Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Diagnóstico e redução da influência da multicolinearidade na estimação de efeitos genéticos aditivos e não-aditivos em uma população de bovinos compostos (*Bos taurus* x *Bos indicus*)

Raphael Antonio Prado Dias

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Raphael Antonio Prado Dias
Licenciado em Matemática

Diagnóstico e redução da influência da multicolinearidade na estimação de
efeitos genéticos aditivos e não-aditivos em uma população de bovinos
compostos (*Bos taurus* x *Bos indicus*)

Orientador:
Prof. Dr. **GERSON BARRETO MOURÃO**

Dissertação apresentada para obtenção do título de
Mestre em Agronomia. Área de concentração: Es-
tatística e Experimentação Agronômica

Piracicaba
2008

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP**

Dias, Raphael Antonio Prado

Diagnóstico e redução da influência da multicolinearidade na estimação de efeitos genéticos aditivos e não-aditivos em uma população de bovinos compostos (*Bos taurus* x *Bos indicus*) / Raphael Antonio Prado Dias. - - Piracicaba, 2008.
93 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2008.
Bibliografia.

1. Análise de regressão e de correlação 2. Bovinos 3. Componentes de variância
4. Melhoramento genético animal 5. Mínimos quadrados I. Título

CDD 636.2082
D541d

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

Dedicatória

A Deus,

Ao meu pai Edson Mesquita Dias,

À minha mãe Deuseli Aparecida Prado Dias,

Às minhas irmãs Aline e Ariane.

AGRADECIMENTOS

Ao meu orientador, Prof. Dr. Gerson Barreto Mourão, pelo conhecimento compartilhado, ao apoio, a amizade e estímulo constante, tornando possível a realização deste trabalho.

Aos Professores e funcionários do Departamento de Ciências Exatas da ESALQ/USP pela atenção, compreensão e a amizade.

Ao grande apoio dos amigos dos cursos de mestrado e doutorado do Departamento de Ciências Exatas da ESALQ/USP.

Aos professores Joanir Pereira Eler, José Bento Sterman Ferraz e Júlio César Carvalho Balieiro do Departamento de Ciências Básicas da FZEA/USP de Pirassununga, pela cedência dos dados e do laboratório de informática.

A Técnica em informática Elisângela Chicaroni de Mattos Oliveira do Departamento de Ciências Básicas da FZEA/USP de Pirassununga, a gentileza e ao apoio junto ao laboratório de informática.

A doutoranda Heloise Patrícia Quirino pela ajuda no ambiente UNIX e no programa PEST.

A graduanda Juliana Petrini, pelas ajudas e a amizade.

Ao Roberto Carvalheiro e ao professor Vanerlei M. Roso pela disponibilização de programas feitos por estes e pela atenção.

Aos amigos Elton Rafael Mauricio da Silva Pereira e Lucas Willian Mendes, a amizade e a boa convivência durante o período de república.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo apoio financeiro em forma de bolsa de estudos por um período de 10 meses.

A todas as pessoas que de alguma forma contribuíram para minha formação profissional e para a realização deste trabalho.

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	9
1 INTRODUÇÃO	14
2 OBJETIVO	16
3 DESENVOLVIMENTO	17
3.1 Revisão Bibliográfica	17
3.1.1 Detecção da Multicolinearidade	27
3.1.2 Como amenizar a multicolinearidade	40
3.1.3 Análise Genética	44
3.1.4 Análise Alternativa	45
3.1.5 Medida do Viés	45
3.2 Material	45
3.3 Métodos	49
3.4 Resultados e Discussão	50
3.4.1 Peso ao Nascimento	50
3.4.2 Peso ao Desmame	58
3.4.3 Perímetro escrotal aos 390 dias	66
3.4.4 Escore de Musculossidade aos 390 dias	73
4 CONCLUSÕES	81
REFERÊNCIAS	83
ANEXOS	86

RESUMO

Diagnóstico e redução da influência da multicolinearidade na estimação de efeitos genéticos aditivos e não-aditivos em uma população de bovinos compostos (*Bos taurus* x *Bos indicus*)

Os efeitos genéticos aditivos e de heterozigoses são importantes na avaliação genética de populações compostas. Quando existem fortes relações lineares entre as variáveis explanatórias, os coeficientes de regressão têm erros-padrão elevados, são sensíveis a mudanças nos dados e a adição ou eliminação de variáveis explicativas no modelo. A alternativa usada na tentativa de diminuir esse problema foi aplicar o método de regressão de cumeeira - *RC*, pois na presença de multicolinearidade, pode permitir a obtenção de estimativas mais estáveis dos efeitos aditivos de origem genética e de heterozigose, em relação às obtidas pelo método dos quadrados mínimos - *QM*. Foram analisados os dados de pesos ao nascimento - PESNAS, ao desmame - PESDES, perímetro escrotal aos 390 dias - CE e escore para musculosidade aos 390 dias - MUSC de bovinos compostos Montana Tropical®, com diferentes composições raciais NABCs, obtidos em várias fazendas brasileiras, relativos aos animais nascidos no período de 1994 a 2008. O modelo incluiu os efeitos aditivos e não aditivos. O grau da multicolinearidade foi obtido através do valor do fator de inflação da variância - *VIF*, dos índices de condição e da decomposição proporcional da variância. Os parâmetros de cumeeira foram obtidos a partir da multiplicação de uma constante, pela razão entre o *VIF* da covariável correspondente e o maior *VIF*. O traço de cumeeira foi utilizado para verificar se as estimativas dos coeficientes se estabilizaram, para o parâmetro de cumeeira obtido para cada variável explicativa. Duas análises foram aplicadas: i) os efeitos foram estimados por quadrados mínimos; ii) os efeitos foram estimados por regressão de cumeeira. Para cada variável resposta foi identificado o número de colinearidades, seus respectivos graus e as variáveis explicativas envolvidas em cada uma. As covariáveis envolvidas no modelo, para peso ao nascimento participaram de uma colinearidade forte e quatro colinearidades fracas; para peso ao desmame e escore de musculosidade aos 390 dias, houve duas relações de quase dependência fortes e três fracas, enquanto que para perímetro escrotal aos 390 dias obteve-se três colinearidades fortes e três fracas. O método que estimou os coeficientes por regressão de cumeeira foi melhor que o método dos quadrados mínimos, para todas as características. A média dos *VIFs* para PESNAS, PESDES, CE e MUSC reduziram de 15,5; 16; 17,5 e 23,9 para 5,8; 5,3; 5,7 e 5,1 respectivamente, após o uso da *RC*. Os erros-padrão diminuíram fornecendo estimativas mais estáveis que as obtidas por quadrados mínimos. Apenas para a covariável *A* sobre a variável resposta peso ao nascimento as soluções obtidas por *QM* e *RC* diferiram em direção, no mais, houve diferenças em magnitude.

Palavras-chave: Avaliação Genética; Quadrados Mínimos; Regressão de Cumeeira; Mestiços

ABSTRACT

Diagnostic and reduction of the influence of multicollinearity in the estimation of genetic additive and non-additive effects in multibreed population of cattle (*Bos taurus* x *Bos indicus*)

The genetic additive and heterozygosity effects are important in the genetic evaluation of multibreed populations. When there is strong linear relation between the explanatory variables, the regression coefficients have large standard errors and are sensitive to changes in the data set and to the addition or removal of explanatory variables in the model. The alternative used to try to reduce this problem was to apply the method of ridge regression - *RC*, which could allow for the estimation of more stable coefficients of direct and maternal breed additive effects of genetic origin and heterozygosity in relation to those obtained by the method of least squares *QM*. The objective is to analyze the data of birth weight - PESNAS, weaning - PESDES, the scrotal perimeter 390 days - CE and scoring for the muscularity 390 days - MUSC of cattle compounds Montana Tropical®, with different racial compositions NABCs, obtained in several Brazilian farms on of animals born from 1994 to 2008. The model included additive and non-additive effects. The degrees of multicollinearity were obtained through the value of the variance inflation factor - *VIF*, the index conditions - *IC* and by proportional decomposition of Variance. The ridge parameters were obtained from the multiplication of a constant to the ratio of the *VIF* from each covariate and the highest *VIF*. For each explanatory variable, the ridge trace was used to verify that the estimated coefficients were stabilized using the ridge parameter. Two different methods were applied: i) the effects were estimated by least squares; ii) the effects were estimated by ridge regression. For each response variable the number of colinearities was identified, their degrees and the variables involved in each. The covariates used in the model for birth weight participated in a strong colinearity and four other weak colinearities; for weaning weight and muscle score for 390 days, there were two strong relations of dependency and three almost weak, while for the perimeter scrotal 390 days it was observed three strong and three weak colinearities. The ridge regression coefficients method was considered better than that of least squares for all factors. The *VIFs* average for PESNAS, PESDES, CE and MUSC reduced from 15.5, 16, 17.5 and 23.9 to 5.8, 5.3, 5.7 and 5.1 respectively, after using the *RC*. The standard errors of the estimators decreased providing estimates more stable than those obtained by least squares. Only for *A* covariate on the response variable weight at birth the solutions obtained by *QM* and *RC* differ in direction, where the other ones differed only in magnitude.

Keywords: Genetic Evaluation; Ordinary Least Square; Ridge Regression; Crossbreeding

LISTA DE FIGURAS

- Figura 1 - Ausência de colinearidade - todos os coeficientes de regressão bem determinados. Uma pequena mudança em qualquer parâmetro do plano causará uma mudança relativamente grande na soma de quadrados residual 19
- Figura 2 - Colinearidade exata - todos os coeficientes de regressão não determinados. Uma mudança simultânea em todos os parâmetros poderá deixar a soma de quadrados residual inalterada 19
- Figura 3 - Colinearidade forte - todos os coeficientes de regressão mal determinados. Uma mudança simultânea em todos os parâmetros pode causar pequena alteração na soma de quadrados residual 20
- Figura 4 - Colinearidade forte - intercepto bem determinado. Mudanças apenas nos coeficientes angulares afetam pouco a soma de quadrados residual 20
- Figura 5 - Colinearidade forte - β_2 bem determinado. Mudanças apenas no intercepto e em β_1 afetam pouco a soma de quadrados residual 21
- Figura 6 - Colinearidade forte, embora a presença de um *outlier* tornou os coeficientes bem determinados 21
- Figura 7 - Comparação do estimador viesado com pequena variância em relação ao estimador não-viesado com grande variância 41

LISTA DE TABELAS

Tabela 1 - Exemplo da decomposição proporcional da variância em relação aos valores singulares	38
Tabela 2 - Exemplo da decomposição proporcional da variância em relação aos valores singulares para uma matriz $\mathbf{X}$ ortogonal	39
Tabela 3 - Exemplo da decomposição proporcional da variância em relação aos valores singulares em que $\mathbf{X}$ possui duas colunas colineares	39
Tabela 4 - Definição das estimativas dos componentes de (co)variância	49
Tabela 5 - Fatores de Inflação da Variância - VIF para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (54) em relação à variável resposta peso ao nascimento	50
Tabela 6 - Autovalores e Índices de Condição - ICs da matriz $\mathbf{X}$ do modelo (54) para a variável resposta peso ao nascimento	51
Tabela 7 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição - IC da matriz $\mathbf{X}$ do modelo (54), para a variável resposta peso ao nascimento	52
Tabela 8 - Valores dos coeficientes de regressão (β) para o modelo (54), estimados pelo método dos quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $	54

Tabela 9 - Soma dos quadrados dos desvios - SQD do modelo (54), máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - QM e regressão de cumeieira - RC e medida do viés para RC - viés (%)	55
Tabela 10 - Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeieira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$	56
Tabela 11 - Estimativa dos coeficientes de regressão das covariáveis data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ em relação à variável resposta peso ao nascimento	57
Tabela 12 - Estimativas dos componentes de (co)variâncias ¹ , para peso ao nascimento, em que o as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cumeieira - RC	57
Tabela 13 - Fatores de Inflação da Variância - VIF para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (55), para a variável resposta peso ao desmame	58
Tabela 14 - Autovalores e Índices de Condição da matriz $\mathbf{X}$ do modelo (55) para peso ao desmame	59
Tabela 15 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (55) para a variável peso ao desmame	60

- Tabela 16 -Valores dos coeficientes de regressão das covariáveis genéticas de β para o modelo (55), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeeira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $|\hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^*|$, para a variável resposta peso ao desmame 62
- Tabela 17 -Soma dos quadrados dos desvios - SQD para o modelo (55) e para a variável resposta peso ao desmame, máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - QM e regressão de cumeeira - RC e medida do viés para RC - viés (%) 63
- Tabela 18 -Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeeira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$, para a variável resposta peso ao desmame 64
- Tabela 19 -Estimativa das covariáveis data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ para a variável resposta peso ao desmame 65
- Tabela 20 -Estimativas dos componentes de (co)variâncias¹ para peso ao desmame, em que as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cumeeira - RC 65
- Tabela 21 -Fatores de Inflação da Variância - VIF para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (56) 66
- Tabela 22 -Autovalores e Índice de Condição da matriz $\mathbf{X}^T\mathbf{X}$ do modelo (56) para a variável resposta perímetro escrotal aos 390 dias 67

Tabela 23 -Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (56), para a variável resposta perímetro escrotal aos 390 dias	68
Tabela 24 -Valores dos coeficientes de regressão (β) para o modelo (56), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeeira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $ para a variável resposta perímetro escrotal aos 390 dias	69
Tabela 25 -Soma dos quadrados dos desvios - SQD para o modelo (56), máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por por quadrados mínimos - QM e regressão de cumeeira - RC e medida do viés para RC - viés (%)	70
Tabela 26 -Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeeira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$	71
Tabela 27 -Estimativas das covariáveis de data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ para a variável resposta perímetro escrotal aos 390 dias	72
Tabela 28 -Estimativas dos componentes de (co)variâncias ¹ , para perímetro escrotal aos 390 dias, em que as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cumeeira - RC	73
Tabela 29 -Fatores de Inflação da Variância - VIF para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (57) sobre a variável resposta escore de musculosidade aos 390 dias	73

Tabela 30 -Autovalores e Índice de Condição da matriz $\mathbf{X}^T \mathbf{X}$ do modelo (57) para a variável resposta score de musculosidade aos 390 dias	74
Tabela 31 -Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (57), para a variável resposta score de musculosidade aos 390 dias .	75
Tabela 32 -Valores dos coeficientes de regressão (β) para o modelo (57), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $	77
Tabela 33 -Soma dos quadrados dos desvios - SQD para o modelo (57) em relação à variável resposta score de musculosidade aos 390 dias, máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - QM e regressão de cumeieira - RC e medida do viés para RC - viés (%)	78
Tabela 34 -Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeieira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$	79
Tabela 35 -Estimativa das covariáveis data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ em relação à variável resposta score de musculosidade aos 390 dias	80
Tabela 36 -Estimativas dos componentes de (co)variâncias ¹ para score de musculosidade aos 390 dias, em que o as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cumeieira - RC .	80

1 INTRODUÇÃO

O melhoramento genético é tradicionalmente realizado através da exploração das diferenças genéticas existentes entre raças e linhagens e também entre indivíduos de uma mesma raça ou linhagem, com o uso da seleção e dos sistemas de cruzamentos entre raças como principais ferramentas.

De acordo com Luchiari e Mourão (2006), os primeiros documentos relativos ao melhoramento genético especificamente em bovinos referem-se aos trabalhos de Robert Bakewell, entre 1725 e 1795, na Inglaterra. Em seguida vieram os trabalhos dos irmãos Colling, que aplicaram os princípios de Bakewell na formação da raça Shorthorn, cujo livro de registro genealógico foi estabelecido em 1822.

Ainda segundo Luchiari e Mourão (2006), no Brasil, a história em melhoramento genético iniciou-se em 1915, com a instalação de um posto pecuário em Nova Odessa no estado de São Paulo, com o objetivo de selecionar animais das raças Mocho Nacional e Caracu.

Atualmente, os programas de melhoramento estão se difundindo amplamente. No entanto, trata-se de um campo científico ainda em desenvolvimento, considerando que o Brasil é líder no mercado mundial de carnes, mas ainda possui apenas cinco por cento de seu rebanho avaliado geneticamente. Há 30 anos, a produção de carne era de apenas 20 kg/ha.ano, porém na última década esse valor ultrapassou 30 kg/ha.ano, além disso, o rebanho atual é da ordem de 207 milhões de cabeças (FAO, 2007), que se caracteriza como o maior rebanho comercial do planeta, correspondendo a 15% da produção mundial.

Diante disto, as características de seleção que outrora eram definidas em termos de coloração de pelagem ou tipo fenótipo foram substituídas por características produtivas como peso e ganho de peso. Buscam-se ainda animais adaptados e mais precoces, que diminuam o tempo de pastagem ou confinamento, trazendo benefícios econômicos aos produtores.

Na tentativa de auxiliar a melhoria desse rebanho, muitos estudos em melhoramento genético se baseiam em obter estimativas de componentes de variância e parâmetros genéticos de características ligadas à produção, assim como avaliar os efeitos de heterose, de fatores ambientais, além de permitir a predição do desempenho de cruzamentos entre diferentes raças e do desempenho animal em ambientes distintos. Assim, na avaliação genética de populações bovinas compostas ou multirraciais, os efeitos genéticos de origem aditiva, de

heterozigose e os epistáticos precisam ser estimados face à sua importância, pois incluem várias raças em sua formação.

Para isso, o método estatístico mais usado para derivar equações de predição é o método dos quadrados mínimos - *QM*. Contudo, quando existem fortes relações lineares entre as covariáveis tem-se um problema denominado multicolinearidade. As estimativas dos coeficientes de regressão por *QM* tendem a ser instáveis, geralmente com grandes erros-padrão e podem induzir à inferências errôneas (BERGMANN; HOHENBOKEN, 1995).

Um dos métodos alternativos de estimação, usado na análise informativa quando existe multicolinearidade (CHATTERJEE; PRICE, 1991; DRAPER; SMITH, 1998), é a regressão de cumeeira - *RC* (HOERL; KENNARD, 1970), que consiste na adição de coeficientes k à diagonal principal da matriz de correlações, visando reduzir ou eliminar as dependências lineares. Tais estimadores de cumeeira são viesados, porém, as estimativas obtidas por este método são mais precisas, ou seja, apresentam menores erros-padrão, além de serem mais estáveis que as obtidas por *QM* na presença de multicolinearidade.

Este projeto é relevante à comunidade acadêmica e a outros interessados do meio agropecuário, uma vez que, auxiliará futuras pesquisas que envolvam efeitos aditivos e não aditivos na presença de multicolinearidade, e ajudará na seleção de bovinos reprodutores geneticamente superiores, colaborando na melhoria do rebanho dessa raça.

2 OBJETIVO

Os objetivos deste estudo com bovinos compostos de várias combinações de tipos biológicos com base no sistema NABC, que caracterizam a população formadora do composto Montana Tropical[®] foram:

- (i) detectar possíveis dependências entre as covariáveis genéticas do modelo e identificar o grau de multicolinearidade na avaliação genética da população de bovinos ;
- (ii) obter estimativas dos efeitos aditivos diretos, maternos, não aditivos e idade (exceto para peso ao nascimento) pelos métodos QM e RC e compará-las;

3 DESENVOLVIMENTO

3.1 Revisão Bibliográfica

Na literatura, não há definição precisa de multicolinearidade. Segundo Belsley et al. (1980), literalmente, duas variáveis são colineares se os vetores de dados representando-as situam-se sobre a mesma reta, isto é, em um subespaço de uma dimensão. De maneira geral, k variáveis são colineares se os vetores que as representam localizam-se em um subespaço de dimensão menor que k , ou seja, se pelo menos um dos vetores é uma combinação linear dos demais. Tal colinearidade exata raramente ocorre na prática, e ela certamente não é necessária para que haja problema.

Assim, para uma definição menos restritiva, duas variáveis são colineares se estão situadas quase que na mesma reta, isto é, se o ângulo entre os vetores de dados é relativamente pequeno. Generalizando para o caso de mais de duas variáveis, pode-se dizer que há colinearidade se existe uma alta correlação múltipla quando é feita a regressão de uma variável em função das demais.

Além do termo *multicolinearidade*, são usados os termos *colinearidade* e *mau condicionamento* para denotar esta mesma situação. Muitos autores preferem a última destas denominações, por estar relacionada ao fato de uma matriz mal condicionada acarretar graves problemas numéricos.

De acordo com a definição acima, nota-se que a colinearidade está relacionada à características específicas da matriz de delineamento $\mathbf{X}$ e não aos aspectos estatísticos do modelo de regressão linear $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Em outras palavras, colinearidade é um problema numérico, não estatístico. De qualquer modo, muitas áreas da ciência aplicam análises de regressão a conjuntos de dados não experimentais, em que dados colineares aparecem e causam problemas. Assim, a colinearidade é um problema de grande importância para a eficácia da estimação de mínimos quadrados.

Segundo Mason et al. (1975), há três fontes de multicolinearidade:

1. Devido a restrições físicas no modelo ou na população. Pode haver alguma razão para que haja uma restrição no modelo. Por exemplo, os conteúdos de certos constituintes em um processo químico podem somar para uma constante ou quase constante.

2. Devido a técnicas amostrais. O pesquisador pode amostrar um subespaço do espaço k -dimensional das variáveis explanatórias. Por exemplo, em operações com plantas isto pode ser devido ao sistema ser necessariamente conduzido em condições quase ótimas.
3. Devido a um modelo com excesso de termos. Por exemplo, pode haver tantas variáveis quantas forem as observações (ou mais), ou o modelo pode estar simplesmente superparametrizado. Pode haver também um modelo desnecessariamente complicado, incluindo por exemplo, muitos termos quadráticos e produtos cruzados.

É possível visualizar geometricamente a natureza da colinearidade, abordagem esta utilizada por Belsley et al. (1980). Nas figuras 1 até 6 encontram-se diversas situações relevantes para o modelo de regressão

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \quad (1)$$

Nas figuras 1 a 6 tem-se a dispersão das n observações. No plano formado pelos eixos x_1 e x_2 estão os pontos (X_1, X_2) , representados por pontos cheios, enquanto que acima está a dispersão dos pontos que resulta quando a dimensão Y é incluída (pontos vazios). Na figura 1 encontra-se o caso ideal em que os dados de X_1 e X_2 não são colineares. A dispersão dos pontos fornece um plano de mínimos quadrados bem definido, ou seja, o plano que minimiza a soma dos quadrados dos erros na direção de Y entre os valores observados Y_i e o plano que contém os valores estimados. O intercepto do plano com o eixo y estima β_0 , enquanto que as inclinações parciais nas direções de x_1 e x_2 , respectivamente, estimam β_1 e β_2 . Como o plano é bem definido, os parâmetros são estimados com precisão.

Na figura 2 tem-se o caso de colinearidade perfeita entre as variáveis X_1 e X_2 . O plano de mínimos quadrados não é bem definido; qualquer plano situado ao longo do “eixo” da dispersão dos pontos resulta na mesma soma de quadrados dos erros. Isto mostra o bem conhecido fato de que a colinearidade perfeita destrói a unicidade do estimador de mínimos quadrados. Na figura 3 encontra-se uma situação de forte, porém não perfeita, colinearidade. Neste caso, o plano de mínimos quadrados é mal definido, no sentido de que inclinando-o ao longo do eixo principal dos pontos resulta em pequena mudança na soma de quadrados residual. Estatisticamente, isto pode ser traduzido no fato de que as estimativas de mínimos quadrados são imprecisas, isto é, elas têm variância elevada.

Figura 1 - Ausência de colinearidade - todos os coeficientes de regressão bem determinados. Uma pequena mudança em qualquer parâmetro do plano causará uma mudança relativamente grande na soma de quadrados residual

Figura 2 - Colinearidade exata - todos os coeficientes de regressão não determinados. Uma mudança simultânea em todos os parâmetros poderá deixar a soma de quadrados residual inalterada

Por outro lado, é importante notar que a colinearidade não prejudica necessariamente todas as estimativas dos parâmetros. Na figura 4, por exemplo, tem-se um caso em que as inclinações parciais estão mal definidas, mas o intercepto permanece bem definido à medida que o plano é inclinado ao longo do eixo de dispersão dos pontos. De maneira

Figura 3 - Colinearidade forte - todos os coeficientes de regressão mal determinados. Uma mudança simultânea em todos os parâmetros pode causar pequena alteração na soma de quadrados residual

Figura 4 - Colinearidade forte - intercepto bem determinado. Mudanças apenas nos coeficientes angulares afetam pouco a soma de quadrados residual

similar (Figura 5), tem-se um caso onde a inclinação parcial na direção de x_2 permanece bem definida. As estimativas de β_0 e β_1 têm precisão baixa, mas a de β_2 não. É interessante notar que, nesta situação, a relação colinear não se dá entre X_1 e X_2 , mas entre X_1 e o intercepto. Por fim, na figura 6 encontra-se uma situação em que há colinearidade entre X_2 e o intercepto, mas um *outlier* faz com que o plano seja bem definido.

Figura 5 - Colinearidade forte - β_2 bem determinado. Mudanças apenas no intercepto e em β_1 afetam pouco a soma de quadrados residual

Figura 6 - Colinearidade forte, embora a presença de um *outlier* tornou os coeficientes bem determinados

Intuitivamente, os danos causados por colinearidades podem ser entendidos notando-se que as variáveis colineares não fornecem informação muito diferente daquela inerente às outras. Assim, torna-se difícil inferir a influência separada de tais variáveis explicatórias na variável resposta.

Efeitos da multicolinearidade

Variáveis explanatórias ortogonais

De acordo com Neter et al. (1990), na análise de regressão linear há alguns pontos de grande interesse: (i) Qual a importância relativa dos efeitos das diferentes variáveis explanatórias? (ii) Qual a magnitude do efeito de uma dada variável explanatória sobre a variável resposta? (iii) Uma variável explanatória pode ser eliminada do modelo por ter um pequeno ou nenhum efeito sobre a variável resposta? (iv) Deve-se considerar a possibilidade de inclusão de alguma variável explanatória ainda não incluída no modelo?

Se as variáveis explanatórias são não correlacionadas entre si e não correlacionadas com outras variáveis explanatórias que são relacionadas à variável resposta mas estão omitidas do modelo, então os efeitos associados às mesmas não mudam quando outras variáveis são incluídas no modelo. De fato, isto constitui um forte argumento para experimentos controlados, sempre que possível, pois o controle experimental permite a obtenção de variáveis não correlacionadas.

Outro aspecto importante diz respeito às somas de quadrados dos erros. Quando duas ou mais variáveis explanatórias são não correlacionadas, a contribuição marginal de uma na redução da soma de quadrados dos erros, quando as outras variáveis estão no modelo, é exatamente a mesma quando esta covariável está sozinha no modelo.

Variáveis explanatórias colineares

Quando as variáveis estão perfeitamente relacionadas e os dados não contêm nenhum componente de erro aleatório, muitas funções conduzirão ao ajuste perfeito. Além disso, os valores ajustados serão os mesmos para quaisquer outras combinações das variáveis explanatórias que seguirem a relação observada na matriz de delineamento. No entanto, essas funções não são as mesmas e conduzirão a valores ajustados diferentes para combinações que não seguirem a relação observada. Duas implicações chave devem ser consideradas:

1. A relação perfeita entre variáveis explanatórias não impede a obtenção de um bom ajuste aos dados.
2. Como muitas funções resultam no mesmo bom ajuste, não se pode interpretar qualquer

conjunto de coeficientes de regressão como refletindo o efeito das diferentes variáveis explanatórias.

Problemas estatísticos

Estatisticamente, o problema da colinearidade em uma matriz de delineamento é a reduzida precisão das estimativas condicionadas aos dados, ou seja, a colinearidade faz com que as variâncias sejam altas. Esta observação pode ser explicada pelo fato de que quando os dados são mal condicionados, alguns dados são praticamente combinações lineares dos demais e portanto, adicionam pouca informação, independente de qual informação estatística adicional possa ser colhida.

Esta questão deve ser avaliada com mais detalhe, o que é feito por Wetherill (1986). Seja k o número de variáveis explanatórias envolvidas e $p = k + 1$ o número de parâmetros, incluindo o intercepto. O modelo de regressão linear pode ser escrito como:

$$E(\mathbf{Y}) = \boldsymbol{\alpha}\mathbf{1} + \mathbf{X}\boldsymbol{\beta} \quad (2)$$

sendo que as colunas de $\mathbf{X}$ ($n \times k$) consistem de variáveis explanatórias centradas, isto é, $\mathbf{1}^T \mathbf{X} = \mathbf{0}$. Assume-se também, por conveniência, que as variáveis explanatórias foram padronizadas, isto é, $\mathbf{x}_i^T \mathbf{x}_i = 1$, em que $\mathbf{x}_i$ é a i -ésima coluna de $\mathbf{X}$. O uso do modelo centrado e padronizado é interessante para mostrar alguns efeitos da colinearidade, mas não é muito adequado para a obtenção das medidas de diagnóstico, conforme será detalhado na próxima seção. Neste ponto, também assume-se que não há dependências exatas nas variáveis explanatórias, mas apenas *quase* dependências lineares. Isto é equivalente a assumir que $\mathbf{X}$ é de posto completo k , de modo que soluções exatas das equações de mínimos quadrados existem.

É possível fazer muitas considerações a respeito das relações entre as multicolinearidades e o comportamento do estimador de mínimos quadrados de $\boldsymbol{\beta}$. Inicialmente, pode-se estabelecer que, se multicolinearidades existem, então ao menos um dos autovalores de $\mathbf{X}^T \mathbf{X}$ é pequeno. De acordo com a definição de dependência linear, pode-se dizer que os k vetores $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, que constituem as colunas de $\mathbf{X}$, exibem *quase* dependência linear

se existe um conjunto de escalares $c_1, c_2, \dots, c_k$ tal que

$$\sum_{i=1}^k c_i \mathbf{x}_i = \boldsymbol{\delta}, \quad (3)$$

em que $\boldsymbol{\delta}$ é suficientemente pequeno. Para consistência, faz-se a restrição $\sum c_i^2 = 1$. Deve-se notar que $\boldsymbol{\delta}$ é um vetor e portanto, não há uma maneira única de definir sua magnitude. No entanto, a norma $\|\boldsymbol{\delta}\|$ é uma medida de tamanho conveniente para a presente dedução, e diz-se que $\boldsymbol{\delta}$ é pequeno se

$$\|\boldsymbol{\delta}\| = (\boldsymbol{\delta}^T \boldsymbol{\delta})^{1/2} < \epsilon, \quad (4)$$

para algum valor pequeno de ϵ . Combinando as equações (3) e (4), tem-se

$$\epsilon > \|\boldsymbol{\delta}\| = \left\| \sum c_i \mathbf{x}_i \right\| = (\mathbf{c}^T \mathbf{X}^T \mathbf{X} \mathbf{c})^{1/2}, \quad (5)$$

em que $\mathbf{c} = (c_1, c_2, \dots, c_k)$, isto é,

$$\mathbf{c}^T \mathbf{X}^T \mathbf{X} \mathbf{c} = \lambda < \epsilon^2. \quad (6)$$

Sejam $\lambda_1, \lambda_2, \dots, \lambda_k$ os autovalores de $\mathbf{X}^T \mathbf{X}$, com os correspondentes autovetores ortonormais $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k$, e seja $\mathbf{V}$ a matriz $k \times k$ cuja i -ésima coluna é $\mathbf{v}_i$. Então, como $\mathbf{X}^T \mathbf{X}$ é uma matriz simétrica não-singular, tem-se

$$\mathbf{V}^T \mathbf{X}^T \mathbf{X} \mathbf{V} = \boldsymbol{\Lambda}, \quad (7)$$

sendo $\boldsymbol{\Lambda}$ uma matriz diagonal com os autovalores de $\mathbf{X}^T \mathbf{X}$ na diagonal. Então, pode-se escolher $\mathbf{c} = \mathbf{V} \boldsymbol{\gamma}$ para algum $\boldsymbol{\gamma}$ apropriado tal que

$$\begin{aligned} \mathbf{c}^T \mathbf{X}^T \mathbf{X} \mathbf{c} &= \boldsymbol{\gamma}^T \mathbf{V}^T \mathbf{X}^T \mathbf{X} \mathbf{V} \boldsymbol{\gamma} \\ &= \boldsymbol{\gamma}^T \boldsymbol{\Lambda} \boldsymbol{\gamma} \\ &= \sum_{i=1}^k \gamma_i^2 \lambda_i \\ &= \lambda. \end{aligned} \quad (8)$$

Agora,

$$\sum_{i=1}^k \gamma_i^2 \lambda_i \geq \text{Min}(\lambda_i) \sum_{i=1}^k \gamma_i^2 = \text{Min}(\lambda_i) \quad (9)$$

em que $\sum \gamma_i^2 = \boldsymbol{\gamma}^T \boldsymbol{\gamma} = \mathbf{c}^T \mathbf{V}^T \mathbf{V} \mathbf{c} = \mathbf{c}^T \mathbf{c} = 1$. Assim,

$$\text{Min}(\lambda_i) \leq \lambda. \quad (10)$$

A igualdade valerá apenas quando $\lambda_i = 1$, $\forall i$, de modo que a matriz $\mathbf{X}$ é ortogonal ou $\mathbf{c}$ é o autovetor correspondendo a $\text{Min}(\lambda_i)$. Então, em geral, o menor autovalor de $\mathbf{X}^T \mathbf{X}$ será menor que o tamanho da combinação linear $\mathbf{X} \mathbf{c}$.

A seguir, investigam-se os efeitos das multicolinearidades no estimador de mínimos quadrados, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, de $\boldsymbol{\beta}$. Da decomposição em autovalores de $\mathbf{X}^T \mathbf{X}$, tem-se que

$$(\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T = \sum \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^T \quad (11)$$

e pode-se escrever

$$\hat{\boldsymbol{\beta}} = \sum \lambda_i^{-1} d_i \mathbf{v}_i, \quad (12)$$

em que $d_i = \mathbf{v}_i^T \mathbf{X}^T \mathbf{Y}$. Por conveniência, assume-se que $\lambda_{\min} = \lambda_k$; então, da equação (12) nota-se que $\hat{\boldsymbol{\beta}}$ tenderá a ser dominado por $\mathbf{v}_k$. Além disso, como $\mathbf{v}_k$ é um autovetor de $\mathbf{X}^T \mathbf{X}$, $\mathbf{X}^T \mathbf{X} \mathbf{v}_k = \lambda_k \mathbf{v}_k$, de tal forma que

$$\mathbf{v}_k^T \mathbf{X}^T \mathbf{X} \mathbf{v}_k = \lambda_k \mathbf{v}_k^T \mathbf{v}_k = \lambda_k \quad (13)$$

e então

$$\|\mathbf{X} \mathbf{v}_k\|^2 = (\mathbf{X} \mathbf{v}_k)^T (\mathbf{X} \mathbf{v}_k) = \lambda_k < \lambda, \quad (14)$$

implicando que $\mathbf{X} \mathbf{v}_k$ é pequeno. Aquelas colunas de $\mathbf{X}$ que correspondem a elementos não desprezíveis de $\mathbf{v}_k$ são, portanto, as variáveis explanatórias originais envolvidas na multicolinearidade. Se há mais de uma multicolinearidade, então haverá mais de um autovalor pequeno de $\mathbf{X}^T \mathbf{X}$ e as covariáveis envolvidas em cada colinearidade podem ser identificadas da mesma maneira explicada acima, usando o autovetor apropriado.

Por fim, considera-se a matriz de variâncias e covariâncias de $\hat{\boldsymbol{\beta}}$. Sabe-se que esta matriz pode ser escrita como

$$\text{V}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \quad (15)$$

Novamente, usando a decomposição em autovalores de $\mathbf{X}^T \mathbf{X}$ pode-se escrever

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= \sigma^2 \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T \\ &= \sigma^2 \sum \lambda_i^{-1} \mathbf{v}_i \mathbf{v}_i^T, \end{aligned} \quad (16)$$

e pode-se notar desta expressão que aqueles elementos de $\hat{\boldsymbol{\beta}}$ que correspondem a elementos não desprezíveis de $\mathbf{v}_k$ terão variâncias e suas covariâncias correspondentes inflacionadas. O quadrado médio do erro $\hat{\boldsymbol{\beta}}$ é:

$$\begin{aligned} \text{EMQ}(\hat{\boldsymbol{\beta}}) &= E \left\{ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \right\} \\ &= \text{tr} \left\{ V(\hat{\boldsymbol{\beta}}) \right\} \\ &= \sigma^2 \text{tr} \left\{ \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T \right\} \\ &= \sigma^2 \sum \lambda_i^{-1}. \end{aligned} \quad (17)$$

Assim, apesar do estimador de mínimos quadrados de $\boldsymbol{\beta}$ ser o estimador linear não viesado de mínima variância, seu quadrado médio residual ainda será grande se existirem colinearidades entre as variáveis explanatórias.

Em suma, nota-se que estimação relativamente imprecisa será obtida na direção dos autovetores que correspondem às multicolinearidades e estimação relativamente precisa será possível na direção dos autovetores remanescentes.

As variâncias inflacionadas são prejudiciais ao uso da regressão como base para testes de hipóteses, estimação e predição. Variâncias demasiadamente elevadas conduzem a testes de significância inconclusivos, assim como a intervalos de confiança amplos. Claramente, esses efeitos podem ser removidos pela adição de dados bem condicionados. No entanto, em muitas aplicações esses dados estão indisponíveis, ou o custo e esforço necessários para sua obtenção são proibitivos. Assim, é aparente a necessidade de ferramentas de diagnóstico que sinalizem a presença de colinearidade e que até mesmo isolem as variáveis envolvidas, pois com elas o pesquisador pode determinar se o esforço em corrigir para colinearidade pode valer a pena.

Neste ponto, é interessante acrescentar que a colinearidade nem sempre é prejudicial. De fato, um grupo de variáveis com alguma relação colinear pode ser ortogonal às demais variáveis, de modo que as estimativas destas últimas não são prejudicadas. Além disso, algumas combinações lineares específicas dos coeficientes de regressão estimados podem ser bem determinadas, mesmo que os coeficientes individuais não o sejam. Adicionalmente, caso a estimativa da variância do erro seja pequena o suficiente, é possível que a variância de algumas estimativas de parâmetros sejam pequenas o bastante para alguns objetivos específicos de testes.

Por outro lado, provar que a colinearidade efetivamente prejudicou a estimação é mais difícil. Para tanto, é preciso fornecer informações de que (1) há fortes *quase* dependências entre as covariáveis, de modo que a colinearidade é problemática, e (2) que as variâncias dos parâmetros de interesse têm uma grande proporção de sua magnitude associadas com a presença de relação(ões) colinear(es), de modo que a colinearidade é potencialmente danosa. Quando essas duas condições são simultaneamente atingidas, os coeficientes de regressão são ditos degradados pela presença de colinearidade. Neste caso, pressupõe-se que intervalos de confiança e de predição, assim como as estimativas pontuais, poderiam ser refinados, caso necessário, pela introdução de dados melhor condicionados.

Fica claro, então, que a capacidade de diagnosticar a colinearidade é muito importante para os usuários de regressão por mínimos quadrados. Tal diagnose consiste de dois elementos relacionados: (1) detectar a presença de relações colineares entre as covariáveis e (2) avaliar a medida em que estas relações degradam os parâmetros estimados. A informação resultante deste diagnóstico permite ao investigador decidir se é necessária e proveitosa alguma ação corretiva, e onde ela deve ser aplicada.

3.1.1 Detecção da Multicolinearidade

Existem muitos procedimentos empregados para detectar as colinearidades, alguns bastante informais, como os seguintes passos, segundo Besley et al. (1991):

1. Verificar a existência de coeficientes de regressão com valores muito altos.
2. Verificar se as variáveis preditoras consideradas importantes tenham valores de t-

Student pequenos para as hipóteses de seus coeficientes.

3. Verificar se a eliminação de uma linha ou coluna de matriz $\mathbf{X}$ produz grandes mudanças no modelo ajustado.
4. Verificar as correlações entre todos os pares de covariâncias para detectar as que são bastante altas.
5. Verificar os sinais esperados para os coeficientes. Por exemplo, um coeficiente aparece com sinal negativo quando um valor positivo era esperado.

Muitas vezes, inclusive, a colinearidade é citada como explicação para essas condições. No entanto, nenhuma dessas condições é *necessária* ou *suficiente* para a existência da colinearidade, sendo necessárias técnicas mais refinadas para sua detecção e avaliação dos danos causados por ela.

O principal procedimento a ser discutido neste trabalho é baseado no fator de inflação da variância do vetor de estimativas β , no índice de condição e na decomposição proporcional da variância dos parâmetros de regressão, proposto por Belsley et al. (1980), porém, outros métodos serão avaliados rapidamente.

Exame da matriz $\mathbf{R}^{-1}$ e dos *Fatores de Inflação de Variância*

Supondo que as variáveis estão centradas e padronizadas, tem-se que $\mathbf{R}^{-1} = (\mathbf{X}^T \mathbf{X})^{-1}$ em que os elementos da diagonal dessa matriz são chamados de fatores de inflação de variância - VIF_i e representam o incremento da variância devido à presença de multicolinearidade. Seu valor como diagnóstico segue da relação:

$$VIF_i = \frac{1}{1 - R_i^2}, \quad (18)$$

sendo R_i^2 o coeficiente de correlação múltipla da regressão linear de X_i em relação às variáveis explanatórias restantes. Se R_i^2 estiver próximo de um, ou seja, existe uma alta correlação entre a variável X_i e as demais, então $(1-R_i^2)$ estará próximo de zero e conseqüentemente o VIF_i assumirá um valor grande, apontando para o envolvimento dessa covariável em colinearidades.

A matriz de variâncias e covariâncias para as estimativas dos coeficientes de regressão padronizados é dada por

$$V(\hat{\beta}) = (\sigma')^2 (\mathbf{X}^T \mathbf{X})^{-1} = (\sigma')^2 \mathbf{R}^{-1}, \quad (19)$$

em que $(\sigma')^2$ é a variância do erro para o modelo transformado. Da equação (19) nota-se que a variância de $\hat{\beta}_i$ ($i = 1, 2, \dots, p-1$) é igual ao produto da variância do erro $(\sigma')^2$ pelo i -ésimo elemento da diagonal da matriz $\mathbf{R}^{-1}$. Assim, tem-se:

$$V(\hat{\beta}_i) = (\sigma')^2 (VIF_i) = \frac{(\sigma')^2}{1 - R_i^2}. \quad (20)$$

Nota-se, então, que esses fatores medem quanto a variância dos coeficientes de regressão estimados são inflacionadas em comparação ao caso em que as variáveis explanatórias não estão linearmente relacionadas (NETER; WASSERMAN; KUTNER, 1990).

O VIF_i é igual a 1 quando $R_i^2 = 0$, isto é, quando X_i não é colinear com as outras variáveis X . Quando $R_i^2 \neq 0$, então VIF_i é maior que 1, indicando uma variância inflacionada para $\hat{\beta}_i$. O maior VIF entre todas as variáveis X é comumente usado como indicador da severidade da multicolinearidade. Um VIF máximo acima de 10 é, geralmente, tomado como indicação de que a colinearidade pode estar influenciando as estimativas de mínimos quadrados.

A média dos VIF s também fornece informação sobre a severidade da colinearidade, em termos de quão distantes os coeficientes de regressão padronizados $\hat{\beta}_i$ s estão dos valores verdadeiros β_i s. Pode-se mostrar que o valor esperado da soma desses erros quadrados $(\hat{\beta}_i - \beta_i)^2$ é dado por:

$$E \left\{ \sum_{i=1}^{p-1} (\hat{\beta}_i - \beta_i)^2 \right\} = (\sigma')^2 \sum_{i=1}^{p-1} (VIF_i). \quad (21)$$

Assim, valores maiores de VIF s resultam, em média, em maiores diferenças entre os coeficientes de regressão padronizados estimados e os verdadeiros.

Quando nenhuma variável X é linearmente relacionada às outras no modelo de regressão, ou seja, $R_i^2 = 0$; tem-se que, $VIF_i = 1$ e conseqüentemente:

$$E \left\{ \sum_{i=1}^{p-1} (\hat{\beta}_i - \beta_i)^2 \right\} = (\sigma')^2 (p-1). \quad (22)$$

A razão dos dois últimos resultados fornece informação útil sobre o efeito da colinearidade na soma de quadrados dos erros:

$$\frac{(\sigma')^2 \sum(VIF_i)}{(\sigma')^2(p-1)} = \frac{\sum(VIF_i)}{(p-1)}. \quad (23)$$

Esta razão é simplesmente a média dos VIF s, denotada por $\overline{VIF}$:

$$\overline{VIF} = \frac{\sum_{i=1}^{p-1}(VIF_i)}{(p-1)}. \quad (24)$$

Valores de $\overline{VIF}$ consideravelmente maiores que 1 são indicativos de sérios problemas com multicolinearidade.

Seu ponto fraco, está relacionado à sua inabilidade em distinguir entre várias *quase* dependências co-existent, além de não haver um limite bem definido para distinguir entre valores de VIF s que podem ser considerados altos e aqueles que podem ser considerados baixos.

Usualmente, valores maiores que 10 indicam a presença de multicolinearidade e podem causar problemas na estimação (CHATTERJEE et al., 2000).

Exame dos autovalores e autovetores da matriz de correlações $\mathbf{R}$

O uso do sistema de autovalores e autovetores da matriz de produtos cruzados $\mathbf{X}^T\mathbf{X}$ ou da sua matriz de correlações $\mathbf{R}$ tem sido empregado para lidar com a colinearidade, tanto para seu diagnóstico quanto para a redução de seus efeitos danosos. De acordo com o que foi dito anteriormente, os autovalores de $\mathbf{X}^T\mathbf{X}$ podem ser usados como a chave para a presença de multicolinearidade: um autovalor pequeno em relação aos demais indica um mau condicionamento da matriz.

Método baseado na Decomposição de Valor Singular

Este método é baseado na união de conceitos desenvolvidos no campo da análise numérica com as medidas baseadas em autovalores. A análise numérica preocupa-se com o condicionamento de uma matriz $\mathbf{A}$ de um sistema de equações $\mathbf{Az} = \mathbf{c}$, que permite que uma solução para $\mathbf{z}$ seja obtida com estabilidade numérica. De fato, esta abordagem está

intimamente relacionada com os problemas causados pela colinearidade, já que o estimador de mínimos quadrados é uma solução para o sistema de equações normais $(\mathbf{X}^T \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y}$ com matriz de covariâncias dada por $\sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$. A colinearidade entre as colunas de $\mathbf{X}$ resulta em uma matriz $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ cujo mau condicionamento faz com que tanto a solução para $\boldsymbol{\beta}$ quanto sua matriz de covariâncias sejam numericamente instáveis.

O objetivo desta metodologia é fornecer um conjunto de índices que sinalizem a presença de uma ou mais *quase* dependências entre as colunas de $\mathbf{X}$. Visa também uma decomposição da variância das estimativas dos coeficientes de regressão, de maneira a descobrir as variáveis envolvidas em *quase* dependências particulares e a avaliar o grau em que os coeficientes estimados estão sendo degradados pela presença da colinearidade.

Decomposição de Valor Singular

Qualquer matriz $\mathbf{X} n \times p$, considerada aqui como uma matriz de n observações em p variáveis, pode ser decomposta como

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T, \quad (25)$$

em que $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V} = \mathbf{I}_p$ e $\mathbf{D}$ é diagonal com elementos μ_k não-negativos na diagonal, chamados de valores singulares de $\mathbf{X}$. Muitos pacotes estatísticos, atualmente, realizam esta decomposição de maneira bastante eficiente e estável. A decomposição acima é válida caso $\mathbf{X}$ tenha ou não sido centrada ou padronizada. No entanto, para os propósitos de diagnóstico de colinearidade a serem detalhados a seguir, é sempre desejável padronizar $\mathbf{X}$ para obter colunas com comprimentos unitários. Além disso, se os dados são relevantes para um modelo com um intercepto, $\mathbf{X}$ deve conter dados não centrados junto com a coluna de 1s. De fato, deve-se evitar o uso da matriz de dados $\mathbf{X}$ centrada, pois esta operação pode mascarar o envolvimento do intercepto em quaisquer *quase* dependências subjacentes e fornecer resultados enganadores. A participação do intercepto em dependências lineares pode ser vista na figura 5.

A decomposição de valor singular está intimamente relacionada com os conceitos de autovalores e autovetores, mas há diferenças importantes. Notando-se que $\mathbf{X}^T \mathbf{X} = \mathbf{V} \mathbf{D}^2 \mathbf{V}^T$, observa-se que $\mathbf{V}$ é uma matriz ortogonal que diagonaliza $\mathbf{X}$ e, portanto, os elementos da diagonal de $\mathbf{D}^2$, os quadrados dos valores singulares, são os autovalores da matriz simétrica $\mathbf{X}^T \mathbf{X}$. Além disto, as colunas ortogonais de $\mathbf{V}$ são os autovetores de $\mathbf{X}^T \mathbf{X}$.

Esta decomposição de valor singular, fornece então informação que abrange aquela dada pelo sistema de autovalores e autovetores de $\mathbf{X}^T \mathbf{X}$. Na prática, no entanto, Belsley et al. (1980) argumentam que o uso da decomposição de valor singular tem razões para ser preferida. Em primeiro lugar, ela aplica-se diretamente à matriz $\mathbf{X}$ que é o foco das preocupações, e não à matriz $\mathbf{X}^T \mathbf{X}$. Além disso, a noção do número de condição de $\mathbf{X}$, que será detalhada posteriormente, é definida corretamente em termos dos valores singulares de $\mathbf{X}$ e não em função das raízes quadradas dos autovalores de $\mathbf{X}^T \mathbf{X}$. Por fim, apesar de as duas abordagens serem matematicamente equivalentes, computacionalmente elas não o são. O cálculo da decomposição de valor singular de $\mathbf{X}$ pode ser realizado de maneira numericamente mais estável do que o cálculo do sistema de autovalores de $\mathbf{X}^T \mathbf{X}$, particularmente no caso em que $\mathbf{X}$ é mal condicionada.

Dependências lineares exatas: deficiência de posto. Inicialmente, assume-se que $\mathbf{X}$ tem dependências exatas entre suas colunas, de modo que $\text{posto}(\mathbf{X}) = r < p$. Como, na decomposição de valor singular de $\mathbf{X}$, $\mathbf{U}$ e $\mathbf{V}$ são ortogonais, e portanto, necessariamente de posto completo, deve-se ter $\text{posto}(\mathbf{X}) = \text{posto}(\mathbf{D})$. Conseqüentemente, haverá tantos elementos nulos na diagonal de $\mathbf{D}$ quanto for a nulidade de $\mathbf{X}$, e a decomposição na equação (25) pode ser particionada como

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T = \mathbf{U} \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}^T, \quad (26)$$

sendo que $\mathbf{D}_{11}$ é $r \times r$ e não-singular. Pós-multiplicando por $\mathbf{V}$ e particionando novamente, obtém-se

$$\mathbf{X} \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (27)$$

em que $\mathbf{V}_1$ é $p \times r$, $\mathbf{U}_1$ é $n \times r$, $\mathbf{V}_2$ é $p \times (p - r)$ e $\mathbf{U}_2$ é $n \times (p - r)$. A equação acima resulta nas duas equações matriciais:

$$\mathbf{X} \mathbf{V}_1 = \mathbf{U}_1 \mathbf{D}_{11} \quad (28)$$

$$\mathbf{X} \mathbf{V}_2 = \mathbf{0}. \quad (29)$$

O interesse está principalmente na equação (29), pois a partir desta, obtém-se todas as dependências lineares de $\mathbf{X}$. A matriz $\mathbf{V}_2$ fornece uma base ortonormal para o espaço

nulo associado com as colunas de $\mathbf{X}$. Se então, $\mathbf{X}$ possuir $p-r$ relações lineares exatas entre suas colunas, haveria exatamente $p-r$ valores singulares iguais a zero em $\mathbf{D}$, e as variáveis envolvidas em cada uma dessas dependências seriam determinadas pelos elementos não nulos de $\mathbf{V}_2$.

Na grande maioria dos casos, as interrelações entre as colunas de $\mathbf{X}$ não são dependências exatas, e os computadores lidam com aritmética finita, não exata. Zeros exatos para os valores singulares ou para os elementos de $\mathbf{V}_2$ raramente ocorrerão. Em geral, então, será difícil determinar a nulidade de $\mathbf{X}$ através de quantidade de μ 's nulos ou as colunas de $\mathbf{X}$ que não participam de relações lineares específicas, zeros em $\mathbf{V}_2$. De qualquer maneira, fica claro que cada *quase* dependência linear entre as colunas de $\mathbf{X}$ resultará na ocorrência de um valor singular pequeno, ou seja, um μ pequeno. A questão que fica, então, é o que pode ser chamado de pequeno. Uma maneira de responder a esta pergunta é através do número de condição de uma matriz $\mathbf{X}$.

O número de condição: Para definir o que vem a ser uma matriz mal condicionada, é comum dizer que a matriz “quase não é de posto completo” ou que “sua inversa quase não existe”. Apesar de estas afirmações parecerem absurdas, elas correspondem ao significado da afirmação que uma matriz mal condicionada é aquela com um determinante pequeno. É preciso notar, no entanto, que um determinante pequeno não tem relação alguma com a invertibilidade da matriz.

Uma maneira intuitiva de definir o condicionamento de uma matriz é suprida pela decomposição de valor singular. A motivação por trás desta técnica é derivada de um método para determinar quando a inversa de uma dada matriz “explode”. É razoável considerar uma matriz $\mathbf{A}$ mal condicionada se o produto de sua norma espectral, definida a seguir, pela norma espectral de $\mathbf{A}^{-1}$, for grande. Esta medida, chamada de *número de condição* de $\mathbf{A}$, fornece informação das potenciais dificuldades a serem encontradas em vários cálculos baseados em $\mathbf{A}$. Quanto maior o número de condição, maior será o mau condicionamento da matriz. Este número está principalmente relacionado aos problemas na obtenção de soluções para sistemas lineares de equações.

A norma Euclidiana de um vetor $\mathbf{z}$ de n elementos, denotada por $\|\mathbf{z}\|$, é

definida como

$$\| \mathbf{z} \| = (\mathbf{z}^T \mathbf{z})^{1/2}. \quad (30)$$

Uma generalização desta norma Euclidiana para uma matriz $\mathbf{A}$ de dimensão $n \times n$ é a norma espectral, denotada $\| \mathbf{A} \|$ e definida como

$$\| \mathbf{A} \| = \sup \| \mathbf{A} \mathbf{z} \| . \quad (31)$$

$$\| \mathbf{z} \| = 1 \quad (32)$$

Pode-se mostrar que $\| \mathbf{A} \| = \mu_{\max}$, isto é, o máximo valor singular de $\mathbf{A}$. De maneira similar, se $\mathbf{A}$ é quadrada, $\| \mathbf{A}^{-1} \| = 1/\mu_{\min}$. Pode-se mostrar também que a norma espectral tem as seguintes propriedades:

1. $\| \lambda \mathbf{A} \| = |\lambda| \cdot \| \mathbf{A} \|$ para todo λ real e qualquer $\mathbf{A}$.
2. $\| \mathbf{A} \| = 0$ se e somente se $\mathbf{A} = \mathbf{0}$.
3. $\| \mathbf{A} + \mathbf{B} \| \leq \| \mathbf{A} \| + \| \mathbf{B} \|$ para quaisquer matrizes $\mathbf{A}$ e $\mathbf{B}$ $m \times n$.
4. $\| \mathbf{A} \mathbf{z} \| \leq \| \mathbf{A} \| \cdot \| \mathbf{z} \|$.
5. $\| \mathbf{A} \mathbf{B} \| \leq \| \mathbf{A} \| \cdot \| \mathbf{B} \|$ para quaisquer $\mathbf{A}$ e $\mathbf{B}$ conformes.

Essa norma espectral é diretamente relevante para a análise do condicionamento de um sistema de equações lineares $\mathbf{A} \mathbf{z} = \mathbf{c}$, $\mathbf{A}$ $n \times n$ e não-singular, com solução $\mathbf{z} = \mathbf{A}^{-1} \mathbf{c}$. O interesse está em saber quanto o vetor solução $\mathbf{z}$ seria alterado ($\delta \mathbf{z}$) se houvesse pequenas mudanças ou perturbações nos elementos de $\mathbf{c}$ ou $\mathbf{A}$, denotados por $\delta \mathbf{c}$ e $\delta \mathbf{A}$. Para o evento em que $\mathbf{A}$ está fixada mas $\mathbf{c}$ muda em $\delta \mathbf{c}$, tem-se $\delta \mathbf{z} = \mathbf{A}^{-1} \delta \mathbf{c}$, ou

$$\| \delta \mathbf{z} \| \leq \| \mathbf{A}^{-1} \| \cdot \| \delta \mathbf{c} \| . \quad (33)$$

Pela propriedade 4 acima, tem-se

$$\| \mathbf{c} \| \leq \| \mathbf{A} \| \cdot \| \mathbf{z} \| ; \quad (34)$$

e multiplicando estas duas últimas expressões obtém-se:

$$\frac{\| \delta \mathbf{z} \|}{\| \mathbf{z} \|} \leq \| \mathbf{A} \| \cdot \| \mathbf{A}^{-1} \| \cdot \frac{\| \delta \mathbf{c} \|}{\| \mathbf{c} \|} . \quad (35)$$

Assim, a magnitude $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ fornece um limite para a mudança relativa no comprimento do vetor solução $\mathbf{z}$ que pode resultar de uma mudança relativa no comprimento de $\mathbf{c}$. Um resultado similar vale para perturbações nos elementos da matriz $\mathbf{A}$. Pode-se mostrar que

$$\frac{\|\delta\mathbf{z}\|}{\|\mathbf{z} + \delta\mathbf{z}\|} \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\| \cdot \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}. \quad (36)$$

Pela sua utilidade neste contexto, a magnitude $\|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ é definida como número de condição da matriz não-singular $\mathbf{A}$ e é denotada por $\kappa(\mathbf{A})$. Essas expressões mostram que $\kappa(\mathbf{A})$ fornece uma medida da sensibilidade da solução de um sistema de equações lineares, como as equações normais de mínimos quadrados, conforme ocorrem mudanças nos elementos de $\mathbf{c}$ e $\mathbf{A}$.

O condicionamento de uma matriz $\mathbf{A}$ pode ser resumido, então, pelo número de condição $\kappa(\mathbf{A})$, definido como o produto do maior valor singular de $\mathbf{A}$, sua norma espectral, e o maior valor singular de $\mathbf{A}^{-1}$. Este conceito pode ser estendido para uma matriz retangular e assim, da decomposição de valor singular, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, pode-se mostrar que a inversa generalizada $\mathbf{X}^+$ de $\mathbf{X}$ pode ser escrita como $\mathbf{V}\mathbf{D}^+\mathbf{U}^T$, em que $\mathbf{D}^+$ é a inversa generalizada de $\mathbf{D}$, que corresponde simplesmente a $\mathbf{D}$ com seus elementos diagonais não-nulos invertidos. Assim, os valores singulares de $\mathbf{X}^+$ são meramente os recíprocos daqueles de $\mathbf{X}$, e o maior valor singular de $\mathbf{X}^+$ é o recíproco do menor valor singular não-nulo de $\mathbf{X}$. Assim, para qualquer matriz $\mathbf{X}$ $n \times p$, pode-se definir seu número de condição como

$$\kappa(\mathbf{X}) = \frac{\mu_{\max}}{\mu_{\min}} \geq 1. \quad (37)$$

Mostra-se que o número de condição de qualquer matriz com colunas ortonormais é unitário, de modo que $\kappa(\mathbf{X})$ atinge seu limite inferior neste caso. Além do mais, $\kappa(\mathbf{X}) = \kappa(\mathbf{X}^+)$, de modo que o número de condição tem a desejável propriedade de fornecer a mesma informação para $\mathbf{X}$ ou para sua inversa generalizada.

Para cada dependência linear exata entre as colunas de $\mathbf{X}$, há um valor singular igual a zero. Analogamente, a presença de *quase* dependências resultará em valores singulares, ou autovalores, pequenos. O grau de mau condicionamento depende de quão pequeno é o menor valor singular em relação ao maior valor, $\mu_{\max}$. Sendo assim, é útil definir

$$\eta_k = \frac{\mu_{\max}}{\mu_k} \quad k = 1, \dots, p \quad (38)$$

como o k -ésimo *índice de condição* da matriz de dados $\mathbf{X}$ $n \times p$ em que $\eta_k \geq 1$ para todo k , com o limite inferior necessariamente ocorrendo para algum k . O maior valor de η_k é também o número de condição da matriz em questão. Um valor singular que é pequeno em relação ao $\mu_{\max}$, tem um alto índice de condição.

Pode-se fazer a seguinte afirmação: há tantas *quase* dependências entre as colunas de $\mathbf{X}$ quantos forem os índices de condição elevados. A ocorrência simultânea de vários η s grandes indica a presença simultânea de mais de uma *quase* dependência linear.

Os valores singulares ou autovalores, tomados isoladamente, não fornecem luz a respeito do condicionamento da matriz; por outro lado, a determinação de quanto um valor singular é pequeno relativo a $\mu_{\max}$ está diretamente relacionado a este problema. Vale notar, no entanto, que não há uma base *a priori* para determinar quão grande um índice de condição deve ser para que haja evidência de colinearidade ou, além disso, para que haja evidência de covariáveis tão colineares que suas presenças estão degradando ou prejudicando as estimativas da regressão. Tais limiares devem ser determinados empiricamente, e a experiência tem mostrado que dependências fracas estão associadas com índices de condição entre 10 e 30, enquanto que relações moderadas a fortes estão associadas com índices de condição variando de 30 a 100.

Decomposição da variância dos coeficientes de regressão

Quando um valor singular qualquer de uma matriz de dados é pequeno em relação a $\mu_{\max}$, este fato é interpretado como indicativo de uma *quase* dependência associada com aquele valor singular. Agora, mostra-se como é possível decompor a variância estimada de cada coeficiente de regressão em uma soma de termos, cada um dos quais associado a um valor singular, fornecendo com isto uma maneira de determinar a intensidade na qual as *quase* dependências, com altos índices de condição, degradam cada variância. Tal decomposição é muito importante, pois permite relacionar o condicionamento da matriz, determinado pela decomposição de valor singular, com a qualidade da análise de regressão, que é determinada pela matriz de variâncias e covariâncias de $\hat{\beta}$.

A matriz de covariâncias do estimador de mínimos quadrados $\hat{\beta}$ é $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$, sendo que σ^2 é a variância comum dos componentes de ε no modelo linear $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$.

Usando a decomposição de valor singular, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, a matriz de covariâncias de $\hat{\boldsymbol{\beta}}$, pode ser escrita como:

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\mathbf{V}\mathbf{D}^{-2}\mathbf{V}^T, \quad (39)$$

ou, para o k -ésimo componente de $\hat{\boldsymbol{\beta}}$,

$$\mathbf{V}(\hat{\beta}_k) = \sigma^2 \sum_j \frac{v_{kj}^2}{\mu_j^2}, \quad (40)$$

em que os μ_j são os valores singulares e $\mathbf{V} = (v_{ij})$.

Nota-se que a equação (40) decompõe a $\mathbf{V}(\hat{\beta}_k)$ em uma soma de componentes, cada um associado a um e apenas um dos p valores singulares μ_j , ou autovalores μ_j^2 . Como os μ_j^2 aparecem no denominador, aqueles componentes associados com *quase* dependências, isto é, com μ_j pequenos, serão grandes em relação aos outros componentes. Isto sugere então, que uma proporção excepcionalmente elevada da variância de dois ou mais coeficientes, concentrada em componentes associados com o mesmo valor singular pequeno, fornece evidências de que a *quase* dependência correspondente está causando problemas.

Para visualizar esta questão em maiores detalhes, define-se a k, j -ésima proporção da decomposição da variância como a proporção da variância do k -ésimo coeficiente de regressão associada com o j -ésimo componente de sua decomposição na equação (40). Para calcular essas proporções, primeiro define-se

$$\phi_{kj} = \frac{v_{kj}^2}{\mu_j^2} \quad \text{e} \quad \phi_k = \sum_{j=1}^p \phi_{kj} \quad k = 1, \dots, p. \quad (41)$$

Então, as proporções da decomposição da variância são

$$\pi_{jk} = \frac{\phi_{kj}}{\phi_k}, \quad k = 1, \dots, p. \quad (42)$$

É interessante e prático organizar essas proporções de decomposição da variância em uma tabela, como mostrado na tabela 1. Em resumo, os π_{jk} fazem uso da informação da decomposição de valor singular sobre *quase* dependências em uma maneira diretamente aplicável ao exame de seus efeitos nas estimativas de regressão.

Na decomposição da variância dada pela equação (40), μ_j pequenos levam a grandes componentes de $\mathbf{V}(\hat{\beta}_k)$. No entanto, é muito importante notar que nem todas as

Tabela 1 - Exemplo da decomposição proporcional da variância em relação aos valores singulares

Valor singular associado	Proporções de			
	$V(\hat{\beta}_1)$	$V(\hat{\beta}_2)$	$\dots$	$V(\hat{\beta}_p)$
μ_1	π_{11}	π_{12}	$\dots$	π_{1p}
μ_2	π_{21}	π_{22}	$\dots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
μ_p	π_{p1}	π_{p2}	$\dots$	π_{pp}

$V(\hat{\beta}_k)$ precisam ser afetadas adversamente por um μ_j pequeno, pois os v_{kj}^2 no numerador podem ser ainda menores. No caso extremo em que $v_{kj} = 0$, $V(\hat{\beta}_k)$ não seria afetada por qualquer dependência entre as colunas de $\mathbf{X}$ que causasse um μ_j . Isto ocorre quando as colunas k e j de $\mathbf{X}$ pertencem a partições mutuamente ortogonais. Assim, os efeitos ruins da multicolinearidade podem ser mitigados para alguns coeficientes pela ortogonalidade, resultando em v_{kj} pequenos.

À primeira vista, parece que a concentração da variância de qualquer um dos coeficientes de regressão em qualquer um dos componentes poderia sinalizar que a colinearidade pode estar causando problemas. No entanto, como duas ou mais variáveis são necessárias para criar uma *quase* dependência, é preciso que duas ou mais variâncias sejam adversamente afetadas por altas proporções da decomposição da variância associadas com um único valor singular.

Isto pode ser mais claramente entendido imaginando-se uma matriz $\mathbf{X}$ com colunas mutuamente ortogonais. Neste caso, a matriz $\mathbf{V}$ da decomposição de valor singular de $\mathbf{X}$ é diagonal, e todo $v_{ij} = 0$ para $i \neq j$. Assim, a tabela associada às proporções da decomposição da variância deve ter a forma mostrada na tabela 2.

Com isso, uma alta proporção de qualquer variância associada com um único valor singular é dificilmente indicativa de colinearidade, pois as proporções de decomposição da variância neste caso são aquelas para uma matriz ortogonal, idealmente condicionada. Refletindo o fato de que duas ou mais colunas de $\mathbf{X}$ devem estar envolvidas em qualquer *quase* dependência, a degradação de uma estimativa de regressão devido à colinearidade pode

Tabela 2 - Exemplo da decomposição proporcional da variância em relação aos valores singulares para uma matriz $\mathbf{X}$ ortogonal

Valor singular associado	Proporções de			
	$V(\hat{\beta}_1)$	$V(\hat{\beta}_2)$	$\cdots$	$V(\hat{\beta}_p)$
μ_1	1	0	$\cdots$	0
μ_2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
μ_p	0	0	$\cdots$	1

ser observada apenas quando um valor singular μ_j pequeno está associado a uma grande proporção da variância de dois ou mais coeficientes. Se, por exemplo, em um caso para $p = 5$, as colunas 4 e 5 de $\mathbf{X}$ são altamente colineares e, fora isso, todas as colunas são mutuamente ortogonais, seria esperado uma tabela da decomposição da variância com a forma mostrada na tabela 3 a seguir.

Tabela 3 - Exemplo da decomposição proporcional da variância em relação aos valores singulares em que $\mathbf{X}$ possui duas colunas colineares

Valor singular associado	Proporções de				
	$V(\hat{\beta}_1)$	$V(\hat{\beta}_2)$	$V(\hat{\beta}_3)$	$V(\hat{\beta}_4)$	$V(\hat{\beta}_5)$
μ_1	1.0	0	0	0	0
μ_2	0	1.0	0	0	0
μ_3	0	0	1.0	0	0
μ_4	0	0	0	1.0	0.9
μ_5	0	0	0	0	0.1

Neste caso, μ_4 tem grande papel tanto na $V(\hat{\beta}_4)$ quanto na $V(\hat{\beta}_5)$.

Uma recomendação de Belsley et al. (1980) é identificar os autovalores (λ_i) que têm índices de condição maiores que 30. Como regra geral, variáveis com proporções da decomposição da variância acima de 0,5 para cada um desses autovalores indicam que as variáveis associadas aos respectivos coeficientes apresentam relação de colinearidade. Este

limiar 0,5 não deve ser interpretado como um valor absoluto, mas sim em termos relativos, isto é, deve ser considerada a razão entre o componente da variância em questão e a fração ainda não explicada da mesma. Medidas de multicolinearidade serão obtidas depois da padronização das variáveis preditoras, como recomendado por Freund e Littell (2000).

3.1.2 Como amenizar a multicolinearidade

A presença de multicolinearidade muitas vezes não afeta a utilidade do modelo, sendo este adequado para fazer inferência sobre a resposta média ou predições, desde que os valores das variáveis explanatórias para as quais a inferência está sendo feita sigam o mesmo padrão de multicolinearidade que os dados nos quais o modelo de regressão é baseado. Assim, uma forma de remediar o problema é restringir o uso do modelo para inferir sobre valores de variáveis explanatórias que seguem o mesmo padrão de multicolinearidade.

Às vezes é possível remover a colinearidade por transformações das variáveis explicativas. Isto é comum no caso de regressão quadrática múltipla em que as interações e as variáveis quadráticas são usadas. Por outro lado, uma solução satisfatória para a multicolinearidade é fazer observações extras no sentido ou sentidos das multicolinearidades.

Um caminho alternativo é deletar covariáveis convenientes do modelo. Desde que a colinearidade seja provocada pela *quase* dependência linear entre as variáveis, a remoção de uma variável de cada um dos conjuntos de colunas envolvidas em cada colinearidade dominará o problema. Portanto, deletar a i -ésima coluna de $\mathbf{X}$ do modelo é efetivamente assumir que o i -ésimo parâmetro de $\boldsymbol{\beta}$, β_i , é zero. Se β_i for diferente de zero, então o estimador de mínimos quadrados de $\boldsymbol{\beta}$ será viesado e com um viés dependendo do tamanho de β_i . Os estimadores de β_j , $j \neq i$, serão também viesados, a menos que a i -ésima coluna de $\mathbf{X}$ seja ortogonal às demais colunas.

Embora o estimador de mínimos quadrados de $\boldsymbol{\beta}$ seja um estimador não viesado de variância mínima, uma melhora no erro médio quadrático de $\boldsymbol{\beta}$ pode ser feita usando os estimadores viesados.

Regressão de Cumeeira

Estimação Viesada

A regressão de cumeeira, proposta por Hoerl e Kennard (1970 a,b), é um dos vários métodos propostos para remediar os problemas de multicolinearidade, alterando o método dos mínimos quadrados para permitir estimadores viesados dos coeficientes de regressão. Quando um estimador tem um viés pequeno e é substancialmente mais preciso que o estimador não viesado, este pode ser escolhido desde que tenha grande probabilidade de estar próximo do valor verdadeiro. A figura 7 ilustra essa situação. O estimador $\hat{\beta}$ é não viesado mas impreciso, enquanto que $\hat{\beta}^*$ é mais preciso, mas tem um pequeno viés. A probabilidade que $\hat{\beta}^*$ caia próximo ao valor verdadeiro de β é maior que para o estimador não viesado $\hat{\beta}$.

Figura 7 - Comparação do estimador viesado com pequena variância em relação ao estimador não-viesado com grande variância

A medida da combinação do efeito do viés e da variação amostral é o valor esperado do quadrado do desvio do estimador $\hat{\beta}^*$ e do valor verdadeiro β . Esta medida é chamada de erro médio quadrático, e pode ser escrita como:

$$E(\hat{\beta}^* - \beta)^2 = V(\hat{\beta}^*) + [E(\hat{\beta}^*) - \beta]^2 \quad (43)$$

Dessa maneira o erro médio quadrático é igual à variância do estimador mais o viés ao quadrado. Note que se o estimador for não viesado, o erro médio quadrático é igual ao estimador da variância.

O modelo comum para uma regressão linear múltipla é:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (44)$$

sendo $\mathbf{y}$ um vetor de observações de ordem $(n \times 1)$; $\mathbf{X}$ uma matriz proposta de ordem $(n \times p)$ e posto p ; $\boldsymbol{\varepsilon}$ um vetor de resíduos de ordem $(n \times 1)$, com $E(\boldsymbol{\varepsilon}) = 0$ e $V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$.

Pelo método *QM*, $\boldsymbol{\beta}$ pode ser estimado como $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ e as estimativas e suas variâncias poderão ser incertas na presença de multicolinearidade. A regressão de cumeeira consiste na adição de coeficientes $k \geq 0$ à diagonal principal da matriz de correlações $(\mathbf{X}^T\mathbf{X})$, causando um decréscimo na variância das estimativas. Dessa maneira, o estimador da regressão de cumeeira de $\boldsymbol{\beta}$ será obtido da seguinte forma:

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}^T\mathbf{y} \quad (45)$$

sendo $\mathbf{K} = \text{diag}(k_1, k_2, \dots, k_p)$, $k_i \geq 0$. Um procedimento bastante usado é $\mathbf{K} = k\mathbf{I}$, $k \geq 0$. Quando $k_i = 0$, para todo i , $\hat{\boldsymbol{\beta}}^*$ se reduz ao estimador de mínimos quadrados.

A matriz composta de variâncias e covariâncias de $\hat{\boldsymbol{\beta}}^*$ será estimada como segue:

$$V(\hat{\boldsymbol{\beta}}^*) = (\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\hat{\sigma}^2 \quad (46)$$

sendo que $\hat{\sigma}^2$ será a estimativa de σ^2 pelo método dos mínimos quadrados.

A regressão de cumeeira é defendida quando a introdução de alguns vieses nas estimativas são equilibrados por um decréscimo substancial na estimação da variância, resultando em um EMQ menor comparado com o obtido pelo método *QM* (HOERL; KENNARD, 1970a).

Os *VIFs* dos coeficientes da regressão de cumeeira são os elementos da diagonal da matriz $(\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \mathbf{K})^{-1}$.

A análise da regressão de cumeeira será feita na forma padronizada do modelo, usando a matriz de correlações. Depois as estimativas serão transformadas e apresentadas nas escalas originais.

Métodos para determinar o valor do parâmetro de cumeeira $\mathbf{K}$

Um valor ideal para o parâmetro de cumeeira $\mathbf{K}$, o qual resulta em um menor EMQ que o obtido no *QM*, depende do vetor parâmetro $\boldsymbol{\beta}$ desconhecido e da variância do

erro σ^2 também desconhecida (HOERL; KENNARD, 1970a). Conseqüentemente, $\mathbf{K}$ precisa ser determinado empiricamente ou obtido dos dados, e não é possível determinar o valor ideal do parâmetro de cumeeira $\mathbf{K}$. Muitos métodos têm sido propostos para obter os valores apropriados, mas não existe um consenso de qual método é o mais adequado. Aqui, o parâmetro de cumeeira $\mathbf{K}$ será estimado através dos seguintes métodos:

Traço de cumeeira e fatores de inflação de variância: O traço de cumeeira é um esboço dos valores dos $p - 1$ coeficientes estimados de regressão de cumeeira padronizados para diferentes valores de k , usualmente entre 0 e 1. Feito o traço de cumeeira, pode-se examinar um valor de k onde as estimativas se estabilizam. Os fatores de inflação de variância, mencionados anteriormente, também podem ser usados na escolha de k tal que o maior VIF seja menor que 10.

O esboço do traço de cumeeira pode ser feito facilmente notando que

$$\boldsymbol{\beta}^* = (\mathbf{X}^T \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y} = \sum_{i=1}^p \frac{1}{\lambda_i + k} \mathbf{v}_i \mathbf{v}_i^T \mathbf{X}^T \mathbf{Y} = \sum_{i=1}^p \frac{1}{\lambda_i + k} d_i \mathbf{v}_i, \quad (47)$$

em que $d_i = \mathbf{v}_i^T \mathbf{X}^T \mathbf{Y}$. Uma vez que os autovetores e autovalores de $\mathbf{X}^T \mathbf{X}$ tenham sido obtidos, estimativas de regressão de cumeeira podem ser facilmente produzidas para algum valor de k .

O traço de cumeeira pode também ser usado para sugerir variáveis para deletar. Algumas variáveis cuja estimativa do parâmetro é instável a cada mudança ou que decresce para zero são candidatos para anulação.

Outro método para estimar $\mathbf{K}$, foi adotado em Roso et al. (2005a), em que os elementos k_i da matriz diagonal $\mathbf{K}$ foram estimados por

$$\hat{k}_i = \theta \frac{VIF_i}{MVIF} \quad (48)$$

sendo que VIF_i é o fator de inflação da variância da i -ésima covariável e $MVIF$ é o maior dos VIF s. O parâmetro θ foi definido como o valor mínimo entre 0 e 1 (incremento de 0.00001) tal que o máximo dos VIF s seja menor que 10.

As estimativas da regressão de cumeeira tendem a ser estáveis no sentido que são pouco afetadas por pequenas mudanças nos dados, e o modelo de regressão de cumeeira estimado pode fornecer boas estimativas da resposta ou predições de novas observações para

níveis das variáveis explicativas fora da região das observações na qual o modelo de regressão foi estimado.

A maior limitação da regressão de cumeieira é que os métodos de inferência não são aplicáveis e as propriedades da distribuição não são conhecidas. Outra limitação é que a escolha das constantes ks , é subjetiva. Embora métodos formais tenham sido desenvolvidos para fazer esta escolha, estes ainda tem suas limitações.

3.1.3 Análise Genética

A análise genética foi baseado no seguinte modelo matemático:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \mathbf{S}\mathbf{c} + \boldsymbol{\varepsilon} \quad (49)$$

em que: $\mathbf{y}$ é o vetor da variável resposta; $\mathbf{X}$ é a matriz de incidência de efeitos genéticos fixos e idade (exceto para a variável resposta peso ao nascimento), associando os elementos de $\boldsymbol{\beta}$ e $\mathbf{y}$; $\boldsymbol{\beta}$ é o vetor de efeitos genéticos fixos e idade; $\mathbf{F}$ é a matriz de incidência de efeitos ambientais fixos, exceto idade; $\mathbf{v}$ é o vetor de efeitos ambientais fixos, exceto idade; $\mathbf{Z}, \mathbf{W}$ e $\mathbf{S}$ são as matrizes de incidência dos efeitos aleatórios, associando respectivamente os elementos de $\mathbf{a}, \mathbf{m}$ e $\mathbf{c}$ à $\mathbf{y}$; $\mathbf{a}$ é o vetor de efeitos aleatórios aditivos direto; $\mathbf{m}$ é o vetor de efeitos aleatórios aditivos maternos; $\mathbf{c}$ é o vetor de efeitos aleatórios não correlacionado do ambiente permanente materno; $\boldsymbol{\varepsilon}$ é o vetor de efeitos do resíduo inerentes a cada observação, $NID(0, \sigma^2)$.

A solução para o modelo (49) foi obtida usando o procedimento descrito abaixo:

Obteve soluções para o $\mathbf{v}$, $\mathbf{a}$, $\mathbf{m}$ e $\mathbf{c}$, usando o modelo seguinte:

$$\mathbf{y}_{(t)} = \mathbf{F}\mathbf{v}_{(t)} + \mathbf{Z}\mathbf{a}_{(t)} + \mathbf{W}\mathbf{m}_{(t)} + \mathbf{S}\mathbf{c}_{(t)} + \boldsymbol{\varepsilon}_{(t)} \quad (50)$$

sendo que t denota a t -ésima iteração e $\mathbf{y}_{(t)} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(t-1)}$. Na primeira iteração, $\hat{\boldsymbol{\beta}}$ foi fixado como os valores obtidos pelo método *QM*.

Usando *QM* ou *RC*, foram obtidas soluções para $\boldsymbol{\beta}$, a partir do modelo seguinte:

$$\mathbf{y}_{(t)} = \mathbf{X}\boldsymbol{\beta}_{(t)} + \boldsymbol{\varepsilon}_{(t)} \quad (51)$$

sendo $\mathbf{y}_{(t)} = \mathbf{y} - \mathbf{F}\hat{\mathbf{v}}_{(t)} - \mathbf{Z}\hat{\mathbf{a}}_{(t)} - \mathbf{W}\hat{\mathbf{m}}_{(t)} - \mathbf{S}\hat{\mathbf{c}}_{(t)}$, e $\hat{\mathbf{v}}$, $\hat{\mathbf{a}}$, $\hat{\mathbf{m}}$ e $\hat{\mathbf{c}}$ as soluções obtidas na primeira etapa da t -ésima iteração.

As etapas anteriores foram repetidas dez vezes e a diferença absoluta entre as duas últimas soluções de $\hat{\beta}$ foram analisadas.

3.1.4 Análise Alternativa

Duas análises foram apresentadas usando o modelo (49): 1) Modelo aditivo-dominante-*QM*, modelo com efeitos aditivos e de heterozigose que foram estimados por *QM*; 2) Modelo aditivo-dominante-*RC*, modelo com efeitos aditivos e de heterozigose que foram estimados por *RC*.

3.1.5 Medida do Viés

Uma relação conhecida entre parâmetro de cuneeira, variância e viés da estimativa da *RC* é que, com o aumento do parâmetro de cuneeira, a variância decresce e o viés aumenta. Dado que $E(\hat{\beta}) = \beta$ e $E(\hat{\beta}^*) = (\mathbf{X}^T \mathbf{X} + \mathbf{K})^{-1} \mathbf{X}^T \mathbf{X} \beta = \mathbf{H} \beta$, uma medida do viés do vetor regressão de cuneeira $\hat{\beta}^*$ será calculada como $1 - \frac{\|\mathbf{H}\|}{\|\mathbf{I}\|} \times 100$, sendo que $\|\cdot\|$ denota a norma euclidiana. Dessa maneira, uma medida de viés próxima de zero para um método *RC* particular, indicará menor viés nas estimativas.

3.2 Material

Os dados para o estudo, que foram pesos ao nascer - PESNAS e ao desmame - PESDES, perímetro escrotal aos 390 dias - CE e escore para musculosidade aos 390 dias MUSC, foram obtidos do banco de dados gerenciado pelo Grupo de Melhoramento Animal e Biotecnologia da Faculdade de Zootecnia e Engenharia de Alimentos da USP, referentes aos animais compostos Montana Tropical® nascidos entre os anos de 1994 e 2008, totalizando informações de 329.737 indivíduos.

Do banco de dados original, eram mantidos apenas os animais com alguma medida válida, a partir da qual foram eliminadas as informações incompletas e inconsistentes, ou seja, somente animais com informações completas de pedigree eram mantidos para a estimação da heterozigose direta e materna.

Devido ao grande número de raças envolvidas na formação da população composta Montana Tropical®, foi usado um sistema denominado NABC (FERRAZ et al., 1999),

no qual as raças foram classificadas em tipos biológicos da seguinte maneira:

Tipo Biológico *N*: animais da raça *Bos taurus indicus*, os zebuínos, incluindo as raças adaptadas no Brasil, como a Gir, Guzerá, Indubrasil, Nelore, Tabapuã e outros zebuínos de origem africana. Estas raças apresentam alta rusticidade, resistência aos parasitas, bom rendimento de carcaça e são representadas principalmente por vacas da raça Nelore.

Tipo Biológico *A*: bovinos de origem não zebuína adaptados aos trópicos, através de seleção natural ou artificial. Animais destas raças têm alto grau de adaptabilidade ao clima tropical e fertilidade, além de algumas boas características relacionadas à qualidade da carne.

Tipo Biológico *B*: raças taurinas de origem britânica, como as raças Angus, Devon, Hereford etc. Estas raças contribuem para com a precocidade sexual e de acabamento, além de características de conformação frigorífica, qualidade de carcaça, de carne e de crescimento.

Tipo Biológico *C*: raças taurinas da Europa continental, incluindo as raças Charolesa, Limousin, Simental etc. Estas raças têm alto potencial de crescimento, de rendimento e de qualidade de carcaça.

Na predição dos efeitos de origem genética aditiva, associados aos tipos biológicos, foram utilizadas as porcentagens dos tipos biológicos, dadas como a proporção de cada raça na composição genética do bezerro e na composição genética da mãe do bezerro, respectivamente. Analogamente, os efeitos das heterozigoses eram obtidos por relação linear aos coeficientes das heterozigoses diretas ($\mathbf{H_D}$) e materna total ($\mathbf{H_M}$), sendo estes obtidos através das seguintes equações:

$$H_D = 1 - \sum_{i=1}^4 S_i \times D_i \quad (52)$$

$$H_M = 1 - \sum_{i=1}^4 MGS_i \times MGD_i \quad (53)$$

sendo quatro o número de tipos biológicos (*N*, *A*, *B*, *C*). S_i , D_i , MGS_i e MGD_i são as frações do *i*-ésimo tipo biológico do pai, mãe, avô materno e avó materna, respectivamente.

A definição de grupos de contemporâneos - GC foi realizada combinando-se a safra de nascimento, o sexo do bezerro, a fazenda e o grupo de manejo do animal nas respectivas fases de mensuração.

Para a variável resposta PESNAS, o modelo incluiu as covariáveis associadas aos

efeitos genéticos aditivos diretos e maternos, as covariáveis associadas aos efeitos genéticos não aditivos das heterozigoses, o efeito da heterozigose materna total - HM , o efeito quadrático de data Juliana - $DTJN$, de grupos de contemporâneos - GC , de classe de idade da mãe ao parto - $CIMP$, além dos efeitos aleatórios aditivos diretos, dos efeitos aleatórios aditivos maternos e dos efeitos aleatórios não correlacionado do ambiente permanente materno, obtendo o seguinte modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \mathbf{S}\mathbf{c} + \boldsymbol{\varepsilon} \quad (54)$$

em que: $\boldsymbol{\beta}$ é o vetor associado a uma constante, aos efeitos genéticos aditivos diretos associado às composições raciais individuais (A , B e C), aos efeitos genéticos aditivos maternos associados às composições raciais maternas (AM , BM , CM), as heterozigoses diretas (NxA , NxB , NxC , AxB , AxC e BxC) e a heterozigose materna total - HM ; $\mathbf{v}$ é o vetor associado aos efeitos $DTJN$ (quadrático), $CIMP$ e GC ; $\mathbf{a}$, $\mathbf{m}$ e $\mathbf{c}$ são os vetores associados aos efeitos aleatórios aditivos diretos, maternos e aos efeitos aleatórios não correlacionado do ambiente permanente materno; $\boldsymbol{\varepsilon}$ é o vetor de resíduos.

Para a variável resposta PESDES, o modelo inclui as covariáveis associadas aos efeitos genéticos aditivos diretos e maternos, as covariáveis associadas aos efeitos genéticos não aditivos das heterozigoses, as covariáveis associadas ao efeito da heterozigose materna total - HM e idade ao desmame - $IDDES$, o efeito cúbico de data Juliana - $DTJN$, de grupos de contemporâneos - GC , de classe de idade da mãe ao parto - $CIMP$, além dos efeitos aleatórios aditivos diretos, dos efeitos aleatórios aditivos maternos e dos efeitos aleatórios não correlacionado do ambiente permanente materno, obtendo o seguinte modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \mathbf{S}\mathbf{c} + \boldsymbol{\varepsilon} \quad (55)$$

em que: $\boldsymbol{\beta}$ é o vetor associado a uma constante inerente à característica, aos efeitos genéticos aditivos diretos associado às composições raciais individuais (A , B e C), aos efeitos genéticos aditivos maternos associados às composições raciais maternas (AM , BM , CM), as heterozigoses diretas (NxA , NxB , NxC , AxB , AxC e BxC), a heterozigose materna total - HM e a $IDDES$; $\mathbf{v}$ é o vetor associado aos efeitos $DTJN$ (cúbico), $CIMP$ e GC ; $\mathbf{a}$, $\mathbf{m}$ e $\mathbf{c}$ são os vetores associados aos efeitos aleatórios aditivos diretos, maternos e aos efeitos aleatórios não correlacionado do ambiente permanente materno; $\boldsymbol{\varepsilon}$ é o vetor de resíduos.

Para a variável resposta CE, o modelo inclui as covariáveis associadas aos efeitos genéticos aditivos diretos e maternos, as covariáveis associadas aos efeitos genéticos não aditivos das heterozigoses, as covariáveis associadas ao efeito da heterozigose materna total - *HM* e idade de mensuração de CE - *IDCE*, o efeito cúbico de data Juliana - *DTJN*, de classe de idade da mãe ao parto - *CIMP*, de grupos de contemporâneos - *GC*, além dos efeitos aleatórios aditivos diretos e dos efeitos aleatórios aditivos maternos, obtendo o seguinte modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \boldsymbol{\varepsilon} \quad (56)$$

em que: $\boldsymbol{\beta}$ é o vetor associado a uma constante inerente à característica, aos efeitos genéticos aditivos diretos associado às composições raciais individuais (*A*, *B* e *C*), aos efeitos genéticos aditivos maternos associados às composições raciais maternas (*AM*, *BM*, *CM*), as heterozigoses diretas (*NxA*, *NxB*, *NxC*, *AxB*, *AxC* e *BxC*), a heterozigose materna total - *HM* e a *IDCE*; $\mathbf{v}$ é o vetor associado aos efeitos de *DTJN* (cúbico), *CIMP* e *GC*; $\mathbf{a}$ e $\mathbf{m}$ são os vetores associados aos efeitos aleatórios aditivos diretos e maternos respectivamente; $\boldsymbol{\varepsilon}$ é o vetor de resíduos.

Para a variável resposta MUSC, o modelo inclui as covariáveis associadas aos efeitos genéticos aditivos diretos e maternos, as covariáveis associadas aos efeitos genéticos não aditivos das heterozigoses, as covariáveis associadas ao efeito da heterozigose materna total - *HM* e idade de mensuração de MUSC - *IDMUSC*, o efeito quadrático de data Juliana - *DTJN*, de classe de idade da mãe ao parto - *CIMP*, de grupos de contemporâneos - *GC*, além dos efeitos aleatórios aditivos diretos e dos efeitos aleatórios aditivos maternos, obtendo o seguinte modelo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{F}\mathbf{v} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{m} + \boldsymbol{\varepsilon} \quad (57)$$

em que: $\boldsymbol{\beta}$ é o vetor associado a uma constante, aos efeitos genéticos aditivos diretos associado às composições raciais individuais (*A*, *B* e *C*), aos efeitos genéticos aditivos maternos associados às composições raciais maternas (*AM*, *BM*, *CM*), as heterozigoses diretas (*NxA*, *NxB*, *NxC*, *AxB*, *AxC* e *BxC*), a heterozigose materna total - *HM* e a *IDMUSC*; $\mathbf{v}$ é o vetor associado aos efeitos de *DTJN* (quadrático), *CIMP* e *GC*; $\mathbf{a}$ e $\mathbf{m}$ são os vetores associados aos efeitos aleatórios aditivos diretos e maternos respectivamente; $\boldsymbol{\varepsilon}$ é o vetor de resíduos.

As estimativas de componentes de (co)variâncias foram obtidas por procedimento de máxima verossimilhança restrita, utilizando o programa MTDFREML (BOLDMAN et al., 1995) e o critério de convergência de variância do simplex de 10^{-9} . No modelo foram usadas todas as 398301 informações disponíveis de pedigree no cálculo da matriz de parentesco. Para todas as variáveis respostas, os BLUPs foram estimados pelo programa PEST (GROENEVELD, 1990). As estimativas dos componentes de (co)variância serão identificadas como na tabela 4.

Tabela 4 - Definição das estimativas dos componentes de (co)variância

(Co)Variância	Notação
Estimativa da variância genética animal	$\hat{\sigma}_a^2$
Estimativa da variância genética materna	$\hat{\sigma}_m^2$
Estimativa da covariância genética entre efeito animal e materno	$\hat{\sigma}_{am}$
Estimativa da variância permanente de ambiente	$\hat{\sigma}_{ep}^2$
Estimativa da variância residual	$\hat{\sigma}_e^2$
Estimativa da variância fenotípica	$\hat{\sigma}_p^2$

3.3 Métodos

Para detectar quais variáveis que estavam envolvidas em *quase* dependências foram utilizados os fatores de inflação da variância - *VIF*. Variáveis cujo o *VIF* eram maiores que 10, foram consideradas envolvidas em colinearidades.

Os índices de condição foram usados afim de descobrir quantas colinearidades existiam no modelo. Há tantas colinearidades entre as variáveis explicativas quantos forem os índices de condição elevados. Índices de condição entre 10 e 30 indicam *quase* dependências fracas, enquanto que relações moderadas a fortes estão associadas com índices de condição variando entre 30 e 100. Foram estudadas apenas as colinearidades fortes, ou seja, cujo *ICs* eram maiores que 30.

Para apontar quais variáveis explicativas que estavam envolvidas em cada colinearidade foi empregado o método da decomposição da variância dos coeficientes de regressão.

A regressão de cumeeira foi utilizada para abrandar o problema da multicolinearidade, o método (48) adotado por Roso et al. (2005a) e o traço de cumeeira foram usados para determinar o valor das constantes k_i .

A solução para o modelo (49) foi obtida, sendo que as estimativas de β foram calculadas por *QM* e *RC*.

3.4 Resultados e Discussão

3.4.1 Peso ao Nascimento

A variável resposta peso ao nascimento possuía 193.135 observações com média igual a 32,7kg, desvio padrão de 4,5kg, peso mínimo de 21kg e peso máximo de 45kg.

Para verificar quais variáveis explicativas da matriz $\mathbf{X}$ do modelo (54) que estavam envolvidas em colinearidades, seus *VIFs* (Tabela 5) foram analisados.

Tabela 5 - Fatores de Inflação da Variância - *VIF* para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (54) em relação à variável resposta peso ao nascimento

Variáveis Explicativas	<i>VIF</i>	Variáveis Explicativas	<i>VIF</i>
<i>A</i>	21,60315	<i>NxB</i>	17,03950
<i>B</i>	15,16588	<i>NxC</i>	20,45722
<i>C</i>	39,39728	<i>AxB</i>	6,84234
<i>AM</i>	7,67965	<i>AxC</i>	16,29217
<i>BM</i>	21,36792	<i>BxC</i>	6,66024
<i>CM</i>	17,13167	<i>HM</i>	2,89265
<i>NxA</i>	8,84698		

De acordo com a tabela 5, observou-se que as variáveis explanatórias que possuíam *VIFs* maiores que 10 foram: os efeitos aditivos diretos associados aos tipos biológicos *A* (21,60315), *B* (15,16588) e *C* (39,39728); os efeitos aditivos maternos *BM* (21,36792) e *CM* (17,13167); e as heterozigoses diretas *NxB* (17,03950), *NxC* (20,45722) e *AxC* (16,29217), indiciando as covariáveis envolvidas em uma ou mais

colinearidades, pois as variâncias das estimativas de seus coeficientes de regressão estavam sendo inflacionadas mais que dez vezes se comparadas com as estimativas obtidas se o sistema fosse ortogonal. A correlação entre efeitos aditivos diretos e maternos relacionados ao mesmo tipo biológico também foi observada por Rodríguez-Almeida, Van Vleck e Gregory (1997), para a diferença de peso entre o nascimento e o desmame em três populações compostas.

Para verificar quantas colinearidades existem entre as covariáveis da matriz $\mathbf{X}$ do modelo (54), seus índices de condição foram analisados (Tabela 6), com seus respectivos autovalores.

Tabela 6 - Autovalores e Índices de Condição - IC s da matriz $\mathbf{X}$ do modelo (54) para a variável resposta peso ao nascimento

Autovalor	Índice de Condição	Autovalor	Índice de Condição
6,37309	1,00000	0,15727	6,36585
2,24066	1,68650	0,10562	7,76781
1,75257	1,90694	0,04839	11,47664
1,18252	2,32151	0,02588	15,69398
0,96852	2,56519	0,01809	18,77038
0,84021	2,75411	0,01285	22,27093
0,26780	4,87830	0,00655	31,19875

Pela tabela 6, observou-se que um autovalor é relativamente pequeno, 0,00655 e o IC 31,19875 associado a este, ficou acima do limiar 30, indicando uma colinearidade forte entre as variáveis explicativas. Os índices de condição 22,27093, 18,77038, 15,69398 e 11,47664, que estavam associados com os respectivos autovalores 0,01285, 0,1809, 0,02588 e 0,04839, ficaram entre 10 e 30, indicando quatro relações de *quase* dependência consideradas fracas.

A partir da decomposição proporcional da variância (Tabela 7), utilizando o limiar 0,5, observou-se na linha associada ao índice de condição igual a 31,19875, que as covariáveis B (0,569), AM (0,574), BM (0,648) e NxB (0,530) apresentaram colinearidade entre si. Notou-se que as covariáveis A (0,468), C (0,473) e CM (0,478) apresentaram valores próximos ao limiar, indicando um possível envolvimento na colinearidade.

Tabela 7 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição - IC da matriz $\mathbf{X}$ do modelo (54), para a variável resposta peso ao nascimento

IC	INT	A	B	C	AM	BM	CM	NxA	NxB	NxC	AxB	AxC	BxC	HM
1,000	0,001	0,000	0,000	0,000	0,001	0,000	0,000	0,001	0,000	0,000	0,001	0,000	0,001	0,003
1,687	0,000	0,000	0,001	0,001	0,000	0,002	0,004	0,000	0,000	0,001	0,006	0,003	0,003	0,000
1,907	0,001	0,001	0,003	0,000	0,001	0,000	0,000	0,002	0,009	0,001	0,001	0,001	0,005	0,002
2,322	0,000	0,000	0,000	0,001	0,001	0,000	0,003	0,000	0,002	0,024	0,002	0,003	0,001	0,000
2,565	0,000	0,000	0,000	0,000	0,059	0,003	0,001	0,000	0,004	0,000	0,003	0,000	0,024	0,000
2,754	0,001	0,000	0,000	0,000	0,042	0,000	0,000	0,005	0,003	0,002	0,000	0,003	0,063	0,001
4,878	0,001	0,001	0,006	0,001	0,004	0,025	0,002	0,057	0,004	0,001	0,000	0,032	0,062	0,004
6,366	0,005	0,002	0,004	0,000	0,010	0,014	0,010	0,060	0,010	0,000	0,124	0,002	0,004	0,305
7,768	0,000	0,009	0,006	0,001	0,006	0,006	0,004	0,010	0,009	0,000	0,318	0,004	0,017	0,562
11,477	0,017	0,001	0,026	0,014	0,003	0,002	0,434	0,001	0,053	0,007	0,004	0,232	0,121	0,056
15,694	0,803	0,000	0,164	0,004	0,061	0,007	0,014	0,224	0,013	0,022	0,043	0,001	0,005	0,003
18,770	0,138	0,499	0,046	0,023	0,195	0,097	0,015	0,198	0,157	0,057	0,474	0,018	0,032	0,017
22,271	0,020	0,016	0,174	0,483	0,043	0,196	0,036	0,050	0,207	0,441	0,017	0,556	0,499	0,017
31,199	0,014	0,468	0,569	0,473	0,574	0,648	0,478	0,392	0,530	0,445	0,007	0,143	0,165	0,030

Pela decomposição proporcional da variância, a variável explicativa AM está envolvida na colinearidade, porém seu VIF foi igual a 7,67965, ou seja, foi menor que o limiar. Com isso o método da decomposição proporcional da variância pareceu ser menos restritivo que os valores dos $VIFs$.

Algumas das variáveis explanatórias que apresentaram $VIFs$ maiores que 10 e não apareceram na colinearidade forte, podem estar envolvidas em colinearidades fracas. Este mesmo comportamento foi observado por Roso et al. (2005a), para a característica ganho de peso pré-desmame de animais de raça pura e cruzados, envolvendo as raças Angus, Blonde d'Aquitaine, Charolesa, Gelbvieh, Hereford, Limousin, Maine-Anjou, Salers, Shorthorn e Simmental.

Após ter sido verificada a existência de uma colinearidade, as etapas (50) e (51) foram realizadas 10 vezes. As estimativas dos coeficientes de regressão do vetor β obtidas por QM e por RC , os elementos (k_i) da matriz diagonal $\mathbf{K}$, assim como a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações estão na tabela 8.

Concluiu-se que a variável explicativa A que tinha efeito negativo em relação à variável peso ao nascimento, por QM , passou a ter efeito positivo por RC , enquanto que as soluções obtidas para B e C , diferiram em magnitude, mas não em direção. Os efeitos aditivos maternos AM , BM e CM foram maiores que NM , tanto pelo método de QM , quanto por RC , porém em magnitudes diferentes. Estes mostram a importância do efeito da raça materna sobre a variável peso ao nascimento. Pelos dois métodos, filhos de mães do tipo biológico B , foram mais pesados ao nascimento em relação aos outros tipos biológicos maternos.

Resultados semelhantes para estes efeitos foram encontrados por Franke et al. (2001), para as raças Angus, Hereford e Charolesa em relação a raça Brahman, com animais criados em ambiente subtropical.

De acordo com a diferença absoluta entre as estimativas dos coeficientes de regressão para as covariáveis genéticas obtidas por RC nas duas últimas iterações (Tabela 8), foi possível verificar a magnitude do erro. Por exemplo, para a estimativa do coeficiente de regressão da variável explicativa A que é 0,00456, a diferença das soluções entre as iterações

foi de 0,00020, conseqüentemente, o erro foi menor que 2×10^{-4} .

Tabela 8 - Valores dos coeficientes de regressão (β) para o modelo (54), estimados pelo método dos quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $|\hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^*|$

Covariáveis	$\hat{\beta}$	$\hat{\beta}^*$	k_i	$ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $
<i>INT</i>	32,48600	31,77778	—	0,00759
<i>A</i>	-0,01103	0,00456	0,01506	0,00020
<i>B</i>	-0,02702	-0,00349	0,01057	0,00072
<i>C</i>	0,00588	0,01266	0,02746	0,00000
<i>AM</i>	0,02599	0,01175	0,00535	0,00034
<i>BM</i>	0,04334	0,01533	0,01489	0,00070
<i>CM</i>	0,03266	0,00787	0,01194	0,00039
<i>NxA</i>	0,01104	0,00608	0,00617	0,00014
<i>NxB</i>	0,01047	0,00255	0,01188	0,00040
<i>NxC</i>	0,00428	0,00541	0,01426	0,00053
<i>AxB</i>	0,00267	0,00479	0,00477	0,00011
<i>AxC</i>	0,00327	0,01367	0,01136	0,00010
<i>BxC</i>	0,00449	0,01084	0,00464	0,00005
<i>HM</i>	0,00417	0,00841	0,00464	0,00036

Da tabela 9 concluiu-se que o método que estimou as covariáveis por RC foi melhor que o método que estimou por QM , pois a soma dos quadrados dos desvios - SQD obtida por RC (1.399.384, 8) foi menor que a obtida por QM (1.490.056, 8).

Segundo Pimentel (2004), quando a análise de regressão é feita com a finalidade de predição, pouca diferença se pode esperar ao usar RC ou QM , em caso de presença de colinearidades. Bergmann e Hohenboken (1995), Freund e Wilson (1998) recomendam a RC para os casos em que o interesse, na análise de regressão, é a interpretação dos coeficientes de regressão individuais. Nesse sentido, observou-se que o maior dos $VIFs$ (9,99855) por RC , ficou abaixo do limiar 10 e a média dos $VIFs$ (5,80979) por RC , ficou mais próxima de 1 se

comparado com a média dos VIF s por QM (15,49051). Conseqüentemente, os coeficientes de regressão obtidos por RC tiveram desvios padrões menores se comparados com os obtidos por QM (Tabela 10). A medida do viés foi igual a 9,62268%. Esta medida de viés foi obtida por Roso et al. (2005b) para a variável resposta peso pré-desmama por dois métodos diferentes de determinar $\mathbf{K}$, com valores iguais a 1,49 e 5,61%, respectivamente.

Tabela 9 - Soma dos quadrados dos desvios - SQD do modelo (54), máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - QM e regressão de cumeira - RC e medida do viés para RC - viés (%)

	QM	RC
SQD	1.490.056,8	1.399.384,8
$MVIF$	39,39729	9,99855
$\overline{VIF}$	15,49051	5,80979
viés (%)	—	9,62268

Tabela 10 - Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeieira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$

	$\hat{V}(\hat{\beta})$	$\hat{s}(\hat{\beta})$	$\hat{V}(\hat{\beta}^*)$	$\hat{s}(\hat{\beta}^*)$
<i>A</i>	166,67046	12,91009	77,13983	8,78293
<i>B</i>	117,00628	10,81694	49,25177	7,01796
<i>C</i>	303,95395	17,43428	49,43428	7,03095
<i>AM</i>	59,24927	7,69735	28,24958	5,31503
<i>BM</i>	164,85564	12,83961	56,51162	7,51742
<i>CM</i>	132,17255	11,49663	60,05936	7,74980
<i>NxA</i>	68,25534	8,26168	40,30680	6,34876
<i>NxB</i>	131,46145	11,46566	54,69373	7,39552
<i>NxC</i>	157,82950	12,56302	32,50956	5,70172
<i>AxB</i>	52,78934	7,26563	40,15101	6,33648
<i>AxC</i>	125,69572	11,21141	51,22992	7,15751
<i>BxC</i>	51,38442	7,16829	22,20233	4,71194
<i>HM</i>	22,31708	4,72410	20,96111	4,57833

As estimativas para as covariáveis $DTJN$, $DTJN^2$ e para variáveis classificatórias $CIMP$ e GC foram obtidas na etapa (50). Apenas as estimativas de $DTJN$, $DTJN^2$ e $CIMP$ estão apresentadas na tabela 11. As estimativas das $CIMPs$ estão como desvio da $CIMP_4$.

Os efeitos das $CIMPs$ foram importantes, alcançando magnitude de ajuste de até $3kg$, diferença observada entre $CIMP_1$ e $CIMP_4$. Houve pouca diferença entre as estimativas obtidas levando em consideração QM e RC , dentro do esperado, pois estes efeitos não foram considerados envolvidos em colinearidades. Valores semelhantes aos obtidos nesse trabalho, foram obtidos por Zampar et al. (2006) para uma população multirracial em relação à variável resposta peso ao nascimento.

Tabela 11 - Estimativa dos coeficientes de regressão das covariáveis data juliana - *DTJN* e da variável classificatória classe de idade da mãe ao parto - *CIMP* em relação à variável resposta peso ao nascimento

	<i>QM</i>	<i>RC</i>
<i>DTJN</i>	-0,05794	-0,04922
<i>DTJN</i> ²	0,00067	0,00051
<i>CIMP</i> ₁	-2,99680	-3,00120
<i>CIMP</i> ₂	-1,40950	-1,39860
<i>CIMP</i> ₃	-0,46190	-0,46420
<i>CIMP</i> ₄	0,00000	0,00000
<i>CIMP</i> ₅	-0,21660	-0,20170
<i>CIMP</i> ₆	-0,74260	-0,72050
<i>CIMP</i> ₇	-0,89040	-0,87530

Para o modelo (54) as estimativas dos componentes das variâncias genéticas animal - $\hat{\sigma}_a^2$ e materna - $\hat{\sigma}_m^2$, a estimativa da covariância entre os efeitos genéticos animal e materno - $\hat{\sigma}_{am}$, as estimativas das variâncias de efeito permanente de ambiente - $\hat{\sigma}_{ep}^2$ e residual - $\hat{\sigma}_e^2$, assim como a fenotípica - $\hat{\sigma}_p^2$ estão na tabela 12. Valores próximos foram obtidos por Mourão et al. (2007), o que era esperado, pois a diferença entre os conjuntos de dados é que nesse caso, o conjunto contém também dados obtidos entre os anos de 2006 até 2008.

Tabela 12 - Estimativas dos componentes de (co)variâncias¹, para peso ao nascimento, em que o as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - *QM* e regressão de cumeieira - *RC*

	$\hat{\sigma}_a^2$	$\hat{\sigma}_m^2$	$\hat{\sigma}_{am}$	$\hat{\sigma}_{ep}^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$
<i>QM</i>	4,32111	0,56634	-0,43207	0,55768	10,32590	14,90689
<i>RC</i>	4,49020	0,58462	-0,49338	0,56892	10,24324	14,89972

¹ definidos na tabela 4

3.4.2 Peso ao Desmame

A variável resposta peso ao desmame possuía 191.036 observações com média igual a 198,4kg, desvio padrão de 36,7kg, peso mínimo de 97kg e peso máximo de 300kg com idade de pesagem entre 82 e 342 dias.

Para verificar quais covariáveis da matriz $\mathbf{X}$ do modelo (55) que estavam envolvidas em colinearidades, seus *VIFs* (Tabela 13) foram analisados.

Tabela 13 - Fatores de Inflação da Variância - *VIF* para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (55), para a variável resposta peso ao desmame

Variáveis Explicativas	<i>VIF</i>	Variáveis Explicativas	<i>VIF</i>
<i>A</i>	22,67645	<i>NxB</i>	19,15956
<i>B</i>	15,45325	<i>NxC</i>	27,00360
<i>C</i>	46,90675	<i>AxB</i>	7,45133
<i>AM</i>	7,46980	<i>AxC</i>	18,88209
<i>BM</i>	21,79265	<i>BxC</i>	7,02547
<i>CM</i>	17,11994	<i>HM</i>	3,06195
<i>NxA</i>	9,33911	<i>IDDES</i>	1,03785

Pela tabela 13, observou-se que os *VIFs* associados aos efeitos aditivos diretos *A* (22,67645), *B* (15,45325) e *C* (46,90675); aos efeitos aditivos maternos *BM* (21,79265) e *CM* (17,11994); e às heterozigotes diretas *NxB* (19,15956), *NxC* (27,00360) e *AxC* (18,88209) estavam acima do limiar 10 e com isso, há indícios de que estas variáveis explicativas estejam envolvidas em uma ou mais colinearidades, pois suas variâncias estavam sendo inflacionadas mais que dez vezes se comparadas com as estimativas obtidas caso os vetores das variáveis explicativas fossem ortogonais. Resultados parecidos foram obtidos para a variável peso ao nascimento. Rodríguez-Almeida, Van Vleck e Gregory (1997) em seu trabalho com população composta de touros e vacas das raças Red Poll, Hereford, Angus, Limousin, Braunvieh, Pinzgauer, Gelbvieh, Simental e Charolesa observaram também a correlação entre efeitos aditivos e maternos para o mesmo tipo biológico.

A partir dos autovalores e de seus respectivos índices de condição (Tabela 14),

foi verificado o número de colinearidades existentes entre as covariáveis da matriz $\mathbf{X}$ do modelo (55).

Tabela 14 - Autovalores e Índices de Condição da matriz $\mathbf{X}$ do modelo (55) para peso ao desmame

Autovalor	Índice de Condição	Autovalor	Índice de Condição
7,22134	1,00000	0,10111	8,45089
2,27721	1,78077	0,04838	12,21739
1,83985	1,98115	0,03662	14,04356
1,21333	2,43960	0,01723	20,47432
0,97395	2,72296	0,01113	25,47055
0,84708	2,91976	0,00706	31,97141
0,25270	5,34574	0,00608	34,47013
0,14693	7,01063		

Como os autovalores 0,00608 e 0,00706 foram relativamente pequenos, seus respectivos índices de condição iguais a 34,47013 e 31,97141 ficaram acima de 30, indicando duas colinearidades fortes entre as covariáveis. Os autovalores 0,01113, 0,01723, 0,03662 e 0,04838 associados aos índices de condição 25,47055, 20,47432, 14,04356 e 12,21739, respectivamente, indicaram quatro colinearidades consideradas fracas.

A partir da decomposição proporcional da variância (Tabela 15), utilizando o limiar 0,5, observou-se na linha associada ao índice de condição igual a 34,47013 que apenas a covariável *BM* (0,565) possuía valor maior que o limiar, porém, as variáveis explicativas *AM* (0,471), *NxB* (0,480) e *NxC* (0,457) estavam próximas de 0,5 e com isso estas covariáveis podem formar uma relação de *quase* dependência considerada forte, enquanto que na linha associada ao índice de condição 31,97141, o intercepto - *INT* (0,766) e a covariável *IDDES* (0,714) estavam envolvidos em uma relação de colinearidade forte.

Tabela 15 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (55) para a variável peso ao desmame

<i>IC</i>	<i>INT</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>AM</i>	<i>BM</i>	<i>CM</i>	<i>NxA</i>	<i>NxB</i>	<i>NxC</i>	<i>AxB</i>	<i>AxC</i>	<i>BxC</i>	<i>HM</i>	<i>IDDES</i>
1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,001	0,000	0,000	0,002	0,000
1,781	0,000	0,000	0,000	0,001	0,000	0,002	0,004	0,000	0,000	0,001	0,005	0,003	0,003	0,000	0,000
1,981	0,000	0,001	0,002	0,000	0,001	0,000	0,000	0,002	0,007	0,001	0,002	0,001	0,002	0,003	0,000
2,440	0,000	0,000	0,000	0,001	0,001	0,000	0,003	0,000	0,002	0,017	0,001	0,002	0,002	0,000	0,000
2,723	0,000	0,000	0,000	0,000	0,062	0,003	0,000	0,000	0,002	0,000	0,003	0,000	0,025	0,000	0,000
2,920	0,000	0,000	0,000	0,000	0,044	0,001	0,000	0,005	0,001	0,001	0,000	0,003	0,063	0,001	0,000
5,346	0,000	0,001	0,007	0,001	0,004	0,028	0,002	0,055	0,002	0,000	0,000	0,028	0,058	0,003	0,001
7,011	0,001	0,003	0,002	0,000	0,010	0,013	0,010	0,050	0,011	0,000	0,093	0,003	0,005	0,394	0,001
8,451	0,000	0,009	0,006	0,000	0,004	0,009	0,007	0,015	0,009	0,000	0,355	0,003	0,014	0,475	0,000
12,217	0,012	0,000	0,003	0,014	0,000	0,001	0,358	0,020	0,079	0,001	0,009	0,185	0,090	0,043	0,025
14,044	0,045	0,009	0,128	0,000	0,038	0,000	0,124	0,155	0,017	0,029	0,002	0,033	0,025	0,009	0,122
20,474	0,001	0,512	0,091	0,013	0,239	0,110	0,012	0,275	0,143	0,025	0,507	0,032	0,020	0,013	0,012
25,471	0,010	0,069	0,276	0,371	0,095	0,258	0,065	0,082	0,243	0,362	0,004	0,448	0,404	0,025	0,022
31,971	0,766	0,063	0,044	0,152	0,029	0,009	0,017	0,006	0,003	0,105	0,016	0,092	0,095	0,000	0,714
34,470	0,164	0,333	0,440	0,446	0,471	0,565	0,396	0,334	0,480	0,457	0,002	0,168	0,194	0,031	0,103

Este valor 0,5 não foi interpretado como um valor absoluto, mas sim em termos relativos, ou seja, foi considerada a razão entre o componente da proporção da variância em questão e o restante da proporção da variância que ainda não tinha sido explicada.

Pela decomposição proporcional da variância, o intercepto e a variável explicativa *IDDES* estavam envolvidos em uma colinearidade, embora o *VIF* associado à covariável *IDDES* tenha sido o menor (1,03785). Isso aconteceu devido ao fato dos *VIFs* terem sido calculados a partir da matriz $\mathbf{X}^T\mathbf{X}$, esta centrada e padronizada, sem a coluna para o intercepto, ou seja, embora a variável explicativa foi quase ortogonal às covariáveis genéticas, em relação ao intercepto ela não foi. Algumas das variáveis explanatórias que possuíam *VIFs* maiores que 10 e que não estavam envolvidas nas colinearidades fortes, podem estar nas quatro colinearidades fracas. O mesmo comportamento aconteceu para a variável resposta peso ao nascimento.

Após ter sido verificada a existência de duas colinearidades fortes e quais covariáveis estavam envolvidas em cada uma delas, as etapas (50) e (51) foram realizadas 10 vezes. As estimativas dos coeficientes de regressão do vetor β obtidas por *QM* e por *RC*, os elementos (k_i) da matriz diagonal $\mathbf{K}$, assim como a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por *RC* nas duas últimas iterações estão na tabela 16.

Os efeitos para as variáveis explicativas *A*, *B* e *C* foram maiores em relação à *N* pelas duas formas de serem estimadas, porém as obtidas por *RC* foram mais precisas que as obtidas por *QM*. Tendência semelhante foi encontrada por Euclides Filho et al. (1998), avaliando peso ao desmame em animais mestiços Europeu-Zebu, observaram efeito aditivo maior para as raças Simental, Charolesa e Chianina em relação à raça Nelore. Os autores apontaram este resultado como esperado, em razão do maior porte das raças européias, que estaria relacionado a maior potencial de crescimento.

Os coeficientes de regressão obtidos para *AM*, *BM* e *CM*, por *QM* e *RC* não diferiram em direção e magnitude. As variáveis explicativas *AM* e *BM* tiveram efeitos menores que o tipo biológico materno - *NM*, enquanto que o efeito de *CM* foi maior. Estes resultados expressam a importância do efeito materno sobre o peso ao desmame. Um trabalho comparável a este é o de Prayaga (2003) em que os efeitos raciais maternos foram iguais a $-3,4kg$ para o tipo biológico Sanga, $-13,3kg$ para as raças britânicas e $-9,0kg$ para as raças

continentais, como desvio da raça Zebu.

De acordo com a diferença absoluta entre as estimativas dos coeficientes de regressão para as variáveis explicativas obtidas por *RC* nas duas últimas iterações (Tabela 16), foi possível verificar o número de casas decimais das estimativas dos β^* que se estabilizaram. Por exemplo, para a estimativa do coeficiente de regressão da variável explicativa *B* que é 0,05660, a diferença das soluções entre as iterações foi de 0,00006. Com isso a estimativa do coeficiente de regressão estabilizou-se até a terceira casa decimal.

Tabela 16 - Valores dos coeficientes de regressão das covariáveis genéticas de β para o modelo (55), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por *RC* nas duas últimas iterações - $|\hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^*|$, para a variável resposta peso ao desmame

Covariáveis	$\hat{\beta}$	$\hat{\beta}^*$	k_i	$ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $
<i>INT</i>	196,93000	76,59297	—	0,00000
<i>A</i>	0,15161	0,03558	0,01618	0,00223
<i>B</i>	0,14454	0,05660	0,01102	0,00006
<i>C</i>	0,12369	0,03551	0,03346	0,00018
<i>AM</i>	-0,24645	-0,23668	0,00533	0,00363
<i>BM</i>	-0,03997	-0,04755	0,01555	0,00307
<i>CM</i>	0,01924	0,02620	0,01221	0,00337
<i>NxA</i>	0,09741	0,09033	0,00666	0,00038
<i>NxB</i>	0,05463	0,06229	0,01367	0,00318
<i>NxC</i>	0,07747	0,05850	0,01926	0,00252
<i>AxB</i>	0,07895	0,09166	0,00532	0,00225
<i>AxC</i>	0,08063	0,06291	0,01347	0,00003
<i>BxC</i>	0,03225	0,00995	0,00501	0,00127
<i>HM</i>	0,06170	0,07216	0,00218	0,00178
<i>IDDES</i>	0,47695	0,50337	0,00074	0,00427

Da tabela 17 concluiu-se que o método que estimou as covariáveis por *RC*

foi melhor que o método que estimou por QM , pois a soma dos quadrados dos desvios - SQD obtida por RC (44.018.682) foi menor à obtida por QM (44.088.623). A pequena diferença entre as SQD era esperada, pois o objetivo na análise de regressão foi a interpretação dos coeficientes de regressão individuais. Nesse sentido, observou-se que o maior dos $VIFs$ (9,99931) por RC , ficou abaixo do limiar 10 e a média dos $VIFs$ (5,25028) por RC , ficou mais próxima de 1 se comparada com a média dos $VIFs$ (16,02713) por QM . Conseqüentemente os coeficientes de regressão obtidos por RC tiveram desvios padrões menores se comparados aos obtidos por QM (Tabela 18), indicando a eficiência do método de regressão de cumeeira no intuito de minimizar a variabilidade dos coeficientes de regressão. Estes, com variância menor em relação às variâncias dos coeficientes obtidos por QM foram estimados por Roso et al. (2005b).

A medida do viés foi de 9,76484%. Carvalheiro et al. (2006), em seu trabalho com bezerros das raças Nelore-Hereford, para ganho de peso a pré-desmama, obteve medida de viés igual a 11,1%.

Tabela 17 - Soma dos quadrados dos desvios - SQD para o modelo (55) e para a variável resposta peso ao desmame, máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - QM e regressão de cumeeira - RC e medida do viés para RC - viés (%)

	QM	RC
SQD	44.088.623	44.018.682
$MVIF$	46,90675	9,99931
$\overline{VIF}$	16,02713	5,25028
viés (%)	—	9,76484

Tabela 18 - Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeieira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$, para a variável resposta peso ao desmame

	$\hat{V}(\hat{\beta})$	$\hat{s}(\hat{\beta})$	$\hat{V}(\hat{\beta}^*)$	$\hat{s}(\hat{\beta}^*)$
<i>A</i>	5.233,42987	72,34245	2.307,71128	48,03864
<i>B</i>	3.566,40836	59,71941	1.441,76519	37,97058
<i>C</i>	10.825,46764	104,04551	1.080,00778	32,86347
<i>AM</i>	1.723,93192	41,52026	801,71576	28,31459
<i>BM</i>	5.029,45941	70,91868	1.624,83275	40,30921
<i>CM</i>	3.951,06035	62,85746	1.771,22054	42,08587
<i>NxA</i>	2.155,34469	46,42569	1.241,83024	35,23961
<i>NxB</i>	4.421,77718	66,49644	1.683,27987	41,02779
<i>NxC</i>	6.232,08026	78,94353	831,08297	28,82851
<i>AxB</i>	1.719,67030	41,46891	1.261,15713	35,51277
<i>AxC</i>	4.357,74118	66,01319	1.437,07521	37,90877
<i>BxC</i>	1.621,38733	40,26645	585,32930	24,19358
<i>HM</i>	706,65875	26,58305	658,62654	25,66372
<i>IDDES</i>	239,52207	15,47650	238,12839	15,43141

As estimativas para as covariáveis *DTJN*, *DTJN*², *DTJN*³ e para variáveis classificatórias *CIMP* e *GC* foram obtidas na etapa (50). Apenas as estimativas de *DTJN*, *DTJN*², *DTJN*³ e *CIMP* estão apresentadas na tabela 19. As estimativas para *CIMP* estão como desvio da *CIMP*₄ que é o nível de maior produtividade, pois a reprodutora atingiu a maturidade.

Os efeitos das *CIMPs* foram importantes, alcançando magnitude de ajuste de até 26kg aproximadamente, diferença verificada entre *CIMP*₁ e *CIMP*₄. Houve pequena diferença entre as estimativas obtidas levando em consideração *QM* e *RC*, o que esperava-se, tendo em vista que essas variáveis não foram consideradas envolvidas em colinearidades. Zampar et al. (2006), obteve valores semelhantes para esta variável classificatória, embora no modelo usado por estes, não haviam os efeitos de data juliana.

Tabela 19 - Estimativa das covariáveis data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ para a variável resposta peso ao desmame

	QM	RC
$DTJN$	1,65700	1,51320
$DTJN^2$	- 0,01209	- 0,00950
$DTJN^3$	- 0,00014	- 0,00015
$CIMP_1$	-25,80500	-25,63200
$CIMP_2$	-14,98400	-14,93600
$CIMP_3$	- 4,99800	- 4,95800
$CIMP_4$	0,00000	0,00000
$CIMP_5$	- 1,74600	- 1,85100
$CIMP_6$	- 5,58300	- 5,75100
$CIMP_7$	-12,13000	-12,33200

Para o modelo (55) as estimativas dos componentes das variâncias genéticas animal - $\hat{\sigma}_a^2$ e materna - $\hat{\sigma}_m^2$, a estimativa da covariância entre os efeitos genéticos animal e materno - $\hat{\sigma}_{am}$, as estimativas das variâncias de efeito permanente de ambiente - $\hat{\sigma}_{ep}^2$ e residual - $\hat{\sigma}_e^2$, assim como a fenotípica - $\hat{\sigma}_p^2$ encontram-se na tabela 20. Os resultados obtidos para as (co)variâncias não diferiram, o que era esperado, pois o método de estimação do vetor β não interfere na variabilidade dos efeitos aleatórios. Estes resultados diferiram dos obtidos por Mourão et al. (2007). Tal fato possivelmente ocorreu devido a diferença no banco de dados e a inclusão dos efeitos de data juliana no modelo.

Tabela 20 - Estimativas dos componentes de (co)variâncias¹ para peso ao desmame, em que as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cuneeira - RC

	$\hat{\sigma}_a^2$	$\hat{\sigma}_m^2$	$\hat{\sigma}_{am}$	$\hat{\sigma}_{ep}^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$
QM	103,28215	54,28327	-11,79966	63,5199	341,73003	539,21601
RC	104,11126	50,96308	-11,91372	66,6410	341,65211	539,54001

¹ definidos na tabela 4

3.4.3 Perímetro escrotal aos 390 dias

A variável resposta perímetro escrotal aos 390 dias possuía 30.113 observações com média igual a 28,418cm, desvio padrão de 3,993cm, valor mínimo de 17,5cm, valor máximo de 39,2cm e idade de mensuração entre 300 e 632 dias.

Para verificar quais covariáveis da matriz $\mathbf{X}$ do modelo (56) estavam envolvidas em colinearidades, seus VIF s (Tabela 21) foram analisados.

Tabela 21 - Fatores de Inflação da Variância - VIF para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (56)

Variáveis Explicativas	VIF	Variáveis Explicativas	VIF
A	16,65405	NxB	13,57339
B	14,22844	NxC	12,70274
C	58,18203	AxB	7,09533
AM	10,77573	AxC	33,82179
BM	26,92378	BxC	10,53494
CM	27,38109	HM	1,55088
NxA	11,04471	$IDCE$	1,03901

De acordo com a tabela 21, observou-se que as variáveis explicativas que possuíam VIF s maiores que 10 foram: os efeitos aditivos diretos associados aos tipos biológicos A (16,65405), B (14,22844) e C (58,18203); os efeitos aditivos maternos AM (10,77573), BM (26,92378) e CM (27,38109); e as heterozigoses diretas NxA (11,04471), NxB (13,57339), NxC (12,70274), AxC (33,82179) e BxC (10,53494). Com isso, houve indícios de que estas covariáveis estavam envolvidas em uma ou mais colinearidades, pois as variâncias das estimativas de seus coeficientes de regressão estavam sendo inflacionadas mais que dez vezes se comparadas com as estimativas obtidas se o sistema fosse ortogonal. A relação de *quase* dependência entre efeitos aditivos diretos e maternos relacionados ao mesmo tipo biológico, também foi observada nas variáveis peso ao desmame e ao nascimento e em trabalhos citados para estas características.

Para verificar colinearidades existentes entre as covariáveis da matriz $\mathbf{X}$ do

modelo (56) para a característica perímetro escrotal aos 390 dias, seus índices de condição com seus respectivos autovalores foram analisados (Tabela 22).

Tabela 22 - Autovalores e Índice de Condição da matriz $\mathbf{X}^T \mathbf{X}$ do modelo (56) para a variável resposta perímetro escrotal aos 390 dias

Autovalor	Índice de Condição	Autovalor	Índice de Condição
8,00345	1,00000	0,08361	9,78405
2,45473	1,80566	0,03808	14,49780
1,44831	2,35076	0,01770	21,26521
1,06783	2,73771	0,00948	29,05586
0,90887	2,96747	0,00746	32,74413
0,64142	3,53237	0,00476	41,00343
0,21136	6,15362	0,00186	65,67646
0,10109	8,89762		

Observou-se que os autovalores 0,00186, 0,00476 e 0,00746 foram relativamente pequenos, conseqüentemente, seus respectivos índices de condição 65,67646, 41,00343 e 32,74413 obtiveram valores acima do limiar, evidenciando três colinearidades fortes entre as variáveis explanatórias. Como os índices de condição 29,05586, 21,26521 e 14,49780 estão entre 10 e 30, há indícios de três colinearidades fracas (Tabela 22).

A partir da decomposição proporcional da variância (Tabela 23), utilizando o valor relativo 0,5 como limiar, observou-se na linha associada ao índice de condição igual a 65,67646 que o intercepto - *INT* (0,986) e a covariável *IDCE* (0,505) apresentaram colinearidade forte entre si. Na linha associada ao índice de condição 41,00343, as covariáveis *A* (0,460), *C* (0,576), *AM* (0,580) e *BM* (0,475) estavam envolvidas em uma relação de *quase* dependência forte. O valor para a covariável *NxC* (0,446) ficou próximo do limiar e esta pode estar envolvida nessa colinearidade. Com relação ao *IC* igual a 32,74413, as covariáveis *C* (0,257), *BM* (0,246), *NxA* (0,260), *NxB* (0,284), *AxC* (0,366) e *IDCE* (0,257) obtiveram valores maiores que o limiar e formaram uma relação de *quase* dependência considerada forte.

Tabela 23 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (56), para a variável resposta perímetro escrotal aos 390 dias

IC	INT	A	B	C	AM	BM	CM	NxA	NxB	NxC	AxB	AxC	BxC	HM	$IDCE$
1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,002	0,000
1,806	0,000	0,000	0,000	0,001	0,000	0,001	0,002	0,000	0,000	0,001	0,004	0,001	0,002	0,000	0,000
2,351	0,000	0,001	0,003	0,000	0,001	0,000	0,000	0,001	0,015	0,002	0,000	0,001	0,008	0,001	0,000
2,738	0,000	0,000	0,000	0,000	0,019	0,000	0,001	0,000	0,001	0,042	0,000	0,001	0,000	0,000	0,000
2,967	0,000	0,000	0,000	0,000	0,046	0,001	0,000	0,000	0,008	0,015	0,002	0,000	0,000	0,000	0,000
3,532	0,000	0,000	0,000	0,000	0,007	0,001	0,000	0,001	0,024	0,007	0,000	0,001	0,059	0,000	0,000
6,154	0,000	0,001	0,016	0,001	0,006	0,015	0,001	0,023	0,008	0,001	0,000	0,014	0,042	0,044	0,000
8,898	0,001	0,006	0,001	0,000	0,007	0,000	0,002	0,024	0,000	0,000	0,000	0,000	0,015	0,852	0,001
9,784	0,000	0,002	0,026	0,000	0,000	0,004	0,014	0,031	0,034	0,000	0,462	0,000	0,002	0,031	0,001
14,498	0,000	0,000	0,089	0,022	0,001	0,030	0,281	0,000	0,080	0,020	0,000	0,084	0,041	0,030	0,001
21,265	0,002	0,141	0,207	0,018	0,116	0,138	0,053	0,169	0,041	0,005	0,474	0,041	0,002	0,008	0,006
29,056	0,006	0,250	0,227	0,118	0,017	0,010	0,007	0,006	0,019	0,313	0,010	0,217	0,241	0,020	0,227
32,744	0,004	0,000	0,010	0,257	0,166	0,246	0,177	0,260	0,284	0,084	0,014	0,366	0,309	0,004	0,257
41,003	0,000	0,460	0,355	0,576	0,580	0,475	0,400	0,282	0,191	0,446	0,016	0,270	0,277	0,008	0,002
65,676	0,986	0,139	0,068	0,007	0,035	0,079	0,062	0,201	0,295	0,065	0,017	0,003	0,001	0,000	0,505

A colinearidade forte entre *INT* e *IDCE* mesmo com o *VIF* da variável explicativa *IDCE* sendo o menor, ocorreu devido ao fato dos *VIFs* terem sido calculados a partir da matriz $\mathbf{X}^T \mathbf{X}$ centrada e padronizada, sem a coluna para o intercepto, ou seja, embora a covariável *IDCE* foi quase ortogonal em relação às covariáveis genéticas, não o é em relação ao *INT*. Algumas das variáveis explanatórias que apresentaram *VIFs* maiores que 10 e não foram relacionadas nas colinearidades fortes, podem estar envolvidas nas colinearidades fracas. O mesmo comportamento foi observado para as variáveis pesos ao nascimento e ao desmame e em trabalhos como o de Roso et al. (2005a).

Tabela 24 - Valores dos coeficientes de regressão (β) para o modelo (56), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por *RC* nas duas últimas iterações - $|\hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^*|$ para a variável resposta perímetro escrotal aos 390 dias

Covariáveis	$\hat{\beta}$	$\hat{\beta}^*$	k_i	$ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $
<i>INT</i>	28,26700	15,43158	—	0,33793
<i>A</i>	0,00778	0,00620	0,00966	0,00010
<i>B</i>	0,03425	0,03074	0,00825	0,00122
<i>C</i>	0,01297	0,00859	0,03374	0,00028
<i>AM</i>	0,01244	0,01217	0,00625	0,00025
<i>BM</i>	0,00372	0,00454	0,01561	0,00047
<i>CM</i>	0,00591	0,00671	0,01588	0,00037
<i>NxA</i>	-0,00034	-0,00144	0,00640	0,00017
<i>NxB</i>	-0,01202	-0,01216	0,00787	0,00107
<i>NxC</i>	-0,00850	-0,00837	0,00737	0,00074
<i>AxB</i>	0,00470	0,00431	0,00411	0,00012
<i>AxC</i>	0,01134	0,01123	0,01961	0,00030
<i>BxC</i>	0,01416	0,01538	0,00611	0,00041
<i>HM</i>	0,00150	0,00154	0,00090	0,00012
<i>IDCE</i>	0,02826	0,02826	0,00060	0,00093

Foram detectadas três possíveis colinearidades fortes, as covariáveis envolvidas em cada uma delas e as etapas (50) e (51) foram realizadas 10 vezes. As estimativas dos coeficientes de regressão do vetor β obtidas por QM e por RC , os elementos (k_i) da matriz diagonal $\mathbf{K}$ assim como a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações para a variável resposta perímetro escrotal aos 390 dias foram obtidas (Tabela 24).

De acordo com os resultados obtidos para os efeitos aditivos diretos e maternos (Tabela 24), concluiu-se que estes efeitos foram maiores em relação à N e NM respectivamente.

Tendência semelhante a este estudo foi obtida por Teixeira (2004), para os efeitos aditivos diretos sobre perímetro escrotal ao desmame, no qual os efeitos aditivos diretos das raças Hereford e Angus foram de 4,1cm e 5,1cm como desvios da raça Nelore.

A partir da diferença absoluta entre as estimativas dos coeficientes de regressão das variáveis estimadas por RC nas duas últimas iterações (Tabela 24), foi possível verificar o número de casas decimais para a qual as estimativas dos β^* estabilizaram-se. Por exemplo, para a estimativa do coeficiente de regressão da variável explicativa $IDCE$ que foi 0,02826, a diferença das soluções nas duas últimas iterações foi de 0,00093. Com isso a estimativa do coeficiente de $IDCE$ estabilizou-se até a segunda casa decimal, ou ainda, o erro foi menor que 10^{-3} .

Tabela 25 - Soma dos quadrados dos desvios - SQD para o modelo (56), máximo fator de inflação da variância - $MVIF$, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por por quadrados mínimos - QM e regressão de cuneeira - RC e medida do viés para RC - viés (%)

	QM	RC
SQD	169.277,92	166.133,49
$MVIF$	58,18203	9,99969
$\overline{VIF}$	17,53628	5,68323
viés (%)	—	9,64823

Da tabela 25 concluiu-se que método que estimou as covariáveis por RC foi

melhor ao método que estimou por QM , visto que a soma dos quadrados dos desvios - SQD encontrada por RC (166.133,49) foi menor que a obtida por QM (169.277,92), embora o interesse desse estudo foi de encontrar estimativas dos coeficientes de regressão para futura interpretação, como recomendado por Freund e Wilson (1998). Nesse sentido, observou-se que o maior dos $VIFs$ (9,99969) por RC , ficou abaixo do limiar 10 e a média dos $VIFs$ (5,68323) por RC , ficou mais próxima de 1 se comparado com a média dos $VIFs$ (17,53628) por QM . Conseqüentemente, os coeficientes de regressão obtidos por RC tiveram desvios padrões menores se comparados com os obtidos por QM (Tabela 26). Valores similares para $MVIFs$ e $\overline{VIF}$ foram estimados por Roso et al. (2005b) sobre ganho de peso a pré-desmama para uma população multirracial na utilização da regressão de cumeeira. A medida do viés ficou próxima das obtidas para as variáveis pesos ao nascimento e ao desmame.

Tabela 26 - Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeeira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$

	$\hat{V}(\hat{\beta})$	$\hat{s}(\hat{\beta})$	$\hat{V}(\hat{\beta}^*)$	$\hat{s}(\hat{\beta}^*)$
A	93,61946	9,67571	56,21252	7,49750
B	79,98409	8,94338	45,96148	6,77949
C	327,06584	18,08496	27,97381	5,28903
AM	60,57496	7,78299	29,07797	5,39240
BM	151,34995	12,30244	49,53866	7,03837
CM	153,92072	12,40648	48,14081	6,93836
NxA	62,08697	7,87953	36,87279	6,07230
NxB	76,30179	8,73509	37,83769	6,15124
NxC	71,40748	8,45029	17,56514	4,19108
AxB	39,88587	6,31553	32,41526	5,69344
AxC	190,12660	13,78864	36,06060	6,00505
BxC	59,22133	7,69554	15,32389	3,91457
HM	8,71814	2,95265	8,47728	2,91158
$IDCE$	5,84070	2,41675	5,81171	2,41075

As estimativas para as covariáveis $DTJN$, $DTJN^2$ e $DTJN^3$ e para as variáveis classificatórias $CIMP$ e GC foram obtidas na etapa (50). Apenas as estimativas de $DTJN$, $DTJN^2$, $DTJN^3$ e $CIMP$ foram apresentadas na tabela 27. As estimativas para $CIMP$ estão como desvio da $CIMP_4$, que é a classe em que a reprodutora atingiu a maturidade.

Os efeitos das $CIMPs$ foram importantes, alcançando magnitude de ajuste de até 1,1cm aproximadamente, diferença encontrada entre $CIMP_1$ e $CIMP_4$. Houve pequena diferença entre as estimativas obtidas levando em consideração QM e RC , o que era esperado, porque essas variáveis não foram consideradas envolvidas em colinearidades. As estimativas deste estudo para $CIMP$ foram concordantes com as obtidas por Mourão (2005).

Tabela 27 - Estimativas das covariáveis de data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ para a variável resposta perímetro escrotal aos 390 dias

	QM	RC
$DTJN$	0,15642	0,11324
$DTJN^2$	-0,00154	-0,00088
$DTJN^3$	-0,00001	-0,00001
$CIMP_1$	-1,09820	-1,15330
$CIMP_2$	-0,70000	-0,72320
$CIMP_3$	-0,23800	-0,24920
$CIMP_4$	0,00000	0,00000
$CIMP_5$	-0,13400	-0,09530
$CIMP_6$	-0,39750	-0,35630
$CIMP_7$	-0,69230	-0,66170

Para o modelo (56) as estimativas dos componentes das variâncias genéticas animal - $\hat{\sigma}_a^2$ e materna - $\hat{\sigma}_m^2$, a estimativa da covariância entre os efeitos genéticos animal e materno - $\hat{\sigma}_{am}$, a estimativa do componente de variância residual - $\hat{\sigma}_e^2$, assim como a fenotípica - $\hat{\sigma}_p^2$, encontram-se na tabela 28.

Tabela 28 - Estimativas dos componentes de (co)variâncias¹, para perímetro escrotal aos 390 dias, em que as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - *QM* e regressão de cumeira - *RC*

	$\hat{\sigma}_a^2$	$\hat{\sigma}_m^2$	$\hat{\sigma}_{am}$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$
<i>QM</i>	1,80492	0,16420	-0,04532	6,92338	8,75654
<i>RC</i>	1,77838	0,16397	-0,03743	6,97025	8,95003

¹ definidos na tabela 4

Mourão et al. (2007) encontrou estimativas das variâncias genéticas animal e materna de 2,03 e 0,50 respectivamente, sendo maiores que as encontradas nesse trabalho.

3.4.4 Escore de Musculosidade aos 390 dias

A variável resposta escore de musculosidade possuía 70.836 observações com média igual a 4,103, desvio padrão de 1,071, valor mínimo de 1 e valor máximo de 6, com idade de mensuração entre 185 e 632 dias.

As variáveis explicativas, da matriz $\mathbf{X}$ do modelo (57) sobre a variável resposta escore de musculosidade aos 390 dias, envolvidas em colinearidades, foram determinadas a partir dos seus fatores de inflação da variância (Tabela 29).

Tabela 29 - Fatores de Inflação da Variância - *VIF* para as variáveis explicativas consideradas na matriz de delineamento $\mathbf{X}$ do modelo (57) sobre a variável resposta escore de musculosidade aos 390 dias

Variáveis Explicativas	<i>VIF</i>	Variáveis Explicativas	<i>VIF</i>
<i>A</i>	25,39443	<i>NxB</i>	51,34070
<i>B</i>	18,83778	<i>NxC</i>	26,22805
<i>C</i>	58,68769	<i>AxB</i>	8,31207
<i>AM</i>	12,30200	<i>AxC</i>	30,82155
<i>BM</i>	35,01483	<i>BxC</i>	8,79666
<i>CM</i>	34,87379	<i>HM</i>	2,43721
<i>NxA</i>	20,36370	<i>IDMUSC</i>	1,04212

Observou-se que as variáveis explanatórias que possuíam VIF s maiores que 10 (Tabela 29) foram: os efeitos aditivos diretos A (25,39443), B (18,83778) e C (58,68769); os efeitos aditivos maternos AM (12,30200), BM (35,01483) e CM (34,87379) e as heterozigotes NxA (20,36370), NxB (51,34070), NxC (26,22805) e AxC (30,82155), indicando as covariáveis que estavam envolvidas em uma ou mais colinearidades, pois as variâncias das estimativas de seus coeficientes de regressão estavam sendo inflacionadas mais que dez vezes se comparadas com as estimativas obtidas caso os vetores formados pelas covariáveis fossem ortogonais. Para as variáveis anteriores, assim como no trabalho de Rodríguez-Almeida, Van Vleck e Gregory (1997), foi observada a correlação entre efeitos aditivos diretos e maternos relacionada ao mesmo tipo biológico.

O número de colinearidades existentes entre as covariáveis da matriz $\mathbf{X}$ do modelo (57), para a variável resposta escore de musculosidade aos 390 dias foi obtido a partir dos índices de condição. Estes estão na tabela 30 com seus respectivos autovalores.

Tabela 30 - Autovalores e Índice de Condição da matriz $\mathbf{X}^T \mathbf{X}$ do modelo (57) para a variável resposta escore de musculosidade aos 390 dias

Autovalor	Índice de Condição	Autovalor	Índice de Condição
7,70138	1,00000	0,06955	10,52328
2,44484	1,77484	0,03006	16,00508
1,63345	2,17136	0,01694	21,32449
1,11615	2,62678	0,00956	28,37883
0,93780	2,86569	0,00646	34,52231
0,75252	3,19908	0,00439	41,86494
0,17499	6,63397	0,00127	77,79170
0,10063	8,74805		

Observou-se (Tabela 30) que os autovalores 0,00127, 0,00439 e 0,00646 foram relativamente pequenos, e assim, seus respectivos índices de condição 77,79170, 41,86494 e 34,52231 estavam acima do limiar 30, indicando três colinearidades fortes entre as variáveis explicativas. Como os índices de condição 28,37883, 21,32449, 16,00508, e 10,52328 estavam entre 10 e 30, estes indicaram quatro relações de *quase* dependência consideradas fracas.

Tabela 31 - Decomposição proporcional da variância das estimativas dos coeficientes de regressão β em relação aos índices de condição da matriz $\mathbf{X}$ do modelo (57), para a variável resposta escore de musculosidade aos 390 dias

IC	INT	A	B	C	AM	BM	CM	NxA	NxB	NxC	AxB	AxC	BxC	HM	$IDMUSC$
1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,000	0,000	0,002	0,000
1,775	0,000	0,000	0,000	0,001	0,000	0,001	0,002	0,000	0,000	0,000	0,004	0,002	0,003	0,000	0,000
2,171	0,000	0,000	0,002	0,000	0,001	0,000	0,000	0,001	0,004	0,001	0,001	0,001	0,003	0,002	0,000
2,627	0,000	0,000	0,000	0,001	0,003	0,000	0,001	0,000	0,000	0,023	0,000	0,001	0,001	0,000	0,000
2,866	0,000	0,000	0,000	0,000	0,054	0,001	0,000	0,000	0,001	0,001	0,002	0,000	0,004	0,000	0,000
3,199	0,000	0,000	0,000	0,000	0,007	0,001	0,000	0,001	0,002	0,001	0,000	0,002	0,078	0,001	0,000
6,634	0,000	0,002	0,008	0,001	0,003	0,020	0,001	0,030	0,001	0,001	0,003	0,019	0,042	0,001	0,000
8,748	0,000	0,004	0,001	0,000	0,012	0,006	0,002	0,009	0,002	0,001	0,005	0,003	0,007	0,812	0,001
10,523	0,000	0,003	0,010	0,000	0,000	0,018	0,019	0,015	0,008	0,000	0,553	0,000	0,002	0,115	0,000
16,005	0,000	0,000	0,073	0,024	0,001	0,004	0,294	0,000	0,042	0,015	0,036	0,137	0,107	0,024	0,001
21,324	0,001	0,196	0,182	0,016	0,130	0,083	0,013	0,149	0,033	0,005	0,333	0,025	0,016	0,003	0,004
28,379	0,001	0,161	0,248	0,230	0,045	0,077	0,042	0,004	0,009	0,310	0,000	0,414	0,367	0,026	0,082
34,522	0,000	0,096	0,039	0,268	0,014	0,041	0,020	0,079	0,101	0,041	0,019	0,225	0,178	0,002	0,573
41,865	0,003	0,362	0,371	0,452	0,535	0,471	0,385	0,235	0,227	0,394	0,013	0,165	0,191	0,011	0,104
77,792	0,994	0,175	0,064	0,006	0,196	0,278	0,220	0,477	0,569	0,206	0,030	0,006	0,000	0,001	0,235

A partir da decomposição proporcional da variância (Tabela 31), utilizando o limiar relativo 0,5, ou seja, foi considerada a razão entre o componente da proporção da variância em questão e o restante da proporção da variância ainda não explicada, observou-se na linha associada ao índice de condição igual a 77,79170 que o intercepto - *INT* (0,994) e a covariável *NxB* (0,569) apresentaram colinearidade entre si. A covariável *NxA* (0,477) possuía valor próximo do limiar e poderia estar envolvida nessa colinearidade. Na linha associada ao índice de condição 41,86494, o intercepto - *INT* (0,003) e as covariáveis *AM* (0,535), *BM* (0,471) e *NxB* (0,227) estavam envolvidas em uma relação de colinearidade, embora os valores para as covariáveis *CM* (0,385) e *NxC* (0,394) estavam próximos do limiar e estas podiam estar envolvidas nessa colinearidade. Com relação ao *IC* igual a 34,522, apenas a covariável *IDMUSC* (0,573) obteve valor maior que o limiar, e como para indicar uma relação de colinearidade é preciso que no mínimo duas variáveis estejam envolvidas, nesse caso, não houve colinearidade. A variável *INT* esteve presente nas duas colinearidades fortes, embora não tem-se o valor do *VIF* para esta variável, pois foram calculados a partir da matriz $\mathbf{X}^T \mathbf{X}$ centrada e padronizada, sem a coluna para o intercepto, ou seja, não foi possível verificar se *INT* estava envolvido em colinearidades pelo fator de inflação da variância. Algumas das variáveis explanatórias que possuíam *VIF*s maiores que 10 e que não foram relacionadas nas colinearidades fortes, podem estar envolvidas nas colinearidades fracas. Este mesmo comportamento foi observado para as variáveis anteriores e em alguns trabalhos mencionados anteriormente.

Após duas colinearidades fortes terem sido detectadas e apostadas as covariáveis envolvidas em cada caso, as etapas (50) e (51) foram realizadas 10 vezes. As estimativas dos coeficientes de regressão do vetor β obtidas por *QM* e por *RC*, os elementos (k_i) da matriz diagonal $\mathbf{K}$, assim como a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por *RC* nas duas últimas iterações estão na tabela 32.

De acordo com a diferença absoluta entre as estimativas dos coeficientes de regressão para as variáveis explicativas obtidas por *RC* nas duas últimas iterações (Tabela 32), foi possível verificar até que casa decimal as estimativas dos β^* se estabilizaram. Por exemplo, para a estimativa do coeficiente de regressão da variável explicativa *AxC* que é 0,00158, a diferença das soluções entre as iterações foi de 0,00001. Com isso a estimativa do

parâmetro estabilizou se até a quarta casa decimal, ou seja, o erro entre as soluções foi de 10^{-5} .

Tabela 32 - Valores dos coeficientes de regressão (β) para o modelo (57), estimados pelo método de quadrados mínimos - $\hat{\beta}$ e por regressão de cumeieira - $\hat{\beta}^*$, valores dos elementos (k_i) da matriz diagonal $\mathbf{K}$ e a diferença absoluta entre as estimativas dos coeficientes de regressão obtidas por RC nas duas últimas iterações - $|\hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^*|$

Covariáveis	$\hat{\beta}$	$\hat{\beta}^*$	k_i	$ \hat{\beta}_{(10)}^* - \hat{\beta}_{(9)}^* $
<i>INT</i>	4,08530	1,03668	—	0,00171
<i>A</i>	0,00442	0,00334	0,01552	0,00014
<i>B</i>	-0,00047	-0,00115	0,01151	0,00006
<i>C</i>	-0,00169	0,00024	0,03586	0,00001
<i>AM</i>	-0,00318	-0,00126	0,00752	0,00000
<i>BM</i>	-0,00231	0,00017	0,02140	0,00012
<i>CM</i>	-0,00116	0,00080	0,02131	0,00008
<i>NxA</i>	-0,00156	-0,00089	0,01244	0,00000
<i>NxB</i>	0,00160	0,00009	0,03137	0,00006
<i>NxC</i>	0,00268	-0,00058	0,01603	0,00015
<i>AxB</i>	0,00174	0,00156	0,00508	0,00000
<i>AxC</i>	0,00303	0,00158	0,01883	0,00001
<i>BxC</i>	0,00263	0,00043	0,00538	0,00001
<i>HM</i>	-0,00033	-0,00078	0,00149	0,00005
<i>IDMUSC</i>	0,00261	0,00719	0,00064	0,00003

Da tabela 33 concluiu-se que o método que estimou as covariáveis por *RC* foi melhor que o método que estimou por *QM*, porque a soma dos quadrados dos desvios - *SQD* obtida por *RC* (30.531, 509) foi menor que a obtida por *QM* (31.614, 964), embora o objetivo na análise de regressão era obter estimativas para os coeficientes de regressão com erros-padrão menores e conseqüentemente mais precisas, para futuras interpretações, como proposto por Bergmann e Hohenboken (1995). Nesse sentido, observou-se que o maior *VIF* (9, 99867) por *RC*, ficou abaixo do limiar 10 e a média dos *VIF*s (5, 05419) por *RC*, ficou mais próxima de 1 se comparado com a média (23, 88947) obtida por *QM*. Conseqüentemente os coeficientes de regressão obtidos por *RC* tiveram desvios padrões menores se comparados aos obtidos por *QM* (Tabela 34). Resultados parecidos foram obtidos por Roso et al. (2005b) para ganho de peso a pré-desmama para uma população multirracial. A medida do viés foi de 13,43259%.

Tabela 33 - Soma dos quadrados dos desvios - *SQD* para o modelo (57) em relação à variável resposta escore de musculosidade aos 390 dias, máximo fator de inflação da variância - *MVIF*, média dos fatores de inflação da variância - $\overline{VIF}$, obtidos por quadrados mínimos - *QM* e regressão de cuneeira - *RC* e medida do viés para *RC* - viés (%)

	<i>QM</i>	<i>RC</i>
<i>SQD</i>	31.614,964	30.531,509
<i>MVIF</i>	58,68769	9,99867
$\overline{VIF}$	23,88947	5,05419
viés (%)	—	13,43259

As estimativas para as covariáveis *DTJN*, *DTJN*² e para as variáveis classificatórias *CIMP* e *GC* foram obtidas na etapa (50). Apenas as estimativas de *DTJN*, *DTJN*² e *CIMP* foram apresentadas (Tabela 35), As estimativas para *CIMP* estão como desvio da *CIMP*₄.

Tabela 34 - Estimativa da variância da estimativa dos coeficientes de regressão β , obtidos por quadrados mínimos $\hat{V}(\hat{\beta})$ e por regressão de cumeieira - $\hat{V}(\hat{\beta}^*)$, com seus respectivos desvios padrões, $\hat{s}(\hat{\beta})$ e $\hat{s}(\hat{\beta}^*)$

	$\hat{V}(\hat{\beta})$	$\hat{s}(\hat{\beta})$	$\hat{V}(\hat{\beta}^*)$	$\hat{s}(\hat{\beta}^*)$
<i>A</i>	11,33384	3,36658	4,46252	2,11247
<i>B</i>	8,40753	2,89957	3,85981	1,96464
<i>C</i>	26,19302	5,11791	1,94985	1,39637
<i>AM</i>	5,49053	2,34319	1,60126	1,26541
<i>BM</i>	15,62754	3,95317	2,15090	1,46660
<i>CM</i>	15,56459	3,94520	3,02280	1,73862
<i>NxA</i>	9,08857	3,01472	2,64467	1,62624
<i>NxB</i>	22,91397	4,78685	2,15671	1,46858
<i>NxC</i>	11,70589	3,42139	1,55177	1,24570
<i>AxB</i>	3,70978	1,92608	2,78400	1,66853
<i>AxC</i>	13,75603	3,70891	2,77883	1,66698
<i>BxC</i>	3,92606	1,98143	1,12357	1,05999
<i>HM</i>	1,08776	1,04296	1,03161	1,01568
<i>IDMUSC</i>	0,46511	0,68199	0,46213	0,67980

Os efeitos das *CIMPs* foram importantes, alcançando magnitude de ajuste de até 0,15 aproximadamente, diferença observada entre *CIMP*₁ e *CIMP*₄. Houve pequena diferença entre as estimativas obtidas levando em consideração *QM* e *RC*, dentro do esperado, visto que esses efeitos não foram considerados envolvidos em colinearidades. Valores semelhantes foram obtidos por Mourão (2005), para uma população multirracial.

Tabela 35 - Estimativa das covariáveis data juliana - $DTJN$ e da variável classificatória classe de idade da mãe ao parto - $CIMP$ em relação à variável resposta escore de musculabilidade aos 390 dias

	QM	RC
$DTJN$	0,02401	0,00307
$DTJN^2$	-0,00037	-0,00004
$CIMP_1$	-0,12208	-0,15092
$CIMP_2$	-0,06504	-0,07468
$CIMP_3$	-0,01816	-0,02356
$CIMP_4$	0,00000	0,00000
$CIMP_5$	0,00402	0,01354
$CIMP_6$	-0,05420	-0,04192
$CIMP_7$	-0,07036	-0,05707

Para o modelo (57) as estimativas dos componentes das variâncias genéticas animal - $\hat{\sigma}_a^2$ e materna - $\hat{\sigma}_m^2$, a estimativa da covariância entre os efeitos genéticos animal e materno - $\hat{\sigma}_{am}$ e residual - $\hat{\sigma}_e^2$, assim como a fenotípica - $\hat{\sigma}_p^2$ estão na tabela 36. Valores semelhantes foram obtidos por Mourão et al. (2007), o que era esperado, pois a diferença entre os conjuntos de dados é que nesse caso, o conjunto contém também dados obtidos entre os anos de 2006 e 2008.

Tabela 36 - Estimativas dos componentes de (co)variâncias¹ para escore de musculabilidade aos 390 dias, em que o as estimativas dos coeficientes de regressão foram obtidas por quadrados mínimos - QM e regressão de cumeeira - RC

	$\hat{\sigma}_a^2$	$\hat{\sigma}_m^2$	$\hat{\sigma}_{am}$	$\hat{\sigma}_e^2$	$\hat{\sigma}_p^2$
QM	0,11692	0,00966	-0,0145	0,54951	0,64708
RC	0,12855	0,01530	-0,0235	0,54658	0,64343

¹ definidos na tabela 4

4 CONCLUSÕES

- A maior parte das variáveis explicativas estudadas apresentaram algum grau de colinearidade entre si. Dentre estas, observou-se que em geral os efeitos aditivos diretos, efeitos aditivos maternos e heterozigotes diretos estavam envolvidas em relações de *quase* dependência. Apenas as variáveis AxB e HM não estavam presentes em nenhuma colinearidade.
- Sobre a variável resposta peso ao nascimento, existe uma colinearidade forte e possivelmente quatro consideradas fracas. Para a variável resposta peso ao desmame, duas relações de *quase* dependência fortes e quatro fracas. Sobre a variável explicativa perímetro escrotal aos 390 dias, houve três colinearidades fortes e possivelmente três fracas e por fim, a variável explicativa escore de musculosidade aos 390 dias obteve duas relações de *quase* dependência consideradas fortes e quatro consideradas fracas. Como apenas as relações de *quase* dependência consideradas fortes foram analisadas, algumas covariáveis que possuíam VIF maiores do que 10 não apareceram nas colinearidades estudadas, indicando que essas podem estar envolvidas em relações de *quase* dependência fracas.
- Nos modelos estudados, as estimativas dos coeficientes de regressão obtidas pela regressão de cumeeira foram mais estáveis se comparadas com as obtidas por quadrados mínimos mínimos e com isso a inclusão de viés nas estimativas dos coeficientes de regressão foi compensada pelo aumento da precisão destas.
- Portanto recomenda-se o uso da regressão de cumeeira para a estimação dos coeficientes de regressão, quando as covariáveis associadas à estes, estão envolvidas em fortes colinearidades, degradando a estimação pelo método de quadrados mínimos

Como continuidade deste estudo, seria interessante:

- Obter a matriz diagonal K por métodos de estimação diferentes.
- Adotar um critério de convergência para as estimativas dos coeficientes de regressão $(\hat{\beta}^*)$, obtidas na etapa (51).
- Desenvolver ferramentas que possibilitem trabalhar com multicolinearidade nas variáveis de efeitos fixos, em modelos mistos.
- Incluir outras variáveis, consideradas importantes na literatura, como epistasia e complementaridade.

REFERÊNCIAS

BELSLEY, D.A.; KUH, E.; WELSCH, R.E. **Regression diagnostics**: Identifying influential data and sources of collinearity. New York: John Wiley & Sons, 1980. 292p.

BELSLEY, D.A. **Conditioning diagnostics**: collinearity and weak data in regression. New York: John Wiley & Sons, 1991. 393p.

BERGMANN, J.A.G.; HOHENBOKEN, W.D. Alternatives to least squares in multiple linear regression to predict production traits. **Journal of Animal Breeding and Genetics**, Berlin, v.112, p.1-16, 1995.

BOLDMAN, K. G.; KRIESE, L. A.; VAN VLECK, L. D.; VAN TASSELL, C. P.; KACHMAN, S. D. **A manual for use of MTDFREML**: a set of programs to obtain estimates of variances and covariances (DRAFT). Lincoln: Department of Agriculture/ Agriculture Research Service, 1995. 120 p.

CHATTERJEE, S.; PRICE, B. **Regression analysis by example**. 2.ed. New York: John Wiley & Sons, 1991. 278p.

CHATTERJEE, S.; HADI, A.S.; PRICE, B. **Regression analysis by example**. 3.ed. New York: John Wiley & Sons, 2000. 359p

DRAPER, N.R.; SMITH, H. **Applied regression analysis**. 3.ed. New York: John Wiley & Sons, 1998. 706p.

EUCLIDES FILHO, K.; FIGUEREDO G.R.; SILVA, L.O.C. Beef cattle preweaning growth and rate of weight gain as affected by genetic group. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 6., 1998, Armidale. **Proceedings...** Armidale: 6WCGALP, 1998. p.197-200.

FERRAZ, J.B.S.; ELER, J.P.; GOLDEN, B.L. Análise genética do composto Montana Tropical. **Revista Brasileira de Reprodução Animal**, Belo Horizonte, v. 23, n. 2, p. 111-113, 1999.

FRANKE, D.E.; HABET, O.; TAWAH, A.L.; WILLIAMS, A.R.; DEROUEN, S.M. Direct and maternal genetic effects on birth and weaning traits in multibreed cattle data and predicted performance of breed crosses. **Journal of Animal Science**, Albany, v. 79, p. 1713-1722, 2001.

FREUND, R.; LITTELL, R.C. **SAS System for Regression** 3. ed. Cary, SAS Inst, 2000

GROENEVELD, E. **PEST Users' Manual**. Neustadt: Institute of Animal Husbandry and Animal Behaviour Federal Research Centre, 1990. 75p.

HOERL, A.E.; KENNARD, R.W. Ridge regression: biased estimation for nonorthogonal problems. **Technometrics**, Alexandria, v.12, n.1, p.55-67, 1970a.

HOERL, A.E.; KENNARD, R.W. Ridge regression: Applications to Nonorthogonal problems. **Technometrics**, Alexandria, v.12, n.1, p.69-82, 1970b.

LUCHIARI FILHO, A.; MOURAO, G. B. **Melhoramento, raças e seus cruzamentos na pecuária de corte brasileira**. Pirassununga: Albino Luchiari Filho, 2006. 142p.

MASON, R.L.; GUNST, R.F.; WEBSTER, J.T. Regression Analysis and Problems of Multicollinearity. **Communications in statistics**, Londres, v.4, n.3, p. 277-292, 1975.

MOURÃO, G.B. **Estimação de efeitos genéticos aditivos diretos e maternos e não aditivos e predição do desempenho de pesos, perímetro escrotal e musculosidade em uma população composta (*Bos indicus* x *Bos taurus*) de bovinos de corte**. 2005. 109p. Tese (Doutorado em Qualidade e Produtividade Animal) - Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, 2005.

MOURÃO, G.B.; FERRAZ, J.B.S.; ELER, J.P.; BALIEIRO, J.C.C. ; BUENO, R.S.; MATTOS, E.C.; FIGUEIREDO, L.G.G. Genetic parameters for growth traits of a Brazilian *Bos taurus* x *Bos indicus* beef composite. **Genetics and Molecular Research**, Ribeirão Preto, v. 6, p. 1190-1200, 2007.

NETER, J.; WASSERMAN, W.; KUTNER, M.H. **Applied linear statistical models: regression, analysis of variance, and experimental designs**. 3ed. Homewood: Irwin, 1990. 1181p.

PIMENTEL, E.C.G. **Alternativas para redução de efeitos de multicolinearidade em modelos de avaliação de efeitos genéticos em bovinos de corte**. 2004. 97p. Dissertação (Mestrado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal, 2005.

PRAYAGA, K.C. Evaluation of beef cattle genotypes and estimation of direct and maternal denetic effects in a tropical environment 1. Growth traits. **Australian Journal Agricultural Research**, East Melbourne, v. 54, p. 1013-1025, 2003.

RODRÍGUEZ-ALMEIDA, F.A.; VAN VLECK, L.D.; GREGORY, K.E. Estimation of direct and maternal breed effects for prediction of expected progeny differences for birth and weaning weights in three multibreed populations. **Journal of Animal Science**, Albany, v. 75, n. 5, p. 1203-1212, 1997.

ROSO, V.M.; SCHENKEL, F.S.; MILLER, S.P.; SCHAEFFER, L.R. Estimation of genetic effects in the presence of multicollinearity in multibreed beef cattle evaluation. **Journal of Animal Science**, Albany, v. 83, p. 1788-1800, 2005a.

ROSO, V.M.; SCHENKEL, F.S.; MILLER, S.P.; Wilton, J.W. Additive, dominance, end epistatic loss effects on preweaning weight gain of crossbreed beef cattle from different . *Bos taurus* breed. **Journal of Animal Science**, Albany, v. 83, p. 1780-1787, 2005b.

TEIXEIRA, G.B. **Efeitos genéticos aditivos diretos e não aditivos que influenciam características reprodutivas em bovinos de corte**. 2004. 83p. Tese (Doutorado em Genética e Melhoramento Animal) - Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista, Jaboticabal, 2004.

WETHERILL, G.B. **Regression analysis with applications**. New York: Chapman and Hall, 1986. 311p.

ZAMPAR, A.; MOURÃO, G.B.; FERRAZ, J.B.S.; ELER, J.P.; BALIEIRO, J.C.C.; BUENO, R.S.; PEDROSA, V.B.; MATTOS, E.C.; FIGUEIREDO, L.G.G. Effects of classes of age of dam at calving on growth traits in a *Bos taurus* x *Bos indicus* composite beef cattle population. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 8., 2006, Belo Horizonte. **Proceedings...** Belo Horizonte: Instituto Prociência, 2006. 1 CD-ROM.

ANEXOS

ANEXO A - Programa (SAS) utilizado na etapa (51) e para ajustar os dados para a etapa (52), em uma das iterações, para a variável resposta perímetro escrotal aos 390 dias.

```
data cimp11;
infile 'C:\Raphael\Programação\Dissertação\CEPEST102' FIRSTOBS=129 OBS=135;
input  cimp sol_cimp; run;

data gc11;
infile 'C:\Raphael\Programação\Dissertação\CEPEST102' FIRSTOBS=139 OBS=1337;
input  gc sol_gc; run;

data f_p11;
infile 'C:\Raphael\Programação\Dissertação\CEPEST102'
FIRSTOBS=1341 OBS=24961;
input  f_p sol_f_p; run;

data animal11;
infile 'C:\Raphael\Programação\Dissertação\CEPEST102'
FIRSTOBS=24965 OBS=343241;
input  animal sol_animal; run;

data sasuser.mont2; set sasuser.mont1;
drop dtnasmae idademaie datanas sexo gcnas_r data12 datace12 datamus12 outxnabc
nracas gp_pd idadegp_pd gc_gp n_pai a_pai b_pai c_pai dif_n montana;
if pesnas_u = 0 then pesnas_u = .; if pesdes_u = 0 then pesdes_u = .;
if pes12_u = 0 then pes12_u = .; if ce12_u = 0 then ce12_u = .;
if musc12_u = 0 then musc12_u = .; run;

data dtjn; set sasuser.mont2;
```



```

keep produto dtjn dtjn2 dtjn3 ef_dtjn ef_dtjn2 ef_dtjn3;
dtjn2=dtjn**2; dtjn3=dtjn**3;ef_dtjn=0.11324*dtjn;
ef_dtjn2=-0.00088*dtjn2; ef_dtjn3= -0.00001*dtjn3; run;

proc sql;
create table ce_102 as select mont2.produto,mont2.ce12_u,
dtjn.ef_dtjn, dtjn.ef_dtjn2, dtjn.ef_dtjn3, cimp11.sol_cimp,
gc11.sol_gc, f_p11.sol_f_p, animal11.sol_animal,
(mont2.ce12_u - dtjn.ef_dtjn - dtjn.ef_dtjn2 - dtjn.ef_dtjn3 -
cimp11.sol_cimp - gc11.sol_gc - f_p11.sol_f_p - animal11.sol_animal)
as ce_102 from sasuser.mont2,WORK.dtjn,WORK.cimp11,WORK.gc11,
WORK.f_p11,WORK.animal11 where mont2.produto eq animal11.animal and
mont2.cimp eq cimp11.cimp and mont2.gcce12 eq gc11.gc and
mont2.mae eq f_p11.f_p and mont2.produto eq dtjn.produto; quit;

data sasuser.ce11;
merge sasuser.mont2 ce_102; by produto;
keep produto a b c a_mae b_mae c_mae nxa nxb nxc axb axc
bxc hetmatnabc idadece12 ce_102; run;

data dados; set sasuser.ce11;
if ce_102=. then delete; run;

data dadosce_orig; set sasuser.mont2;
keep ce12_u; if ce12_u=. then delete; run;

ods html;
ods graphics on;
proc reg data=dados outest=betaQM ;
model ce_102 = a b c a_mae b_mae c_mae nxa nxb nxc axb axc bxc hetmatnabc

```

```

idadece12/vif collin; run; quit;

proc reg data=dados outest=ridge outvif outstb ridge=0.01 to 0.05 by .001;
    model ce_102 = a b c a_mae b_mae c_mae nxa nxb nxc axb axc bxc
        hetmatnabc idadece12;
    plot / ridgeplot; data new; set ridge;
    if _type_='RIDGESTB' or _type_='RIDGEVIF';
    proc sort; by _type_; proc gplot; by _type_;
    plot (a b c a_mae b_mae c_mae nxa nxb nxc axb axc bxc hetmatnabc
idadece12)*_RIDGE_/overlay vref= 10 cvref=blue lvref=1 ; run;

proc iml;
use dados;read all into dados;
use dadosce_orig; read all into dadosce_orig;
*use b_ce ; *read all into dados2;
    * MATRIX X E VETOR Y;
X = dados[,3:5]||dados[,12:14]||dados[,6:11]||dados[,15]||dados[,2];
y = dados[,16]; produto = dados[,1]; n=nrow(X);

meanX=X[+,]/n; * Vai criar um vetor linha com as médias para cada coluna;
Xc=X-repeat(meanX,n,1); * centralizei X;
cssX=Xc[##,];* Soma dos quadrados dos elementos de cada coluna;
sX=Xc*diag(1/(sqrt(cssX))); * padronizo Xc; XX=T(sX)*sX;

meany = y[+,1]/n;
yc=y-repeat(meany,n,1);
Xy = t(sX)*yc;

    * obtendo soluções por QM ;
betar=inv(XX)*Xy;

```

```

beta_QM=diag(1/(sqrt(cssX)))*betar;

FIV_QM=vecdiag(inv(XX));
mFIV_QM=FIV_QM/max(FIV_QM);
sFIV_QM=sum(FIV_QM)/ncol(XX);
maxFIV_QM=max(FIV_QM);
print beta_QM FIV_QM;
print sFIV_QM maxFIV_QM;

/* obtendo soluções por RR */
do teta=.00001 to 1 by .00001;
    K=diag(teta*mFIV_QM);
    XXIk=XX+K;
    iXXIk=inv(XXIk);
H=iXXIk*XX;
H2=H##2;
iBIASp=(1-((sqrt(sum(H2))/sqrt(ncol(XX))))) *100;

FIV_RC=vecdiag(iXXIk*XX*iXXIk);
sFIV_RC=(sum(FIV_RC))/ncol(XX);
maxFIV_RC=max(FIV_RC);
    if maxFIV_RC<=10 then do;
        solr=iXXIk*Xy;
        betaRR=diag(1/(sqrt(cssX)))*solr;
        vK=vecdiag(K);
        print iBIASp; print betaRR FIV_RC vK;
        print sFIV_RC maxFIV_RC ; teta=1;
    end;
end;

```

```

create betaRR10 from betaRR;
    append from betaRR;

    * CRIANDO Y AJUSTADO;
y_orig=dadosce_orig;
KCOND=y-X*betaRR;
int=KCOND[+,]/n;
print int;
ce_aj11=y_orig-X*betaRR;

D = ce_aj11||produto;
create D1 from D;
    append from D;
run; quit;
ods graphics off;
ods html close;

data tres (rename=(col1=ce_aj11 col2=produto));
set D1; run;

data sasuser.mont_cell1;
merge sasuser.mont1 tres;
by produto;
keep produto mae cimp ce_aj11 gcce12 dtjn;
if ce_aj10=. then ce_aj10=0;
format produto f7.0 mae f7.0 cimp f1.0 ce_aj11 f12.9 gcce12 f5.0; run;

```

ANEXO B - Programa (PEST) utilizado na etapa (50), em uma das iterações, após as variâncias e covariâncias terem sido estimadas no MTDFREML, para a variável resposta perímetro escrotal aos 390 dias

COMMENT

DATA

INFILE = '../dados/ce_10'

INPUT	[	VAR_NAME	MAXLEVEL	START_COLUMN	VAR_LENGTH	DECIMAL	]
		animal	400000	1	7		
		f_p	400000	9	7		
		cimp	8	17	1		
		gcce12	1250	22	5		
		dtjn	0	19	2		
		ce_10	0	28	12	9	

RELATIONSHIP

rel_for animal

infile = '../dados/pedlimpo'

input

animal 1 7

m_p 9 7

f_p 17 7

MODEL

ce_10 = int dtjn dtjn*dtjn dtjn*dtjn*dtjn cimp gcce12 f_p animal

TRANSFORMATION

TREATED_AS_MISSING

ce_10 none 0. none

VE

6.97025

VG

vg_for f_p

0.16397

```
vg_for animal
1.77838
PRINTOUT
  OUTFILE = cepest102
  page = 772073
system_size
  non_zero = 5000000
  total=30000000
solver
  iod_for f_p [max_iter_iod = 10000]
  iod_gs_for animal [max_iter_iod_gs = 10000]
```