ISSN: 2538-2152; ECOPERSIA 2024;12(1):39-53.

Determining the Minimum Sample Size Required to Create Regression Model in Plant Ecology (Case Study: Cover-Production Relationship)

ARTICLEINFO

Article Type
Original Research

Authors
Moslem Rostampour, *Ph.D.*[1*]
Maedeh Yousefian, *Ph.D.*[2]
Reza Yari, *Ph.D.*[3]

How to cite this article
Rostampou M., Yousefian M., Yari R. Determining the Minimum Sample Size Required to Create Regression Model in Plant Ecology (Case Study: Cover-Production Relationship). ECOPERSIA 2024;12(1): 39-53.

DOI:
10.22034/ecopersia.12.1.39

[1]Ph.D. Assistant Prof., Department of Rangeland and Watershed Management and Research Group of Drought and Climate Change, Faculty of Natural Resources and Environment, University of Birjand, Birjand, Iran
[2]Ph.D. Research Assistant Prof., Forest and Rangelands Research Department, Mazandaran Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Sari, Iran.
[3]Ph.D. Research Assistant Prof., Khorasan Razavi Agricultural and Natural Resources Research and Education Center, Agricultural Research, Education and Extension Organization (AREEO), Mashhad, Iran.

**** Correspondence***
Address: Department of Rangeland and Watershed Management and Research Group of Drought and Climate Change, Faculty of Natural Resources and Environment, University of Birjand, Birjand, Iran.
Telephone: +989151637869
Fax: +985632202517
E-mail: rostampour@birjand.ac.ir,

Article History
Received: September 22, 2023
Accepted: February 5, 2024
Published: March 31, 2024

ABSTRACT

Aims: Regression analysis is one of the most widely used statistical tests in vegetation evaluation. However, determining more than the sample size for the model validation is required in the plant ecology literature. Indirect methods of estimating forage production always require regression analysis. The fundamental question in such research is that at least a few pairs of samples are required to achieve a valid regression equation. Therefore, this study determines the sample size required to estimate the production of *Haloxylon persicum* Bunge, *Artemisia sieberi* Besser, and *Stipagrostis pennata* (Trin.) De Winter uses plant dimensions, including cover percentage, plant area, height, and volume.

Materials & Methods: The study was conducted in Shahrakht Plain, Zirkouh in South Khorasan province. The research focused on three indicator species in the study area: *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*. The study utilized the relationship between plant cover and dimensions to estimate forage production. In each habitat, 25 plots were established. After computing the power of the correlation and regression tests, the minimum data pair required for the study was estimated, aiming for a power of 80% at the significance level of 0.05. The effect size and power analysis methods were employed to determine the sample size and were then compared with the coefficient of determination (R^2) and thumb rules methods.

Findings: The results of correlation analysis between cover percentage and production show that in *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata* species, the correlation coefficients are 0.54 ($p \leq 0.01$), 0.76 ($p \leq 0.001$) and 0.40 ($p \leq 0.05$) respectively. The correlation power analysis results indicate that with a sample size of 25 pairs of numbers, the effect size of the correlation coefficient is large, and the power ranges between 52% and 99%. The regression power analysis results indicate that with a sample size of 25 pairs of numbers, the effect size of the regression coefficient is significant for some species and medium for others, with power ranging from 78% to 97%. To achieve a test power of 80%, the recommended number of pairs for regression analysis in the three species would be around 30, 12, and 56, respectively.

Conclusion: The results showed that for regression analysis and the statistical importance of the equation and regression coefficients between the cover and production for *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*, about 30, 12, and 56 pairs were proposed, respectively. This study did not examine the relationship between all plant species and production dimensions, but only the relationship between cover and production. Although the correlation test results showed significant relationships between plant dimensions and production, it is not necessarily a good predictor of production (valid regression equations were not obtained).

Keywords: Cover; Power Analysis; Rangeland; Regression; Sample Size.

CITATION LINKS

[1] Motamedi J., ... [2] Bonham C.D. ... [3] Olsoy P. J., ... [4] Yao X., Yang G., ... [5] Attaeian B. Estimation of ... [6] Paruelo J. M., ... [7] Ebrahimi M. Effects of sem ... ... [8] Ghorbani A., Moameri M., D ... [9] Arzani H., Abedi M. Rangel ... [10] Clark P. E., Hardegree S. ... [11] Zarekia S., Jafari A. A., ... [12] Motaharfard E., Mahdavi A. ... [13] Bados R., ... [14] Gholinejad B., ... [15] Grinath J.B. Comparing pre ... [16] Liu H., Dahlgren R.A., Lar ... [17] Rojo V., Arzamendia Y., Pé ... [18] Tsutsumi M., Itano S., and ... [19] Coulloudon B., Eshelman K. ... [20] Flombaum P., ... [21] Jan S.L., Shieh G. Sample ... [22] Bonett D. G., Wright T. A. ... [23] Mousaei Sanjerehei M. Samp ... [24] Jenkins D.G., ... [25] Button KS, ... [26] Bujang M. A., Sa'at N., Si ... [27] Cohen J. Statistical power ... [28] R Core Team. R: A language ... [29] Algina J., Olejnik S. Dete ... [30] Gregory T.K., ... [31] Hair J., Black W.C., Babin ... [32] Harris R. J. A Primer of M ... [33] Hemphill J. F. Interpretin ... [34] Franklin J., Miller J. A. ... [35] Gotelli N. J., Ellison A. ... [36] Schweiger A. H., Irl S. D. ... [37] Green S.B. How Many Subjec ... [38] Kutner M.H., Nachtsheim C. ... [39] Tabachnick B.G., Fidell LS ... [40] Khamis H. J., Kepler M. Sa ... [41] Rossi J.S. Statistical pow ... [42] Bujang M. A. ... [43] Cohen J., Cohen P., West S ... [44] Arzani H., Dehdari S., Kin ... [45] Aliloo F., ... [46] Arzani H., ... [47] Hoseini S., Mesdaghi M., P ... [48] Sadeghi Nia M., Arzani H., ... [49] Mohammadi Golrang B., Gaza ...

[DOI: 10.22034/ecopersia.12.1.39]

Introduction

Rangeland production holds significant importance when assessing rangelands, and its yearly estimation plays a crucial role in rangeland science. It helps determine various factors such as rangeland grazing capacity, rangeland condition, ecosystem well-being, water and soil preservation, and the capacity for carbon sequestration [1, 2, 3, 4, 5]. The assessment of forage production is critical in managing natural resources as it directly influences the quantity of available forage for livestock and wildlife [6]. Various techniques have been developed to measure forage production. One of the most common methods for estimating production is the clipping and weighing [7, 8]. However, due to the extensive size of rangelands, directly estimating forage production annually becomes impractical, time-consuming, and costly. So, direct methods have been replaced by indirect methods for estimation [9].

An effective indirect method for estimating production involves utilizing canopy cover information [2]. Canopy cover and production are commonly observed vegetation characteristics that correlate strongly when analyzing rangeland vegetation [10]. Plant height, canopy cover, and crown diameter are the influential factors in plant production [11, 12]. While measuring vegetation is relatively straightforward, accurately estimating production can be challenging. Therefore, combining cover and production information can offer a more suitable approach for estimating production. Rangeland science researchers have long utilized plant dimensions (such as height, volume, and surface area) to estimate production, comparing different methods in accuracy, time efficiency, and cost [13, 14, 15, 16, 17]. Tsutsumi et al. (2007) conducted a study on estimating plant production through cutting and weighing. According to their findings, when aiming for a 10% acceptable error and 95% confidence level, approximately 200, 77, and 9 plots 50 × 50 cm are necessary for rangelands with very high, medium, and very low heterogeneity, respectively [18]. In certain studies, it is worth noting that some sources inaccurately refer to the estimation of forage production based on cover percentage as the "double sampling" method. The genuine double sampling method involves estimating forage production across numerous plots, with a subset of these plots being randomly selected for cutting and weighing to obtain precise measurements [9].

The fundamental question in the production estimation method, which relies on cover information, is how many plots should undergo cutting and weighing. The method involves measuring plant dimensions in a specific and limited number of plots, and within those same plots, production is directly determined through cutting and weighing. According to Coulloudon et al. (1999), a general rule in the double sampling method is to cut at least one plot for every seven estimated plots. This cut plots to estimated plots ratio corresponds to approximately 15% [19].

In any case, all of these methods utilize regression tests to examine relationships and establish the equation [20]. Regression analysis is widely employed across various scientific disciplines [21], and determining the appropriate sample size is critical to regression analysis [22, 23]. A research study requires adequate statistical power and sample size to detect scientifically valid effects. Despite multiple linear regression being a widely recognized statistical tool, the issue of sample size for model validation has yet to receive sufficient attention in the literature [21]. In biological sciences, the sample size is relatively small [24]. Jenkins and Quintana-Ascencio (2020) found that approximately 64% of ecological studies related to disturbance and degradation and about 80% of biogeographical studies related to species-area relationships had sample sizes of less than 25 [24]. In contrast, disciplines such as psychology and economics in the humanities tend to have very high sample sizes, and there might be a general relationship between research budget and sample size across different fields [24]. The implications of having a low sample size can

[DOI: 10.22034/ecopersia.12.1.39]

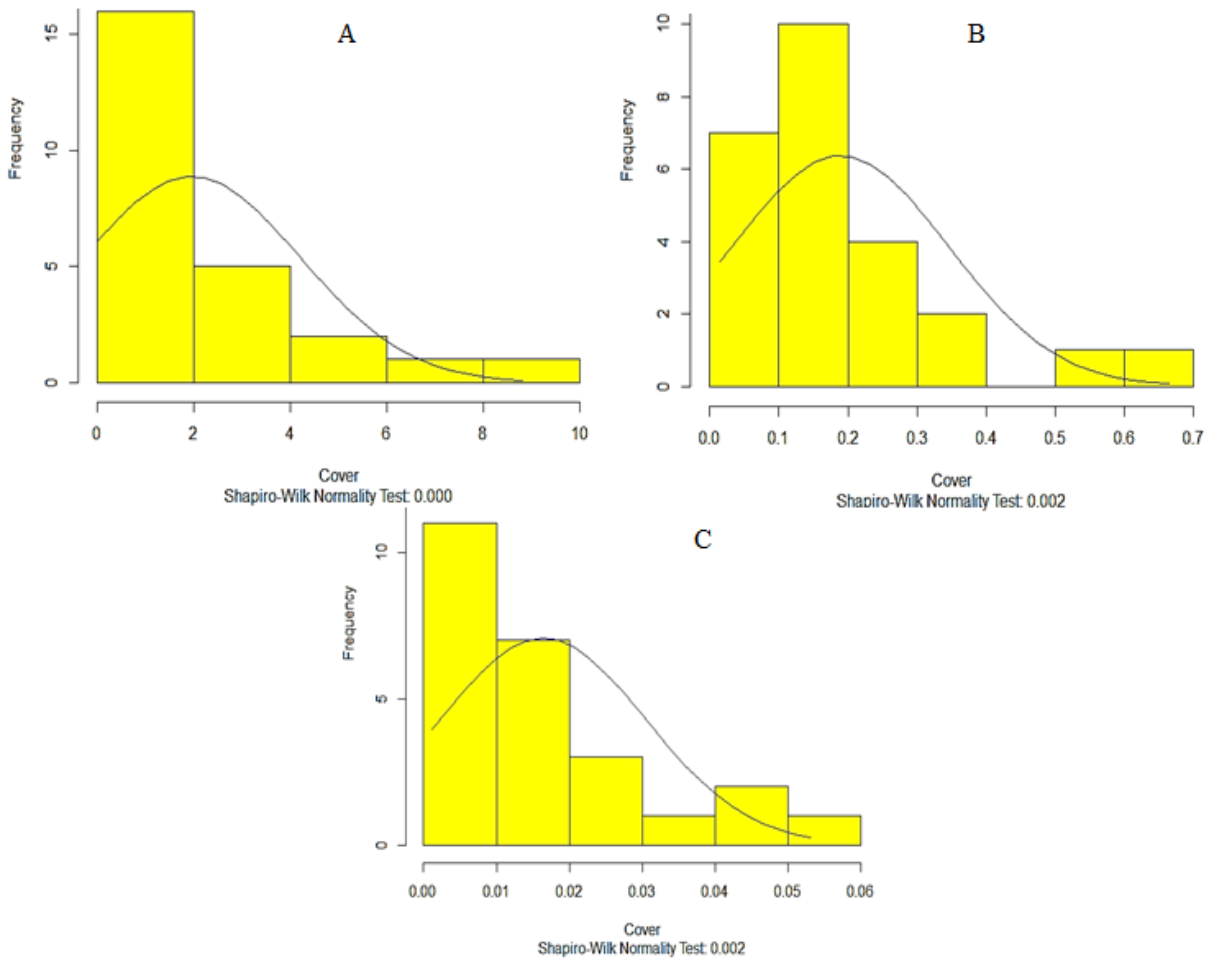

Figure 1) shows the histogram curve and the significance level of the Shapiro-Wilk test of the cover percentage in *Haloxylon persicum* (A), *Artemisia sieberi* (B), and *Stipagrostis pennata* (C).

be detrimental to the credibility of research, leading to wasted time and resources and the production of inconclusive or contradictory results. Conversely, having a high sample size may increase the workload and cost [25].

Indeed, the primary objective of most statistical analyses is inference, which makes it essential to establish the required sample size before conducting any analysis [26]. For regression analysis, this necessity becomes twofold because the researcher aims to determine the relationships between the dependent variable and independent variable(s) and predict and assess the contribution of each independent variable to the dependent variable using regression analysis.

This study tested two hypotheses:

1. There is a significant correlation between plant dimensions and production.
2. The 25 pairs of samples provide a valid regression equation to predict production from the plant dimensions.

Therefore, this study focuses on determining the appropriate sample size for correlation and regression analysis when investigating the relationship between the percentage of vegetation and forage production in three species: *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*. The research aims to compare its findings regarding sample size determination with other proposed methods, thereby ensuring the robustness and reliability of the statistical analysis performed.

[DOI: 10.22034/ecopersia.12.1.39]

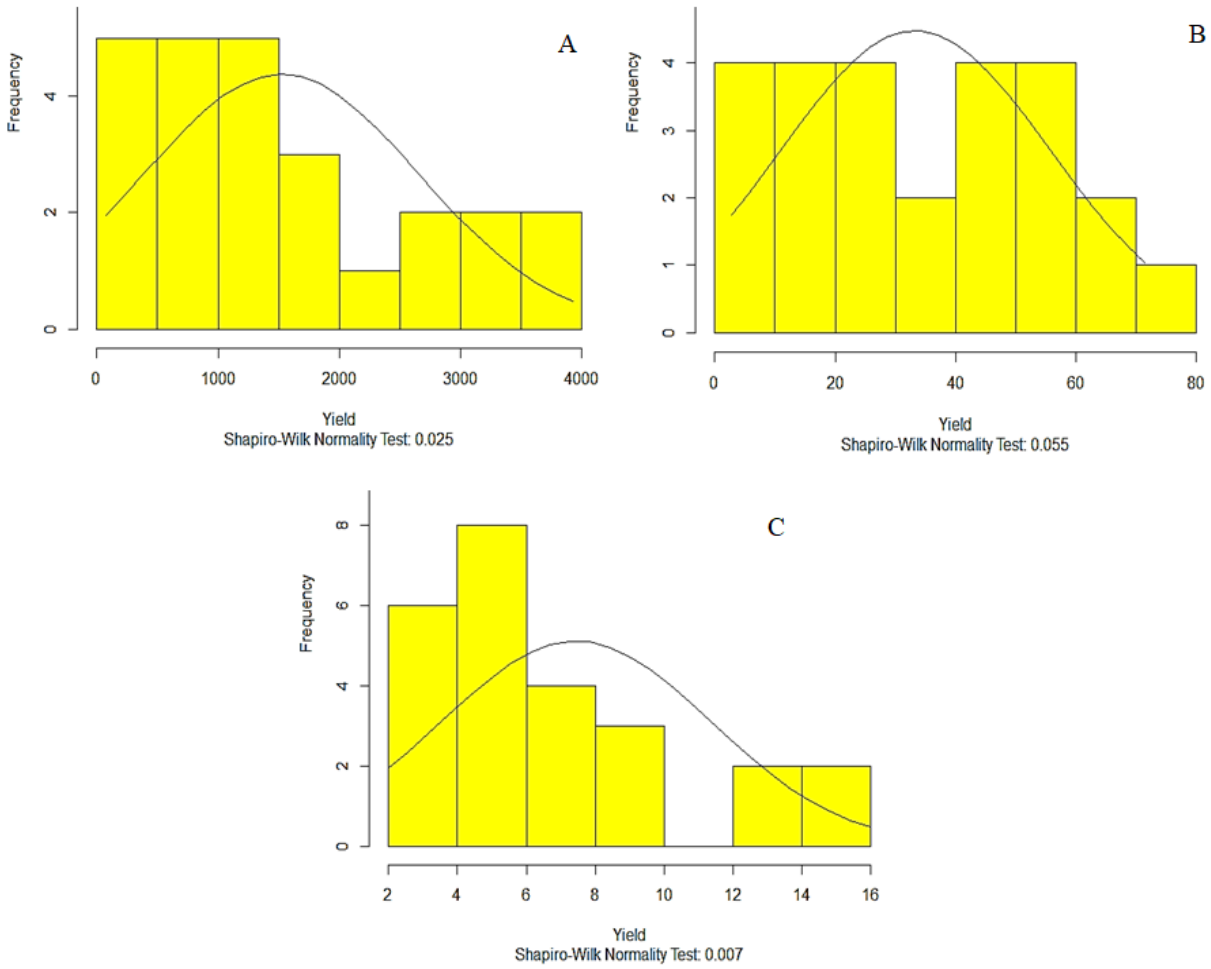

Figure 2) The histogram curve and the significance level of the Shapiro-Wilk test of production in *Haloxylon persicum* (A), *Artemisia sieberi* (B), and *Stipagrostis pennata* (C).

Materials & Methods

The study was conducted in Shahrakht Plain, which is located approximately 25 km from Haji Abad, Zirkouh, in South Khorasan province, at coordination of E 60°30' to 60°56' and N 33°10'to 33°30'. The mean elevation is about 802 m above the sea level. The mean annual rainfall is 147.75 mm, and the mean annual temperature is 20.53 ° C. Sampling occurred in June 2022 across three distinct habitats. The research focused on three indicator species in the study area: *Haloxylon persicum* Bunge, *Artemisia sieberi* Besser, and *Stipagrostis pennata* (Trin.) De Winter. The study utilized the relationship between plant cover and dimensions to estimate forage production. In each habitat, 25 plots were established randomly, ensuring that each plot contained at least one of the individual species. In cases where multiple species were present within a plot, one species was randomly selected, and the cover percentage of the target species was determined using the plot method. Measurements of plant height and large and small diameter were taken using meters and determined using trigonometric relationships based on the surface and volume of the specific plant species. Subsequently, the forage production (the current year's growth) of the selected species was directly determined in all plots. For this purpose, the green parts of the aerial sections and branches were carefully cut using gardening scissors and placed in

[DOI: 10.22034/ecopersia.12.1.39]

separate paper envelopes. These samples were then dried in a shaded environment to achieve the dry weight of the species, which was estimated for each plant individual.

After recording cover percentage, dimensions, and plant production data, the histogram curve and the Shapiro-Wilk test assessed the data's normality. Because the data did not meet the normality assumption, Spearman's correlation coefficient was employed to investigate the relationship between plant dimensions and production using a correlation test.

As all three studied species exhibited a significant relationship between the percentage of cover and forage production, the cover percentage was treated as an independent variable. In contrast, production was considered the dependent variable in the simple linear regression analysis. Even though all traits were initially included as independent variables, the analysis revealed high correlations among these variables, and the variance inflation factor (VIF) values, determined using the VIF function from the car package (version 3.0-11), indicated the presence of collinearity between them. As the primary goal of the research was to identify the minimum data required for the cover-production relationship, only simple linear regression was employed.

Several methods are available to determine sample size in regression analysis, with the four most commonly used ones being the coefficient of determination (R^2), effect size (d and f2), power analysis (1-β), and Rules of Thumb. Each method has its advantages and disadvantages, which this article will thoroughly discuss. The effect size (Cohen's f2) measures the strength of correlation and regression slope coefficients. The probability of making a type II error (failing to reject the null hypothesis when it is false) is called β (beta). Power is the probability of avoiding a Type II error. The higher the statistical power of a test, the lower the risk of making a Type II error. [27]

After conducting correlation and regression

Table 1) The effect size of correlation and regression analyses and the limit suggested by Cohen [27].

	Small	Medium	Large
Correlation Analysis	0.1	0.3	0.5
Regression Analysis	0.02	0.15	0.35

Table 2) Correlation analysis between vegetation percentage and forage production of the three studied plant species.

	Correlation Coefficient	Confidence Limits	Upper Limit	Lower Limit
H. persicum	0.54	0.63	0.15	0.78
A. sieberi	0.76	0.42	0.48	0.90
S. pennata	0.40	0.71	-0.01	0.69

Table 3) Comparison of the correlation coefficients between the percentage of vegetation cover and the forage production of the three studied plant species.

	Z	P-value
Artemisia sieberi - Haloxylon persicum	1.3	0.19
Stipagrostis pennata - Haloxylon persicum	0.6	0.55
Stipagrostis pennata - Artemisia sieberi	1.9	0.06

[DOI: 10.22034/ecopersia.12.1.39]

analyses, the research proceeded to assess power analysis, specifically Type II error or β, and effect size using the "pwr.r.test" and "pwr.f2.test" functions from the pwr package (version 0-1.3), respectively. To calculate the effect size for both correlation (d) and regression (f2), the "r_to_d" function from the effect size package (version 0.5) was utilized. Furthermore, the effect sizes were expressed based on Cohen's criterion using the "cohen.ES" function from the pwr package, which categorized them into three levels: small, medium, and large (Table 1). After computing the power of the correlation and regression tests, the minimum data pair required for the study was estimated, aiming for a power of 80% at the significance level of 0.05. The effect size and power analysis methods were employed to determine the sample size and were then compared with the coefficient of determination (R^2) and thumb rules methods. All statistical analyses were conducted using the R program (version 4.2.2) [28].

Findings

The histogram curve results and the significance level of the Shapiro-Wilk test ($p \geq 0.05$) indicated that both the cover percentage and forage production of all three species exhibit non-normal distributions (Figures 1 and 2). The histogram charts visually display the data's lack of symmetry, positively skewed or skewed to the right. Therefore, a non-parametric correlation test (Spearman's coefficient) was used.

The results of the scatter diagram show that in *Haloxylon persicum* and *Artemisia sieberi*, all traits have a significant positive correlation with each other and production at the 0.01 level (Figures 3 and 4). There was a strong, positive correlation between cover and production of *Haloxylon persicum* and *Artemisia sieberi*, which was statistically significant (r (25)= 0.54, p = 0.01 and r (25)= 0.76, p = 0.001, respectively).

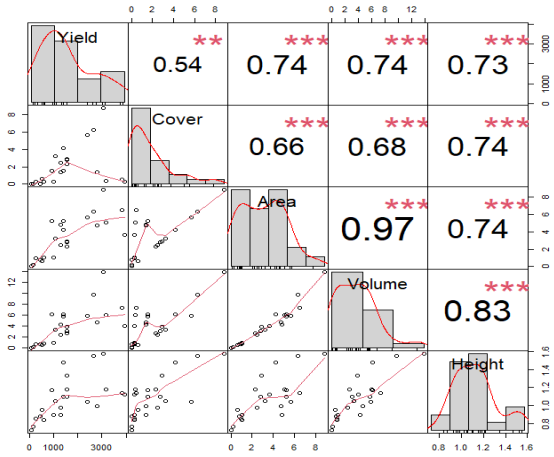

Figure 3) Corrplot of the studied characteristics of *Haloxylon persicum.*

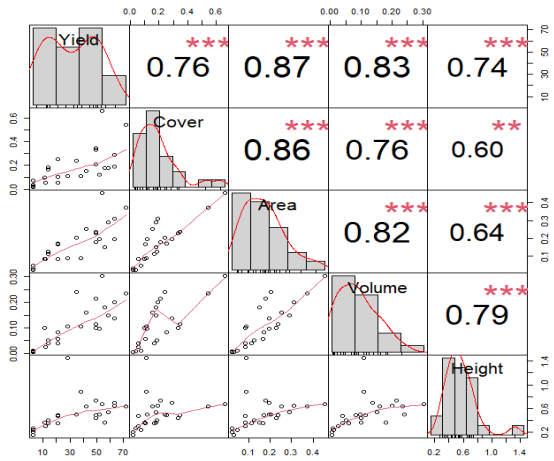

Figure 4) Corrplot of the studied characteristics of *Artemisia sieberi*.

Table 4) Variance inflation factor values of three general regression equations calculated in multiple regression.

Variance Inflation Factor (VIF)				
	Cover Percentage	Area	Volume	Height
H. persicum	3.04	21.29	36.68	5.03
A. sieberi	4.02	5.97	4.24	1.35
S. pennata	5.13	25.97	20.54	3.13

[DOI: 10.22034/ecopersia.12.1.39]

In *the Stipagrostis pennata* species, there is a significant positive correlation between the percentage of cover and production (at the 0.05 level) and the percentage of cover, area, and volume of the plant (at the 0.01 level) (Figure 5). Cover and production were moderately positively correlated, $r(25) = 0.40$, $p = 0.03$.

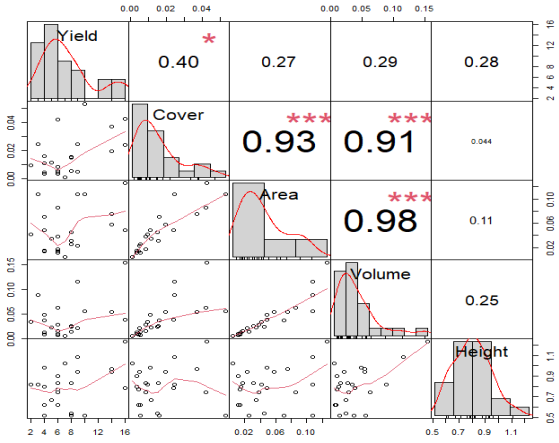

Figure 5) Corrplot of the studied characteristics of *Stipagrostis pennata*.

The results of correlation analysis between cover percentage and production show that in *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata* species, the correlation coefficients are 0.54 ($p \leq 0.01$), 0.76 ($p \leq 0.001$), and 0.40 ($p \leq 0.05$) respectively. Confidence limits of correlation coefficients are presented in Table 2.

The t-test results for the standard values (z and P.value) indicate no significant difference between cover percentage and production correlation coefficients among the three studied plant species (Table 3). The correlation coefficients of 0.76 in *Artemisia sieberi* and 0.40 in *Stipagrostis pennata* do not exhibit a significant difference ($p \geq 0.01$), implying that it is not possible to conclude that the correlation between these two traits is higher in *Artemisia sieberi* compared to *Stipagrostis pennata*. The lack of significance in the t-test suggests that the correlations are comparable across the three species.

The strong correlation between plant area, volume, and height attributes indicates the possibility of collinearity, a prerequisite for multiple regression analysis. The variance inflation factor (VIF) was calculated for the regression model involving production and the dimensions of the three species to assess collinearity. VIF values exceeding 5 indicate a high degree of collinearity. As depicted in Table 4, *Haloxylon persicum* and *Stipagrostis pennata* species display significant linearity between area and volume attributes with other traits. Consequently, including all four factors in the regression model is not feasible. Although a regression elimination method could be employed, where the least effective trait is eliminated, followed by recalculating VIF to redo the regression, this study focused solely on the cover-production relationship. Therefore, the details of all regression relationships were omitted, and the findings apply solely to simple regression analyses with an independent variable.

After establishing the direction and strength of the relationship between the cover percentage and forage production for the three studied species, a simple linear regression analysis was employed to derive the equation. The analysis outcomes revealed that the three species' coefficient of determination (R^2) is 0.29, 0.58, and 0.16, respectively (Tables 5, 6, and 7).

The analysis of variance for the model, as indicated by the F statistic, demonstrates that the linearity of the relationship was confirmed at the 0.05 level for all three species. However, while the models were significant, it was determined that they were not valid for *Haloxylon persicum*. This is because the slope of the regression line, represented by the coefficient of the cover percentage in the equation, was insignificant ($p \geq 0.05$) for this species.

The correlation power analysis results indicate that with a sample size of 25 pairs of numbers, the effect size of the correlation

[DOI: 10.22034/ecopersia.12.1.39]

coefficient is large, and the power ranges between 52% and 99%. The Type II error rate is approximately 18%, 1%, and 48%, respectively (Table 8). To achieve a test power of 80%, the recommended number of pairs for correlation analysis in the three species would be around 25, 11, and 47, respectively.

The regression power analysis results indicate that with a sample size of 25 pairs of numbers, the effect size of the regression coefficient is significant for some species and medium for others, with power ranging from 78% to 97%. Additionally, the Type II error rate is approximately 22%, 3%, and 8%, respectively (as shown in Table 9). To achieve a test power of 80%, the recommended number of pairs for regression analysis in the three species would be around 30, 12, and 56, respectively.

The question is whether the coefficients of determination of the three equations are valid. In this study, with 25 pairs of samples, the R^2 should be approximately 0.38 to reach a significant level of 0.05 (Figure 6). Thus, although the linearity of the relationship was significant (F in Tables 5 and 7), the coefficient of determination for *Haloxylon persicum* and *Stipagrostis pennata* is less than 0.38, so it was not significant.

Table 5) Regression analysis of the relationship between the cover percentage and forage production of *Haloxylon persicum*.

Model Results				
Coefficient of Determination		F-Statistic		
0.29		3.82		
Model Coefficients				
Variable	Regression Coefficient	SE	t-value	p-value
Constant	1149.87	287.28	4.003	0.00***
Cover	190.85	97.63	1.955	0.06

*** $p < 0.00$

Table 6) Regression analysis of the relationship between the cover percentage and forage production of *Artemisia sieberi.*

Model Results				
Coefficient of Determination		F-Statistic		
0.58		18.51		
Model Coefficients				
Variable	Regression Coefficient	SE	t-value	p-value
Constant	15.37	5.37	2.86	0.00***
Cover	94.87	22.05	4.30	0.00***

*** $p < 0.00$

[DOI: 10.22034/ecopersia.12.1.39]

Table 7) Regression analysis of the relationship between the cover percentage and forage production of *Stipagrostis pennata.*

Model Results				
Coefficient of Determination		F-Statistic		
0.58		18.51		
Model Coefficients				
Variable	Regression Coefficient	SE	t-value	p-value
Constant	15.37	5.37	2.86	0.00***
Cover	94.87	22.05	4.30	0.00***

*** $p < 0.00$

Table 8) Analysis of the correlation power at the 0.05 level (two-tailed) and the number of suggested pairs based on the 80% test power.

Plant Species	Number of Initial Pairs	Cohen Statistics	Effect Size	Power	Number of Suggested Pairs
H. persicum	25	1.28	Large	0.82	25
A. sieberi	25	2.34	Large	0.99	11
S. pennata	25	0.87	Large	0.52	46

Table 9) Regression power analysis at the 0.05 level (two-tailed) and the number of suggested pairs based on 80% test power.

Plant Species	Number of Initial Pairs	Cohen Statistics	Effect Size	Power	Number of Suggested Pairs
H. persicum	25	0.41	Large	0.78	30
A. sieberi	25	1.38	Large	0.97	12
S. pennata	25	0.19	Medium	0.92	56

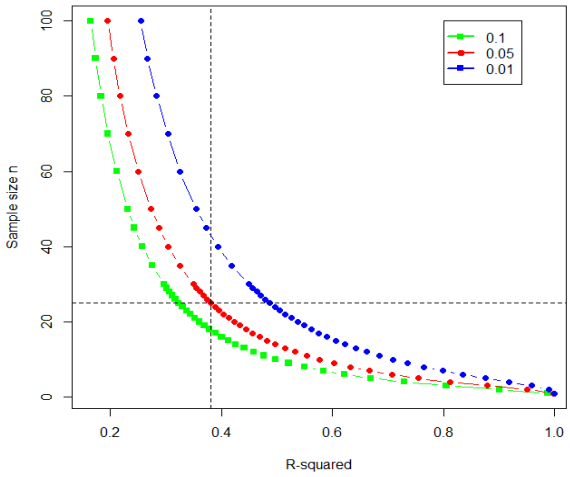

Figure 6) Sample size estimation for regression analysis. In the present study, with 25 pairs of samples, R^2 =0.38 is significant at the significance level of 5%

Discussion

Estimating the required sample size and statistical power for research is an essential part of research design, and calculating the appropriate sample size and power analysis have become important topics in the research. In this paper, the power and sample size of the regression were studied to investigate the relationships between the cover and the production of the dominant rangeland species of Shahrakht in Iran. The initial phase of this research involved gathering 25 pairs of data samples associated with plant dimensions (such as area and volume) and the production of three rangeland plants: *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*.

[DOI: 10.22034/ecopersia.12.1.39]

Table 10) Sample size for multiple linear regression based on the coefficient of determination (R^2) and the number of variables studied (significance level: 0.05 and power: 0.80)

Variable Number	R^2						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
1	91	44	28	20	15	12	9
2	90	42	26	18	14	11	9
3	103	48	30	21	16	12	10
4	113	53	33	24	18	14	12

Table 11) Some rules of thumb in determining regression sample size in different sciences

The Field of Study	Sample Size	Description	Ref.
Statistics	N > 50+ m	Step-by-step regression	31, 32, 33
	N =10-15 m	--	31
Biostatistics	N ≥ 25	Data with high variance	24
	N = 8	Data with minimal variance	24
	N =100+50 m	Logistic regression	26
	N =10 m	Logistic regression	24
Ecology	N >20× m	Species abundance distribution modeling	34
	N =10-15 m	--	35
	N = 30-45	Environmental gradient analysis studies	36
Humanities	N ≥ 50 + 8 m	Testing the entire model	37
	N ≥104+ m	Testing the significance of the coefficients	37
	N > 10× m	Stepwise regression and hierarchical regression	38, 39
	N > 40× m	Inter regression	39
	N > 5× m	The normality of the residuals	39
	N = 20+5 m	--	40

In the current research, based on the power analysis method, about 25, 11, and 46 number pairs were proposed to analyze the correlation between cover and production for *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*, respectively. The equation and regression coefficients were proposed for *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata* for regression analysis and significance, about 30, 12, and 56 number pairs, respectively.

Comparing the Results of Power Analysis with the Coefficient of Determination (R^2)

Algina and Olejnik (2000) suggested utilizing the coefficient of determination (R^2) to determine the appropriate sample size for multiple regression analysis [29]. However, it is essential to consider that the value of R^2 is heavily influenced by the size of the initial sample (n) and the nature of the data under examination. Moreover, R^2 does not attain statistical significance when dealing with a large sample size. Referring to Table 10, it is evident that having a minimum of one independent variable necessitates approximately 28, 12, and 44

[DOI: 10.22034/ecopersia.12.1.39]

pairs of samples, respectively. Notably, as the coefficient of determination (R^2) decreases, the sample size increases rapidly [30].

Comparison of Power Analysis Results with Rules of Thumb

Determining the minimum number of sample pairs in regression analysis involves various rules of thumb, with some applied in the humanities and medical fields and others in ecology and rangeland sciences. Table 11 provides insights into this matter, revealing that the required number of data pairs for at least one independent variable can range from 5 to 104 samples, depending on factors such as the type of data (normal or not, low or high variance), test purpose, and regression type. The range of these values spans approximately 100 sample pairs, indicating a significant disparity in the rules of thumb for calculating sample size. This wide range emphasizes the considerable variation in approaches across different disciplines.

In ecology, it is generally recommended to have a sample size between 10 and 45 samples [35,36]. However, specific rules for other regressions, such as estimating production based on cover, are only sometimes observed in rangeland science. In the double sampling method, which is a type of regression, a rule of 1 to 7 (ratio of cut plot to estimated plot) has been suggested [19]. For instance, in a plant species that is studied with around 30 to 60 plots, somewhere between 4 and 8 plots should be cut directly, which is relatively low. It is important to note that this rule may not lead to a valid regression equation when dealing with a small number of initial plots (less than 30 plots) and only four pairs of estimated and interrupted production data. The mentioned rule becomes more applicable when the initial plots increase to 100 to 200 samples, where 14 and 28 pairs of numbers should be cut. Interestingly, this closely aligns with the results obtained from the current research, suggesting a similar range for sample pairs when the number of initial plots is in that range.

Green (1991) provides a comprehensive overview of the procedures used to determine regression sample sizes. He suggests N > 50 + 8 m (where m is the number of IVs) for testing the multiple correlations and N > 104 + m for testing individual predictors [37]. After evaluating the approach utilized in this study and comparing it with other rules, it can be deduced that employing Green's formula is not advisable in the area under investigation. When at least one independent variable is involved, a substantial number of sample pairs, ranging from approximately 58 to 105, are necessary. Considering the vast extent of the rangelands, this requirement could be more practical regarding cost, time, and environmental damage. Currently, the issues surrounding rules of thumb for estimating sample size in multiple regression analysis are subject to critical discussions [41]. While some researchers initially questioned the scientific rigor of the rule-of-thumb approach compared to precise calculations, it remains a viable and commonly used method [42]. Nevertheless, there are criticisms regarding the need for more empirical evidence and the arbitrary nature of these recommendations, as they often vary significantly. For instance, Khamis and Kepler (2010) expressed their reservations about rules of thumb, suggesting that they are difficult to substantiate and somewhat subjective [40]. To address this concern, Khamis and Kepler (2010) proposed an alternative formula based on the confidence criterion, resulting in the 20+5m formula, which appears to produce outcomes akin to the present study's findings.

In the current study, the regression effect size was significant for two species, *Haloxylon persicum* and *Artemisia sieberi*, and medium for *Stipagrostis pennata*. Cohen's general rule indicates that approximately 25 pairs of samples are required for a large effect size and around 54 for a medium effect size, with a desired power level of 80% (Table 12).

The power of the regression tests conducted in this research was between 78% and 97%.

[DOI: 10.22034/ecopersia.12.1.39]

Table 12) Comparison of the sample size studied in this research with other rules and tests

Rules of Thumb	Minimum Pair of Samples	Ref.
Green's law (whole model)	58	37
Green's law (significance of coefficients)	105	37
Khamis-Kepler formula	25	40
Power analysis (large effect size)	25	27
Power analysis (medium effect size)	54	27
Power analysis (small effect size)	403	27
Current research	12-56	

Table 13) Comparison of the sample size studied in the present study with another research conducted in Iran

The Dominant Vegetative Form	Pair of Samples	Ref.
Grass and forb	8-12	44
Bush	15	45, 46
Bush and forb	16	47
Bush	7-11	48
Bush and forb	30	49
Bush and forb	25	14
Bush and shrub	12-56	Current research

Based on this information, the recommended number of sample pairs falls between 15 and 25, which aligns with the study's power analysis results. The current study's findings demonstrate similarity with the results obtained from the power analysis conducted for a large effect size. According to Cohen's general rule, a sample size of approximately 400 pairs is advised for detecting a small effect size (Table 12). However, when there is a low correlation between two variables, resulting in a small effect size, it is not recommended to use relational and causal hypothesis tests. In this research, another important finding is the presence of high collinearity among the independent variables, precisely the percentage of cover, volume, surface area, and plant height. This violation of the assumption of independence among the variables makes it inappropriate to use multiple regression analysis. To address this issue, employing a single regression analysis using a plant dimension or trait that exhibits a strong correlation and a high coefficient of determination with production is suggested. Alternatively, it is possible to include traits related to production, but these traits should be independent when used in the multiple regression equation.

The current study's findings suggest that power analysis can be utilized to determine the sample size needed for regression analysis, given that the effect size, a crucial parameter in power analysis, is accurately computed. Bujang (2021) also advocates for researchers to employ methods for determining sample size that incorporate effect size [42]. However, several criticisms are concerned with using Cohen's effect size in power analysis. Cohen (2003 and

[DOI: 10.22034/ecopersia.12.1.39]

2013) pointed out that these guidelines only apply when estimates related to the specific research area of interest are unknown [27, 43]. Brydges (2019) criticized the use of Cohen's effect size in sample size determination, asserting that these guidelines lack a basis in quantitative estimates [44]. Furthermore, the interpretation of Cohen's effect size does not rely on a formal statistical analysis of the data, and the distribution of effect sizes may differ across various disciplines [33].

Comparison of Power Analysis Results with Experimental Research Conducted in Iran

In the studies within Iran, double sampling methods are generally used, in which between 7 and 30 pairs (an average of 15 pairs) are recommended (Appendix, Table 13), which is less than the number of pairs in the present study (12 to 56).

Conclusion

The results showed that for regression analysis and the statistical importance of the equation and regression coefficients between the cover and production for *Haloxylon persicum*, *Artemisia sieberi*, and *Stipagrostis pennata*, about 30, 12, and 56 pairs were proposed, respectively. This study did not examine the relationship between all plant species and production dimensions, but only the relationship between cover and production. For the first hypothesis, sample data provided enough evidence to reject the null hypothesis; therefore, there are significant relationships between plant dimensions and production, but it is not necessarily a good predictor of production (valid regression equations were not obtained). The null hypothesis was confirmed for the second hypothesis, and 25 pairs of samples are not enough to use regression equations to estimate production from the cover percentage of *Haloxylon persicum* and *Stipagrostis pennata*.

Acknowledgments

The authors of this article express their sincere gratitude to the Vice-Chancellor of Research, Technology, and Innovation of the University of Birjand for the financial support of this research.

Ethical Permission: None declared by Authors

Conflict of interest: The author declares no conflict of interest regarding the paper's publication by ECOPERSIA.

Funding/Support: University of Birjand supported the study that produced this research

References

1. Motamedi J., Afradi J., Sheidai Karkaj E., Alijanpour A., Emadodin I., Banej Shafiei S . Environmental Factors Affecting the Structural Trials and Biomass of *Onobrychis aurea* Bioss. ECOPERSIA 2020; 8(4):247-259.
2. Bonham C.D. Measurements for Terrestrial Vegetation, 2nd Edition. John Wiley Sons, New York, NY. 2013; 260 p.
3. Olsoy P. J., Glenn N. F., Clark P. E. Estimating sagebrush biomass using terrestrial laser scanning. Rangel. Ecol. Manag. 2014; 67(2): 224–228.
4. Yao X., Yang G., Wu B., Jiang L., Wang F. Biomass Estimation Models for Six Shrub Species in Hunshandake Sandy Land in Inner Mongolia, Northern China. Forests. 2021; 12(2): 167.
5. Attaeian B. Estimation of Aboveground Biomass Carbon Sequestration Potential in Rangeland Ecosystems of Iran. ECOPERSIA 2016; 4(1):1283-1294.
6. Paruelo J. M., Lauenroth W. K., Roset P. A. Technical note: Estimating aboveground plant biomass using a photographic technique. J. Range. Manage. 2000; 53(2): 190-193.
7. Ebrahimi M. Effects of semi-circular bunds on plant vegetation and soil properties of Naroon and Neron rangelands in Sistan and Balochistan. ECOPERSIA 2022; 10(1):1-11.
8. Ghorbani A., Moameri M., Dadjou F., Seyedi Kaleybar S., Pournemati A., Asghari S. Determinization of Environmental Factors Effects on Plants Production in QezelOzan-Kosar Rangelands, Ardabil Province Factors Effect on Rangelands Production. ECOPERSIA 2020; 8(1):47-56.
9. Arzani H., Abedi M. Rangeland Assessment Survey and Monitoring. University of Tehran Press. 2013; 305p.
10. Clark P. E., Hardegree S. P., Moffet C. A., Pierson F. B.. Point sampling to stratify biomass variability in sagebrush steppe vegetation. Rangel. Ecol. Manag. 2008;61(6):614-622.
11. Zarekia S., Jafari A. A., Mirhaji T. Assessment of Planting Season Effects on Vegetation Parameters of Astragalus effusus and Astragalus

[DOI: 10.22034/ecopersia.12.1.39]

brachyodontus Accessions. ECOPERSIA 2016;4(1):1225-1237.
12. Motaharfard E., Mahdavi A., Iranmanesh Y., Jafarzadeh A. Effect of Land Uses on Aboveground Biomass and Carbon Pools in Zagros Forests, Iran. ECOPERSIA 2019; 7 (2):105-114.
13. Bados R., Esteban L. S., Esteban J., Fernández-Landa A., Sánchez T., Tolosana, E. Biomass equations for rockrose (Cistus laurifolius L.) shrublands in North-central Spain. For. Syst. 2021; 30(3): e015.
14. Gholinejad B., PourBabaei H., Farajollahi, A. Parvane E. Assessment and Comparison of Different Methods for Estimating Forage Production (Case Study: Rangeland of Kurdistan Province). J. Rangel. Sci. 2012; 2(2): 483-489.
15. Grinath J.B. Comparing predictive measures and model functions for estimating plant biomass: lessons from a sagebrush–rabbitbrush community. J. Plant. Ecol. 2019; 220(6): 619-632.
16. Liu H., Dahlgren R.A., Larsen R.E., Devine S.M., Roche L.M., O' Geen A.T., Wong A.J.Y., Covello, S., Jin Y. Estimating Rangeland Forage Production Using Remote Sensing Data from a Small Unmanned Aerial System (sUAS) and PlanetScope Satellite. Remote Sens-Basel. 2019; 11(5):595.
17. Rojo V., Arzamendia Y., Pérez C., Baldo J.L., Vilá B. Double Sampling Methods in Biomass Estimates of Andean Shrubs and Tussocks. Rangel. Ecol. Manag. 2017; 70(6): 718 - 722.
18. Tsutsumi M., Itano S., and Shiyomi M. Number of samples required for estimating herbaceous biomass. Rangel. Ecol. Manag. 2007; 60(4): 447-452.
19. Coulloudon B., Eshelman K., Gianola J., Habich N., Hughes L., Johnson C., Pellant M., Podborny P., Rasmussen A., Robles B. Sampling vegetation attributes: interagency technical reference. Technical Reference 1734-4, USDI Bureau of Land Management. Second Revision. Denver, CO, USA: National Applied Resource Sciences Center.1999, 163 pp.
20. Flombaum P., Sala O. E. A non-destructive and rapid method to estimate biomass and aboveground net primary production in arid environments. J. Arid. Environ. 2007; 69(2): 352-358.
21. Jan S.L., Shieh G. Sample size calculations for model validation in linear regression analysis. BMC Med. Res. Methodol. 2019: 19(1):1-9.
22. Bonett D. G., Wright T. A. Sample size requirements for multiple regression interval estimation. J. Organ. Behav. 2011; 32(6): 822–830.
23. Mousaei Sanjerehei M. Sample Size Calculations for Vegetation Studies. MJEE. 2021;23(2):85–97.
24. Jenkins D.G., Quintana-Ascencio P.F. A solution to minimum sample size for regressions. PLoS ONE. 2020; 15(2): e0229345.
25. Button K.S, Ioannidis J.P., Mokrysz C., Nosek B.A., Flint J., Robinson E.S. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 2013; 14(5): 365-376.
26. Bujang M. A., Sa'at N., Sidik T. M. I. T. A. B., Joo L. C. Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. Malays. J. Med. Sci. 2018; 25(4): 122–130.
27. Cohen J. Statistical power analysis for the behavioral sciences. 3rd ed. Hillsdale: Erlbaum. 2013; 579 p.
28. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021. URL
29. Algina J., Olejnik S. Determining sample size for accurate estimation of the squared multiple correlation coefficient. Multivar. Behav. Res. 2000; 35(1): 119–137.
30. Gregory T.K., Daniel M. Sample Sizes When Using Multiple Linear Regression for Prediction. Educ. Psychol. Meas. 2008; 68(3):431-442.
31. Hair J., Black W.C., Babin B.J., Anderson R.E. Multivariate Data Analysis, 8th Edition. United Kingdom: Cengage Learning. 2018; 832 p.
32. Harris R. J. A Primer of Multivariate Statistics. 3rd ed. Mahwah, NJ: Lawrence Erlbaum. 2001; 634 p.
33. Hemphill J. F. Interpreting the magnitudes of correlation coefficients. Am. Psychol. 2003; 58(1): 78–79.
34. Franklin J., Miller J. A. Mapping Species Distributions: Spatial Inference and Prediction. Cambridge; New York: Cambridge University Press. 2009; 320 p.
35. Gotelli N. J., Ellison A. M. A Primer of Ecological Statistics. Sunderland, MA: Sinauer Associates, Inc. 2004; 510 p.
36. Schweiger A. H., Irl S. D. H., Steinbauer M. J., Dengler J., Beierkuhnlein C. Optimizing sampling approaches along ecological gradients. Methods. Ecol. Evol. 2016; 7(4): 463–471.
37. Green S.B. How Many Subjects Does It Take to Do a Regression Analysis? Multivar. Behav. Res. 1991; 26 (3): 499-510.
38. Kutner M.H., Nachtsheim C.J., Neter J., Li, W. Applied Linear Statistical Models. 5th ed. The McGraw-Hill/Irwin Series Operations and Decision Sciences. Boston: McGraw-Hill Irwin. 2005; 1415 p.
39. Tabachnick B.G., Fidell LS Using multivariable statistics. 6th ed. Boston: Pearson Education. 2013; 983 p.
40. Khamis H. J., Kepler M. Sample size in multiple

[DOI: 10.22034/ecopersia.12.1.39]

regression: 20+ 5K. J. Appl. Statist. Sci. 2010; 17(4): 505-517.
41. Rossi J.S. Statistical power analysis. In J.A. Schinka and W.F. Velicer (Eds.), Handbook of psychology. Volume 2: Research methods in psychology. John Wiley and Sons, Inc. 2013; 71–108pp.
42. Bujang M. A. Step-by-Step Process on Sample Size Determination for Medical Research. Malays. J. Med. Sci. 2021, 28(2):15-27.
43. Cohen J., Cohen P., West S.G., Aiken, L.S. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed. Mahwah: Erlbaum. 2003; 734 p.
44. Arzani H., Dehdari S., King, G. Models for estimating range production by cover measurement. Iran. J. Range Desert Res. 2011;18(1): 1-16.
45. Aliloo F., Keyvan behjou F., Moetamedi, J. Study presentation feasibility of statistical models for estimating rangeland plants species of *Artemisia aucheri* and *Agropyron trichophorum* (Case Study: Dizaj Batchi and Ghotor Ranglands of Khoy). Iranian Journal of Range and Desert Research, 2016;22(4):625-638.
46. Arzani H., Adnani S. M, Bashari H, Azimi M.A.S., Baghri H, Akbarzadeh M., Kaboli S. H. Assessment of vegetation covers and production variation in rangelands of Qum province (2000-2005). Iran. J. Range Desert Res. 2007; 25(4): 296- 313.
47. Hoseini S., Mesdaghi M., Pambokhchyan C. Comparing 3 methods of forage estimation in summer rangelands (Case study: Sar-Aliabad rangelands of Golestan province). Iran. J. Range Desert Res. 2012;18(4):637-651.
48. Sadeghi Nia M., Arzani H., Baghestani N. Comparison of different production estimation methods for some important shrub plants (the case study in Yazd and Isfahan Provinces). J. Pajohesh Sazandegi. 2001; 61(1): 28-32.
49. Mohammadi Golrang B., Gazanchian G., Ramzani Moghadam R., Falahati H., Rouhani H., Mashayekhi M. Estimation of forage productions of some range plant species by plant height and diameter measurements. Iran. J. Range Desert Res.2008;15(2):158-178.

[DOI: 10.22034/ecopersia.12.1.39]