Statistics/ Original Article

Reference sample size for multiple regression in corn

Abstract – The objective of this work was to determine the number of plants required to model corn grain yield (Y) as a function of ear length (X_1) and ear diameter (X_2), using the multiple regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The Y, X_1, and X_2 traits were measured in 361, 373, and 416 plants, respectively, of single-, three-way, and double-cross hybrids in the 2008/2009 crop year; and in 1,777, 1,693, and 1,720 plants, respectively, of single-, three-way, and double-cross hybrids in the 2009/2010 crop year, totaling 6,340 plants. Descriptive statistics were calculated, and frequency histograms and scatterplots were created. The sample size (number of plants) for the estimate of the β_0, β_1, and β_2 parameters, of the residual standard error, the coefficient of determination, the variance inflation factor, and the condition number between the explanatory traits of the model (X_1 and X_2) were determined by resampling with replacement. Measuring 260 plants is sufficient to adjust precise multiple regression models of corn grain yield as a function of ear length and ear diameter. The $Y = -229.76 + 0.54X_1 + 6.16X_2$ model is a reference for estimating corn grain yield.

Index terms: *Zea mays*, descriptive statistics, hybrids, modeling, resampling.

Alberto Cargnelutti Filho[1 ✉] 🔾 and Marcos Toebe[2] 🔾

[1] Universidade Federal de Santa Maria, Departamento de Fitotecnia, Avenida Roraima, nº 1.000, Cidade Universitária, Camobi, CEP 97105-900 Santa Maria, RS, Brazil.
E-mail: alberto.cargnelutti.filho@gmail.com

[2] Universidade Federal de Santa Maria, Departamento de Ciências Agronômicas e Ambientais, Campus de Frederico Westphalen, Linha 7 de Setembro, s/nº, BR 386, Km 40, CEP 98400-000 Frederico Westphalen, RS, Brazil.
E-mail: m.toebe@gmail.com

✉ Corresponding author

Tamanho de amostra-referência para regressão múltipla em milho

Resumo – O objetivo deste trabalho foi determinar o número de plantas necessário para modelar a produtividade de grãos de milho (Y) em função do comprimento de espiga (X_1) e do diâmetro de espiga (X_2), por meio do modelo de regressão múltipla $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Os caracteres Y, X_1 e X_2 foram mensurados em 361, 373 e 416 plantas, respectivamente, de híbridos simples, triplo e duplo no ano agrícola 2008/2009; e em 1.777, 1.693 e 1.720 plantas, respectivamente, de híbridos simples, triplo e duplo no ano agrícola 2009/2010, tendo-se totalizado 6.340 plantas. Foram calculadas estatísticas descritivas, e confeccionados histogramas de frequência e diagramas de dispersão. O tamanho de amostra (número de plantas) para a estimação dos parâmetros β_0, β_1 e β_2, do erro-padrão residual, do coeficiente de determinação, do fator de inflação da variância e do número de condição entre os caracteres explicativos do modelo (X_1 e X_2) foram determinados por reamostragem, com reposição. A mensuração de 260 plantas é suficiente para ajustar modelos de regressão múltipla precisos para produtividade de grãos de milho, em função do comprimento de espiga e do diâmetro de espiga. O modelo $Y = -229,76 + 0,54X_1 + 6,16X_2$ é referência para estimar a produtividade de grãos de milho.

Termos para indexação: *Zea mays*, estatística descritiva, híbridos, modelagem, reamostragem.

Introduction

Corn (*Zea mays* L.) is the cereal with the highest production volume worldwide according to the United States Department of Agriculture (Usda, 2019), with an estimated production of 1,099.61 million tons for the 2018/2019 crop in an area of 189.31 million hectares. Brazil is the third largest corn producer, with an estimated productivity of 5.40 tons per hectare and a total production of 94.50 million tons in an area of 17.50 million hectares (Usda, 2019).

Numerous bi- and multivariate techniques, such as linear correlation coefficients (Toebe et al., 2015), canonical correlation (Alves et al., 2016), and path analysis (Toebe et al., 2017), have been applied to identify the direction and magnitude of the associations between corn traits. Multiple linear regression has also been used to predict the behavior of one principal variable as a function of two or more explanatory variables in corn. Laurie et al. (2004), for example, found, via simulations, that multiple linear regression was the most effective method to detect quantitative trait loci in a cross between high- and low-selection lines for oil concentration in corn. Ge & Wu (2019) used multiple linear regression to predict corn price fluctuation, considering production-consumption and import and export volume as independent variables. Mohammadi (2007) verified, via multiple linear regression, that relative growth rate and specific leaf area were the best predictors of the competitiveness of corn cultivars against weeds.

In some of these bi- and multivariate techniques, sample sizing was performed for different precision levels. Toebe et al. (2015) recommended 195 corn plants to estimate correlation coefficients, whereas, in a specific path analysis scenario, Toebe et al. (2017) suggested 120 corn plants to estimate direct effects. Using a multivariable prediction model, Riley et al. (2019) recommended, based on four criteria of sample sizing, at least 918 subjects in a model with 25 predictor parameters. For multiple linear regression and the analysis of covariance, Bujang et al. (2017) suggested a minimum sample size of 300 or more to generate an approximation of estimates with parameters in a clinical survey. In order to obtain a reliable regression model to predict leaf area, Antunes et al. (2008) recommended, at least, 200 leaves for two coffee species – *Coffea arabica* L. and *Coffea canephora* Pierre ex A.Froehner; Pompelli et al.

(2012), 415 leaves for physic nut (*Jatropha curcas* L.); Cargnelutti Filho et al. (2015), 200 leaves for jack bean [*Canavalia ensiformis* (L.) DC.]; and Cargnelutti Filho et al. (2018), 240 leaves for velvet bean (*Stizolobium cinereum* Piper & Tracy).

According to Knofczynski & Mundfrom (2008) and Bujang et al. (2017), in multiple linear regression, sample size varies according to effect size and the number of independent variables. Knofczynski & Mundfrom (2008) found a negative exponential relationship between the squared multiple correlation coefficient and the minimum sample size, i.e., as the squared multiple correlation coefficient decreases, the sample size increases. Furthermore, Kelley (2008) showed how the population squared multiple correlation coefficients, desired confidence interval width, and number of regressor variables affected the necessary sample size for multiple linear regression. Hanley (2016) highlighted differences in sample size for Y regressions as a function of controlled (exposure) or uncontrolled (nonexperimental) X values in multiple linear regression.

In the sampling design used to determinate the squared multiple correlation (ρ^2) in multiple linear regression, Bonett & Wright (2011, 2014) emphasized the importance of adopting sample size planning formulas to obtain an acceptably accurate estimate of ρ^2. In addition, Shieh (2013) showed the importance of computationally intensive and simulation-based methods to determine this statistic. According to Knofczynski & Mundfrom (2008) and Bonett & Wright (2014), the different sample size recommendations for ρ^2 and/or multiple linear regression are associated with the different criteria adopted by each researcher. However, there are no know studies in the literature on the sample size recommended for multiple linear regression in corn.

The objective of this work was to determine the number of plants required to model corn grain yield (Y) as a function of ear length (X_1) and ear diameter (X_2), using the multiple regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Materials and Methods

Two experiments with corn were carried out in an area located in the municipality of Santa Maria, in the state of Rio Grande do Sul, Brazil (29º42'S, 53º49'W, at 95 m altitude). The first was conducted

in the 2008/2009 crop year, and the second, in the 2009/2010 crop year. According to Köppen-Geiger's classification, the climate of the region is Cfa, subtropical humid (Alvares et al., 2013). The soil is an Argissolo Vermelho Distrófico arênico (Santos et al., 2013), i.e., a dystrophic sandy Argisol.

In the first experiment, sowing was performed on 12/26/2008. Four plots were sown with the P32R21 single-cross hybrid, four with the DKB566 three-way cross hybrid, and four with the DKB747 double-cross hybrid. In the second experiment, sowing was carried out on 10/26/2009. Sixteen plots were sown with the 30F53 single-cross hybrid, 16 with the DKB566 three-way cross hybrid, and 16 with the DKB747 double-cross hybrid.

Each plot consisted of four 6.0-m rows, 0.8 m apart, with density adjusted to five plants per row meter, representing a density of 62,500 plants per hectare. Therefore, each plot consisted of 120 plants, totaling: 1,440 plants in the first experiment, with 3 hybrids × 4 plots per hybrid × 120 plants per plot; and 5,760 plants in the second, with 3 hybrids × 16 plots per hybrid × 120 plants per plot. In each crop year, plots of the single-, three-way, and double-cross hybrids were randomized in the experimental area. In both experiments, basic fertilization was 750 kg ha^{-1} of the 3-24-18 (N-P$_2$O$_5$-K$_2$O) formula, and topdressing was 300 kg ha^{-1} urea with 45% N. The other cultural practices were performed according to the recommendations for corn (Fancelli & Dourado Neto, 2004).

In the first experiment, 361, 373, and 416 plants were assessed, respectively, for single-, three-way, and double-cross hybrids. In the second, 1,777, 1,693, and 1,720 plants were evaluated, respectively, for single-, three-way, and double-cross hybrids. Therefore, a total of 6,340 plants were measured for the following traits: ear length (X_1, in mm), ear diameter (X_2, in mm), and grain yield (Y, in grams per plant). Since only plants that presented the three traits were assessed, the final number of plants varied between plots and hybrids.

For each trait (X_1, X_2, and Y) of each hybrid in each experiment and for all hybrids and experiments (overall, n=6,340 plants), the following statistics were calculated: mean, median, minimum, maximum, standard deviation (SD), coefficient of variation (CV), skewness, and kurtosis. Pearson's linear correlation matrix between traits also was estimated.

From the overall data set of 6,340 plants, frequency histograms and scatterplots were created. Then, Y was adjusted as a function of X_1 and X_2 by the multiple regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where β_0, β_1, and β_2 are the regression parameters; and ε is the residue or error of regression. The decision to use all plants (n=6,340 plants) was based on the similarity between hybrids and experiments (six cases) regarding the measures of central tendency and variability and the coefficients of skewness, kurtosis, and correlation, and also on the aim to increase the representativeness of the data set and sample size.

The sample size (number of plants) required to adjust Y as a function of X_1 and X_2 in the multiple regression model was determined through resampling with replacement. For resampling, 991 sample sizes were planned, with an initial sample size of 10 plants, considered as a reference, i.e., the minimum size required for model adjustment. The other sizes were obtained in increments of one unit, until reaching 1,000 plants; therefore, sample sizes of 10 to 1,000 plants were planned.

For each planned sample size, 3,000 resamples with replacement were obtained. For each resample, the estimates of the β_0, β_1, and β_2 parameters of the used multiple regression model, the residual standard error (RSE), and the coefficient of determination (R^2) were calculated. The degree of multicollinearity between the explanatory traits of the model (X_1 and X_2) was evaluated based on the variance inflation factor (VIF) and condition number (CN). The VIF was obtained by: $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is the multiple determination coefficient of X_i over the other explanatory traits. The CN was calculated by the ratio between the highest (λ_{max}) and lowest eigenvalue (λ_{min}) of the correlation matrix between the explanatory traits ($CN = \lambda_{max}/\lambda_{min}$). Multicollinearity between traits is considered: low, when $CN \leq 100$; moderate to high, when $100 < CN < 1,000$; and severe, when $CN \geq 1,000$; when the VIF is greater than 10, multicollinearity is also considered severe (Montgomery et al., 2012). Therefore, for each sample size, 3,000 estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN were obtained, and the 2.5% percentile ($P_{2.5\%}$), mean, and 97.5% percentile ($P_{97.5\%}$) were determined. The amplitude of the 95% confidence interval was calculated by the expression: $ACI = P_{97.5\%} - P_{2.5\%}$.

It should be interpreted that the smaller the ACI, the more accurate are the estimates of β_0, β_1, β_2, RSE,

R^2, VIF, and CN, which would allow determining the number of plants required to achieve the desired ACI values for these parameters. However, there are no values for β_0, β_1, β_2, RSE, R^2, VIF, and CN that can be taken as a reference. Therefore, the following statistical criterion was used to define sample size: initially, the ACI obtained with the smaller sample size of 10 plants (ACI_{10}) was considered as a reference for β_0, β_1, β_2, RSE, R^2, VIF, and CN; that is, it was considered as 100% (maximum ACI and, therefore, with minimum accuracy in the estimates of these parameters). The accuracy gain (AG_i, in %) was then calculated with the addition of ith plants (i = 1, 2, ..., 990 plants, respectively, for sample sizes 11, 12, ..., 1,000 plants), using the expression: $AG_i = 100 - (ACI_i/ ACI_{10}) \times 100$, where ACI_i is the amplitude of the 95% confidence interval of the sample sizes of 11, 12, ..., 1,000 plants.

Sample size (number of plants) was considered as the one in which the gain in accuracy for β_0, β_1, β_2, RSE, R^2, VIF, and CN was at least 80%. This minimum value was determined because, above it, accuracy gains became less expressive and tended to stabilize, requiring a high investment for the evaluation of a larger number of plants and indicating a low accuracy gain. The results obtained in the present study can be used by other researchers to define sample size according to the desired accuracy gains.

The 2.5% percentile, mean, 97.5% percentile, and accuracy gain of the sample sizes of β_0, β_1, β_2, RSE, R^2, VIF, and CN were plotted in graphs for a better visual representation. The ACI and accuracy gain were presented at an interval of 20 plants, to reduce the dimensionality of the results, still keeping them sufficiently informative. The statistical analysis was performed using Microsoft Office Excel and the R software (R Core Team, 2019).

Results and Discussion

The minimum and maximum values of X_1 were similar between the six experimental cases (28 $\leq$ minimum $\leq$ 56; 211 $\leq$ maximum $\leq$ 281) (Table 1), and a similar pattern was observed for X_2 and Y. The values of the SD and CV of each trait were also similar among the six cases, oscillating between 26.34 $\leq$ SD $\leq$ 41.80 and 16.70% $\leq$ CV $\leq$ 26.35% for X_1, 3.52 $\leq$ SD $\leq$ 4.90 and 7.73% $\leq$ CV $\leq$ 12.23% for X_2, and

40.52 $\leq$ SD $\leq$ 55.84 and 31.86% $\leq$ CV $\leq$ 46.91% for Y; however, among traits, SD and CV increased in the following order: X_2, X_1, and Y. In all cases, for the three traits, the values of skewness and kurtosis were close to zero and the median and mean were similar, indicating good adherence of these data to the normal distribution curve.

In the six cases, Pearson's linear correlation coefficients (r) between the pairs of traits were positive and similar, oscillating within the following limits: 0.77 $\leq$ r $\leq$ 0.91 for $Y \times X_1$; 0.81 $\leq$ r $\leq$ 0.86 for $Y \times X_2$; and 0.56 $\leq$ r $\leq$ 0.76 for $X_1 \times X_2$ (Table 1). These coefficients revealed that larger ears, i.e., ears with greater length and greater diameter, presented higher grain yield and vice versa. In this sense, Toebe et al. (2017), in the path analysis, pointed out the importance of measuring ear length and ear diameter to predict corn grain yield.

As previously mentioned, the use of the overall data set of 6,340 plants as a sample size is justified by the similar pattern observed between hybrids and experiments (six cases) for measures of central tendency and variability and for the coefficients of skewness, kurtosis, and correlation, as well as by the better representativeness of the sample. The data set of 6,340 plants allows visualizing the reflex of the similarity between the six cases in relation to data variability and distribution and to the linear relationship between traits (Table 1 and Figure 1).

The r between $Y \times X_1$ (r = 0.71) and $Y \times X_2$ (r = 0.84) (Table 1) and the scatterplots between these pairs of traits (Figure 1) showed a linear association pattern. This is an indicative of the adequacy of the adopted multiple regression model. Moreover, the positive linear association between $X_1 \times X_2$ (r = 0.53) indicated that it is necessary to investigate the degree of multicollinearity in the correlation matrix of these explanatory traits. Regarding the CVs, the obtained values for X_1, X_2, and Y were 21.96, 12.41, and 44.15%, respectively. High CV values are important for modeling, since they show a wide variability among corn ears in the dataset (n=6,340 plants), increasing the representativity of the multiple regression model of Y as a function of X_1 and X_2.

Based on 6,340 plants, the estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN were -229.76, 0.54, 6.16, 22.13, 0.80, 1.39, and 3.25, respectively. For the 3,000 samples of 10 plants (smaller size used), the ACI was 329.86, 1.33, 8.77, 25.86, 0.43, 4.16, and 17.49, and the average of the 3,000 samples was -251.79, 0.60, 6.47, 19.76, 0.85,

1.85, and 5.05, respectively, for the estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN (Table 2 and Figure 2). For the 3,000 samples of 1,000 plants (largest size used), the ACI was 31.87, 0.11, 0.81, 2.72, 0.05, 0.22, and 0.97, and the average of the 3,000 samples was -229.87, 0.54, 6.16, 22.11, 0.80, 1.39, and 3.27, respectively, for the estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN. Visually, it was observed that, with the increase in the number of plants, the mean of the 3,000 estimates of the assessed parameters stabilizes and approaches the averages obtained with the 6,340 plants. This suggests a possible bias in the estimates of the mean in the case

of sample insufficiency. A similar result was also presented graphically by Toebe et al. (2017) for the estimate of the direct effect of ear insertion height on corn grain yield, using the path analysis.

The highest amplitude was observed for the confidence interval of the β_0, β_1, β_2, RSE, R^2, VIF, and CN from 10 plants, when compared with 1,000 plants. Therefore, with 10 plants, the estimates of the parameters of the model were less accurate, which may result in inaccurate estimates of grain yield and in bias when the sample is insufficient. Therefore, it can be inferred that models fitted from a small number

Table 1. Mean, median, minimum, maximum, standard deviation (SD), coefficient of variation (CV), skewness, and kurtosis of three traits measured in corn (*Zea mays*) hybrids, as well as Pearson's linear correlation matrix between traits.

Traits[1]	Mean	Median	Minimum	Maximum	SD	CV (%)	Skewness	Kurtosis	Correlation[1]		
									X_1	X_2	Y
P32R21 single-cross hybrid (n=361 plants) in the 2008/2009 crop year											
X_1	146.66	148.00	56	211	29.79	20.31	-0.42	0.04	1.00	0.56	0.78
X_2	48.10	49.00	31	56	4.14	8.60	-1.18	1.71	0.56	1.00	0.83
Y	131.44	135.00	2	276	55.84	42.48	-0.08	-0.37	0.78	0.83	1.00
DKB566 three-way cross hybrid (n=373 plants) in the 2008/2009 crop year											
X_1	157.74	162.00	50	226	26.34	16.70	-0.84	1.55	1.00	0.66	0.83
X_2	46.44	47.00	30	55	4.02	8.67	-1.03	1.99	0.66	1.00	0.83
Y	153.25	155.00	2	273	54.45	35.53	-0.42	-0.13	0.83	0.83	1.00
DKB747 double-cross hybrid (n=416 plants) in the 2008/2009 crop year											
X_1	165.62	171.00	55	226	31.26	18.87	-0.68	0.42	1.00	0.61	0.84
X_2	45.56	46.00	31	54	3.52	7.73	-0.72	1.31	0.61	1.00	0.81
Y	144.85	149.00	10	259	46.15	31.86	-0.42	0.00	0.84	0.81	1.00
30F53 single-cross hybrid (n=1,777 plants) in the 2009/2010 crop year											
X_1	147.95	152.00	40	221	31.80	21.49	-0.48	-0.04	1.00	0.76	0.91
X_2	42.74	43.00	17	53	4.28	10.02	-1.13	2.39	0.76	1.00	0.86
Y	115.68	117.00	2	249	44.16	38.18	-0.15	-0.36	0.91	0.86	1.00
DKB566 three-way cross hybrid (n=1,693 plants) in the 2009/2010 crop year											
X_1	154.61	157.00	40	250	27.87	18.03	-0.71	1.05	1.00	0.62	0.77
X_2	41.24	42.00	18	52	4.90	11.89	-1.24	2.42	0.62	1.00	0.86
Y	116.62	121.00	2	245	48.49	41.58	-0.24	-0.51	0.77	0.86	1.00
DKB747 double-cross hybrid (n=1,720 plants) in the 2009/2010 crop year											
X_1	158.66	163.00	28	281	41.80	26.35	-0.34	-0.19	1.00	0.62	0.77
X_2	39.05	40.00	15	51	4.78	12.23	-0.98	1.85	0.62	1.00	0.82
Y	86.37	87.00	2	219	40.52	46.91	0.02	-0.49	0.77	0.82	1.00
Overall (n=6,340 plants)											
X_1	154.29	157.00	28	281	33.89	21.96	-0.40	0.35	1.00	0.53	0.71
X_2	42.05	43.00	15	56	5.22	12.41	-0.79	1.44	0.53	1.00	0.84
Y	113.00	114.00	2	276	49.89	44.15	0.02	-0.40	0.71	0.84	1.00

[1]X_1, ear length, in millimeters; X_2, ear diameter, in millimeters; and Y, grain yield, in grams per plant.
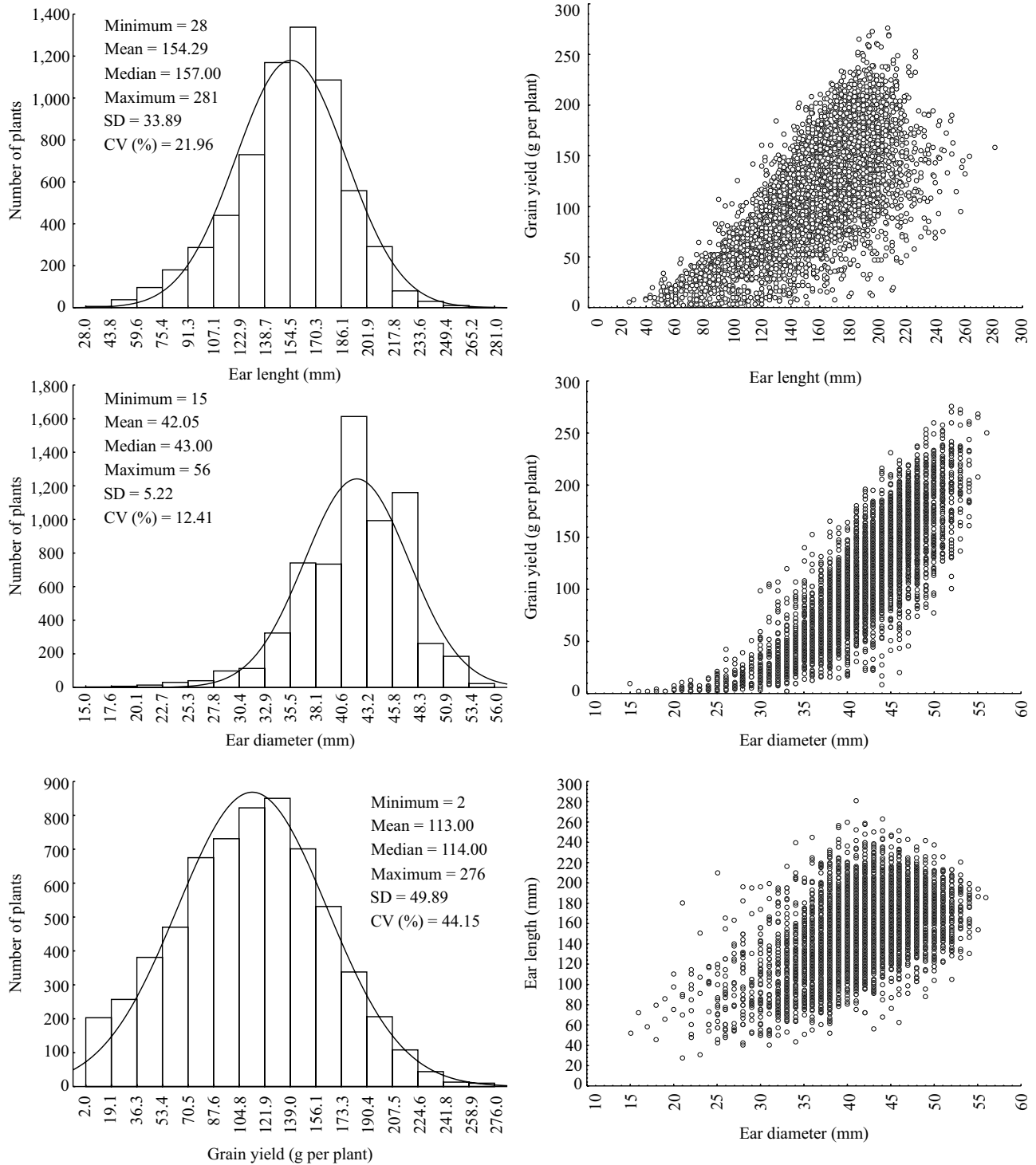
Figure 1. Frequency histograms (on the left side) and scatterplots (on the right side) of the three evaluated traits measured in 6,340 corn (*Zea mays*) hybrid plants. In histograms, the line represents the normal distribution curve. The 6,340 plants are composed of 361 P32R21 hybrids, 373 DKB566 hybrids, and 416 DKB747 hybrids in the 2008/2009 crop year; and of 1,777 30F53 hybrids, 1,693 DKB566 hybrids, and 1,720 DKB747 hybrids in the 2009/2010 crop year. SD, standard deviation; and CV, coefficient of variation.

Table 2. Amplitude of the 95% confidence interval (ACI$_i$) and accuracy gain (AG$_i$, %) of the estimates of the β_0, β_1, and β_2 parameters of the multiple regression model for grain yield (Y, g per plant) as a function of ear length (X$_1$, mm) and ear diameter (X$_2$, mm), as well as residual standard error (RSE), coefficient of determination (R^2), variance inflation factor (VIF), and condition number (CN) between the explanatory traits of the model (X$_1$ and X$_2$), considering sample sizes of 10 to 1,000 corn (*Zea mays*) plants of the P32R21, DKB566, and DKB747 hybrids in the 2008/2009 crop year and of the 30F53, DKB566, and DKB747 hybrids in the 2009/2010 crop year.

Plants	β_0		β_1		β_2		RSE		R^2		VIF		CN	
	ACI$_i$	AG$_i$[(1)]	ACI$_i$	AG$_i$	ACI$_i$	AG$_i$	ACI$_i$	AG$_i$	ACI$_i$	AG$_i$	ACI$_i$	AG$_i$	ACI$_i$	AG$_i$
10	329.86	-	1.33	-	8.77	-	25.86	-	0.43	-	4.16	-	17.49	-
30	182.82	44.58	0.71	46.91	4.84	44.77	15.13	41.50	0.25	41.32	1.55	62.73	6.81	61.05
50	143.35	56.54	0.52	60.99	3.81	56.54	11.71	54.70	0.20	53.36	1.02	75.44	4.50	74.25
70	118.90	63.95	0.43	67.46	3.19	63.64	9.99	61.37	0.16	62.65	0.90	78.29	3.97	77.28
90	105.20	68.11	0.39	70.90	2.79	68.14	8.91	65.54	0.16	63.81	0.76	81.62	3.37	80.72
110	93.92	71.53	0.35	73.98	2.45	72.02	7.79	69.86	0.13	69.35	0.67	83.88	2.96	83.06
130	86.10	73.90	0.31	76.48	2.29	73.83	7.27	71.88	0.12	71.30	0.60	85.51	2.66	84.78
150	82.28	75.06	0.29	78.22	2.15	75.48	6.61	74.42	0.12	73.00	0.57	86.27	2.53	85.53
170	74.95	77.28	0.27	79.69	2.05	76.59	6.39	75.27	0.11	75.37	0.54	86.93	2.40	86.30
190	72.58	78.00	0.26	80.62	1.89	78.42	5.99	76.82	0.11	75.72	0.51	87.72	2.26	87.11
210	70.57	78.61	0.24	81.65	1.84	78.97	6.02	76.70	0.10	77.31	0.48	88.34	2.14	87.77
230	68.26	79.31	0.23	82.44	1.75	80.04	5.63	78.21	0.09	78.42	0.46	88.99	2.02	88.46
250	61.69	81.30	0.23	82.72	1.63	81.36	5.18	79.98	0.09	79.71	0.44	89.42	1.94	88.92
270	60.85	81.55	0.21	83.92	1.57	82.13	5.16	80.05	0.09	80.18	0.42	89.92	1.85	89.40
290	59.45	81.98	0.21	84.19	1.57	82.13	4.91	81.01	0.08	80.89	0.39	90.52	1.74	90.06
310	56.11	82.99	0.20	85.03	1.51	82.76	4.71	81.78	0.08	82.07	0.40	90.49	1.74	90.03
330	57.05	82.70	0.20	85.20	1.45	83.40	4.56	82.38	0.08	81.82	0.38	90.76	1.70	90.31
350	53.62	83.74	0.19	85.51	1.41	83.93	4.48	82.69	0.07	83.05	0.37	91.20	1.61	90.78
370	51.31	84.45	0.19	85.77	1.36	84.48	4.32	83.29	0.07	82.73	0.37	91.11	1.63	90.67
390	50.50	84.69	0.19	86.12	1.30	85.14	4.31	83.33	0.07	82.76	0.36	91.37	1.58	90.95
410	51.43	84.41	0.18	86.86	1.30	85.21	4.20	83.77	0.07	83.66	0.35	91.69	1.53	91.27
430	48.05	85.43	0.17	87.34	1.27	85.57	3.95	84.73	0.07	84.07	0.33	92.09	1.45	91.70
450	48.36	85.34	0.17	87.09	1.26	85.67	3.92	84.86	0.07	84.47	0.33	92.08	1.45	91.69
470	46.05	86.04	0.17	87.52	1.19	86.46	3.91	84.90	0.07	84.56	0.30	92.72	1.34	92.35
490	46.84	85.80	0.16	87.87	1.22	86.06	3.82	85.24	0.06	85.51	0.32	92.22	1.43	91.84
510	43.12	86.93	0.16	88.21	1.14	87.01	3.68	85.78	0.06	85.32	0.31	92.62	1.35	92.26
530	43.88	86.70	0.16	88.33	1.17	86.70	3.66	85.83	0.06	85.74	0.30	92.81	1.32	92.46
550	44.12	86.62	0.15	88.54	1.15	86.89	3.48	86.54	0.06	86.47	0.29	92.95	1.29	92.60
570	41.78	87.33	0.15	88.93	1.12	87.25	3.49	86.52	0.06	86.47	0.28	93.16	1.26	92.82
590	40.60	87.69	0.15	88.74	1.08	87.64	3.34	87.08	0.06	86.51	0.29	92.95	1.29	92.62
610	40.83	87.62	0.15	89.11	1.06	87.95	3.43	86.75	0.06	86.51	0.29	93.09	1.27	92.75
630	40.48	87.73	0.14	89.56	1.03	88.20	3.28	87.31	0.06	86.97	0.27	93.57	1.18	93.25
650	40.77	87.64	0.14	89.24	1.05	88.06	3.32	87.18	0.06	87.06	0.27	93.55	1.18	93.24
670	38.91	88.20	0.14	89.52	1.03	88.21	3.19	87.67	0.05	87.65	0.27	93.39	1.21	93.07
690	38.10	88.45	0.14	89.79	0.97	88.93	3.13	87.90	0.05	87.62	0.26	93.82	1.13	93.52
710	38.25	88.40	0.13	90.00	1.00	88.57	3.16	87.79	0.05	87.83	0.26	93.75	1.15	93.43
730	36.38	88.97	0.13	89.89	0.98	88.82	3.08	88.10	0.05	87.92	0.25	93.95	1.11	93.66
750	35.38	89.27	0.13	90.08	0.96	89.10	3.05	88.20	0.05	87.85	0.24	94.18	1.07	93.89

Continuation...

Table 2. Continuation...

Plants	β_0		β_1		β_2		RSE		R^2		VIF		CN	
	ACI_i	$AG_i^{(1)}$	ACI_i	AG_i	ACI_i	AG_i	ACI_i	AG_i	ACI_i	AG_i	ACI_i	AG_i	ACI_i	AG_i
770	35.40	89.27	0.13	90.25	0.94	89.30	3.02	88.34	0.05	87.79	0.25	94.09	1.08	93.80
790	34.99	89.39	0.13	90.49	0.92	89.49	2.92	88.72	0.05	88.66	0.24	94.20	1.06	93.92
810	34.18	89.64	0.12	90.73	0.93	89.39	2.93	88.66	0.05	88.48	0.25	94.08	1.09	93.79
830	36.10	89.06	0.12	90.68	0.91	89.56	2.91	88.73	0.05	88.47	0.23	94.38	1.03	94.09
850	35.00	89.39	0.13	90.49	0.92	89.49	2.90	88.80	0.05	88.59	0.24	94.24	1.06	93.96
870	34.15	89.65	0.12	90.91	0.92	89.52	2.83	89.04	0.05	89.29	0.23	94.36	1.03	94.09
890	34.08	89.67	0.12	90.96	0.90	89.79	2.88	88.87	0.05	89.08	0.23	94.51	1.01	94.24
910	33.48	89.85	0.12	91.25	0.88	90.01	2.79	89.20	0.05	89.04	0.22	94.62	0.99	94.35
930	33.64	89.80	0.12	91.24	0.86	90.23	2.75	89.35	0.05	89.09	0.24	94.32	1.04	94.04
950	31.48	90.46	0.11	91.41	0.85	90.32	2.72	89.49	0.05	89.29	0.23	94.49	1.01	94.22
970	31.70	90.39	0.11	91.42	0.84	90.44	2.65	89.75	0.04	89.65	0.21	94.85	0.95	94.59
990	32.29	90.21	0.11	91.57	0.84	90.47	2.56	90.12	0.05	89.47	0.22	94.78	0.96	94.52
1,000	31.87	90.34	0.11	91.65	0.81	90.75	2.72	89.49	0.05	89.47	0.22	94.71	0.97	94.45

[1]$AG_i = 100 - (ACI_i/ACI_{10}) \times 100$, where ACI_i is the amplitude of the 95% confidence interval of the sample sizes of 10, 30, ..., 1,000 plants, and ACI_{10} is the amplitude of the 95% confidence interval for the reference sample size of 10 plants.

of plants should not be used in studies of grain yield prediction, showing the importance and need to set the reference sample size for precise model adjustments.

The ACI of the estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN decreased gradually with the increase in the number of plants (Table 2 and Figure 2). This result was expected and indicates that increasing the number of plants improves the accuracy of estimates and, consequently, the reliability of the models, as already verified for Pearson's linear correlations (Toebe et al., 2015) and the path analysis (Toebe et al., 2017) in corn. However, a sharp decrease in the ACI to approximately 260 plants was also observed (Figure 2), becoming less marked afterwards, which indicates that measuring more plants would result in inexpressive benefits in the accuracy of model parameter estimates. Therefore, for the estimates of β_0, β_1, β_2, RSE, R^2, VIF, and CN, it can be suggested visually that 260 plants would be sufficient to fit the multiple regression model. In other bi- and multivariate techniques, the variability of sample size was considered a function of the magnitude of associations and of combinations of variables, years, hybrids, and pre-established levels of precision. Toebe et al. (2015) recommended from 120 to 375 plants, depending on the level of precision, for the estimation of Pearson's linear correlations in corn harvest and hybrids. Toebe et al. (2017) suggested 10 to 530 plants to estimate the direct effects of the path

analysis, depending on the type of hybrid, harvest, scenario, path analysis, and explanatory variable.

In multiple linear regression, according to Knofczynski & Mundfrom (2008), the sample size increased more quickly for models with larger numbers of predictor variables than for those with fewer predictor variables, as the squared multiple correlation coefficient decreased. The authors also concluded that the sample size for an excellent prediction level and two predictor variables ranged from 15 to 950 observations, depending on the population squared multiple correlation coefficients. Boutilier et al. (2016), testing four statistical models, recommended more than 200 samples to achieve consistent model predictions for all metrics. Bujang et al. (2017) suggested 300 or more subjects to generate an approximation of estimates with parameters. The sample sizes recommended by Boutilier et al. (2016) and Bujang et al. (2017) were similar to those obtained in the present work. Riley et al. (2019) suggested at least 36.7 subjects per predictor parameter, whereas Kelley (2008) found the need for up to 3,653 observations in multiple linear regression, depending on the effect of the population squared multiple correlation coefficient, desired confidence interval width, and number of variables.

When increasing the number of plants from 10 to 30, there were accuracy gains of 44.58, 46.91, 44.77, 41.50, 41.32, 62.73, and 61.05% for the estimates of β_0, β_1, β_2,
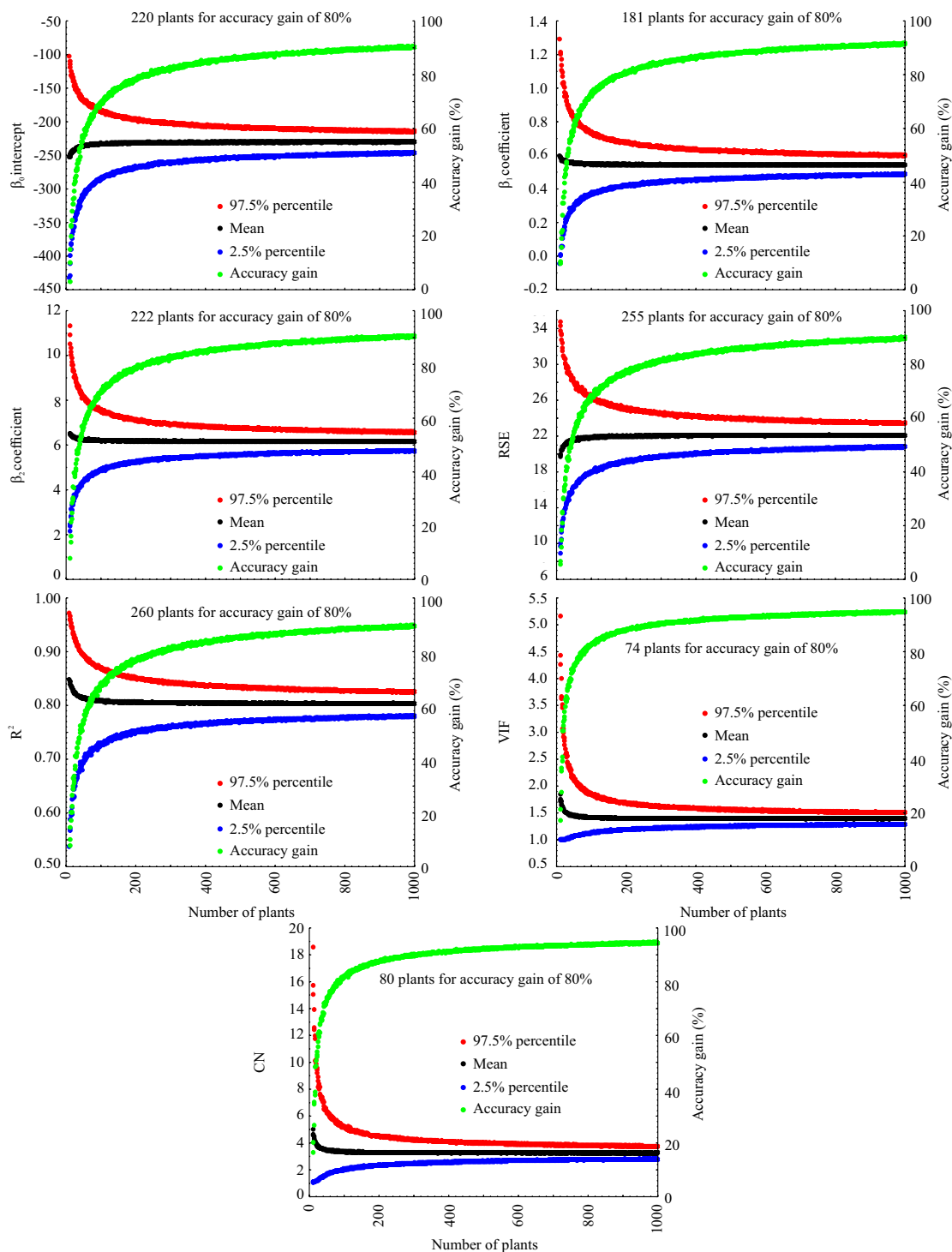
Figure 2. 2.5% percentile, 97.5% percentile, and mean (on the left Y-axis), as well as accuracy gain (on the right Y-axis) for 3,000 estimates of parameters β_0, β_1, β_2, RSE, R^2, VIF, and CN in the 2008/2009 and 2099/2010 crop years. On the X-axis, the number of corn plants ranges from 10 to 1,000. Plants of the P32R21, DKB566, and DKB747 hybrids were evaluated in the 2008/2009 crop year, and of the 30F53, DKB566, and DKB747 hybrids in the 2009/2010 crop year. β_0, β_1, β_2, regression parameters; RSE, residual standard error; R^2, coefficient of determination; VIF, variance inflation factor; and CN, condition number.

RSE, R^2, VIF, and CN, respectively (Table 2). From 10 to 50 plants, the gains in accuracy were, respectively, 56.54, 60.99, 56.54, 54.70, 53.36, 75.44, and 74.25%. Therefore, accuracy gains, with the increase in the number of plants, were of similar magnitudes for the estimates of β_0, β_1, β_2, RSE, and R^2, and relatively superior for those of VIF and CN.

In addition, gains in accuracy were more expressive from 10 to 30 plants than from 30 to 50 plants, and so on, successively (Figure 2). Gains over 80% (β_0 = 81.63%; β_1 = 82.89%; β_2 = 81.90%; RSE = 80.05%; R^2 = 80.24%; VIF = 89.70%; and CN = 89.19%) were obtained for 10 to 261 plants. Although estimates from the largest possible number of plants should be sought in order to ensure reliable models, the obtained results are indicative that the studied model parameters may be estimated with 260 corn plants; however, from this number of plants, accuracy gains were inexpressive. Sample sizes (number of leaves) similar to this one were recommended for the adjustment of leaf area models: 200 leaves by Antunes et al. (2008) for two species of coffee, 415 leaves by Pompelli et al. (2012) for physic nut, 200 leaves by Cargnelutti Filho et al. (2015) for jack bean, and 240 leaves by Cargnelutti Filho et al. (2018) for velvet bean. In this sense, it is important to recommend sample sizes that can be evaluated, because, as already shown by Toebe et al. (2015, 2017), Kelley (2008) and Knofczynski & Mundfrom (2008), in situations of excellent prediction level, in general, impractical sample sizes (n > 1,000) are necessary.

Models adjusted from small samples – less than 260 plants in the present study – should be avoided due to the imprecision of the obtained estimates, whereas those adjusted from larger samples – equal to or greater than 260 plants – should be encouraged. It should be noted that, from a given sample size (number of plants), gains are negligible in relation to the costs for measuring plant traits. Considering the obtained results and the inferences mentioned above, it is reasonable to accept that 260 plants are sufficient to adjust corn grain yield (Y) as a function of ear length (X_1) and ear diameter (X_2) by the multiple regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

Conclusions

1. Measuring 260 plants is sufficient to adjust precise multiple regression models of corn (*Zea mays*) grain yield (Y, in g per plant) as a function of ear length (X_1, in mm) and ear diameter (X_2, in mm).

2. The model $Y = -229.76 + 0.54X_1 + 6.16X_2$ is a reference for estimating corn grain yield.

References

ALVARES, C.A.; STAPE, J.L.; SENTELHAS, P.C.; GONÇALVES, J.L. de M.; SPAROVEK, G. *Köppen's climate classification map for Brazil.* **Meteorologische Zeitschrift**, v.22, p.711-728, 2013. DOI: https://doi.org/10.1127/0941-2948/2013/0507.

ALVES, B.M.; CARGNELUTTI FILHO, A.; TOEBE, M.; BURIN, C. Linear relations among phenological, morphological, productive and protein-nutritional traits in early maturing and super-early maturing maize genotypes. **Journal of Cereal Science**, v.70, p.229-239, 2016. DOI: https://doi.org/10.1016/j.jcs.2016.06.013.

ANTUNES, W.C.; POMPELLI, M.F.; CARRETERO, D.M.; DAMATTA, F.M. Allometric models for non-destructive leaf area estimation in coffee (*Coffea arabica* and *Coffea canephora*). **Annals of Applied Biology**, v.153, p.33-40, 2008. DOI: https://doi.org/10.1111/j.1744-7348.2008.00235.x.

BONETT, D.; WRIGHT, T. Sample size planning for multiple correlation: reply to Shieh (2013). **Psicothema**, v.26, p.391-394, 2014. DOI: https://doi.org/10.7334/psicothema2013.309.

BONETT, D.G.; WRIGHT, T.A. Sample size requirements for multiple regression interval estimation. **Journal of Organizational Behavior**, v.32, p.822-830, 2011. DOI: https://doi.org/10.1002/job.717.

BOUTILIER, J.J.; CRAIG, T.; SHARPE, M.B.; CHAN, T.C.Y. Sample size requirements for knowledge based treatment planning. **Medical Physics**, v.43, p.1212-1221, 2016. DOI: https://doi.org/10.1118/1.4941363.

BUJANG, M.A.; SA'AT, N.; SIDIK, T.M.I.T.A.B. Determination of minimum sample size requirement for multiple linear regression and analysis of covariance based on experimental and non-experimental studies. **Epidemiology Biostatistics and Public Health**, v.14, e12117, 2017. DOI: https://doi.org/10.2427/12117.

CARGNELUTTI FILHO, A.; TOEBE, M.; BURIN, C.; ALVES, B.M.; NEU, I.M.M. Number of leaves needed to model leaf area

in jack bean plants using leaf dimensions. **Bioscience Journal**, v.31, p.1651-1662, 2015. DOI: https://doi.org/10.14393/BJ-v31n6a2015-26135.

CARGNELUTTI FILHO, A.; TOEBE, M.; BURIN, C.; NEU, I.M.M; ALVES, B.M. Número de folhas para modelar a área foliar de mucuna cinza por dimensões foliares. **Revista de Ciências Agroveterinárias**, v.17, p.571-578, 2018. DOI: https://doi.org/10.5965/223811711732018571.

FANCELLI, A.L.; DOURADO NETO, D. **Produção de milho**. Guaíba: Agropecuária, 2004. 360p.

GE, Y.; WU, H. Prediction of corn price fluctuation based on multiple linear regression analysis model under big data. **Neural Computing and Applications**, p.1-13, 2019. DOI: https://doi.org/10.1007/s00521-018-03970-4.

HANLEY, J.A. Simple and multiple linear regression: sample size considerations. **Journal of Clinical Epidemiology**, v.79, 112-119, 2016. DOI: https://doi.org/10.1016/j.jclinepi.2016.05.014.

KELLEY, K. Sample size planning for the squared multiple correlation coefficient: accuracy in parameter estimation via narrow confidence intervals. **Multivariate Behavioral Research**, v.43, p.524-555, 2008. DOI: https://doi.org/10.1080/00273170802490632.

KNOFCZYNSKI, G.T.; MUNDFROM, D. Sample sizes when using multiple linear regression for prediction. **Educational and Psychological Measurement**, v.68, p.431-442, 2008. DOI: https://doi.org/10.1177/0013164407310131.

LAURIE, C.C.; CHASALOW, S.D.; LEDEAUX, J.R.; MCCARROLL, R.; BUSH, D.; HAUGE, B.; LAI, C.; CLARK, D.; ROCHEFORD, T.R.; DUDLEY, J.W. The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. **Genetics**, v.168, p.2141-2155, 2004. DOI: https://doi.org/10.1534/genetics.104.029686.

MOHAMMADI, G.R. Growth parameters enhancing the competitive ability of corn (*Zea mays* L.) against weeds. **Weed Biology and Management**, v.7, p.232-236, 2007. DOI: https://doi.org/10.1111/j.1445-6664.2007.00261.x.

MONTGOMERY, D.C.; PECK, E.A.; VINNING, G.G. **Introduction to linear regression analysis**. 5th ed. New York: J.Wiley & Sons, 2012. 672p.

POMPELLI, M.F.; ANTUNES, W.C.; FERREIRA, D.T.R.G.; CAVALCANTE, P.G.S.; WANDERLEY-FILHO, H.C.L.; ENDRES, L. Allometric models for non-destructive leaf area estimation of *Jatropha curcas*. **Biomass and Bioenergy**, v.36, p.77-85, 2012. DOI: https://doi.org/10.1016/j.biombioe.2011.10.010.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2019. Available at: <http://www.R-project.org>. Accessed on: Mar. 8 2019.

RILEY, R.D.; SNELL, K.I.E.; ENSOR, J.; BURKE, D.L.; HARRELL JR, F.E.; MOONS, K.G.M.; COLLINS, G.S. Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. **Statistics in Medicine**, v.38, p.1262-1275, 2019. DOI: https://doi.org/10.1002/sim.7993.

SANTOS, H.G. dos; JACOMINE, P.K.T.; ANJOS, L.H.C. dos; OLIVEIRA, V.A. de; LUMBRERAS, J.F.; COELHO, M.R.; ALMEIDA, J.A. de; CUNHA, T.J.F.; OLIVEIRA, J.B. de. **Sistema brasileiro de classificação de solos**. 3.ed. rev. e ampl. Brasília: Embrapa, 2013. 353p.

SHIEH, G. Sample size requirements for interval estimation of the strength of association effect sizes in multiple regression analysis. **Psicothema**, v.25, p.402-407, 2013. DOI: https://doi.org/10.7334/psicothema2012.221.

TOEBE, M.; CARGNELUTTI FILHO, A.; LOPES, S.J.; BURIN, C.; SILVEIRA, T.R. da; CASAROTTO, G. Sample size in the estimation of correlation coefficients for corn hybrids in crops and accuracy levels. **Bragantia**, v.74, p.16-24, 2015. DOI: https://doi.org/10.1590/1678-4499.0324.

TOEBE, M.; CARGNELUTTI FILHO, A.; STORK, L.; LÚCIO, A.D. Sample size for estimation of direct effects in path analysis of corn. **Genetics and Molecular Research**, v.16, gmr16029523, 2017. DOI: https://doi.org/10.4238/gmr16029523.

USDA. United States Department of Agriculture. **World Agricultural Production**. 2019. 33p. (USDA. Circular Series WAP 11-19). Available at: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf>. Accessed on: Mar. 8 2019.