TADEU VILELA DE SOUZA

ASPECTOS ESTATÍSTICOS DA ANÁLISE DE TRILHA (*PATH ANALYSIS*) APLICADA EM EXPERIMENTOS AGRÍCOLAS

LAVRAS - MG

2013

TADEU VILELA DE SOUZA

ASPECTOS ESTATÍSTICOS DA ANÁLISE DE TRILHA (*PATH ANALYSIS*) APLICADA EM EXPERIMENTOS AGRÍCOLAS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador
Dr. João Domingos Scalon

LAVRAS - MG
2013

Ficha Catalográfica Elaborada pela Divisão de Processos Técnicos da Biblioteca da UFLA

Souza, Tadeu Vilela de.

Aspectos estatísticos da análise de trilha (path anlysis) aplicada em experimentos agrícolas / Tadeu Vilela de Souza. – Lavras : UFLA, 2013.

82 p. : il.

Dissertação (mestrado) – Universidade Federal de Lavras, 2013.

Orientador: João Domingos Scalon.

Bibliografia.

1. Multicolinearidade. 2. Diagrama de trilha. 3. Análise exploratória. 4. Teste de hipótese. I. Universidade Federal de Lavras. II. Título.

CDD - 519.535

TADEU VILELA DE SOUZA

ASPECTOS ESTATÍSTICOS DA ANÁLISE DE TRILHA (*PATH ANALYSIS*) APLICADA EM EXPERIMENTOS AGRÍCOLAS

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 21 de fevereiro de 2013.

Dr. Joel Augusto Muniz	UFLA
Dr. José Airton Rodrigues Nunes	UFLA

Dr. João Domingos Scalon
Orientador

LAVRAS - MG
2013

A meus pais Maria Anália e Geiel,
pela dedicação, apoio, amor e educação.
A minha irmã Elaine,
pelo incentivo, carinho e conselhos.

DEDICO

AGRADECIMENTOS

A Deus, pela oportunidade de estudar e pela força dada para aguentar firme os momentos difíceis.

Aos maiores merecedores da minha gratidão, minha mãe Maria Anália, meu pai Geiel e minha irmã Elaine, pessoas a quem dedico incondicionalmente meus agradecimentos.

À minha namorada, Brunna, pela compreensão, apoio e tolerância em todos os momentos.

A todos os meus familiares e amigos, pelo apoio, carinho e momentos de alegrias inesquecíveis passados juntos.

Ao meu orientador João Domingos Scalon, pelos conhecimentos e esclarecimentos intelectuais, pela confiança em mim, pela paciência e compreensão das minhas dificuldades e por me aceitar como seu orientando.

Aos professores membros da minha banca pelas importantes contribuições nesta dissertação, por serem receptivos e gentis ao me receberem em seus gabinetes e por participarem da minha qualificação e defesa.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Ciências Exatas (DEX), por oferecer estrutura e acolhimento nessa oportunidade de cursar o mestrado.

Aos professores do programa de pós-graduação em estatística e experimentação agropecuária da UFLA, pelas importantes e úteis contribuições na minha formação durante as suas disciplinas.

Às funcionárias do DEX: Edila, Josiane Cristina, Josiane Oliveira, Kelly, Maria, Miriam e Selma, pela dedicação ao depertamento e às pessoas que o frequentam.

À Fundação de Amparo à Pesquisa do estado de Minas Gerais (FAPEMIG),

pela bolsa de estudos que tornou financeiramente possível a realização do mestrado.

A todos que contribuíram de forma direta ou indireta para a realização deste trabalho.

RESUMO

A análise de trilha é um importante recurso da estatística multivariada, onde correlações entre caracteres são desdobradas em efeitos diretos e indiretos que medem a influência de uma variável, independente das demais, sobre a outra. Essa técnica vem sendo bastante utilizada em muitos campos de pesquisa. Objetiva-se neste trabalho abordar e discutir aspectos estatísticos necessários para o uso dessa metodologia e utilizá-la na análise de dados provenientes de dois experimentos agrícolas. Especificamente, discuti-se técnicas estatísticas importantes, como análise exploratória, análise de multicolinearidade, método de estimação de parâmetros e etc, e necessárias para a realização das duas etapas básicas da análise de trilha, que são: i) a formulação do diagrama de trilha; ii) estimação dos coeficientes de trilha. A estimação das correlações, coeficientes de trilha, assim como os testes estatísticos, foram feitos usando funções desenvolvidas e/ou disponíveis no *software* R. As várias técnicas discutidas são aplicadas em duas análises. O primeiro conjunto de dados é provenientes de um experimento conduzido no Laboratório de Biotecnologia Vegetal da Embrapa Mandioca e Fruticultura, em Cruz das Almas, Bahia, que foi conduzido por Faria (2008). Nessa análise desdobram-se as correlações referentes a características do cultivo *in vitro* da planta de maracujazeiro da espécie *Passiflora giberti* N.E. Brown em efeitos diretos e indiretos sobre o tamanho da plântula. A variável número de gemas foi a principal determinante da variação no tamanho da plântula. A segunda análise foi aplicada em dados oriundos de um experimento conduzido por Ribeiro (2012) na fazenda experimental da Universidade Federal de Lavras, em Lavras, onde é estudada a relação de características morfológicas e componentes de rendimento da planta de milho (*Zea mays* L.) em efeitos diretos e indiretos sobre sua produtividade. O uso da análise de trilha nesse experimento mostrou que o peso de 100 grãos foi o componente que apresentou o maior efeito direto sobre a produção de grãos (PROD), sendo os mais indicados para seleção indireta para PROD.

Palavras-chave: Multicolinearidade. Diagrama de trilha. Análise exploratória. Teste de hipótese.

ABSTRACT

The path analysis is an important tool of multivariate statistics, where correlations between variables are unfolded into direct and indirect effects measuring the influence of one variable, independent of the others, on the other. This technique has been widely used in many fields of research. The aim of this work is to address and discuss statistical aspects required to use this methodology and to apply in the analysis of data from two agricultural experiments. Specifically, it is discussed important statistical techniques, such as exploratory analysis, analysis of multicollinearity, parameter estimation, etc. which are necessary to the achievement of the two basic steps of the path analysis: i) the formulation of the path diagram, ii) estimation of the path coefficients. The estimation of the correlation coefficients, path coefficients and statistical tests were made by using both available and hand made functions under the R software. The techniques are applied in two data sets. The first set are from an experiment conducted in the Laboratory of Plant Biotechnology at Embrapa Mandioca e Fruticultura, in Cruz das Almas, Bahia, which was conducted by Faria (2008). This analysis unfold the correlations of *in vitro* cultivation characteristics of the plant species of passion fruit *Passiflora giberti* N.E. Brown in direct and indirect effects on the size of the seedling. The variable number of buds was the main determinant of variation in the size of the seedling. The second analysis was carried out in a data set from an experiment conducted by Ribeiro (2012) at the Experimental Farm of the Federal University of Lavras, where was studied the relationship of morphological characteristics and yield components of maize plants (*Zea mays* L.) in direct and indirect effects on yield. The use of path analysis in this experiment showed that the weight of 100 grains was the component that had the highest direct effect on grain yield, being the most suitable for indirect selection for grain yield.

Keywords: Multicollinearity. Path diagram. Exploratory analysis. Hypothesis test.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O estudo das correlações entre variáveis é aplicável em praticamente todos os campos de pesquisa. A correlação simples permite apenas avaliar a magnitude e o sentido da associação entre duas variáveis, mas não fornece as informações necessárias sobre os efeitos diretos e indiretos de um grupo de variáveis independentes em relação a uma variável dependente. A análise de trilha (*path analysis*) permite o estudo dos efeitos diretos e indiretos de várias variáveis independentes sobre uma variável dependente (básica), cujas estimativas são obtidas por meio de equações de regressão em que as variáveis são primeiramente padronizadas (CRUZ; REGAZZI; CARNEIRO, 2004).

O sucesso da análise de trilha se baseia na formulação mais consistente do relacionamento causa-efeito entre as variáveis. Além disso, o desdobramento de correlações é dependente do conjunto de variáveis estudadas, que normalmente é estabelecido a partir do conhecimento prévio de sua importância pelo pesquisador e de possíveis inter-relações expressas em diagramas de trilha (CRUZ; CARNEIRO, 2003). Esse sucesso da análise de trilha também pode ser medido pelo grande número de artigos em que a técnica vem sendo empregada em áreas de conhecimento tão diversas quanto ciências sociais (LOEHLIN, 2004) e agrárias (NUNES et al., 2004). Entretanto, o que se observa na literatura é que, em geral, a análise de trilha apresentada nesses trabalhos consta apenas de duas partes: construção do diagrama de trilha e estimação dos coeficientes de trilha. Além disso, esses trabalhos apresentam apenas os resultados produzidos por algum *software* sem explicitar vários aspectos estatísticos envolvidos na análise, o que pode comprometer todas as conclusões baseadas nesses resultados. Somente para exemplificar alguns problemas estatísticos observados em alguns trabalhos, pode-se mencionar: omissão de testes de hipóteses para verificar a suposição de normalidade, independência

e homocedasticidade dos resíduos da regressão; ausência de análise exploratória para detectar *outliers* e multicolinearidade nas variáveis; critérios para a escolha dos caminhos não é informada; omissão sobre a informação do método de estimação dos parâmetros; entre outros.

Do exposto anteriormente, o objetivo da dissertação será apresentar e discutir aspectos estatísticos envolvidos na análise de trilha, tais como: análise exploratória, detalhamento teórica da técnica, técnicas para a escolha dos caminhos, métodos de estimação dos parâmetros e verificação dos pressupostos para aplicação da metodologia. Os métodos estatísticos apresentados para conduzir a análise de trilha serão utilizados para analisar dados provenientes de dois diferentes experimentos. A primeira análise explora a relação entre diversas características do cultivo *in vitro* da planta de maracujazeiro da espécie *Passiflora giberti* N.E. Brown através de seus desdobramentos em efeitos diretos e indiretos sobre o tamanho da plântula. E a segunda análise estuda a relação de características morfológicas e componentes de rendimento da planta de milho em efeitos diretos e indiretos sobre sua produtividade.

2 REFERENCIAL TEÓRICO

Nesta seção são apresentados os vários apectos envolvidos no uso da análise de trilha, desde a definição do diagrama até o processo final de estimação e interpretação dos coeficientes de trilha.

2.1 Correlações

Em várias áreas de estudo, muitas vezes se faz necessário medir a existência e/ou a intensidade da interação entre caracteres. Por exemplo, em produção vegetal o estudo das relações entre as variáveis envolvidas no melhoramento genético é um dos aspectos mais importantes a se considerar, pois possibilita a obtenção de ganhos para caracteres de interesse por meio da manipulação de outras características correlacionadas.

A análise de correlação fornece um valor que representa a variação conjunta entre duas variáveis, e também mede a intensidade e a direção da relação linear ou não-linear entre duas variáveis (CHARNET et al., 2008). Para Lira (2004), o coeficiente de correlação é um indicador que atende à necessidade de se estabelecer a existência ou não de uma relação entre essas variáveis sem que, para isso, seja preciso o ajuste de uma função matemática. Não existe a distinção entre a variável independente (X) e a variável dependente (Y), ou seja, o grau de variação conjunta entre X e Y é igual ao grau de variação conjunta entre Y e X. Neste sentido, o conhecimento do coeficiente de correlação é importante pois possibilita ao melhorista saber como a seleção para um caráter está relacionado a expressão de outros caracteres (FREIRE FILHO, 1988).

O coeficiente de correlação linear assume que há uma relação linear entre duas variáveis, ou seja, que a mudança de uma variável sempre envolve a mudança

constante no valor médio de outra variável. Este coeficiente, portanto, reflete o grau de associação entre duas variáveis, e o valor desse coeficiente é positivo quando ocorre aumento (ou diminuição) nas duas variáveis e é negativo quando uma variável aumenta e a outra diminui. Quanto mais próximo de um, mais forte é a correlação entre as variáveis. Assim, pode-se afirmar que o coeficiente de correlação linear é adimensional e seu valor absoluto não ultrapassa à unidade, sendo que quando for igual a zero reflete a falta de relação linear (CHARNET et al., 2008).

A correlação linear simples permite avaliar a magnitude e o sentido das relações entre dois caracteres, sendo de grande utilidade no melhoramento, por permitir avaliar a viabilidade da prática da seleção indireta, que, em alguns casos pode levar a progressos mais rápidos que a seleção do caráter desejado (CRUZ; CARNEIRO, 2003).

Antes de calcular o coeficiente de correlação faz-se necessário uma análise de *outliers*, pois o coeficiente de correlação é fortemente afetado pela presença deles. A presença de *outliers* pode comprometer fortemente as estimativas do coeficiente de correlação levando, inclusive, o pesquisador a cometer erros do tipo I ou do tipo II (OSBORNE; WATERS, 2002). Onde o erro tipo I acontece quando se rejeita uma hipótese sendo ela verdadeira, e o erro tipo II ocorre quando se aceita uma hipótese falsa. Também faz-se necessária a independência das observações, ou seja, a ocorrência de uma observação $(X1, Y1)$ não influencia a ocorrência de outra observação $(X2, Y2)$. Segundo Osborne e Waters (2002), a violação desta orientação implica risco do coeficiente produzir correlações espúrias.

2.1.1 Coeficiente de correlação linear de Pearson

Existem, na literatura, diversos estimadores para a correlação linear. O coeficiente de correlação linear de Pearson, também chamado de "coeficiente de correlação produto-momento" mede a força e a direção da correlação (positiva ou negativa) entre duas variáveis aleatórias X e Y de escala métrica (intervalar ou de razão). Este coeficiente de correlação é dado pela expressão:

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X).V(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \cdot \sigma_Y}$$

Em que: $\sigma_{X,Y}$ é a covariância entre as variáveis aleatórias X e Y; σ_X é o desvio padrão da variável aleatória X; e σ_Y é o desvio padrão da variável aleatória Y.

Em Casella e Berger (2001) é provado que o coeficiente de correlação de Pearson sempre está no intervalo $[-1, 1]$. Quanto mais o valor absoluto do coeficiente de correlação se aproxima de 1 mais forte é a correlação entre as variáveis. O sinal indica se a direção da correlação é positiva ou negativa. O sinal é positivo quando há uma relação direta entre as variáveis, e negativo quando há uma associação inversa entre elas, isto é, valores altos de uma variável estão associados a valores baixos da outra variável, e vice-versa.

Sob a suposição de normalidade bivariada, pode-se construir um teste de hipótese para a correlação nula entre as variáveis aleatórias X e Y (COSTA NETO, 2009). Essa hipótese é testada pela seguinte estatística de significância t_c:

$$t_c = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

Em que: r_{XY} é a correlação linear entre X e Y, e n é o número de ele-

mentos da amostra. A hipótese de nulidade H_0 ($H_0 : \rho = 0$, a correlação é estatisticamente igual a zero) não é rejeitada se $|t_c| < t_{(\frac{\alpha}{2};n-2)}$ (COSTA NETO, 2009).

2.1.2 Coeficiente de correlação múltipla

O coeficiente de correlação linear múltipla mede o grau de relacionamento entre as variáveis independentes (X_i, $i = 1, 2, ..., k$) e a variável dependente Y de um modelo. Os princípios gerais deste método constituem apenas uma extensão direta dos conceitos e raciocínios apresentados para o coeficiente de correlação linear de Pearson.

Considerando, para efeito ilustrativo, duas variáveis independentes X_1 e X_2, e uma variável dependente Y. Dessa forma, a estimativa do coeficiente de correlação múltipla entre três variáveis é obtido através da expressão:

$$r_{Y,X_1,X_2} = \sqrt{\frac{r^2_{X_1Y} + r^2_{X_2Y} - 2r_{X_1Y}r_{X_2Y}r_{X_1X_2}}{1 - r^2_{X_1X_2}}}$$

Uma vez que existe relação entre a análise de correlação múltipla e regressão múltipla, também é possível, através da segunda, obter-se o coeficiente de correlação múltipla pela raíz quadrada do coeficiente de determinação da regressão múltipla r^2:

$$r_{Y,X_1,X_2} = \sqrt{\frac{SQ_{Reg}}{SQ_{Total}}}$$

Em que: SQ_{Reg} é a soma de quadrados da regressão linear múltipla e SQ_{Total} é a soma de quadrados total.

Para a realização do teste de significância do coeficiente de correlação

múltipla utiliza-se, como estatística de teste, a razão F_c (LIRA, 2004):

$$F_c = \frac{r^2/k}{(1-r^2)/(n-k-1)}.$$

Onde: r^2 é o coeficiente de determinação, n é o tamanho da amostra e k é o número de variáveis independentes. A hipótese de nulidade H_0 ($H_0 : \rho = 0$, a correlação é estatisticamente igual a zero) é aceita se $F_c < F_{(\frac{\alpha}{2};k;n-k-1)}$.

2.1.3 Coeficiente de correlação parcial

Enquanto a correlação simples mede a associação linear entre duas variáveis, o coeficiente de correlação parcial mede a associação entre duas variáveis após controlar os efeitos de uma ou mais variáveis adicionais. Nessa situação, supondo que se deseja estudar a relação entre três variáveis X_1, X_2 e X_3, para isso pode-se calcular o coeficiente de correlação de Pearson para cada par de variáveis. Mas, por exemplo, o coeficiente de correlação mensurado entre as variáveis X_1 e X_2 (r_{12}) compreende também os efeitos que X_3 possa ter causado no comportamento delas. Ou seja, a utilização do coeficiente de correlação linear simples somente entre duas variáveis, desconsiderando o efeito que uma terceira variável possa provocar sobre elas, pode acarretar um resultado impreciso. Já o coeficiente de correlação parcial é usado quando o objetivo é conhecer a correlação entre duas variáveis quaisquer, controlando o efeito das outras variáveis envolvidas, ou seja, desconsiderando seus efeitos. Para representar a correlação entre as variáveis X_1 e X_2, controlando X_3, utiliza-se a notação $r_{12.3}$ que pode ser estimada por:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Existe uma ordem associada às correlações parciais. Essa ordem indica

quantas variáveis estão sendo controladas. O coeficiente $r_{12.3}$ é um coeficiente de correlação parcial de primeira ordem, pois controla o efeito de uma variável adicional. Já um coeficiente de segunda ordem $r_{12.34}$ controla o efeito de duas variáveis, o coeficiente de terceira ordem $r_{12.345}$ controla o efeito de três outras variáveis, e assim por diante. São calculados de modo análogo os coeficientes parciais de ordem mais elevada, para a correlação parcial de *(n+1)-ésima* ordem substitui-se os coeficientes de correlação simples do membro direito da equação acima pelos coeficientes parciais de *n-ésima* ordem.

Para a utilização desta análise é necessário observar, que os dados sigam uma distribuição normal multivariada, mais as outras suposições da correlação linear de Pearson (CARVALHO et al., 2004).

O teste de significância para a correlação parcial é semelhante ao utilizado para a correlação linear de Pearson. A estatística utilizada é:

$$t_c = \frac{r\sqrt{n-k-2}}{\sqrt{1-r^2}}$$

onde: r é a correlação parcial estimada entre duas variáveis; n é o número de observações da qual o coeficiente de correlação simples foi calculado; k é o número de variáveis independentes. A hipótese de nulidade H_0 ($H_0 : \rho = 0$, a correlação é estatisticamente igual a zero) é aceita se $|t_c| < t_{(\frac{\alpha}{2};n-n_x-2)}$ (CARVALHO et al., 2004).

2.2 Análise de regressão

A análise de regressão estuda o relacionamento entre uma variável chamada variável dependente e outra variável chamada variável independente através de um modelo matemático. As análises de regressão e de correlação são duas técnicas estreitamente relacionadas. Essas técnicas analisam dados amostrais, para mostrar

como duas ou mais variáveis estão relacionadas, uma com a outra, em certa população. Enquanto a correlação dá o número que resume a magnitude, ou o grau de relacionamento entre as variáveis, a análise de regressão fornece um modelo matemático, que descreve esse relacionamento. Esse modelo pode ser utilizado para estimar ou predizer valores futuros de uma variável, quando se conhecem, ou se supõe conhecidos, os valores de outras variáveis (COSTA NETO, 2009).

O modelo matemático denominado modelo de regressão linear simples se define a partir de uma relação linear entre a variável dependente e uma variável independente. Se existirem várias variáveis independentes, o modelo passa a denominar-se modelo de regressão linear múltipla (CHARNET et al., 2008).

2.2.1 Análise de regressão linear múltipla

A regressão linear múltipla envolve uma única variável dependente e duas ou mais variáveis independentes. A análise tem por objetivo encontrar um modelo, que possa ser utilizado para predizer valores de Y dado valores das diversas variáveis independentes.

O modelo de regressão linear múltiplo, com k variáveis independentes e β_j ($j = 0, ..., k$) parâmetros, pode ser escrito como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i \quad i = 1, 2, 3, ..., n$$

em que os resíduos são independente e identicamente distribuídos e $\varepsilon_i \sim N(0, \sigma^2)$.

O modelo é escrito na forma matricial como:

$$Y = X\beta + \varepsilon.$$

Sendo Y a matriz $(n \times 1)$ das observações aleatórias, X a matriz $(n \times k)$

das variáveis independentes, β é matriz ($k \times 1$) dos coeficientes de regressão e ε é a matriz ($n \times 1$) dos erros aleatórios ($\varepsilon \sim N(0, I\sigma^2)$). Ou seja:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & X_{31} & X_{32} & \cdots & X_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

A estimação dos parâmetros ($\beta_0, \beta_1, \beta_2, ..., \beta_k$) pode ser realizada pelo método dos mínimos quadrados (ordinários ou ponderados) e também pode-se utilizar o método de máxima verossimilhança. Após a estimação dos parâmetros, há a necessidade de testar a significância da regressão e realizar um teste individual para cada coeficiente, a fim de validar o modelo. Para utilizar o método de máxima verossimilhança e realizar as inferências estatísticas, deve-se supor que ε tenha distribuição normal com média zero e variância constante (CHARNET et al., 2008; MONTGOMERY; PECK, 1992). Assim, análise da pressuposições dos resíduos é de fundamental importância para realizar inferências. No ajuste de modelos de regressão linear múltipla deve-se detectar a presença de variáveis independentes multicolineares (ver seção 2.3.3). A violação desta orientação implica risco de inferências espúrias (MONTGOMERY; PECK,1992). A qualidade do modelo ajustado pode ser feita utilizando diversas medidas como o coeficiente de determinação múltiplo (R^2), soma de quadrado de resíduos, etc. (CHARNET et al., 2008; DRAPER; SMITH, 1998; HAIR et al., 1998; MONTGOMERY; PECK, 1992).

2.2.1.1 Estimação dos parâmetros de regressão

Existem diversos métodos para a estimação dos parâmetros de um modelo de regressão linear múltiplo, são apresentados nessa seção o processo de estimação pelos métodos de mínimos quadrados ordinários, mínimos quadrados ponderados e máxima verossimilhança.

1 - Método dos mínimos quadrados ordinários

Considerando um conjunto com k variáveis independentes e uma variável dependente Y, onde a relação entre a variável dependente e as variáveis independentes pode ser representada da seguinte forma:

$$Y = X\beta + \varepsilon.$$

O método de mínimos quadrados para obtenção o dos estimadores dos parâmetros $\beta's$ consiste em minimizar a soma dos quadrados dos erros, ou seja, deve-se encontrar o vetor dos estimadores de mínimos quadrados $\hat{\beta}$ que minimiza a expressão:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta).$$

Pode-se, então, escrever L da seguinte forma:

$$L = Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta = Y'Y - 2\beta'X'Y + \beta'X'X\beta.$$

Para encontrar $\hat{\beta}$ que minimiza $L = \varepsilon'\varepsilon$, calcula-se a diferencial de L em relação a $\hat{\beta}$ e depois iguala-se essa diferencial a zero:

$$\left.\frac{\partial L}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0.$$

Onde, simplificando, obtém-se o sistema de equações normais:

$$X'X\hat{\beta} = X'Y.$$

Para resolver esta equação, sabendo que $X'X$ é uma matriz não-singular, basta multiplicar ambos os lados da igualdade pela sua inversa $(X'X)^{-1}$. Dessa forma, os estimadores de mínimos quadrados para os coeficientes são dados por:

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

A esperança e a variância desse estimador são: $E\left[\hat{\beta}\right] = \beta$ e $V\left[\hat{\beta}\right] = \sigma^2(X'X)^{-1}$. Desta forma, $\hat{\beta}$ é um estimador linear não tendencioso e com variância mínima, entre todos os estimadores lineares não tendenciosos.

2 - Método dos mínimos quadrados ponderados

A presença de variâncias desiguais é uma das violações mais comuns, em que a matriz de covariância dos erros não é da forma $I\sigma^2$, e sim uma matriz diagonal com elementos desiguais σ_i^2, ou seja, a estrutura dos erros viola a pressuposição de homogeneidade de variâncias. Quando isso ocorre, o estimadores de mínimos quadrados ponderados são eficientes, pois pertencem à classe dos estimadores lineares não tendenciosos e produzem um vetor de erros com variância constante.

A utilização desse método consiste em fazer transformações apropriadas na variável dependente Y e nas variáveis independente X_i ($i = 1, 2, ..., k$), de forma que ao estimar os parâmetros do modelo por mínimos quadrados $\hat{\beta} = (X'X)^{-1}X'Y$, produzam um vetor novo de erros, u, com média zero e variância constante ($u \sim N(0, I\sigma^2)$). Seja o seguinte modelo na forma matricial com

erro heterocedástico:

$$Y = X\beta + \varepsilon, \tag{2.1}$$

onde:

$E(\varepsilon) = 0$ e $Var(\varepsilon) = H(X)\sigma^2 = W^{-1}\sigma^2$, ou seja, $\varepsilon \sim N(0, W^{-1}\sigma^2)$. Em que $H(X)$ é uma função das variáveis explicativas, que determinam a heterocedasticidade. Essa função compõe uma matriz simétrica $(n \times n)$ e, para facilitar a interpretação de cálculos adiante, será denominada W^{-1}, e W^{-1} é uma matriz diagonal, positiva definida cujos elementos da diagonal são os pesos que ponderam a variância. Através da fatoração de Cholesky, é possível escrever a matriz W como função de uma matriz triangular superior P, de forma que $W = P'P$ ou $W^{-1} = P^{-1}(P')^{-1}$.

Multiplicando ambos os lados da equação 2.1 por P, obtém-se o modelo com as variáveis transformadas:

$$PY = PX\beta + P\varepsilon.$$

Que pode ser escrito como:

$$Z = Q\beta + u.$$

Como $u = P\varepsilon$, tem-se que $u \sim N(0, I\sigma^2)$, pois:

$$E(u) = E(P\varepsilon) = PE(\varepsilon) = P0 = 0$$

e

$$Var(u) = E(uu') - [E(u)]^2 = E(P\varepsilon\varepsilon'P') - 0 = P.E(\varepsilon\varepsilon').P' = P.\{Var(\varepsilon) + [E(\varepsilon)]^2\}.P' = P.(W^{-1}\sigma^2).P' = PP^{-1}(P')^{-1}P'.\sigma^2 = I\sigma^2.$$

Portanto, o estimador dos parâmetros por mínimos quadrados ponderados

é:

$$\hat{\beta} = (X'P'PX)^{-1}(X'P'PY) = (X'WX)^{-1}(X'WY).$$

Esse é um estimador não tendencioso, pois sua esperança é $E\left[\hat{\beta}\right] = \beta$, e com variância constante igual a $V\left[\hat{\beta}\right] = \sigma^2(X'WX)^{-1}$.

Os resíduos que devem ser analisados são estimados através da equação $\hat{u} = P(Y - \hat{Y})$, e sabendo que $\hat{Y} = X\hat{\beta}$, têm-se:

$$\hat{u} = P(Y - X((X'WX)^{-1}(X'WY))).$$

2 - Método da máxima verossimilhança

Para utilização desse método é necessário que $\varepsilon \sim N(0, I\sigma^2)$ e $Y \sim N(X\beta, I\sigma^2)$. A estimação dos parâmetros de regressão múltipla através da máxima verossimilhança consiste em encontrar valores para os $\beta's$ que maximizem a função de verossimilhança $L(Y_i|X, \beta, \sigma^2)$, dada por:

$$L(Y_i|X, \beta, \sigma^2) = \prod_{i=1}^{n} f(Y_i|X, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}$$

E considerando o modelo na forma matricial $Y = X\beta + \varepsilon$, a função (acima) $L(Y_i|X, \beta, \sigma^2)$ pode ser escrita da seguinte forma:

$$L(Y_i|X, \beta, \sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n}\, e^{-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)} \tag{2.2}$$

Para a estimação dos parâmetros de 2.2 primeiro é feita a log-verossimilhança dessa função:

$$\ln L(Y_i|X, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\left[\frac{(Y - X\beta)'(Y - X\beta)}{\sigma^2}\right].$$

Simplificando, tem-se:

$$\ln L(Y_i|X,\beta,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(Y'Y - 2Y'X\beta + \beta'X'X\beta).$$

Agora, derivando em relação a β:

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{2\sigma^2}(-2X'Y + 2X'X\beta) = \frac{1}{\sigma^2}(X'Y - X'X\beta).$$

Igualando a zero pode ser obter a estimativa $\hat{\beta}$:

$$\frac{1}{\sigma^2}(X'Y - X'X\hat{\beta}) = 0 \;\Rightarrow\; X'X\hat{\beta} = X'Y \;\Rightarrow\; \hat{\beta} = (X'X)^{-1}X'Y.$$

Ou seja, o sistema resultante é resolvido da mesma forma que nos métodos de mínimos quadrados. Os estimadores $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ são iguais aos obtidos pelo método de mínimos quadrados, com $E\left[\hat{\beta}\right] = \beta$ e $V\left[\hat{\beta}\right] = \sigma^2(X'X)^{-1}$.

2.3 Análise de trilha

Os coeficientes de correlação simples entre caracteres não permitem que sejam tiradas conclusões sobre relações de causa e efeito entre eles, ou seja, não compreende os efeitos diretos e indiretos de caracteres sobre uma variável básica. Dada à importância dessas relações, Wright (1921) desenvolveu o método da análise de trilha, que quantifica essas relações de causa e efeito. É possível ver de forma detalhada o método dos coeficientes de trilha em Wright (1934).

A análise de trilha consiste no desdobramento das correlações em efeitos diretos e indiretos, permitindo medir a influência direta de uma variável, independentemente das demais, sobre a outra, onde as estimativas (coeficientes de trilha ou caminho) que quantificam esses efeitos são obtidas por meio de equações de

regressão, em que as variáveis são previamente padronizadas (CRUZ; REGAZZI; CARNEIRO, 2004). Essas estimativas são obtidas a partir do método de mínimos quadrados. Quando se considera um único modelo casual, a análise de trilha trata-se de uma análise de regressão linear múltipla padronizada.

A execução da análise de trilha é, em geral, realizada em duas partes:

1. construção de um diagrama de caminho que, embora não seja essencial para a análise numérica, é muito útil para exibir graficamente o padrão de hipótese das relações de causa e efeito entre um conjunto de variáveis, ou seja, estabelece uma relação de causa e efeito entre as variáveis;

2. decomposição das correlações observadas em um conjunto de coeficientes (coeficientes de caminho) que indica o efeito direto de uma variável hipoteticamente tomada como causa sobre uma variável tratada como um efeito.

Deve-se observar que nessa dissertação a análise de trilha é aplicada de tal maneira que é possível abordar diversos métodos estatísticos envolvidos nessas duas etapas da análise que são, em geral, ignoradas por alguns pesquisadores.

Dentre as utilidades da análise de trilha, uma das principais é possibilitar o conhecimento dos efeitos diretos e indiretos que variáveis explicativas exercem sobre uma variável principal permitindo, assim, estabelecer qual estratégia será mais eficiente na seleção, para incrementar o melhoramento genético.

Negreiros et al. (2007) estudaram o efeito de cinco caracteres (comprimento e diâmetro equatorial do fruto, peso do fruto, peso da casca e da polpa, espessura da casca, rendimento e relação comprimento/diâmetro) sobre o rendimento de polpa de maracujá-amarelo. Na estimação dos coeficientes de trilha, primeiramente, utilizou-se um diagrama de caminho apresentando as relações causa/efeito, partindo-se da associação entre a variável básica, peso do fruto, e seus componentes primários, peso da casca e peso da polpa com seus componentes secundários

comprimento e diâmetro do fruto, espessura de casca e relação entre o comprimento e o diâmetro (comprimento/diâmetro). O segundo diagrama causal indicou a inter-relação da variável básica rendimento e seus componentes primários, comprimento e diâmetro do fruto, espessura de casca. Usando a análise foi possível evidenciar que a seleção dos frutos com maior diâmetro equatorial possibilita a obtenção de maracujás mais pesados e com maior rendimento de polpa, uma vez que o diâmetro tem maior efeito direto sobre o peso da polpa e rendimento, e que rendimento da polpa também pode ser selecionado indiretamente, com base na menor espessura da casca.

Nunes et al. (2004) avaliaram, usando a análise de trilha, a importância das características físicas e químicas na determinação do teor de vitamina C em frutos de aceroleira. As características físicas foram: peso da polpa por peso do fruto (RPF), diâmetro do fruto, comprimento do fruto, peso do fruto, peso da polpa e peso das sementes; e químicas: sólidos solúveis totais, acidez (pH) e acidez total titulável. Calculando os coeficientes de trilha descobriu-se que a característica acidez total titulável foi a principal determinante, com alto efeito direto, do teor de vitamina C.

Como a análise de trilha constitui-se numa expansão da regressão múltipla, quando são envolvidas inter-relações complexas e, ou, vários diagramas causais, a confiabilidade dos coeficientes de trilha pode ser afetada pelos efeitos de multicolinearidade existentes entre os caracteres que compõem o diagrama causal em razão das elevadas variâncias associadas aos seus estimadores (KLINE, 1991).

Quando a multicolinearidade aumenta, a habilidade de definir quaisquer efeitos das variáveis diminui. Também se observa que alguns estimadores atingem valores muito altos, evidenciando uma estimativa pouco confiável (HAIR et al., 1998). Uma correção recomendada é retirar uma, ou mais, variáveis indepen-

dentes, altamente correlacionadas. A seleção dessas variáveis pode ser feita pelo método *Stepwise* de seleção de variáveis, apesar desse procedimento ser, em alguns casos, questionável (KOSACK; AZEVEDO, 2011). Uma outra maneira de retirar variáveis multicolineares é usar a regressão em componentes principais. Nesta técnica, o procedimento consiste na exclusão de variáveis por intermédio dos componentes principais correspondentes aos autovalores. Neste caso, variáveis com autovalores próximos de zero são removidas da análise e o método dos mínimos quadrados é aplicado aos componentes restantes (MONTGOMERY; PECK, 1992).

Existem ainda, quando não se deseja retirar variáveis, métodos alternativos à estimação de mínimos quadrados para contornar os efeitos da multicolinearidade e aumentar a estabilidade dos coeficientes de regressão. Embora esses métodos fornecerem estimadores tendenciosos, Gunst e Mason (1977) afirmam que eles apresentam melhor desempenho quando comparados aos estimadores de mínimos quadrados.

O método de regressão em crista proposto por Hoerl e Kennard (1970a, 1970b) é o método alternativo aos estimadores de mínimos quadrados mais usado para combater os problemas proporcionados pela multicolineridade. Segundo Cruz e Carneiro (2003), o método da regressão em crista consiste em obter estimativas dos coeficientes de regressão a partir de uma versão ligeiramente modificada das equações normais. Maiores detalhes sobre como identificar e trabalhar com multicolinearidade serão apresentados nos capítulos 2.3.4 e 2.3.5.

2.3.1 Escolha do diagrama de trilha

O êxito da análise de trilha está ligado intrinsecamente a formulação do diagrama de trilha (diagrama causal). A construção gráfica desse esquema causal

possibilita a obtenção de um conjunto de equações simultâneas que tem o objetivo de explicar as possíveis inter-relações de causa e efeito entre as variáveis. Porém, este é talvez o passo mais difícil da análise de trilha, pois a especificação desse diagrama deveria ser feita de modo que explicasse as verdadeiras relações entre as variáveis explicativas com a variável básica, e às vezes isso é muito difícil e até impossível. Para Vasconcelos, Almeida e Nobre (1998), existe uma falta de metodologias úteis nesta fase, que é a base de toda a estimativa inicial, como testes e procedimentos de validação.

A escolha do diagrama causal na maioria das vezes é feita baseada numa revisão da literatura específica, onde busca-se informações para se estabelecer as inter-relações possíveis entre variáveis. Este processo permite a construção de um modelo preliminar de que é testado pelos dados empíricos, utilizando os procedimentos da análise de trilha. De acordo com Li (1956), o diagrama causal é baseado em um conhecimento prévio de relações causais entre as variáveis pelo pesquisador, ou através de uma relação hipotética de causa efeito que o pesquisador escolhe para ser testada.

Vasconcelos, Almeida e Nobre (1998) citam duas abordagens estatísticas para a escolha do diagrama e formulação do modelo que foram identificados: o método da correlação parcial, e o método de decomposição hierárquica.

2.3.1.1 Método da correlação parcial

A comparação entre a correlação parcial e correlação de ordem zero proposta por Goldsmith em 1977 pode ser utilizada para assistência na especificação do diagrama de trilha (VASCONCELOS; ALMEIDA; NOBRE, 1998). De acordo com esse critério, se a diferença entre r_{ij} (correlação linear simples entre as variáveis X_i e X_j) e $r_{ij.k}$ (correlação parcial linear entre X_i e X_j, controlando para

X_k) é alta, então pode ser aceito que X_k participa da trilha causal entre X_i e X_j. Essa diferença é escrita da seguinte forma:

$$|r_{ij} - r_{ij.k}| = \Delta_{ij} \tag{2.3}$$

A significância estatística de 2.3 pode ser avaliada pela transformação de Fisher (1932):

$$z = \frac{1}{2} ln(\frac{1+r}{1-r}) \tag{2.4}$$

que é aproximadamente normal com $\sigma(z) = \sqrt{1/(n-3)}$, onde n é o número de observações independentes.

Segundo Vasconcelos, Almeida e Nobre (1998), a equação 2.4 permite a classificação dos Δ_{ij}, de acordo com intervalos de desvio padrão. Na prática, se $\Delta_{ij} < \sigma$ implica que X_k não influencia a relação entre X_i e X_j, enquanto $\Delta_{ij} > 2\sigma$ implica que X_k é uma parte importante dessa relação.

2.3.1.2 Método da decomposição hierárquica

Brooks (1980) propôs um outro critério para a especificação do diagrama de caminho, apropriado para situações onde as variáveis independentes apresentam um elevado grau de multicolinearidade. Este critério, comum na análise de regressão múltipla, é baseado em métodos de seleção "para trás e para frente", e é constituído por duas etapas.

No primeiro passo, todas as informações disponíveis *a priori* sobre as variáveis são usadas para ordenar as variáveis independentes em uma sequência, de acordo com sua hipótese de relevância causal, para uma variável dependente. No segundo passo, uma informação *posteriori* é usada para decidir quais as variáveis independentes devem ser mantidas (VASCONCELOS; ALMEIDA; NOBRE,

1998). Este passo é realizado retirando-se pelo menos uma variável independente da equação de regressão múltipla, onde retira-se a variável considerada mais distante ("menos importante") na sequência causal e observa-se o coeficiente de correlação múltipla da regressão. Se esse coeficiente diminuir acentuadamente significa que a variável independente em questão contribui significativamente para a causa da variável dependente. Se, por outro lado, a correlação múltipla não diminui muito, então a variável independente sob consideração não contribui diretamente para a causa da variável dependente e, portanto, deve ser retirada da equação.

O critério de decisão para o segundo passo é fornecido pelo método de decomposição hierárquica e verifica que somente variáveis com contribuição estatisticamente significante para o coeficiente de determinação múltipla R^2 devem ser mantidas na equação. Para Brooks (1980), o mérito desta abordagem está em proporcionar um limite de tolerância para a multicolinearidade, permitindo a inclusão de uma variável ou conjunto de variáveis quando se adiciona uma informação relevante.

2.3.2 Desdobramento das correlações e estimação dos coeficientes de trilha

Considerando-se uma variável básica Y e um conjunto de variáveis explicativas (X_1, X_2, X_3) que apresentam o seguinte diagrama de caminho, tem-se:

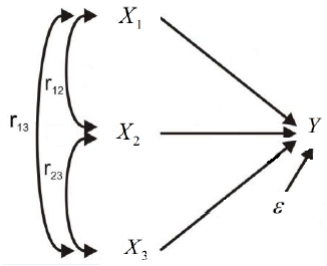

Figura 1 Diagrama causal ilustrativo dos efeitos das variáveis explicativas (X_1, X_2, X_3) e residual (ε) sobre a variável básica (dependente) Y

Dessa forma, as variáveis se relacionam através de um único modelo de regressão múltipla:

$$Y = b_{YX_1}(X_1) + b_{YX_2}(X_2) + b_{YX_3}(X_3) + \varepsilon \tag{2.5}$$

A padronização das variáveis em 2.5, conforme descrita por Li (1975), é feita da seguinte forma:

1. Subtraindo-se a média de cada variável:

$$Y - \bar{Y} = b_{YX_1}(X_1 - \bar{X}_1) + b_{YX_2}(X_2 - \bar{X}_2) + b_{YX_3}(X_3 - \bar{X}_3) + (\varepsilon - \bar{\varepsilon})$$

2. Dividindo ambos os membros pelo desvio padrão da variável básica:

$$\frac{Y - \bar{Y}}{\sigma_Y} = \frac{b_{YX_1}(X_1 - \bar{X}_1)}{\sigma_Y} + \frac{b_{YX_2}(X_2 - \bar{X}_2)}{\sigma_Y} + \frac{b_{YX_3}(X_3 - \bar{X}_3)}{\sigma_Y} + \frac{(\varepsilon - \bar{\varepsilon})}{\sigma_Y}$$

3. Multiplicando-se e dividindo-se cada termo do 2° membro pelo respectivo desvio-padrão da variável associada a esse termo:
$\frac{Y-\bar{Y}}{\sigma_Y} = \frac{b_{YX_1}(X_1-\bar{X}_1)}{\sigma_Y} \cdot \frac{\sigma_{X_1}}{\sigma_{X_1}} + \frac{b_{YX_2}(X_2-\bar{X}_2)}{\sigma_Y} \cdot \frac{\sigma_{X_2}}{\sigma_{X_2}} + \frac{b_{YX_3}(X_3-\bar{X}_3)}{\sigma_Y} \cdot \frac{\sigma_{X_3}}{\sigma_{X_3}} + \frac{(\varepsilon-\bar{\varepsilon})}{\sigma_Y} \cdot \frac{\sigma_\varepsilon}{\sigma_\varepsilon}$

Dessa expressão obtém-se:

$$y = p_{yx_1}x_1 + p_{yx_2}x_2 + p_{yx_3}x_3 + p_{\varepsilon}u \tag{2.6}$$

em que:

$y = \frac{Y-\bar{Y}}{\sigma_Y}$ é a variável básica padronizada;

$x_i = \frac{X_i-\bar{X}_i}{\sigma_{X_i}}$ é a variável explicativa padronizada;

$u = \frac{(\varepsilon-\bar{\varepsilon})}{\sigma_\varepsilon}$;

$p_\varepsilon = \frac{\sigma_\varepsilon}{\sigma_Y}$ é o coeficiente da variável residual na análise de trilha; e

$p_{yx_i} = \frac{b_{YX_i}\sigma_{X_i}}{\sigma_Y}$ é o coeficiente da variável explicativa na análise de trilha.

Da expressão 2.6 tem-se que:

$$V(y) = V(p_{yx_1}x_1)+V(p_{yx_2}x_2)+V(p_{yx_3}x_3)+V(p_\varepsilon u)+2Cov(p_{yx_1}x_1, p_{yx_2}x_2)$$
$$+ 2Cov(p_{yx_1}x_1, p_{yx_3}x_3) + 2Cov(p_{yx_1}x_1, p_\varepsilon u) + 2Cov(p_{yx_2}x_2, p_{yx_3}x_3)$$
$$+ 2Cov(p_{yx_2}x_2, p_\varepsilon u) + 2Cov(p_{yx_3}x_3, p_\varepsilon u)$$

Essa expressão pode ser escrita de maneira mais simplificada da seguinte forma:

$$V(y) = p^2_{yx_1}V(x_1)+p^2_{yx_2}V(x_2)+p^2_{yx_3}V(x_3)+p^2_\varepsilon V(u)+2p_{yx_1}p_{yx_2}Cov(x_1,x_2)+$$
$$2p_{yx_1}p_{yx_3}Cov(x_1,x_3) + 2p_{yx_1}p_\varepsilon Cov(x_1,u) + 2p_{yx_2}p_{yx_3}Cov(x_2,x_3)$$
$$+ 2p_{yx_2}p_\varepsilon Cov(x_2,u) + 2p_{yx_3}p_\varepsilon Cov(x_3,u)$$

Como as variáveis foram padronizadas tem-se que:

1. $V(x_i) = V\left(\frac{X_i-\bar{X}_i}{\sigma_{X_i}}\right)$
   Dessa forma: $V(x_i) = \frac{V(X_i)}{\sigma^2_{X_i}} = 1$. De maneira análoga, $V(y) = 1$ e $V(u) = 1$.

2. $Cov(y, x_i) = Cov\left(\frac{Y-\bar{Y}}{\sigma_Y}, \frac{X_i-\bar{X}_i}{\sigma_{X_i}}\right)$

Então: $Cov(y, x_i) = \frac{1}{\sigma_Y \sigma_{X_i}}[Cov(Y, X_i) - Cov(Y, \bar{X}_i) - Cov(\bar{Y}, X_i) + Cov(\bar{Y}, \bar{X}_i)] = \frac{1}{\sigma_Y \sigma_{X_i}} Cov(Y, X_i) = r_{YX_i}$.

3. $Cov(x_i, x_j) = Cov\left(\frac{X_i - \bar{X}_i}{\sigma_{X_i}}, \frac{X_j - \bar{X}_j}{\sigma_{X_j}}\right)$
   Então: $Cov(x_i, x_j) = \frac{1}{\sigma_{X_i}\sigma_{X_j}}[Cov(X_i, X_j) - Cov(X_i, \bar{X}_j) - Cov(\bar{X}_i, X_j) + Cov(\bar{X}_i, \bar{X}_j)] = \frac{1}{\sigma_{X_i}\sigma_{X_j}} Cov(X_i, X_j) = r_{X_i X_j}$.

4. $Cov(u, x_i) = Cov\left(\frac{\varepsilon - \bar{\varepsilon}}{\sigma_\varepsilon}, \frac{X_i - \bar{X}_i}{\sigma_{X_i}}\right)$
   Então: $Cov(u, x_i) = \frac{1}{\sigma_\varepsilon \sigma_{X_i}}[Cov(\varepsilon, X_i) - Cov(\varepsilon, \bar{X}_i) - Cov(\bar{\varepsilon}, X_i) + Cov(\bar{\varepsilon}, \bar{X}_i)] = 0$.

Portanto, é possível observar as seguintes relações:

$$V(y) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23} + p_\varepsilon^2.$$

$$V(y) = V(\hat{y}) + p_\varepsilon^2.$$

$$V(\hat{y}) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23}.$$

Essas relações permitem estimar a correlação (r_{yx_1}) da seguinte forma:

$$Cov(y, x_1) = r_{yx_1} = Cov(p_{yx_1}x_1 + p_{yx_2}x_2 + p_{yx_3}x_3 + p_\varepsilon u, x_1) = p_{yx_1}Cov(x_1, x_1) + p_{yx_2}Cov(x_1, x_2) + p_{yx_3}Cov(x_1, x_3) + p_\varepsilon Cov(u, x_1)$$

$$\Rightarrow r_{yx_1} = p_{yx_1} + p_{yx_2}r_{12} + p_{yx_3}r_{13}$$

De maneira análoga tem-se as outras correlações, obtendo portanto:

$$Cov(y, x_1) = r_{yx_1} = p_{yx_1} + p_{yx_2}r_{12} + p_{yx_3}r_{13}$$

$$Cov(y, x_2) = r_{yx_2} = p_{yx_1}r_{12} + p_{yx_2} + p_{yx_3}r_{23}$$

$$Cov(y, x_3) = r_{yx_3} = p_{yx_1}r_{13} + p_{yx_2}r_{23} + p_{yx_3}$$

Através da expressão:

$$V(\hat{y}) = p_{yx_1}^2 + p_{yx_2}^2 + p_{yx_3}^2 + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23}$$

estima-se o coeficiente de determinação do modelo causal ($R^2_{0.123}$), que mede os efeitos das variáveis explicativas (X_1, X_2, X_3) sobre a variável principal (Y).

O coeficiente ($R^2_{0.123}$) é dado por:

$$R^2_{0.123} = \frac{SQRegressão}{SQTotal}$$

Sabendo que:
$SQRegressão = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = V(\hat{y})$ e $SQTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2 = V(y)$.
O coeficiente de determinação pode ser representado por:

$$R^2_{0.123} = \frac{V(\hat{y})}{V(y)}.$$

Mas, como $V(y) = 1$, então:

$R^2_{0.123} = V(\hat{y}) = p^2_{yx_1} + p^2_{yx_2} + p^2_{yx_3} + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23}$

Pode-se estimar também o efeito da variável residual p_ε sobre a variável principal. Temos que:

$V(y) = p^2_{yx_1} + p^2_{yx_2} + p^2_{yx_3} + 2p_{yx_1}p_{yx_2}r_{12} + 2p_{yx_1}p_{yx_3}r_{13} + 2p_{yx_2}p_{yx_3}r_{23} + p^2_\varepsilon$

Então:

$V(y) = R^2_{0.123} + p^2_\varepsilon$

Mas, sabendo que $V(y) = 1$, tem-se que:

$1 = R^2_{0.123} + p^2_\varepsilon \Rightarrow p^2_\varepsilon = 1 - R^2_{0.123} \Rightarrow$

$$p_\varepsilon = \sqrt{1 - R^2_{0.123}}$$

Os valores p_{yx_i} $(i = 1, 2, 3)$ que aparecem no modelo 2.6 são os coeficientes de trilha. Estes valores permitem que, na decomposição das correlações, por exemplo $Cov(y, x_1) = r_{yx_1} = p_{yx_1} + p_{yx_2}r_{12} + p_{yx_3}r_{13}$, sejam calculados os efeitos direto da variável x_1 sobre y, expresso por p_{yx_1}, e efeitos indiretos de

x_1 sobre y, via as outras variáveis explicativas x_2 e x_3 correlacionadas com x_1 , expressos respectivamente por $p_{yx_2}r_{12}$ e $p_{yx_3}r_{13}$.

As estimativas desses coeficientes ($\hat{p}_{yx_i}$) são obtidas pela resolução do sistema linear de equações normais $\dot{Y} = \dot{X}\hat{P}$. Neste sistema $\dot{X}$ é uma matriz não singular das correlações entre as variáveis explicativas; $\hat{P} = (\dot{X})^{-1}\dot{Y}$ um vetor coluna contendo as estimativas dos coeficientes de trilha; e $\dot{Y}$ é um vetor coluna das correlações entre a variável principal e cada variável explicativa do modelo. Tem-se então:

$$\dot{Y} = \begin{bmatrix} r_{YX_1} \\ r_{YX_2} \\ r_{YX_3} \end{bmatrix}, \ \dot{X} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \text{ e } \hat{P} = \begin{bmatrix} p_{yx_1} \\ p_{yx_2} \\ p_{yx_3} \end{bmatrix}$$

E o sistema linear de equações normais $\dot{Y} = \dot{X}\hat{P}$ fica da seguinte forma:

$$\begin{bmatrix} r_{YX_1} \\ r_{YX_2} \\ r_{YX_3} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} \cdot \begin{bmatrix} p_{yx_1} \\ p_{yx_2} \\ p_{yx_3} \end{bmatrix}$$

Estes resultados podem ser estendidos para k variáveis explicativas ($X_1, X_2, X_3, ..., X_k$), o que resultaria no seguinte sistema:

$$\begin{bmatrix} r_{YX_1} \\ r_{YX_2} \\ r_{YX_3} \\ \vdots \\ r_{YX_k} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ r_{31} & r_{32} & 1 & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \cdots & 1 \end{bmatrix} \cdot \begin{bmatrix} p_{yx_1} \\ p_{yx_2} \\ p_{yx_3} \\ \vdots \\ p_{yx_k} \end{bmatrix}$$

E também para a estimação dos coeficientes $\hat{p}_{yx_i}$ pode-se recorrer a algum

dos métodos descritos na seção 2.2.1.1, considerando na estimação as matrizes derivadas do modelo $y = p_{yx_1}x_1 + p_{yx_2}x_2 + ... + p_{yx_k}x_k + p_{\varepsilon}u$.

Como critério para a interpretação da análise de trilha, segundo Singh e Chaudhary (1979 apud GOMES, 1996), quando o coeficiente de trilha (efeito direto) de uma variável explicativa for, em módulo, menor que o efeito variável residual, mas o coeficiente de correlação (efeito total) for maior que o efeito da variável residual, ou seja,

$$|p_{yx_i}| < p_{\varepsilon} < r_{yx_i},$$

significa que essa variável explicativa influencia a variável principal apenas indiretamente, sendo sua importância só em conjunto. Se o coeficiente de trilha for, em módulo, maior que o coeficiente da variável residual, isto é,

$$|p_{yx_i}| > p_{\varepsilon},$$

indica que existe efeito direto da variável explicativa sobre a principal.

2.3.2.1 Análise de trilha em cadeia (mais de um modelo causal)

Quando, na análise de trilha, existem uma ou mais variáveis que são ao mesmo tempo variáveis básicas e explicativas significa que existem mais de um modelo causal, desse modo tem-se uma análise de trilha em cadeia (KLINE, 1991).

Considerando que existem numa análise uma variável básica (Y), duas variáveis primárias (X_1 e X_2), que são básicas e explicativas e duas variáveis secundárias (Z_3 e Z_4), tem-se o seguinte diagrama de trilha:

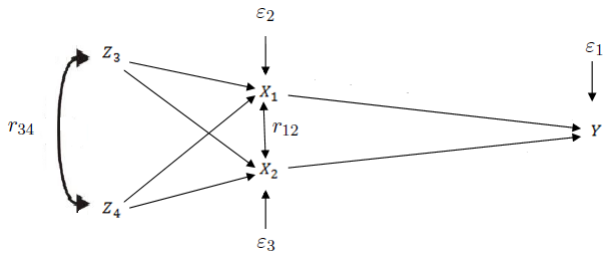

Figura 2 Diagrama em cadeia ilustrativo dos efeitos das variáveis explicativas primárias e secundárias sobre a variável básica

A análise de trilha nessa situação deve ser realizada em partes, analisando cada diagrama causal separadamente da seguinte maneira:

a) Para a influência de X_1 e X_2 sobre Y, tem-se:

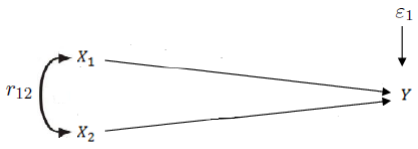

Figura 3 Diagrama ilustrativo do primeiro modelo causal

Modelo:

$$y = p_{yx_1}x_1 + p_{yx_2}x_2 + p_{\varepsilon_1}u_1$$

Correlações:

$$\text{cov}(y, x_1) = r_{yx_1} = p_{yx_1} + p_{yx_2}r_{12}$$

$$\text{cov}(y, x_2) = r_{yx_2} = p_{yx_1}r_{12} + p_{yx_2}$$

As estimativas dos coeficientes de trilha são obtidas através do seguinte

sistema:

$$\begin{bmatrix} r_{YX_1} \\ r_{YX_2} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} . \begin{bmatrix} \hat{p}_{yx_1} \\ \hat{p}_{yx_2} \end{bmatrix}$$

b) Para a influência de Z_3 e Z_4 sobre X_1, tem-se:

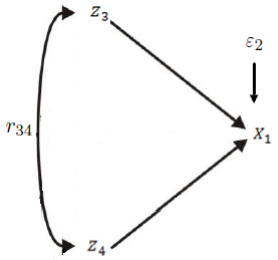

Figura 4 Diagrama ilustrativo do segundo modelo causal

Modelo:

$$x_1 = p_{x_1 z_3} z_3 + p_{x_1 z_4} z_4 + p_{\varepsilon_2} u_2$$

Correlações:

$$\text{cov}(x_1, z_3) = r_{13} = p_{x_1 z_3} + p_{x_1 z_4} r_{34}$$

$$\text{cov}(x_1, z_4) = r_{14} = p_{x_1 z_3} r_{34} + p_{x_1 z_4}$$

As estimativas dos coeficientes de trilha são obtidas através do seguinte sistema:

$$\begin{bmatrix} r_{X_1 Z_3} \\ r_{X_1 Z_4} \end{bmatrix} = \begin{bmatrix} 1 & r_{34} \\ r_{34} & 1 \end{bmatrix} . \begin{bmatrix} \hat{p}_{x_1 z_3} \\ \hat{p}_{x_1 z_4} \end{bmatrix}$$

c) Para a influência de Z_3 e Z_4 sobre X_2, tem-se:

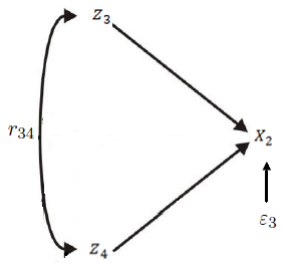

Figura 5 Diagrama ilustrativo do terceiro modelo causal

Modelo:

$$x_2 = p_{x_2 z_3} z_3 + p_{x_2 z_4} z_4 + p_{\varepsilon_3} u_3$$

Correlações:

$$\text{cov}(x_2, z_3) = r_{23} = p_{x_2 z_3} + p_{x_2 z_4} r_{34}$$

$$\text{cov}(x_2, z_4) = r_{24} = p_{x_2 z_3} r_{34} + p_{x_2 z_4}$$

As estimativas dos coeficientes de trilha são obtidas através do seguinte sistema:

$$\begin{bmatrix} r_{X_2 Z_3} \\ r_{X_2 Z_4} \end{bmatrix} = \begin{bmatrix} 1 & r_{34} \\ r_{34} & 1 \end{bmatrix} . \begin{bmatrix} \hat{p}_{x_2 z_3} \\ \hat{p}_{x_2 z_4} \end{bmatrix}$$

d) Para a influência de Z_3 e Z_4 sobre Y, tem-se:

Modelo:

$$y = p_{y z_3} z_3 + p_{y z_4} z_4$$

Correlações:

- $\text{cov}(y, z_3) = r_{y z_3} = \text{cov}(p_{y x_1} x_1 + p_{y x_2} x_2 + p_{\varepsilon_1} u_1, z_3) = p_{y x_1} r_{13} + p_{y x_2} r_{23} = p_{y x_1}(p_{x_1 z_3} + p_{x_1 z_4} r_{34}) + p_{y x_2}(p_{x_2 z_3} + p_{x_2 z_4} r_{34}) =$
  $= p_{y x_1} p_{x_1 z_3} + p_{y x_1} p_{x_1 z_4} r_{34} + p_{y x_2} p_{x_2 z_3} + p_{y x_2} p_{x_2 z_4} r_{34}$

- $\text{cov}(y, z_4) = r_{yz_4} = \text{cov}(p_{yx_1}x_1 + p_{yx_2}x_2 + p_{\varepsilon_1}u_1, z_4) = p_{yx_1}r_{14} + p_{yx_2}r_{24} = p_{yx_1}(p_{x_1z_3}r_{34} + p_{x_1z_4}) + p_{yx_2}(p_{x_2z_3}r_{34} + p_{x_2z_4}) =$
  $= p_{yx_1}p_{x_1z_4} + p_{yx_1}p_{x_1z_3}r_{34} + p_{yx_2}p_{x_2z_4} + p_{yx_2}p_{x_2z_3}r_{34}$

Uma outra abordagem da análise de trilha usando a metodologia de regressão aleatória, que teoriza desde a padronização das variáveis até o processo de estimação e desdobramento das correlações, pode ser entendida e encontrada em Rencher e Schaalje (2008).

2.3.3 Multicolinearidade

O termo multicolinearidade foi criado por Ragnar Frisch em 1934, e significava, originalmente a existência de uma relação linear exata entre duas ou mais variáveis, ou seja, um dos vetores sendo uma combinação linear dos outros. Segundo Neter et al. (2005), a multicolinearidade ocorre quando existe algum nível de inter-relação entre as variáveis independentes do modelo de regressão linear múltipla. Como a correlação exata raramente ocorre, o termo multicolinearidade é utilizado com frequência nos casos em que a correlação entre as variáveis é muito alta. Na análise de trilha o interesse está na multicolinearidade das variáveis independentes. Conforme foi mencionado anteriormente, a colinearidade entre as variáveis independentes é extremamente nociva no ajuste de modelos de regressão múltipla. Como a análise de trilha está calcada no modelos de regressão linear múltipla, a presença de multicolinearidade nas variáveis explicativas pode comprometer os resultados da análise de trilha.

2.3.4 Diagnóstico de Multicolinearidade

Existem várias propostas para diagnosticar a presença de multicolinearidade, sendo característica desejável de um procedimento de diagnóstico aquelas que, além de refletirem diretamente o grau do problema de multicolinearidade, forneçam informações úteis na determinação de quais variáveis estão envolvidas (MONTGOMERY; PECK, 1992).

2.3.4.1 Análise da matriz de correlação

Este procedimento consiste na análise dos elementos não diagonais, (r_{ij}, tal que $i \neq j$), da matriz de correlação $\dot{X}$.

Se as variáveis independentes X_i e X_j apresentam dependência linear aproximada, então a correlação linear entre elas, em valor absoluto ($|r_{ij}|$), será aproximadamente igual a 1. Um alto coeficiente de correlação indica multicolinearidade, mas a ausência de alta correlação entre duas variáveis não implica ausência de multicolinearidade (KMENTA, 1971). Existe a possibilidade de que três ou mais variáveis independentes apresentem uma relação de multicolinearidade mesmo que qualquer par dessas variáveis não apresente um coeficiente de correlação alto, ou seja, a condição de alta correlação para existir multicolinearidade é somente suficiente, mas não necessária quando o número de variáveis independentes é maior do que dois.

2.3.4.2 Teste do determinante da matriz de correlação

Como $\dot{X}$ é uma matriz de correlação, o seu determinante pode ser usado como avaliador de multicolinearidade. Como a matriz é na forma de correlação, o

seu determinante varia de zero a um, ou seja:

$$0 \leq det(\dot{X}) \leq 1.$$

Na avaliação da multicolinearidade, quanto mais o valor do determinante se aproxima de zero, $det(\dot{X}) \rightarrow 0$, mais intensa é a multicolinearidade. Conforme Montgomery e Peck (1992), este método é útil e de fácil execução na avaliação da multicolineridade, porém ele não fornece informações sobre a origem dessa multicolinearidade por não permitir a identificação das variáveis causadoras.

2.3.4.3 Análise dos autovalores e autovetores da matriz de correlação

Segundo Belsey et al. (1980) e Silvey (1969) apud Carvalho e Cruz (1996), as raízes características ou autovalores da matriz de correlação $\dot{X}$, denotados por $\lambda_1, \lambda_2, ..., \lambda_p$ podem ser usados no diagnóstico de multicolinearidade. Pois, quando uma ou mais dependências lineares aproximadas, um ou mais autovalores serão pequenos. Baseado nisto, Montgomery e Peck (1992) propuseram o método da análise dos autovalores associados à matriz de correlação $\dot{X}$, onde a multicolinearidade é diagnosticada pelo número de condições (NC) matriz de correlação $\dot{X}$, que é a relação entre o maior e o menor autovalor da matriz de correlação, ou seja:

$$NC = \frac{\lambda_{max}}{\lambda_{min}}.$$

Quando $NC < 100$, a multicolinearidade é considerada fraca e não constitui problema para a análise; se $100 < NC < 1000$, é considerada de moderada a forte; e se $NC > 1000$, a multicolinearidade é considerada severa.

A análise dos autovalores também pode ser utilizada para identificar a natureza da dependência linear, aproximada, existente entre as variáveis. A matriz

$\dot{X}$ pode ser decomposta como:

$$\dot{X} = T\Lambda T'.$$

em que:

Λ: matriz diagonal $p \times p$, cujos elementos da diagonal são os autovalores λ_j ($j = 1, 2, ..., p$) da matriz $\dot{X}$. Assim:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

T: matriz ortogonal $p \times p$, cujas colunas $(t_1, t_2, ..., t_p)$ são os autovetores normalizados de $\dot{X}$. Ou seja:

$$T = \begin{bmatrix} t_1 & t_2 & \cdots & t_p \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ t_{p1} & t_{p2} & \cdots & t_{pp} \end{bmatrix}$$

Sabendo que $T'T = I$ e que $\dot{X} = T\Lambda T'$, então $T'(\dot{X})T = \Lambda$, ou ainda $t'_j(\dot{X})t_j = \lambda_j$, e $t'_j(\dot{X})t_k = 0$ se $j \neq k$.

Se um autovalor λ_j é próximo de zero, indicando uma dependência linear entre as observações, os elementos do autovetor associado a esse autovalor descrevem a natureza desse dependência linear.

2.3.4.4 Fatores de inflação da variância

O fator de inflação da variância representa o quanto da variância do coeficiente está inflacionada em comparação ao que seria se a variável não estivesse correlacionada com qualquer outra do modelo.

Segundo Marquard (1970), os elementos da diagonal principal de $C = (X'X)^{-1}$ são os fatores de inflação da variância (VIF's) quando a matriz $X'X$ é na forma de correlação. Estes fatores são úteis para detectar a multicolinearidade. Os elementos diagonais da matriz C podem ser escritos como $C_{jj} = (1 - R_j^2)^{-1}$ ($j = 1, 2, ..., p$), onde R_j^2 é o coeficiente de determinação múltipla da regressão de X_j sobre as outras variáveis explicativas.

Como a variância do $j - simo$ coeficiente de regressão ($\hat{\beta}_j$) de mínimos quadrados ($\hat{\beta}_j$) é $V(\hat{\beta}_j) = (1-R_j^2)^{-1}\sigma^2 = C_{jj}\sigma^2$, pode-se, assim, considerar C_{jj} o fator que aumenta a variância de $\hat{\beta}_j$ quando existe dependência linear entre as variáveis. Na avaliação da multicolinearidade, de acordo com Neter et al. (2005), se qualquer valor VIF (C_{jj}) for maior do que 10 (VIF$>$ 10), há indicativo de que a multicolinearidade pode estar influenciando indevidamente as estimativas de mínimos quadrados.

Apesar do VIF ser o mais utilizado para diagnóstico de multicolinearidade, existem limitações do uso dessa ferramenta isoladamente, devido a inabilidade em distinguir entre as quase-dependências coexistentes, aliando-se ao fato de não haver um limite bem definido para distinguir entre os valores de VIF críticos. Por isso, se fazem necessárias comparações e análises conjuntas com o exame do número de condição e da matriz de autovalores.

2.3.4.5 Teste de Farrar e Glauber

Farrar e Glauber (1967) propuseram um processo para determinar a multicolinearidade, onde a hipótese de nulidade ($H_0 : \nexists\ multicolinearidade$) de que não existe multicolinearidade entre as variáveis independentes do modelo de regressão linear múltipla é testada usando-se a seguinte estatística:

$$\chi^2_* = -\left[(n-1) - \frac{1}{6}(2k+5)\right] \cdot \ln\left|\dot{X}\right|, \tag{2.7}$$

em que: n é o número de observações, k é o número de variáveis independentes e $\dot{X}$ é a matriz de correlação das variáveis independentes.

A estatística 2.7 tem distribuição aproximadamente qui-quadrado com $k(k-1)/2$ gruas de liberdade, ou seja, $\chi^2_* \sim \chi^2_{k(k-1)/2}$. Dessa forma, a hipótese H_0 é aceita se $\chi^2_* < \chi^2_{\frac{k(k-1)}{2};\alpha}$.

2.3.5 Métodos alternativos de estimação quando existe multicolinearidade

Segundo Montgomery e Peck (1992), várias técnicas tem sido propostas para solucionar os problemas acarretados pela multicolinearidade, tais como, a obtenção de dados adicionais, a reespecificação do modelo e o uso de outros métodos de estimação de mínimos quadrados que são especificamente planejados para combater os problemas advindos da multicolinearidade.

Esses mesmos autores observam, porém, que o uso desses procedimentos nem sempre é possível e viável. Quando a multicolinearidade é devido a restrições sobre o modelo ou sobre a população, a coleta de dados adicionais é uma solução pouco recomendável. Em relação a eliminação de variáveis, apesar de ser uma técnica geralmente efetiva, poderá não estabelecer uma solução satisfatória se as variáveis retiradas do modelo tiverem uma grande poder explicativo em relação à resposta, prejudicando o poder de predição do modelo.

Os métodos alternativos ao de mínimos quadrados, regressão em crista e regressão em componentes principais, são especificamente planejados para combater os problemas da multicolinearidade (MONTGOMERY; PECK, 1992).

2.3.5.1 Regressão em crista

O estimador nesse procedimento é obtido aumentando-se os elementos da diagonal principal da matriz $X'X$ para uma escala dos valores de uma constante c escolhida arbitrariamente. Denominado estimador em cristas ($\hat{\beta}^*$), ele é definido como a solução para:

$$(X'X + Ic)\hat{\beta}^* = X'Y$$

ou

$$\hat{\beta}^* = (X'X + Ic)^{-1}X'Y,$$

onde $0 \leq c \leq 1$, uma vez que $X'X$ se encontra na forma de correlações.

Esse é um estimador tendencioso, pois

$$E\left[\hat{\beta}^*\right] = \left(X'X + Ic\right)^{-1}X'X\beta \neq \beta,$$

porém conforme Gunst e Mason (1977), na presença de multicolinearidade eles apresentam melhor desempenho do que os estimadores de mínimos quadrados superando o problema da inflação da variância e da instabilidade das estimativas dos coeficientes de regressão.

Hoerl e Kennard (1970a) provaram que sempre é possível encontrar um valor positivo da constante c para o qual os estimadores dos coeficientes tornam-se estáveis, não variam, de modo que o quadrado médio do erro usando o estimador

em cristas, embora seja ele tendencioso, seja menor do que o quadrado médio do erro do estimador de mínimos quadrados.

Para Montgomery e Peck (1992), se a multicolinearidade é severa, será evidente a instabilidade nos coeficientes de regressão pelo traço de crista. A medida que o valor de c aumenta, algumas estimativas em crista irão variar bastante, e para algum valor de c, a estimativa em crista $\hat{\beta}^*$ será estável. Então, o objetivo é selecionar um valor razoavelmente pequeno de c, cujas estimativas em crista são estáveis, o que certamente produzirá um conjunto de estimativas com um quadrado médio do erro $QME(\hat{\beta}^*)$ menor que as estimativas de mínimos quadrados. Os mesmos autores salientam que a escolha correta da constante c tem sido objeto de muitas discussões acerca do emprego da regressão em crista, com diferentes procedimentos propostos por vários autores.

Hoerl e Kennard (1970b) sugerem que um valor apropriado de c pode ser determinado pela inspeção do traço de crista. O traço de crista é um diagrama dos elemento de $\hat{\beta}^*$ por c, para os valores de c normalmente no intervalo $[0, 1]$. Especificamente, é um gráfico bidimensional do valor de cada coeficiente versus c, mostrando como os valores de $\hat{\beta}^*$ variam em função dos valores de c. Por meio desse gráfico pode-se analisar os efeitos da multicolinearidade sobre as estimativas dos parâmetros, mas a principal finalidade de sua construção é a escolha do valor da constante com o qual se obtém a regressão estimada. Dessa forma, a inspeção desse gráfico permite escolher um valor de c que estabilize as estimativas dos parâmetros produzindo um quadrado médio do erro menor que as estimativas de mínimos quadrados.

2.3.5.2 Regressão em componentes principais

Os componentes principais podem ser obtidos, segundo Montgomey e

Peck (1992), considerando o modelo na forma canônica

$$y = Z\alpha + \varepsilon,$$

em que: $Z = XT$, $\alpha = T'\beta$, $T'X'XT = Z'Z = \Lambda$ e $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ é uma matriz diagonal $(p \times p)$ dos autovalores de $X'X$, e T é uma matriz ortogonal $(p \times p)$, cujas colunas são os autovetores associados a $\lambda_1, \lambda_2, ..., \lambda_p$. As colunas da matriz $Z = [Z_1, Z_2, ..., Z_p]$, que definem um novo conjunto de variáveis ortogonais, são denominadas componentes principais.

O estimador de mínimos quadrados de α é:

$$\hat{\alpha} = (Z^{'}Z)^{-1} Z^{'}y = \Lambda^{-1}Z^{'}y.$$

E a matriz de covariância de $\hat{\alpha}$ é:

$$V(\hat{\alpha}) = \sigma^2 (Z^{'}Z)^{-1} = \sigma^2 \Lambda^{-1}.$$

Assim, um autovalor pequeno de $X'X$ significa que variância do coeficiente de regressão ortogoal correspondente será grande. Como:

$$Z^{'}Z = \sum_{i=1}^{p} \sum_{j=1}^{p} Z_i Z_j^{'} = \Lambda,$$

frequentemente, o autovalor λ_j é referido como a variância *j-ésimo* componente principal. E a matriz de covariância dos coeficientes de regressão padronizados $\hat{\beta}$ é:

$$V(\hat{\beta}) = V(T\hat{\alpha}) = T\Lambda^{-1}T'\alpha^2.$$

Para a obtenção do estimador dos componentes principais, as variáveis

independentes são consideradas em ordem decrescente de seus autovalores, isto é, $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. Os últimos "s" desses autovalores, com $s < p$, são considerados como sendo aproximadamente iguais a zero. As colunas da matriz correspondente a esses autovalores próximos de zero são excluídas da análise, e a regressão em componentes principais é, então, obtida pela aplicação do método dos mínimos quadrados aos componentes restantes. Isto é:

$$\hat{\alpha}_{CP} = T\hat{\alpha},$$

onde: $t_1 = t_2 = ... = t_{p-s} = 1$ e $t_{p-s+1} = t_{p-s+2} = ... = t_p = 0$. Então, o estimador em componentes principais é:

$$\hat{\alpha}_{CP} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_{p-s} \\ --- \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

ou em termos de variáveis padronizadas:

$$\hat{\beta}_{CP} = T\hat{\alpha}_{CP} = \sum_{j=1}^{p-s} \lambda_j^{-1} t_j^{-1} X' y t_j.$$

Segundo Montgomery e Peck (1992), um estudo de simulação feito por Gunst e Mason (1977) mostrou que a regressão em componentes principais ofer-

ece considerável melhoria sobre os mínimos quadrados quando os dados são mal-condicionados.

3 METODOLOGIA

Divide-se este capítulo em dois assuntos: descrição dos dados experimentais e descrição das análises estatísticas.

3.1 Dados experimentais

Os dados são provenientes de dois diferentes experimentos. No primeiro experimento utilizou-se a espécie de maracujazeiro *Passiflora giberti* N.E. Brown, e no segundo usou-se a planta de milho.

3.1.1 Experimento 1 - Maracujá (*Passiflora giberti* N.E. Brown)

Esse experimento foi conduzido no Laboratório de Biotecnologia Vegetal da Embrapa Mandioca e Fruticultura, em Cruz das Almas, Bahia, utilizando como material vegetal segmentos nodais de plantas de *Passiflora giberti* N. E. Brown, do Banco Ativo de Germoplasma da Embrapa Mandioca e Fruticultura, cultivadas *in vitro*.

Esses explantes foram dispostos em *magentas* contendo 20 mL do meio de cultura MS, suplementado com 10, 20 e 40 gL^{-1} de sorbitol combinados com 0, 15 e 30 gL^{-1} de sacarose, mais uma testemunha, contendo 30 gL^{-1} de sacarose, gelificado com 2 gL^{-1} de *phytagel*, ajustado a um pH de 5,8 e sem adição de fitoreguladores. O cultivo foi realizado sob condições de fotoperíodo de 16 horas, temperatura de 20 ± 1 °C e densidade de fluxo de fótons 22 $\mu Em^{-2}s^{-1}$.

O delineamento experimental foi o inteiramente casualizado, com 20 repetições, sendo os tratamentos dispostos em esquema fatorial (3 x 3)+1, sendo três concentrações de sacarose, três concentrações de sorbitol, mais uma testemunha, totalizando 10 tratamentos. Cada unidade experimental foi constituída de 1 explante por *magenta*.

Aos 150 dias de cultivo, foi realizada uma análise destrutiva para avaliação das seguintes variáveis: comprimento da plântula (cm), peso seco da plântula, peso com água, número de explantes para micropropagação e número de gemas. O experimento e a coleta de dados foram conduzidos por Faria (2008).

3.1.2 Experimento 2 - Milho

O experimento foi conduzido na fazenda experimental da Universidade Federal de Lavras, em Lavras, a 951 m de altitude, nas coordenadas $44^{o}58'$ longitude Oeste e $21^{o}12'$ latitude Sul.

Foram escolhidas cinco linhagens de milho do programa de melhoramento da UFLA, obtidas pela autofecundação de híbridos existentes no mercado, obtendo-se 15 tratamentos genéticos. O critério da escolha das mesmas foi o tamanho dos grãos, sendo três de grãos grandes e duas de grãos pequenos. Os cruzamentos foram realizados seguindo um esquema de dialelo completo na safra de 2009/2010. O delineamento utilizado foi o de blocos completos ao acaso com quatro repetições. As parcelas eram constituídas de duas linhas de 2,0 m de comprimento, com espaçamento entre linhas de 0,6 m e quatro plantas por metro linear.

No final do período do florescimento masculino, foram realizadas medições de cinco plantas competitivas no interior da parcela para altura de planta (AP), em metros, considerando a distância do solo ao ponto de inserção da folha bandeira, altura de espiga (AE), em metros, do ponto de inserção da espiga superior for-

mada no colmo e o diâmetro do colmo (DC), em centímetros, a partir do primeiro entrenó do colmo acima do solo, utilizando-se um paquímetro digital.

No momento da colheita, foram colhidas individualmente as espigas de cinco plantas na parcela para a obtenção da produtividade de grãos (PROD). De cada planta, após a trilha das espigas, foram obtidos o peso de 100 grãos, peso total de grãos (PT) e o número de grãos por planta, o qual foi estimado por regra de três a partir dos dados do P100 e PROD. O experimento e a coleta de dados foram conduzidos por Ribeiro (2012).

3.2 Análises estatísticas

Todas análises estatísticas foram realizadas utilizando funções desenvolvidas no software R (R DEVELOPMENT CORE TEAM, 2012), e outras funções disponíveis nas bibliotecas *agricolae*, *MASS*, *ppcor* e *car* desse mesmo software.

3.2.1 Análise exploratória

Através de uma análise exploratória baseada na contrução de gráficos *box-plot* para cada variável, foi verificada a existência ou não de *outliers*. A presença de *outliers* é um fator relevante que pode causar a mensuração de correlações pouco confiáveis, impossibilitando resultados consistentes usando a análise de trilha. A verificação da normalidade das variáveis pode ser feita por meio de testes de hipótese consagrados, tais como Kolmogorov-Smirnov, Lilliefors e Shapiro-Wilk. Neste trabalho, através do teste de Shapiro-Wilk verificou-se a hipótese de normalidade para cada variável, que é uma das exigências para se medir a correlação linear simples entre duas variáveis X e Y. Nesse teste, aceita-se a hipótese H_0 (H_0: *A variável é normal*), a um nível α de significância, se o valor-p obtido pelo teste é maior que α ($valor - p > \alpha$). Para esse teste, adotou-se $\alpha = 5\%$. Depois

de verificada essas suposições avançou-se para a construção do diagrama.

3.2.2 Escolha do diagrama causal

Sendo indispensável na análise de trilha, um diagrama é construído para especificar a natureza exata da estrutura proposta. O diagrama é muito útil para exibir graficamente o padrão de hipótese das relações de causa e efeito entre um conjunto de variáveis, ou seja, estabelece uma relação de causa e efeito entre as variáveis. No diagrama, as setas unidirecionais indicam os efeitos diretos de cada variável explicativa (independente) sobre uma variável básica (dependente), enquanto as bidirecionais representam a interdependência das variáveis explicativas (KLINE, 1991). A escolha do diagrama foi feita com a combinação de dois critérios: o conhecimento *a priori* da relações entre as variáveis em estudos ou hipótese considerada de causa e efeito nessas variáveis e, o método da correlação parcial (GOLDSMITH, 1977).

Depois de estabelecidas as relações de causa e efeito entre as variáveis foi feita a construção gráfica do diagrama causal para possibilitar um melhor entendimento das equações que darão origem aos coeficientes de trilha. Após essa pré-definição do diagrama, foi aplicado o método da correlação parcial nas variáveis envolvidas no modelo ou nos modelos causais (caso se escolha uma análise em cadeia) com o objetivo de verificar a consistência dessa formulação de causas e efeitos.

3.2.3 Estimação e desdobramento das correlações

Depois das variáveis serem padronizadas seguindo os passos da seção (2.3.2), foram ajustados modelos de acordo com os diagramas causais estabele-

cidos. Os parâmetros dos modelos foram estimados pelo método dos mínimos quadrados quando não foi constatada a multicolinearidade entre as variáveis independentes. Mas, sendo a análise de trilha uma forma de regressão múltipla, com base em matrizes de correlação, a presença de multicolinearidade entre as variáveis independentes ocasiona problemas nas estimativas dos coeficientes de trilha, impossibilitando a utilização dos estimadores de mínimos quadrados. Nessa situação, diversos autores, como Carvalho et al. (1999), Espósito (2010), Oliveira et al. (2010) e Rios et al. (2012), utilizaram a regressão em crista para a estimação dos coeficientes de trilha e obtiveram resultados satisfatórios, conseguindo contornar os efeitos da multicolinearidade. Dessa forma, na existência de multicolinearidade entre as variáveis independentes, optou-se pela regressão em crista para eliminar os efeitos acarretados pela multicolinearidade, e os parâmetros foram estimados. Portanto, antes de se realizar a estimação foi feito um diagnóstico de multicolinearidade, para decidir qual a estratégia que será usada na estimação desses coeficientes. Para diagnosticar a multicolinearidade foi feita a análise dos autovalores e autovetores da matriz de correlação (MONTGOMERY; PECK, 1992).

Para que os resultados obtidos pelos estimadores dos parâmetros possam ser utilizados para se fazer algum tipo de inferência sobre o desdobramento das correlações é necessário que alguns pressupostos sobre os resíduos do modelo sejam atendidos. A análise de resíduos é muito importante para verificar a adequabilidade do modelo. Caso algum dos pressupostos não seja atendido, o modelo não é adequado e esta quebra de suposição deve ser corrigida ou incorporada ao modelo. Essas pressuposições iniciais sobre os resíduos (normalidade, independência, homocedasticidade) foram testadas e avaliadas por testes de hipóteses. E a detecção de existência de pontos influentes foi feita por análise gráfica através da distância de *cook*. Todos esses testes estão implementados no software R.

Para testar a normalidade residual foi aplicado o teste de Shapiro-Wilk. Onde aceita-se a hipótese H_0 (H_0:*Existe normalidade nos resíduos*) a um nível α de significância se o valor-p obtido pelo teste é maior que α ($valor - p > \alpha$). Utilizou-se os níveis de significância $\alpha = 10\%$ e $\alpha = 5\%$. Mais informações e detalhes podem ser encontrados em Shapiro e Wilk (1965).

A independência ou autocorrelação dos resíduos foi avaliada pelo teste de Durbin Watson, onde testa-se a hipótese H_0 (H_0:*Os resíduos são independentes*). A hipótese H_0 é aceita a um nível α de significância se o valor-p obtido pelo teste é maior que α ($valor - p > \alpha$). Considerou-se os níveis de significância $\alpha = 10\%$ e $\alpha = 5\%$ (MONTGOMEY; PECK, 1992).

A homocedasticidade foi testada usando-se o teste de Breusch-Pagan, que tem como hipótese nula a homocedasticidade dos resíduos. Este teste segue uma distribuição de qui-quadrado e o valor calculado é comparado com a tabela desta distribuição, considerando 1 grau de liberdade. Mais detalhes sobre este teste podem ser obtidos em Breusch e Pagan (1979). Nesse teste, aceita-se a hipótese H_0 (H_0:*As variâncias são homogêneas*) a um nível de α de significância se o valor-p obtido pelo teste é maior que α ($valor - p > \alpha$). Utilizou-se os níveis de significância $\alpha = 10\%$ e $\alpha = 5\%$.

Depois de testado cada modelo de regressão, se a análise constituir uma análise de trilha em cadeia, os desdobramentos das correlações foram realizados. A correlação simples é aquela mensurada diretamente entre dois caracteres, os quais são obtidos a partir da avaliação de uma determinada quantidade de indivíduos de uma população. Assim, neste estudo, estimou-se e desdobrou-se a correlação simples. As estimativas foram obtidas usando o estimador da correlação linear de Pearson e, considerou-se como valores nas variáveis a média de cada tratamento dos experimentos. Por fim, foram feitas todas as inferências possíveis

analisando os efeitos diretos e indiretos das variáveis explicativas sobre a variável básica.

4 RESULTADOS E DISCUSSÃO

Nesta seção estão os resultados obtidos pela aplicação da análise de trilha nos dados dos experimentos em estudo.

Experimento I - Maracujá (*Passiflora giberti* N.E. Brown)

Através da análise do gráfico *boxplot* das variáveis em estudo, verificou-se a não existência de "pontos aberrantes", ou seja, *outliers* entre as observações dessas variáveis, e todas as variáveis atenderam a suposição de normalidade pelo teste de Shapiro-Wilk ($valor - p > 0,05$). Atendida a suposição de normalidade e verificado a não existência de *outliers*, aplicou-se a análise de trilha.

A análise foi realizada considerando-se como hipótese um único diagrama causal, Figura 6. Esse diagrama teve como objetivo desdobrar as correlações em efeitos direto e indiretos de variáveis tomadas como explicativas sobre a variável principal (básica). Considerou-se o comprimento da plântula (CPL) como a variável principal e como variáveis explicativas o peso seco da plântula (PSPL), número de explantes para micropropagação (EXPL), número de gemas (NG) e peso da plântula com água (PA).

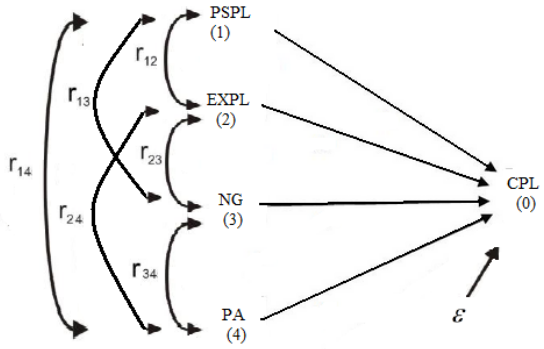

Figura 6 Diagrama causal, onde tem-se o comprimento da plântula (CPL) como variável básica e como variáveis explicativas tem-se o peso seco da plântula (PSPL), número de explantes para micropropagação (EXPL), número de gemas (NG) e peso da plântula com água (PA)

Na Tabela 1 estão os valores das correlações entre as variáveis. Pode-se notar que a variável que apresentou a correlação mais forte e positiva (0,841) com a variável básica CPL foi a variável NG. Ainda, existem correlações fortes e significativas entre algumas variáveis explicativa, como entre NG e EXPL (0,919), entre PA e PSPL (0,836).

Tabela 1 Correlações simples entre as cinco variáveis relativas à planta de maracujá.

	PSPL	EXPL	NG	PA	CPL
PSPL	1				
EXPL	$0,711^{*}$	1			
NG	$0,751^{*}$	$0,919^{**}$	1		
PA	$0,836^{**}$	$0,816^{**}$	$0,775^{*}$	1	
CPL	$0,663^{*}$	$0,807^{*}$	$0,841^{**}$	$0,729^{*}$	1

*$(valor - p < 0,05)$ e **$(valor - p < 0,01)$.

Utilizando o método da correlação parcial para testar e comprovar a im-

portância que cada variável tem no relacionamento escolhido, Figura 6, foram obtidos os seguintes resultados, apresentados na Tabela 2:

Tabela 2 Resultado do método da correlação parcial para o modelo.

Cor. Parcial	$\Delta_{0.i}(i = 1, 2, 3, 4)$	Resultados
$r_{01.234} = -0,059$	0,725	$> 2\sigma$
$r_{02.134} = 0,057$	0,749	$> 2\sigma$
$r_{03.124} = 0,015$	0,825	$> 2\sigma$
$r_{04.123} = 0,015$	0,714	$> 2\sigma$

$\sigma = 0.356$

Esses resultados do método da correlação parcial aplicado no modelo causal escolhido *a priori* reforçam que todas as variáveis que compõem o primeiro modelo causal desempenham uma parte importante nessa relação. Dessa forma, para se obter as estimativas dos coeficientes de trilha desse relacionamento, foi ajustado o seguinte modelo:

$$y = p_{01}x_1 + p_{02}x_2 + p_{03}x_3 + p_{04}x_4 + p_{\varepsilon}u \tag{4.1}$$

Em que:

y é a variável básica CPL padronizada;

x_i são as variáveis explicativas PSPL, EXPL, NG e PA padronizadas ($i = 1, 2, 3, 4$);

p_{0i} são os coeficientes de trilha ou efeitos diretos das variáveis explicativas PSPL, EXPL, NG e PA sobre a básica CPL (i=1,2,3,4);

p_{ε} é o efeito da variável residual sobre a variável principal; e

u é o erro padronizado.

A análise de trilha constitui uma extensão da regressão múltipla, onde os parâmetros da regressão são os coeficientes de trilha. Dessa forma, para que as estimativas desses coeficientes sejam confiáveis, uma vez que altas correlações

entre as variáveis explicativas podem tornar as estimativas sem credibilidade, é necessário que não exista multicolinearidade entre essas variáveis. Se existir multicolinearidade, métodos alternativos devem ser usados para se estimar os coeficientes de trilha.

Foi aplicado o teste dos autovalores e autovetores da matriz de correlação das variáveis explicativas do modelo. O número de condições desse modelo foi $NC = 94$, o que revela uma multicolinearidade fraca, ou seja, um nível aceitável. Dessa forma, sem a necessidade de usar algum método alternativo para contornar os efeitos da multicolinearidade, os parâmetros foram estimados usando o método de mínimos quadrados.

Como os $p_{0i}(i = 1, 2, 3, 4)$ são coeficientes de um modelo de regressão múltipla, para que os seus valores possam ser confiáveis, quando estimados pelo método de mínimos quadrados, é necessário que os resíduos gerados obedeçam alguns pressupostos (normalidade, independência e homocedasticidade) e também que não existam *outliers* que sejam pontos influentes. Depois de estimado os coeficientes pelo método dos mínimos quadrados ordinários, foi feita a análise dos resíduos, onde a independência foi confirmada pelo teste de Durbin Watson ($valor - p > 0,01$); normalidade pôde ser verificada usando o teste de normalidade Shapiro-Wilk ($valor - p > 0,01$); e os resíduos são homocedásticos ($valor - p > 0,01$). A não existência de pontos influentes foi verificada pela na distância de Cook (MONTGOMEY; PECK, 1992).

Como foram atendidas todas as pressuposições, a análise de trilha foi aplicada e, considerando o modelo 4.1, verificam-se as seguintes relações:

$$Cov(y, x_1) = r_{01} = p_{01} + p_{02}r_{12} + p_{03}r_{13} + p_{04}r_{14}$$

$$Cov(y, x_2) = r_{02} = p_{01}r_{12} + p_{02} + p_{03}r_{23} + p_{04}r_{24}$$

$$Cov(y, x_3) = r_{03} = p_{01}r_{13} + p_{02}r_{23} + p_{03} + p_{04}r_{34}$$

$$Cov(y, x_4) = r_{04} = p_{01}r_{14} + p_{02}r_{24} + p_{03}r_{34} + p_{04}$$

Os coeficientes de trilha, efeitos diretos e indiretos, da variáveis explicativas sobre o comprimento da plântula encontram-se na Tabela 3. Pode-se verificar pelo coeficiente de determinação ($R^2_{0.1234}$) que estas variáveis explicaram $78,3\%$ da variação do tamanho da plântula. A variável NG foi a mais influente, com estimativa de efeito direto maior do que o efeito residual. Assim, pode-se dizer que essa variável é a principal determinante na variação da variável CPL. Segundo Faria (2008), o número de gemas está estritamente relacionado com o tamanho da plântula, onde plântulas com maior números de gemas apresentam maior crescimento, reforçando a idéia de que NG é determinante na variação da CPL. Também foram observados efeitos indiretos maiores que o efeito residual entre EXPL e CPL, e entre PA e CPL.

Verifica-se que as correlações entre EXPL e CPL, e entre PA e CPL, embora relativamente altas e positivas, ocorreram por influência da variável NG, pois essas duas variáveis explicativas, EXPL e PA, apresentaram altos efeitos indiretos sobre CPL via NG, e baixos efeitos diretos sobre CPL. Esses resultados reforçam uma relação de causa e efeito, onde a variável NG é a principal determinante nas alterações da CPL.

O efeito direto da variável PSPL sobre CPL foi negativo e bastante baixo, indicando uma baixa contribuição dessa variável para a CPL. Dessa forma, possivelmente, a correlação moderada entre PSPL e CPL está sendo causada pelos efeitos indiretos via NG e PA.

Tabela 3 Estimativas dos efeitos diretos e indiretos das variáveis consideradas como explicativas sobre a variável básica.

Variáveis primárias	Vias de associação	Estimador	Estimativa
PSPL	Efeito direto sobre CPL	$\hat{p}_{01}$	-0,048
	Efeito indireto via EXPL	$\hat{p}_{02}r_{12}$	0,086
	Efeito indireto via NG	$\hat{p}_{03}r_{13}$	0,461
	Efeito indireto via PA	$\hat{p}_{04}r_{14}$	0,160
	Total	r_{01}	0,66
EXPL	Efeito direto sobre CPL	$\hat{p}_{02}$	0,121
	Efeito indireto via PSPL	$\hat{p}_{01}r_{12}$	-0,034
	Efeito indireto via NG	$\hat{p}_{03}r_{23}$	0,566
	Efeito indireto via PA	$\hat{p}_{04}r_{24}$	,156
	Total	r_{02}	0,81
NG	Efeito direto sobre CPL	$\hat{p}_{03}$	0,615
	Efeito indireto via PSPL	$\hat{p}_{01}r_{13}$	-0,036
	Efeito indireto via EXPL	$\hat{p}_{02}r_{23}$	0,111
	Efeito indireto via PA	$\hat{p}_{04}r_{34}$	0,148
	Total	r_{03}	0,84
PA	Efeito direto sobre CPL	$\hat{p}_{04}$	0,190
	Efeito indireto via PSPL	$\hat{p}_{01}r_{14}$	-0,040
	Efeito indireto via EXPL	$\hat{p}_{02}r_{24}$	0,09
	Efeito indireto via NG	$\hat{p}_{03}r_{34}$	0,480
	Total	r_{04}	0,73
$R^2_{0.1234}$			0,783
Efeito residual ($\hat{p}_{\varepsilon}$)			0,465

Experimento II - Milho

Realizou-se um diagnóstico para verificar a existência ou não de *outliers* através da construção de gráficos do tipo *boxplot* para cada uma das variáveis. Pelas análises desses gráficos foi concluído que nenhuma das variáveis em estudo continha algum *outlier*. A suposição de normalidade das variáveis foi avaliada pelo teste de Shapiro-Wilk, onde foi constatada a normalidade de todas as variáveis ($valor - p > 0.05$). Verificada essas premissas procedeu-se a análise de trilha.

Segundo Ribeiro (2012), a produtividade de grãos de milho, provavelmente, é afetada por praticamente todos os demais caracteres da planta, ou seja, em tese, a maioria dos genes da planta contribui para a expressão da produtividade. Entretanto, existe uma hierarquia na influência dos caracteres na produtividade. Assim, o número de grãos por planta (NGP) e o peso de 100 grãos (P100) são os que estão mais diretamente associados à produtividade de grãos. Esses caracteres são denominados componentes primários da produção.

Dessa forma, considerou-se, para esse segundo experimento, uma análise de trilha em cadeia, Figura 7. Foi tomada como variável dependente a produção de grãos (PROD) como variáveis primárias o peso total de grãos (PT), peso de 100 grãos (P100) e número de grãos por planta (NGP), e como variáveis secundárias a altura da planta (AP), altura de espiga (AE) e diâmetro do colmo (DC).

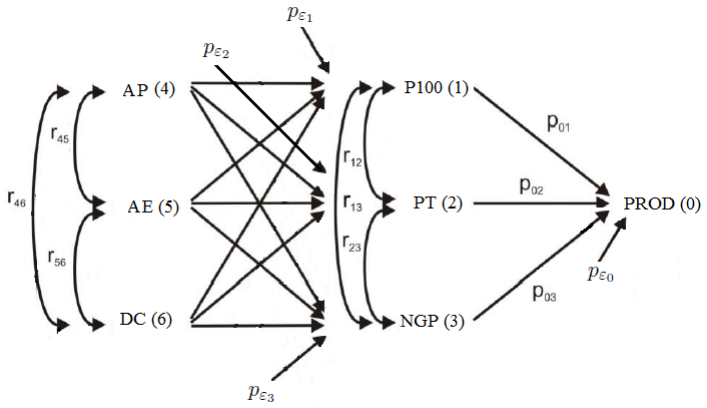

Figura 7 Diagrama causal em cadeia, onde a produção de grãos (PROD) é a variável básica, o peso de 100 grãos (P100), peso total de grãos (PT), e número de grãos por planta (NGP) são as variáveis primárias, e a altura da planta (AP), altura de espiga (AE) e diâmetro do colmo (DC) são as variáveis secundárias

Na Tabela 4, estão os valores das correlações das variáveis independentes. Percebe-se que as variáveis primárias apresentaram fortes correlações positivas e significativas com a variável básica PROD. Esse era esperado uma vez que os componentes primários são aqueles diretamente relacionados com a produção de grãos (LENG, 1954). Altas correlações também foram encontradas entre as variáveis primárias indicando a possível existência de multicolinearidade entra essas variáveis. As variáveis secundárias apresentaram correlações moderadas com a variável básica, e algumas correlações não significativas entre elas, não indicando a existência de multicolinearidade

Tabela 4 Correlações simples entre as sete variáveis do relativas à produção de milho.

	P100	PT	NGP	AP	AE	DC	PROD
P100	1						
PT	$0,918^{**}$	1					
NGP	$0,722^{**}$	$0,932^{*}$	1				
AP	$0,601^{*}$	$0,631^{*}$	$0,555^{*}$	1			
AE	$0,564^{*}$	$0,520^{*}$	$0,433$	$0,340$	1		
DC	$-0,409$	$-0,543^{*}$	$-0,618^{*}$	$-0,315$	$-0,127$	1	
PROD	$0,922^{**}$	$0,945^{**}$	$0,831^{**}$	$0,582^{*}$	$0,608^{*}$	$-0,584^{*}$	1

$^{*}(valor - p < 0,05)$ e $^{**}(valor - p < 0,01)$.

Primeiramente estudou-se a relação entre as variáveis primárias e a variável básica, diagrama da Figura 8.

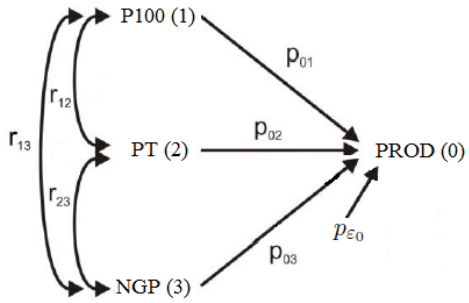

Figura 8 Primeiro diagrama causal da análise de trilha em cadeia

Analisando os resultados do método da correlação parcial, Tabela 5, foi possível verificar que todas as variáveis que compõem o primeiro modelo causal (Figura 8) desempenham uma parte importante nessa relação.

Tabela 5 Resultado do método da correlação parcial para o primeiro modelo.

Cor. Parcial	$\Delta_{0.i}(i = 1, 2, 3)$	Resultados
$r_{01.23} = 0,093$	0,826	$> 2\sigma$
$r_{02.13} = 0,223$	0,716	$> 2\sigma$
$r_{03.12} = -0,076$	0,907	$> 2\sigma$

$\sigma = 0.288$

Dessa forma, para se obter as estimativas dos coeficientes de trilha do primeiro diagrama causal, foi ajustado o seguinte modelo:

$$y = p_{01}x_1 + p_{02}x_2 + p_{03}x_3 + p_{\varepsilon}u \tag{4.2}$$

Em que:

y é a variável básica PROD padronizada;

x_i são as variáveis explicativas P100, PT e NGP padronizadas ($i = 1, 2, 3, 4$);

p_{0i} são os coeficientes de trilha ou efeitos diretos das variáveis explicativas P100,

PT e NGP sobre a básica PROD (i=1,2,3);

p_ε é o efeito da variável residual sobre a variável principal; e

u é o erro padronizado.

Avaliando a multicolinearidade, usando o teste dos autovalores e autovetores da matriz de correlação, nas variáveis independentes desse primeiro modelo, foi obtido o número de condições igual a 813 ($NC = 813$), ou seja, uma multicolinearidade de moderada a severa. Dessa forma, como não foi interessante retirar uma variável dessa relação para tentar se evitar a multicolinearidade, foi utilizado a regressão em crista (seção 2.3.5.1) para a estimação dos coeficientes do modelo e, assim, contornar os efeitos da multicolinearidade. O valor adequado referente à constante c foi determinado, neste ensaio, pelo exame do traço da crista (HOERL; KENNARD, 1970a). O traço da crista foi obtido plotando os parâmetros estimados (coeficientes de trilha) em função dos valores de c no intervalo de $0 < c < 1$, Figura 9. O menor valor de c capaz de estabilizar a maioria dos estimadores dos coeficientes de trilha foi empregado.

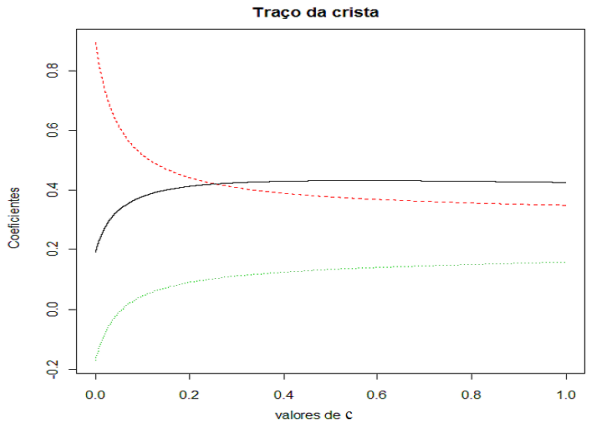

Figura 9 Gráfico do traço da crista, que representa a variação no valor dos coeficientes da regressão com diversos valores de c

Adotando-se o valor $c = 0,4$, foi possível verificar uma diminuição da soma de quadrado dos resíduos, em relação ao modelo com os coeficientes estimados pelo método dos mínimos quadrados ordinários.

Depois de estimado os coeficientes, foi feita a análise dos resíduos, onde a independência foi confirmada pelo teste de Durbin Watson ($valor - p > 0,05$); a normalidade pôde ser verificada usando o teste de normalidade Shapiro-Wilk ($valor - p > 0,05$); e também se verificou a homocedasticidade ($valor - p > 0,05$). Pela distância de cook foi verificada a não existência de pontos influentes.

Como foram atendidas todas as pressuposições, a análise de trilha foi aplicada e, considerando o modelo 4.2, verificam-se as seguintes relações:

$$Cov(y, x_1) = r_{01} = p_{01} + p_{02}r_{12} + p_{03}r_{13}$$

$$Cov(y, x_2) = r_{02} = p_{01}r_{12} + p_{02} + p_{03}r_{23}$$

$$Cov(y, x_3) = r_{03} = p_{01}r_{13} + p_{02}r_{23} + p_{03}$$

Os coeficientes de trilha, diretos e indiretos, da variáveis explicativas sobre produção encontram-se na Tabela 6. Pode-se verificar pelo coeficiente de determinação ($R^2_{0.123}$) que estas variáveis explicaram $89,11\%$ da variação da produção de grãos. A variáveis P100 e PT apresentaram efeitos diretos semelhantes, 0,446 e 0,404 respectivamente, e maiores que o efeito residual sobre a variável básica PROD sendo, dessa forma, as principais determinantes sobre a produção de grãos. Foram também essas variáveis as que apresentaram as maiores correlações com PROD, indicando que são realmente as mais relacionadas com PROD. Segundo Ribeiro (2012), o número de grãos por planta (NGP) e o peso de 100 grãos (P100) são os que estão mais diretamente associados à produtividade de grãos. Essa constatação corrobora com Lopes et al. (2007), que obtiveram resultados semelhantes,

onde espigas com maior P100 tiveram efeito direto sobre o aumento da produtividade de grãos.

O efeito indireto de P100 via PT e o efeito indireto de PT via P100 sobre PROD também foram altos e maiores do que o efeito residual, reforçando ainda mais a importância dessas duas variáveis em relação a variação da produção de grãos. O que está de acordo com Ottaviano e Camussi (1981 apud IVANOVIC; ROSIC, 1985), que em seu trabalho, através da análise de trilha, verificaram existir elevado efeito de componentes de rendimento sobre a produtividade de grãos em milho.

A variável NGP apresentou uma razoável correlação positiva com a variável básica, porém seu efeito direto não foi considerável. Dessa forma, verifica-se que essa correlação ocorreu por influência de P100 e PT, pois os efeitos indiretos via essas duas variáveis foram altos.

Tabela 6 Estimativas dos efeitos diretos e indiretos das variáveis primárias sobre a variável básica produção de grãos (PROD).

Variáveis primárias	Vias de associação	Estimador	Estimativa
P100	Efeito direto sobre PROD	$\hat{p}_{01}$	0,446
	Efeito indireto via PT	$\hat{p}_{02}r_{12}$	0,367
	Efeito indireto via NGP	$\hat{p}_{03}r_{13}$	0,110
	Total	r_{01}	0,92
PT	Efeito direto sobre PROD	$\hat{p}_{02}$	0,404
	Efeito indireto via P100	$\hat{p}_{01}r_{12}$	0,408
	Efeito indireto via NGP	$\hat{p}_{03}r_{23}$	0,128
	Total	r_{02}	0,94
NGP	Efeito direto sobre PROD	$\hat{p}_{03}$	0,128
	Efeito indireto via P100	$\hat{p}_{01}r_{13}$	0,327
	Efeito indireto via PT	$\hat{p}_{02}r_{23}$	0,375
	Total	r_{03}	0,83
$R^2_{0.123}$			0,8911
Efeito residual ($\hat{p}_{\varepsilon}$)			0,329

Existe também grande interesse em verificar os efeitos dos componentes considerados como secundário sobre os primários (RODRIGUES et al., 2010). Dessa forma, foi analisada a segunda parte do diagrama causal em cadeia, Figura 10, que é composta por três modelos, onde as variáveis primárias são consideradas como variáveis dependentes.

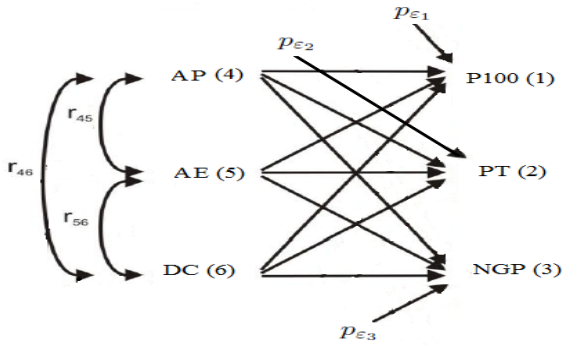

Figura 10 Segundo diagrama causal da análise de trilha em cadeia

Pelo método da correlação parcial, Tabela 7, percebe-se que as variáveis que compõem os três modelos não justificam os relacionamentos propostos em cada um desses modelos, ou seja, as variáveis secundárias muito possivelmente não explicam a variação de cada uma das variáveis primárias. Dessa forma, com objetivo comprobatório, os três modelos forma ajustados, e foram estimados seus os efeitos diretos e indiretos.

Tabela 7 Resultados do método da correlação parcial considerando os três modelos, onde as variáveis P100, PT e NGP são variáveis dependentes, e as variáveis AP, AE e DC são variáveis explicativas.

Var. dependente	Cor. parcial	$\Delta_{1.i}(i=4,5,6)$	Resultados
P100	$r_{14.56}=0,467$	0,132	$<2\sigma$
	$r_{15.46}=0,494$	0,065	$<2\sigma$
	$r_{16.45}=-0,318$	0,091	$<2\sigma$
		$\Delta_{2.i}(i=4,5,6)$	
PT	$r_{24.56}=0,507$	0,122	$<2\sigma$
	$r_{25.46}=0,462$	0,057	$<2\sigma$
	$r_{26.45}=-0,504$	0,035	$<2\sigma$
		$\Delta_{3.i}(i=4,5,6)$	
NGP	$r_{34.56}=0,400$	0,149	$<2\sigma$
	$r_{35.46}=0,363$	0,066	$<2\sigma$
	$r_{36.45}=-0,584$	0,025	$<2\sigma$

$\sigma = 0.288$

Como exemplo, os efeitos diretos e indiretos das variáveis secundárias 4, 5, 6 (AP, AE, DC) sobre a variável primária 1 (P100) são obtidos a partir do seguinte modelo:

$$x_1 = p_{14}x_4 + p_{15}x_5 + p_{16}x_6 + p_{\varepsilon}u \tag{4.3}$$

Em que:

x_1 é a variável primária P100 padronizada;

x_i são as variáveis explicativas AP, AE e DC padronizadas ($i = 4, 5, 6$);

p_{1i} são os coeficientes de trilha ou efeitos diretos das variáveis explicativas AP, AE e DC sobre a básica PROD (i=4,5,6);

p_{ε} é o efeito da variável residual sobre a variável principal; e

u é o erro padronizado.

A multicolinearidade entre as variáveis independentes foi testada pelo número de condições da matriz de correlação, sendo encontrada uma multicolinearidade fraca ($NC = 3$), assim os coeficientes de cada modelo que compõem

esse diagrama foram estimados pelo método dos mínimos quadrados.

Depois de ajustados todos esses três modelos foram verificados, pelos testes estatísticos já mencionados, que todos os pressupostos da análise de resíduos foram satisfeitos ($valor - p > 0,05$), e também foi verificada a não existência de *outliers*.

As estimativas dos efeitos diretos e indiretos das variáveis secundárias sobre as variáveis primárias se encontram na Tabela 8. Percebe-se, pelo coeficiente de determinação de cada modelo, que as variáveis secundárias não explicaram de maneira satisfatória a variação de cada variável primária, confirmando que o teste da correlação parcial foi pertinente. Dessa forma, os efeitos diretos e indiretos da variáveis secundárias sobre as primárias não foram interessantes.

Tabela 8 Efeitos diretos e indiretos das variáveis secundárias sobre as variáveis primárias.

Variáveis secundárias	Vias de associação	Variáveis primárias		
		P100	PT	NGP
AP	Efeito direto sobre	0,390	0,397	0,316
	Efeito indireto via AE	0,134	0,114	0,088
	Efeito indireto via DC	0,076	0,118	0,155
	Total	0,60	0,63	0,55
AE	Efeito direto sobre	0,396	0,336	0,259
	Efeito indireto via AP	0,132	0,136	0,108
	Efeito indireto via DC	0,040	0,048	0,063
	Total	0,56	0,52	0,43
DC	Efeito direto sobre	-0,233	-0,369	-0,484
	Efeito indireto via AP	-0,124	-0,127	-0,101
	Efeito indireto via AE	-0,051	-0,044	-0,033
	Total	-0,40	-0,54	-0,61
R^2		0,5522	0,6256	0,5894
Efeito residual ($\hat{p}_\varepsilon$)		0,66	0,61	0,64

Por último, foram estimados os efeitos diretos e indiretos dos componentes secundários sobre a variável principal. Como exemplo, a seguir é apresentado o desdobramento da correlação entre a variável principal PROD e o componente secundário AP:

$$r_{04} = p_{01}r_{14} + p_{02}r_{24} + p_{03}r_{34} = p_{01}(p_{14} + p_{15}r_{45} + p_{16}r_{46})$$

$$+p_{02}(p_{24} + p_{25}r_{45} + p_{26}r_{46}) + p_{03}(p_{34} + p_{35}r_{45} + p_{36}r_{46})$$

Onde, definem-se os seguintes efeitos:

a) Efeito direto do componente secundário AP via componentes primários:

- via P100 é dado por: $\hat{p}_{01}\hat{p}_{14}$
- via PT é dado por: $\hat{p}_{02}\hat{p}_{24}$
- via NGP é dado por: $\hat{p}_{03}\hat{p}_{34}$

b) Efeito indireto do componente secundário AP via outros componentes secundários e primários:

- via AE por P100: $\hat{p}_{01}\hat{p}_{15}r_{45}$

  por PT: $\hat{p}_{02}\hat{p}_{25}r_{45}$

  por NGP: $\hat{p}_{03}\hat{p}_{35}r_{45}$
- via DC por P100: $\hat{p}_{01}\hat{p}_{16}r_{46}$

  por PT: $\hat{p}_{02}\hat{p}_{26}r_{46}$

  por NGP: $\hat{p}_{03}\hat{p}_{36}r_{46}$

Quando se analisou os efeitos diretos e indiretos dos componentes secundários sobre a variável principal, Tabela 9, verificou-se que as variáveis secundárias AP e AE apresentaram efeitos diretos sobre a produção de grãos aproximadamente iguais, e efeitos diretos baixos. A variável AE foi a que apresentou a maior correlação com PROD, podendo ser considerada a mais determinante na

variação da produção de grãos. Churata e Ayala-Osuna (1996) em seu trabalho, também encontraram um resultado parecido, onde a altura de espiga foi uma das variáveis explicativas mais fortemente relacionada com a produção, com efeito direto de 0,39.

Tabela 9 Efeitos diretos e indiretos das variáveis secundárias sobre a variável básica.

Variáveis secundárias	Vias de associação	Variáveis primárias			Total dos Efeitos
		P100	PT	NGP	
AP	Efeito direto	0,174	0,16	0,04	0,375
	Efeito indireto via AE	0,06	0,046	0,011	0,117
	Efeito indireto via DC	0,032	0,047	0,02	0,099
	r_{04}				0,59
AE	Efeito direto	0,176	0,145	0,054	0,375
	Efeito indireto via AP	0,07	0,064	0,023	0,157
	Efeito indireto via DC	0,023	0,029	0,008	0,06
	r_{05}				0,59
DC	Efeito direto	-0,114	-0,15	-0,081	-0,345
	Efeito indireto via AP	-0,064	-0,07	-0,022	-0,156
	Efeito indireto via AE	-0,032	-0,037	-0,004	-0,073
	r_{06}				-0,57

5 CONCLUSÃO

Esta dissertação comprovou que diversos aspectos estatísticos devem ser considerados quando se utiliza a análise de trilha, principalmente, no que se refere aos diversos pressupostos necessários para conduzir a análise.

O uso do método da correlação parcial, para a construção do diagrama de trilha, foi eficiente na escolha das variáveis.

O procedimento usado para diagnosticar a multicolinearidade entre as variáveis explicativas mostrou-se eficiente para detectar e quantificar a intensidade com que a multicolinearidade se manifesta.

A estimação dos parâmetros usando regressão em crista mostrou-se uma alternativa confiável e concisa na presença de multicolinearidade.

No experimento com maracujazeiros, o número de gemas foi a variável mais correlacionada com o comprimento da plântula (CPL) e foi também a que apresentou o maior efeito direto sobre CPL, sendo, dessa forma, a principal característica na variação da CPL.

No experimento com milho, o peso de 100 grãos foi o componente primário que apresentou o maior efeito direto sobre a produção de grãos (PROD), sendo assim, o mais indicado para seleção indireta para a PROD. A altura de espiga foi o componente secundário mais influente na variação da PROD.

REFERÊNCIAS

BREUSCH, T.; PAGAN, A. Teste simples para heterocedasticidade e coeficiente de variação aleatória Econométrica. **Sociedade Econométrica**, Rio de Janeiro, v. 47, p. 1287-1294, 1979.

BROOKS, C. H. Social, economic, and biologic correlates of infant mortality in city Neighborhoods. **Journal of Health and Social Behavior**, Cleveland, v. 21, n. 1, p. 2-11, Mar. 1980.

CARVALHO, C. G. P. et al. Análise de trilha sob multicolinearidade em pimentão. **Pesquisa Agropecuária Brasileira**, Brasília, v. 34, n. 4, p. 603-613, abr. 1999.

CARVALHO, F. I. F. et al. **Estimativas e implicações da correlação no melhoramento vegetal**. Pelotas: UFPel, 2004. 142 p.

CARVALHO, P. C.; CRUZ, C. D. Diagnosis of multicollinearity: assement of the condition of correlation matrices used in genetic studies. **Brazilian Journal of Genetics**, Ribeirão Preto, v. 19, n. 3, p. 479-484, 1996.

CASELLA, G.; BERGER, R. **Inferência estatística**. 2^{nd} ed. São Paulo: C. Learning, 2001. 588 p.

CHARNET, R. et al. **Análise de modelos de regressão linear**. Campinas: Unicamp, 2008. 357 p.

CHURATA, B. G. M.; AYALA-OZUNA, J. T. Correlações genotípica, fenotípica e de ambiente e análise de trilha em caracteres avaliados no composto de milho (*Zea mays*) arquitetura. **Revista Ceres**, Viçosa, MG, v. 43, n. 249, p. 628-636, 1996.

COSTA NETO, P. L. O. **Estatística**. 2. ed. São Paulo: E. Blücher, 2009. 280 p.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa, MG: UFV, 2003. v.2, 585 p.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**. 3. ed. Viçosa, MG: UFV, 2004. 480 p.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. 3^{rd} ed. New York: J. Wiley, 1998. 706 p.

ESPÓSITO, D. C. **Análise de trilha em dados de produção e tecnológicos de cana-de-açúcar**. 2010. 102 p. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2010.

FARIA, G. A. **Tamanho ótimo de parcelas experimentais para experimentos in vitro com maracujazeiro**. 2008. 101 p. Tese (Doutorado em Agronomia) - Universidade Estadual Paulista "Julio de Mesquita Filho", Ilha Solteira, 2008.

FARRAR, D. E.; GLAUBER, R. R. Multicollinearity in regression analysis: the problem revisited. **The Review of Economics and Statistics**, Cambridge, v. 49, n. 1, p. 92-107, 1967.

Felipe de Mendiburu (2010). **agricolae**: Statistical Procedures for Agricultural Research. R package version 1.0-9.
http://CRAN.R-project.org/package=agricolae.

FISHER, R. A. **Statistical methods for research workers**. 4^{th} ed. London: Oliver e Boyd, 1932. 307 p.

FREIRE FILHO, F. R. Genética no feijão-caupi. In: ARAÚJO, J. P. P.; WATT, E. E. (Ed.). **O Feijão-caupi no Brasil**. Brasília: IITA/EMBRAPA-CNPAF, 1988. p. 159-229.

GOLDSMITH, J. R. Paths of association in epidemiological analysis: appliconfounding factor, since no statistical test or procedure cation to health effects of environmental exposures. **International Journal of Epidemiology**, Oxford, v. 6, n. 4, p. 391-399, 1977.

GOMES, T. C. A. **Análise de trilha no estudo de fatores físicos e químicos relacionados ao adensamento e, ou, à compactação em dois solos do norte de Minas Gerais**. 1996. 105 p. Dissertação (Mestrado em Curso de Solos e Nutrição de Plantas) - Universidade Federal de Viçosa, Viçosa, MG, 1996.

GUNST, R. F.; MASON, R. L. Advantages of examining multicollinearities in regression analisys. **Biometrics**, Washington, v. 33, p. 249-260, 1977.

HAIR, J. F. et al. **Multivariate data analysis**. 5^{th} ed. New Jersey: Prentice-Hall, 1998. 730 p.

HOERL, A. E.; KENNARD, R. W. Ridge regression: applications to nonorthogonal problems. **Technometrics**, Washington, v. 12, n. 1, p. 69-82, 1970a.

_____. Ridge regression: biased estimation for nonorthogonal problems. **Technometrics**, Washington, v. 12, n. 1, p. 55-67, 1970b.

IVANOVIC, M.; ROSIC, K. Path coefficient analysis for three stalk traits and grain yield in maize (*Zea mays* L.). **Maydica**, Bergamo, v. 30, p. 233-239, 1985.

John Fox and Sanford Weisberg (2011). **An R Companion to Applied Regression**, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

KLINE, R. B. Latent variable path analysis in clinical research: a beginner's tour guide. **Journal of Clinical Psychology**, Montreal, v. 47, n. 4, p. 471-484, July 1991.

_____. **Principles and practice of structural equation modeling**. 3^{rd} ed. New York: The Guilford, 2011. 427 p.

KMENTA, J. **Elements of econometrics**. New York: MacMillan, 1971. 655 p.

KOSAK, M.; AZEVEDO, R. A. Does using stepwise variable selection to build sequential path analysis models make sense? **Physiologia Plantarum**, Copenhagen, v. 141, n.3, p. 197-200, Mar. 2011.

LENG, E. R. Effects of heterosis on the major components of grain yield in corn. **Agronomy Journal**, Madison, v. 46, n. 11, p. 502-506, 1954.

LI, C. C. Concept of path coefficient and its impact on population genetics. **Biometrics**, Washington, v. 12, p. 190-210, 1956.

_____. **Path analysis**: a primer. Pacific Grover: Boxwood, 1975. 346 p.

LIRA, S. A. **Análise de correlação**: análise teórica e de construção dos coeficientes com a plicações. 2004. 196 p. Dissertação (Mestrado em Métodos Numéricos em Engenharia dos Setores de Ciências Exatas e de Tecnologia) - Universidade Federal do Paraná, Curitiba, 2004.

LOEHLIN, J. C. **Latent variable models**: an introduction to factor, path, and structural equation analysis. 4^{th} ed. Mahwah: L. Erlbaum, 2004. 317 p.

LOPES, S. J. et al. Relações de causa e efeito em espigas de milho relacionadas aos tipos de híbridos. **Ciência Rural**, Santa Maria, v. 37, n. 6, p. 1536-1542, nov./dez. 2007.

MARQUARD, D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. **Technometrics**, Washington, v. 12, n. 3, p. 591-612, Aug. 1970.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to linear regression analysis**. 2^{nd} ed. New York: J. Wiley, 1992. 544 p.

NEGREIROS, J. R. S. et al. Relação entre característica físicas e o rendimento de polpa de maracujá-amarelo. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 29, n. 3, p. 546-549, 2007.

NETER, J. et al. **Applied linear statistical models**. 5^{th} ed. New York: McGraw-Hill/Irwin, 2005. 1396 p.

NUNES, E. S. et al. Importância das características físicas e químicas na determinação do teor de vitamina C em frutos de aceroleira. **Revista Ceres**, Viçosa, MG, v. 51, n. 297, p. 657-662, set./out. 2004.

OLIVEIRA, E. D. et al. Correlações genéticas e análise de trilha para número de frutos comerciais por planta em mamoeiro. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 8, p. 855-862, ago. 2010.

OSBORNE, J.; WATERS, E. Four assumptions of multiple regression that researchers should always test. **Practical Assessment, Research e Evaluation**, Washington, v. 8, n. 2, p. 1-5, 2002.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2012. Disponível em: <http://www.r-project.org>. Acesso em: 12 out. 2012.

RENCHER, A. C.; SCHAALJE, G.B. **Linear models in statistics**. 2^{nd} ed. New Jersey: J. Wiley, 2008. 672 p.

RIBEIRO, C. B. **Caracteres que explicam a heterose na produtividade de grãos de milho**. 2012. 64 p. Dissertação (Mestrado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2012.

RIOS, S. A. et al. Análise de trilha para carotenoides em milho. **Revista Ceres**, Viçosa, MG, v. 59, n. 3, p. 368-373, mai./jun. 2012.

RODRIGUES, G. B et al. Análise de trilha de componentes de produção primários e secundários em tomateiro do grupo Salada. **Pesquisa Agropecuária Brasileira**, Brasília, v. 45, n. 2, p. 155-162, fev. 2010.

Seongho Kim (2011). **ppcor**: Partial and Semi-partial (Part) correlation. R package version 1.0. http://CRAN.R-project.org/package=ppcor.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality. **Biometrika**, Cambridge, v. 52, n. 3/4, p. 591-611, 1965.

VASCONCELOS, A. G. G.; ALMEIDA, M. V. A.; NOBRE, F. F. The path analysis approach for the multivariate analysis of infant mortality data. **Annals of Epidemiology**, New York, v. 8, n. 4, p. 262-271, May 1998.

Venables, W. N. e Ripley, B. D. (2002) **Modern Applied Statistics with S**. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, Washington, v. 20, n. 7, p. 557-585, Jan. 1921.

_____. The method of path coefficients. **Annals of Mathematical Statistics**, Stanford, v. 5, n. 3, p. 161-215, Sep. 1934.