Scientific Notes

Interference of sample size on multicollinearity diagnosis in path analysis

Bruno Giacomini Sari[1], Alessandro Dal'Col Lúcio[1], Tiago Olivoto[1], Dionatan Ketzer Krysczun[1], André Luís Tischler[1] and Lucas Drebes[1]

[1]Universidade Fedral de Santa Maria, Departamento de Fitotecnia, Avenida Roraima, nº 1.000, CEP 97105-900 Santa Maria, RS, Brazil. E-mail: brunosari@hotmail.com, adlucio@ufsm.br, tiagoolivoto@gmail.com, diona1994@hotmail.com, andreluistischler@gmail.com, lucasdrebes@hotmail.com

Abstract – The objective of this work was to evaluate the interference of sample size on multicollinearity diagnosis in path analysis. From the analyses of productive traits of cherry tomato, two Pearson correlation matrices were obtained, one with severe multicollinearity and the other with weak multicollinearity. Sixty-six sample sizes were designed, and from the amplitude of the bootstrap confidence interval, it was observed that sample size interfered on multicollinearity diagnosis. When sample size was small, the imprecision of the diagnostic criteria estimates interfered with multicollinearity diagnosis in the matrix with weak multicollinearity.

Index terms: *Solanum lycopersicum*, bootstrapping, multivariate analysis, sampling.

Interferência do tamanho de amostra no diagnóstico de multicolinearidade em análise de trilha

Resumo – O objetivo deste trabalho foi avaliar o impacto do tamanho da amostra sobre o diagnóstico da multicolinearidade em análise de trilha. A partir das análises de variáveis produtivas do tomate-cereja, foram obtidas duas matrizes de correlação de Pearson, uma com multicolinearidade severa e outra com multicolinearidade fraca. Sessenta e seis tamanhos de amostra foram delineados e, a partir da amplitude do intervalo de confiança "bootstrap", verificou-se a existência de interferência do tamanho da amostra na multicolinearidade. Quando o tamanho da amostra foi pequeno, a imprecisão das estimativas dos critérios de diagnóstico interferiu na conclusão quanto à multicolinearidade da matriz com multicolinearidade fraca.

Termos para indexação: *Solanum lycopersicum*, bootstrap, análise multivariada, amostragem.

The study of the linear relationship among traits can be performed by using the Pearson correlation coefficient; however, it does not determine how many other variables indirectly interfere with this relation. When information of several traits is available in an experiment, the logical consequences of a causal relationship hypothesis on a system of correlated traits can be evaluated through path analysis (Rodrigues et al., 2010; Sari et al., 2017). Path analysis decomposes the correlation coefficient into direct and indirect effects, and it determines whether the cause and effect relationship between two traits is true and independent of the other trait. The relationship of cause and effect is shown by the comparison (in magnitude and sign) between the direct effect and the Pearson correlation coefficient (Cruz et al., 2012).

In plant breeding, the evaluation of multicollinearity is fundamental, especially for indirect selection, when the traits of interest show low heritability, and when there is difficulty to measure a trait, both for the early selection of plants and simultaneous selection for more than one character (Cruz & Regazzi, 1997). The multicollinearity among explanatory traits is one of the assumptions that should be fulfilled for the path coefficients to be estimated with accuracy. When multicollinearity is severe, the estimated coefficients may show illogical values, with inflated or negative variances, and without biological sense (Toebe & Cargnelutti Filho, 2013b). Coimbra et al. (2005) found direct effects of high-magnitude (greater than 1) and illogical relationships between traits in canola crop. In the path analysis, the diagnosis of multicollinearity is

performed in the Pearson correlation matrix (sample survey data), in the genotype matrices (in competition trials of cultivars), or in the phenotypic matrix (delineated trials). The diagnosis of multicollinearity should always be carried out, and measures should be taken to adjust this problem (Toebe & Cargnelutti Filho, 2013a, 2013b).

Several studies have shown that the sample size is related to the precision of the Pearson correlation coefficient estimate. The sample size is inversely proportional to the magnitude of the coefficient, that is, the smaller the Pearson correlation coefficient, the larger the sample size needed for its estimation to be precise (Cargnelutti Filho et al., 2010, 2011, 2012). Therefore, it is easy to recognize that the correlation matrix configuration is related to the sample size and that, therefore, it can affect the multicollinearity diagnosis in path analysis. The works found in the literature are concerned only about determining the sample size needed to estimate the correlation coefficients among traits, but none is concerned about the interference of this sample size on the multicollinearity diagnosis.

The objective of this work was to evaluate the interference of sample size on the diagnosis of multicollinearity in path analysis.

Data used in the present study come from a cherry tomato uniformity trial. The experiment was carried out in a greenhouse in the spring-summer of 2014 growing season. A tomato genotype (*Solanum lycopersicum* var. cerasiforme 'Lilli') was tutored with two stems, spaced at 1.00 m between rows and 0.5 m between plants. The raceme and fruit of 347 plants were harvested and, from them, the following agronomic traits were obtained: average fruit length per plant (AFL, cm), which corresponds to the transversal measure of the fruit; average fruit width per plant (AFWI, in cm), which corresponds to the longitudinal measure of the fruit; average fruit weight per plant (AFWE, in g); number of racemes per plant (NRP); number of fruit per raceme (NFR), obtained by the ratio between the total number of fruit and the total number of racemes; number of fruit per plant (NFP); and total weight of fruit per plant (TWF, in g). The measurements were performed with the aid of calipers and scales at 0.01 cm resolution.

The TWF was considered as the dependent trait, and the other variables as independent traits. Pearson correlation matrix of the independent trait $X'X_{6\times6}$ was calculated, and the multicollinearity diagnosis was performed, by interpreting the values of the matrix determinant (MD), calculated by the sum of the eigenvalue products, and the condition number (CN), calculated by the ratio between the larger and the smaller eigenvalue of the matrix (Montgomery & Peck, 1982). The matrix determinant was very close to zero and CN was higher than 1,000, indicating a severe multicollinearity. The problem was circumvented by eliminating the NFR variable, reducing the CN value to 35.82. Thus, two configurations of matrices were obtained, one with a severe multicollinearity ($X'X_{6\times6}$), and the other with a weak multicollinearity ($X'X_{5\times5}$).

Subsequently, based on the 347 observations, 66 sample sizes were designed. The lowest-sample size was composed of 20 plants, and the others were obtained with the addition of five plants (20, 25, 30, ..., 345 plants). For each of the 66 designed sample sizes, 3,000 Pearson correlation matrices were obtained from the independent traits, in the two matrix configurations, by using bootstrap resampling. The MD values of the 3,000 matrices were estimated. For the matrices with weak multicollinearity, beyond MD, the eigenvalues and CN were also estimated. Finally, for each dataset of these measures (3,000 bootstrap resampling), the following descriptive statistics were obtained: minimum, percentile 2.5% (LI2.5%); mean, percentile 97.5% (LS 97.5%); and the maximum and 95% confidence interval (CI 95%) were calculated by the difference between the percentiles 2.5 and 97.5%. The 95% CIs were used to compare the mean values of the 3,000 MD and CN estimates with the reference values of 0 and 100, respectively, considering a 5% error probability. All analyses were performed using the R software (R Core Team, 2014).

The mean value of MD of the 3,000 Pearson correlation matrices of the independent traits ($X'X_{6\times6}$) was very close to zero, which was already expected, since the matrix has a severe multicollinearity. Regardless of the sample size, the MD did not differ statistically from zero, since the AIC 95% encompassed this value (Figure 1). As a consequence, it can be inferred that the matrix is singular and cannot be inverted, preventing the obtaining of the path coefficients. It should be noticed that because the MD was considered statistically equal to zero,

originating a singular matrix, the eigenvalues and the CNs of the 3,000 matrices simulated by resampling were not calculated. Therefore, the sample size does not interfere with the multicollinearity diagnosis in this case, as the matrix is singular in all cases. The elimination of variables is one of the most effective methods to circumvent this problem, and allows of the estimation of the most accurate path coefficients (Toebe & Cargnelutti Filho, 2013a, 2013b; Olivoto et al., 2017).

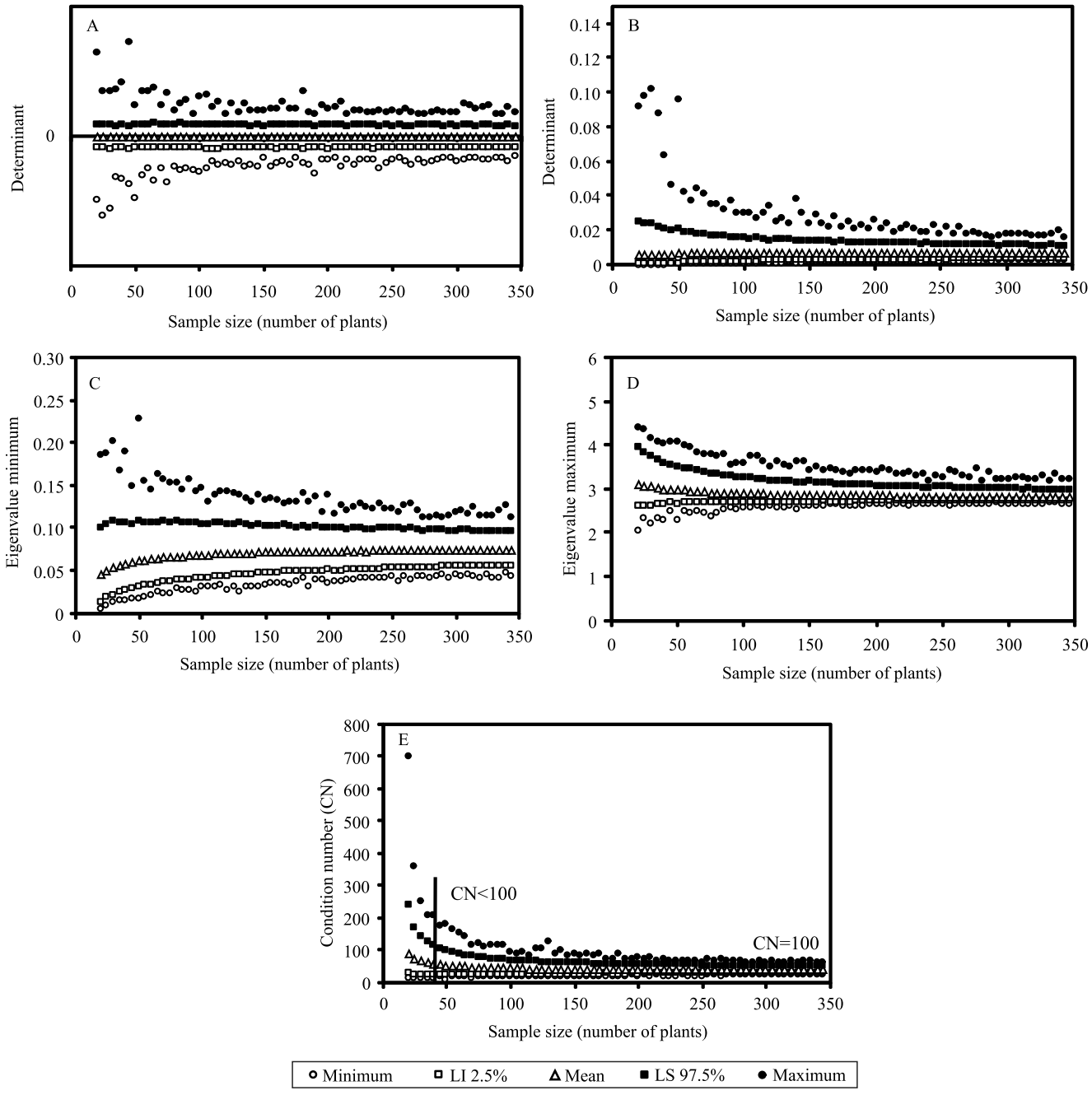

Figure 1. Values of the matrix determinant of independent traits with severe (A) and weak (B) multicollinearity: minimum; percentile 2.5% (LI 2.5%); mean; percentile 97.5% (LS 97.5); and maximum; and values for the eigenvalues of the matrix with weak multicollinearity: minimum eigenvalue (C); maximum eigenvalue (D); and condition number (E) for each one of the 66 sample sizes designed.

When the NFR trait was eliminated, the mean value of the determinant of the 3,000 estimates increased, and statistically differed from zero in all designed sample sizes (Figure 1). Therefore, it can be inferred that the Pearson correlation matrix of the independent variables ($X'X_{5\times5}$) became nonsingular and invertible, and the estimation of the path coefficient was possible. It has been noticed, however, that when the sample size is small, the precision of the MD estimate is extremely low. For instance, in the sample size composed of 20 plants, the 95% CI of MD ranged from 0.0002 to 0.025. Thus, when the sample size is reduced, the diagnosis of multicollinearity based on the MD can lead to contradictory conclusions.

The same fact was observed for the maximum and minimum eigenvalues and the CN calculated in the correlation matrix with weak multicollinearity (Figure 1). In addition to being less precise, the mean value of the minimum eigenvalue tends to be smaller, and the maximum eigenvalue to be higher, when the sample size is small. As a consequence, the value of CN (which is the ratio between the maximum and minimum eigenvalues) tends to be somewhat higher. The precision of the CD estimation is also extremely low when the sample size is small, and it increases as the sample size increases. When the sample size is equal to or less than 45 plants, the CN mean values of the 3,000 estimates did not differ from zero, and the matrix was erroneously considered to have a moderate multicollinearity. According to Montgomery & Peck (1982), when multicollinearity is moderate, adequate procedures should be taken aiming at reducing it.

Multicollinearity diagnosis in matrices with weak multicollinearity can be biased when the sample size is small, resulting either in the elimination of variables, or in the use of unnecessary statistical procedures such as ridge regression analysis. Thus, the number of estimated coefficients would be reduced (when eliminating traits), or show biased estimates (when using ridge path analysis). Thus, in addition to influencing the precision of correlation coefficient estimates, sample size also affects the diagnosis of multicollinearity, and may, therefore, interfere with the conclusion on cause and effect relationship between traits. To avoid the sampling influence on the multicollinearity diagnosis among productive variables of cherry tomato (with 5% error probability), a sample size of greater than 45 plants (12.96% of the population) is required. In sample sizes greater than or equal to 135 (38.90% of the population), the probability of sampling interference with the multicollinearity diagnosis is zero, regardless of the probability of error of the estimates.

This approach shows that the effects of sample size on multicollinearity diagnosis can be used in the planning of future studies aiming at a precise estimation of multicollinearity indicators.

References

CARGNELUTTI FILHO, A.; LOPES, S.J.; BRUM, B.; TOEBE, M.; SILVEIRA, T.R. da; CASAROTTO, G. Tamanho de amostra para a estimação do coeficiente de correlação linear de Pearson entre caracteres de mamoneira. **Semina: Ciências Agrárias**, v.33, p.953-962, 2012. DOI: 10.5433/1679-0359.2012v33n3p953.

CARGNELUTTI FILHO, A.; LOPES, S.J.; TOEBE, M.; SILVEIRA, T.R. da; SCHWANTES, I.A. Tamanho de amostra para estimação do coeficiente de correlação de Pearson entre caracteres de *Crambe abyssinica*. **Revista Ciência Agronômica**, v.42, p.149-158, 2011. DOI: 10.1590/S1806-66902011000100019.

CARGNELUTTI FILHO, A.; TOEBE, M.; BURIN, C.; SILVEIRA, T.R. da; CASAROTTO, G. Tamanho de amostra para estimação do coeficiente de correlação linear de Pearson entre caracteres de milho. **Pesquisa Agropecuária Brasileira**, v.45, p.1363-1371, 2010. DOI: 10.1590/S0100-204X2010001200005.

COIMBRA, J.L.M.; BENIN, G.; VIEIRA, E.A.; OLIVEIRA, A.C. de; CARVALHO, F.I.F; GUIDOLIN, A.F.; SOARES, A.P. Conseqüências da multicolinearidade sobre a análise de trilha em canola. **Ciência Rural**, v35, p.347-352, 2005. DOI: 10.1590/S0103-84782005000200015.

CRUZ, C.D.; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético**. 2.ed. Viçosa: Ed. da UFV, 1997. 390p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 4.ed. Viçosa: Ed. da UFV, 2012. 514p.

MONTGOMERY, D.C.; PECK, E.A. **Introduction to linear regression analysis**. New York: J. Wiley, 1982. 504p. (Wiley Series in Probability and Mathematical Statistics).

OLIVOTO, T.; SOUZA, V.Q. de; NARDINO, M.; CARVALHO, I.R.; FERRARI, M.; PELEGRIN, A.J. de; SZARESKI, V.J.; SCHMIDT, D. Multicollinearity in path analysis: a simple method to reduce its effects. **Agronomy Journal**, v.109, p.131-142, 2017. DOI: 10.2134/agronj2016.04.0196.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2014.

RODRIGUES, G.B.; MARIM, B.G.; SILVA, D.J.H. da; MATTEDI, A.P.; ALMEIDA, V. de S. Análise de trilha de componentes de produção primários e secundários em tomateiro do grupo Salada. **Pesquisa Agropecuária Brasileira**, v.45, p.155-162, 2010. DOI: 10.1590/S0100-204X2010000200006.

SARI, B.G.; LÚCIO, A.D.; SANTANA, C.S.; LOPES, S.J. Linear relationships between cherry tomato traits. **Ciência Rural**, v.47, e20160666, 2017. DOI: 10.1590/0103-8478cr20160666.

TOEBE, M.; CARGNELUTTI FILHO, A. Multicollinearity in path analysis of maize (*Zea mays* L.). **Journal of Cereal Science**, v.57, p.453-462, 2013a. DOI: 10.1016/j.jcs.2013.01.014.

TOEBE, M.; CARGNELUTTI FILHO, A. Não normalidade multivariada e multicolinearidade na análise de trilha em milho. **Pesquisa Agropecuária Brasileira**, v.48, p.466-477, 2013b. DOI: 10.1590/S0100-204X2013000500002.

Received on April 4, 2017 and accepted on September 11, 2017