD, E. R.; HELMS, B. P. (1982), *Detecting Multicollinearity*. The American Statistician, vol. 36, number 3, p. 158-160.
32. MASON, R. I.; GUNST, R. F. e WEBSTER, J. T. (1975), *Regression Analysis and Problems of Multicollinearity*. Communications in Statistics A, vol. 4, nº 3, p. 277-292.

33. MELA, C.F.; KOPALLE, P.K. (2002), *The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations*. Applied Economics, vol. 34, p. 667-677.
34. MONTGOMERY, D. C.; PECK, E.; VINNING, G. G. (2001), *Introduction to Linear Regression Analysis*. Nova York: John Wiley & Sons, p. 289-290
35. MONTGOMERY, D. C. (2005), *Design and Analysis of Experiments*, 6th ed., John Wiley and Sons, Inc. New York, NY.
36. NARULA, S. C. (1979), *Orthogonal Polynomial Regression*. International Statistical Review, vol. 47, p. 31-36.
37. ROBSON, D.S. (1958), *A simple method for constructing orthogonal polynomials, when the independent variable is unequally spaced*. Biometrics, vol. 15, p. 187-191.
38. SEBER, G. A. F. (1977), *Linear Regression Analysis*. New York: John Wiley & Sons.
39. SHACHAM, M.; BRAUNER, N. (1997), *Minimizing the effects of Collinearity in polynomial regression*. Ind. Eng. Chem. Res., vol 36, Number 10, p. 4405-4412.
40. SILVERMAN, B. W. (1985), *Some aspects of the spline smoothing approach to parametric regression curve fitting*. J. R. Statist. Soc., number B 47, p. 1-52.
41. SILVERMAN, B. W. (1984), *A fast and efficient cross-validation method for smoothing parameter choice in spline regression*. Am. Statist. Assoc., number 79, p. 584-589.
42. SILVEY, S.D. (1968), *Multicollinearity and Imprecise Estimation*. Journal of the Royal Statistical Society, vol. 38, p. 539-552.
43. STEWART, G. W. (1987), *Collinearity and Least Squares Regression*. Statistical Science, vol. 2, number 1, p. 68-100.

44. SUN, J. G. (1987), *Matrix Perturbation Analysis*. Beijing - China: Science Press.
45. WAHBA, G. (1987), *Spline models for observational data*. Ohio: University at Columbus, Applied Mathematics.
46. WISHART, J.; METAKIDES, T. (1953), *Orthogonal Polynomial fitting*. Biometrika, vol. 40, p. 361-369.