https://doi.org/10.31298/sl.148.3-4.5

Prethodno priopćenje – Preliminary communication
Šumarski list, 3–4 (2024): 161–167

LINEFIT.XLS: A MICROSOFT EXCEL TEMPLATE FOR FITTING 11 REGRESSION MODELS TO Y-X DATA

LINEFIT.XLS: PREDLOŽAK PROGRAMA MICROSOFT EXCEL ZA PRILAGODBU 11 REGRESIJSKIH MODELA PODACIMA Y-X

Kyriaki KITIKIDOU[1]*, Elias MILIOS[1]

SUMMARY

An essential challenge in any environmental field (forestry, agriculture, etc.), which places an emphasis on data analysis for the purpose of decision-making and problem-solving, is the estimation of a dependent environmental variable (Y) through an independent one (X). In this work, a Microsoft Excel template is proposed for assessing a set of eleven popular regression Y-X models. Any researcher could use LineFit.xls as a modeling tool for assessing these eleven regression models and selecting the one that best fits their data by running tests on all regression assumptions and comparing models using the most common fitting comparison criteria. Microsoft Excel, being a widely used and user-friendly program, makes it easy to update, expand, and personalize the tests to meet specific needs.

KEY WORDS: data fitting, data modeling, fit comparison, MS Office software

INTRODUCTION
UVOD

All environmental sciences rely heavily on methods that estimate the value of a dependent variable (Y) using a value for a related independent variable (X).

For instance, Freese (1964) noted that in the field of forestry, the volume of a tree (Y) may be described as a function of its breast height diameter (X), the strength of wood (Y) as a function of its specific gravity (X), and the cost of logging (Y) as a function of its proximity to hard-surfaced roads (X).

In agriculture, for example, soil organic matter (Y) may be expressed as a function of time (X), net photosynthesis (Y) can be related to irradiance (X), and respiration (Y) can be described as a function of temperature (X) (Archontoulis and Miguez, 2015).

Environmentalists might greatly benefit from a modeling tool that allows them to compare and assess a set of common regression models using the most frequently used fitting comparison criteria and conduct tests on all regression assumptions to see which one best fits their data.

Given Excel's popularity and ease of use, it would be an excellent option to provide a template for assessing a collection of such regression models.

MATERIALS AND METHODS
MATERIJALI I METODE

Within the LineFit.xls template, we suggest conducting tests on 11 different regression models (SPSS, 2007), as shown in Table 1

[1] Prof. Dr. Kyriaki Kitikidou, PhD, Prof. Dr. Elias Milios, PhD, Democritus University, Department of Forestry and Management of the Environment and Natural Resources, 68200, Orestiada, Greece
* Corresponding author: Kyriaki Kitikidou, kkitikid@fmenr.duth.gr

Table 1. Models assessed with the LineFit template
Tablica 1. Modeli ocijenjeni pomoću predloška LineFit

No. *Br.*	Model *Model*	Y-X *Y-X*
1	Linear	$Y = b_0 + b_1X$
2	Logarithmic	$Y = b_0 + b_1 \ln X$
3	Inverse	$Y = b_0 + \frac{b_1}{X}$
4	Quadratic	$Y = b_0 + b_1X + b_2X^2$
5	Cubic	$Y = b_0 + b_1X + b_2X^2 + b_3X^3$
6	Compound	$Y = b_0 b_1^X$
7	Power	$Y = b_0X^{b_1}$
8	S-curve	$Y = e^{b_0 + \frac{b_1}{X}}$
9	Growth	$Y = e^{b_0 + b_1X}$
10	Exponential	$Y = b_0e^{b_1X}$
11	Logistic	$Y = \frac{1}{\frac{1}{u} + b_0 b_1^X}$

b_i: regression coefficients.
u: upper boundary value of Y variable.

Linear regression is a useful statistical approach for estimating the value of a dependent variable (Y) through an independent variable (X). However, the following five assumptions must be satisfied before we are able to assess regression models' fit to our data:

1. Linearity: b_i regression coefficients should be statistically significantly different from zero.
2. Independence: Residuals should be independent, non-correlated.
3. Homoscedasticity: Residuals should have constant (homogeneous) variance across all X_i values.
4. Zero error: Residuals should have zero average.
5. Normality: Residuals should approximate a normal distribution.

Table 2 summarizes the five aforementioned assumptions along with the relevant literature.

Table 3 provides three statistical comparison criteria and corresponding literature, for selecting the best fitted regression model for a given set of data

Table 2. Regression assumptions examined with the LineFit template
Tablica 2. Pretpostavke regresije ispitane u skladu s predloškom LineFit

No *Br.*	Regression assumption *Regresijska pretpostavka*	Test *Test*	It should be: *Trebala bi biti:*	Reference *Referenca*
1	Linearity	One-sample t-test for b_i regression coefficients, critical value is zero.	p-value$<$0.05	Student, 1908
2	Independence	Durbin-Watson DW value for residuals	$1 \leq DW \leq 3$	Durbin and Watson, 1950; Durbin and Watson, 1951. Field (2009) suggests that DW values under 1 or more than 3 are a definite cause for concern.
3	Homoscedasticity	Koenker-Bassett (generalized Breusch–Pagan) homoscedasticity (homogeneity of variance) test for residuals	p-value$>$0.05	Koenker and Bassett, 1982
4	Zero error	One-sample t-test for residuals, critical value is zero.	p-value$>$0.05	Student, 1908
5	Normality	Jarque-Bera test	p-value$>$0.05	Jarque, 2011

Table 3. Statistical criteria for the comparison of regression models calculated with the LineFit template.
Tablica 3. Statistički kriteriji za usporedbu regresijskih modela izračunanih pomoću predloška LineFit.

No *Br.*	Criterion *Kriterij*	Formula *Formula*	Optimum value *Optimalna vrijednost*	Reference *Referenca*
1	Coefficient of determination	$R^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$	1	Draper and Smith, 1997; Kitikidou, 2005
2	Standard error of the estimate	$SEE = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - Y_i)^2}{n - p}}$	min	Wackerly et al., 1996; Ezekiel and Fox, 1959; Mathews, 1987; Draper and Smith, 1997; Kitikidou, 2005
3	Root of the mean squared error	$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n}}$	min	Draper and Smith, 1997; Kitikidou, 2005

Y_i: observed i-th value of Y
$\hat{Y}_i$: estimated i-th value of *Y* from the regression model
p: number of regression coefficients b_i
n: number of observations

RESULTS
REZULTATI

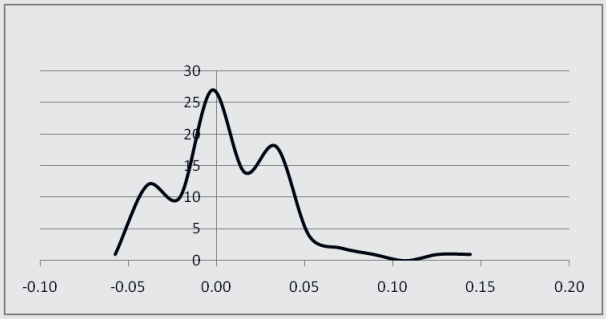
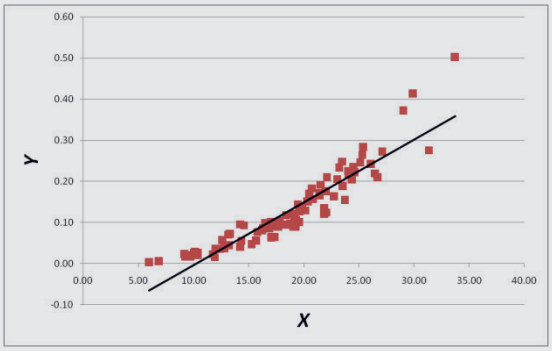
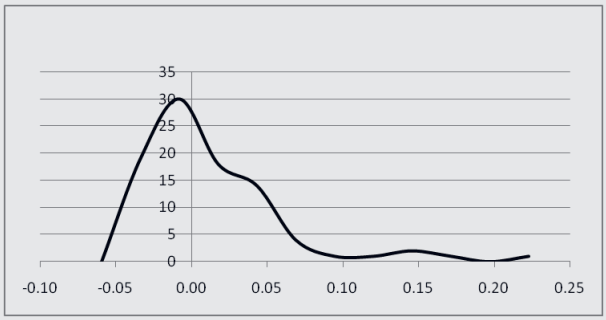
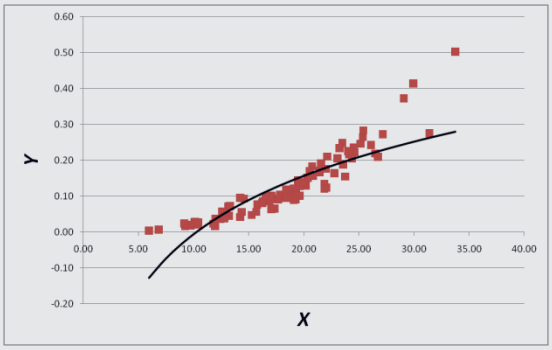
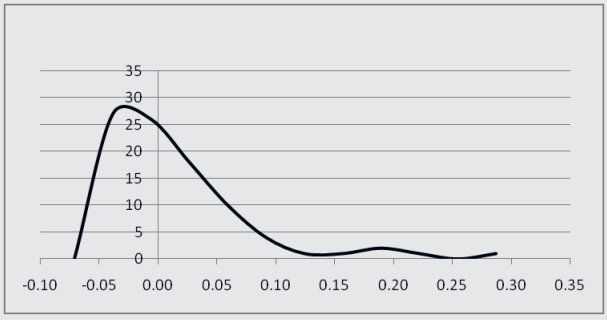
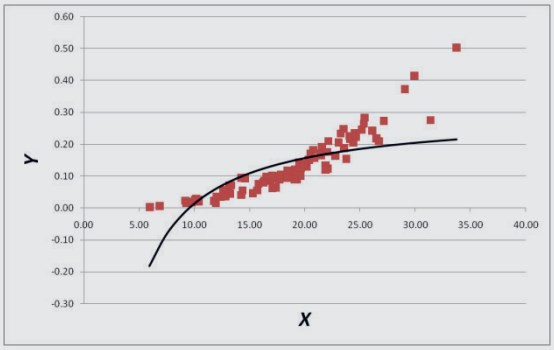
LineFit.xls includes 13 spreadsheets. In the "data" spreadsheet, one can enter their own Y-X data in columns A and B, respectively (highlighted in blue). The amount of data can be up to 65535 entries (pairs of Y-X data). Summary statistics (count, mean, standard deviation, min, and max values) are given in cells D1:I4.

For demonstration purposes, columns A and B in the "data" spreadsheet are filled with a random number generator. By pressing the F9 key, one can observe all changes throughout the template.

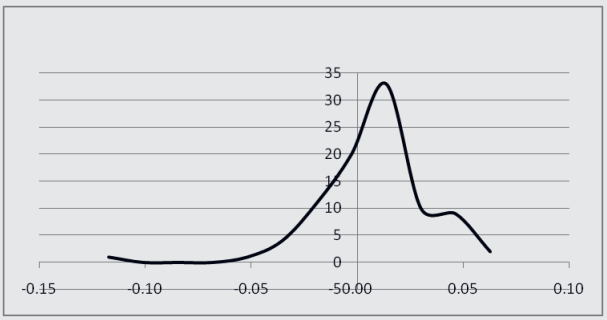
In the spreadsheets from "1" up to "11", the statistics of Tables 2 and 3 are calculated, for each of the 11 regression models of Table 1. The linearity test is calculated across cells C5:F5, the independence test in cell M2, the homoscedasticity test in cell AO2, the zero error test in cell AM5, and the normality test in cell AM8. When the regression assumption is not satisfied, the font in these cells turns red. In addition, the graph of the residuals' distribution and the Y-X scatterplot, along with the fitted line, are given for each of the 11 regression models.

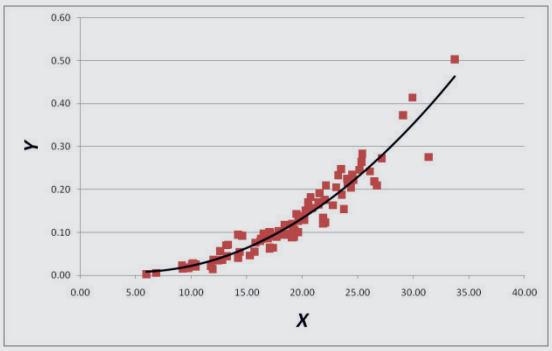
An example of these graphs produced from volume (V, m^3) – diameter at breast height (DBH, cm) data (Y-X data) (Softa, 2023) is given in Table 4.

Table 4. Graphs produced with the LineFit template from sample Y-X data
Tablica 4. Grafikoni proizvedeni pomoću predloška LineFit iz uzorka podataka Y-X

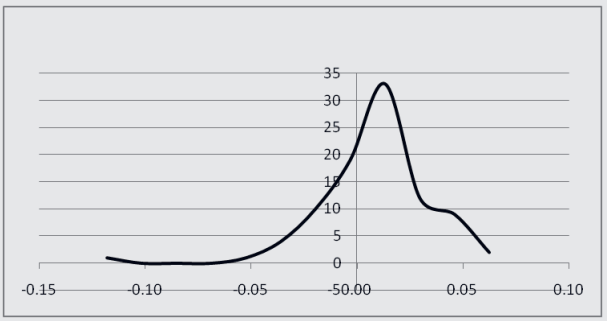
No *Br*	Residuals' distribution *Distribucija reziduala*	Y-X scatterplot with fitted line *Raspršeni grafikon Y-X s prilagođenom linijom*
1		
2		
3		

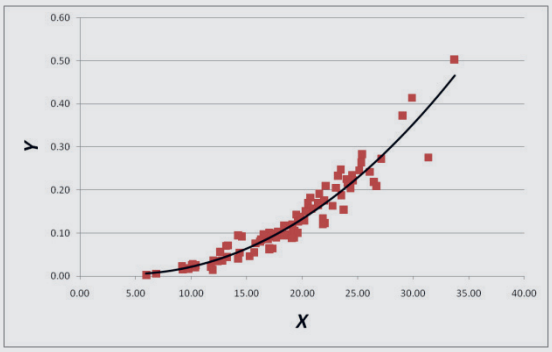
4


35
30
25
20
15
10
5
0
-0.15
-0.10
-0.05
-50.00
0.05
0.10

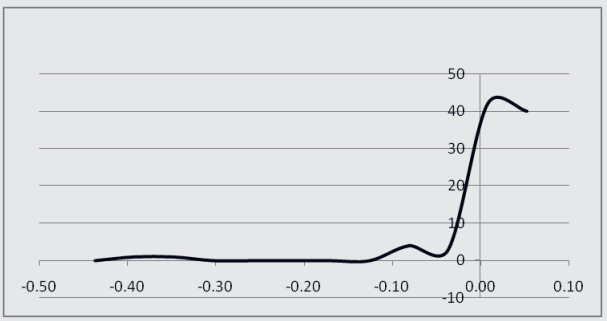

0.60
0.50
0.40
0.30
0.20
0.10
0.00
Y
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
X

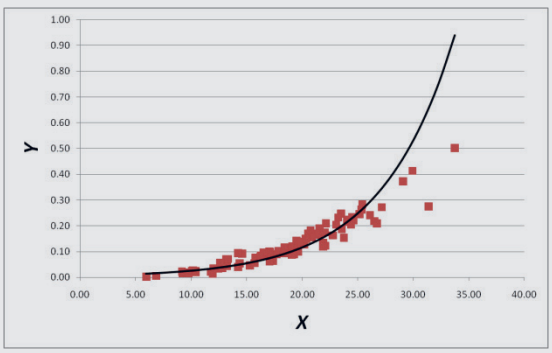
5


35
30
25
20
15
10
5
0
-0.15
-0.10
-0.05
-50.00
0.05
0.10

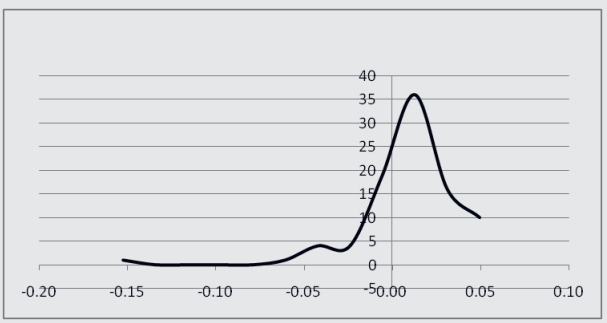

0.60
0.50
0.40
0.30
0.20
0.10
0.00
Y
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
X

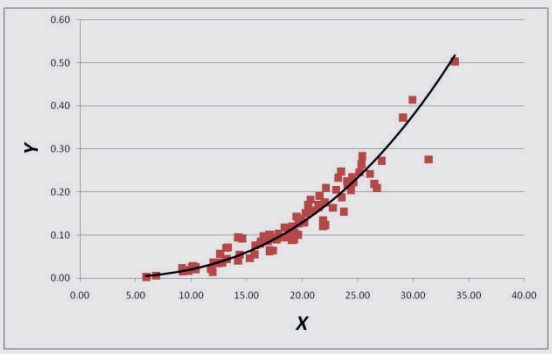
6


50
40
30
20
10
0
-10
-0.50
-0.40
-0.30
-0.20
-0.10
0.00
0.10

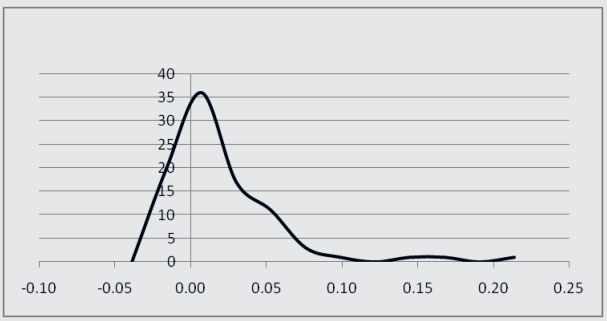

1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Y
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
X

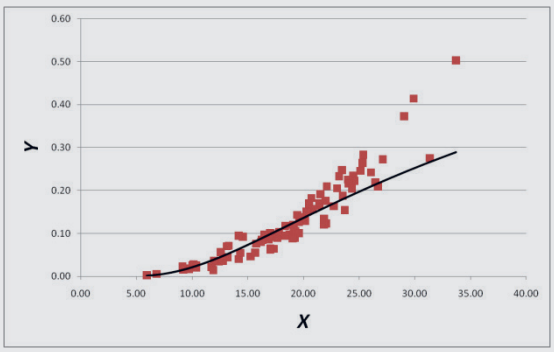
7

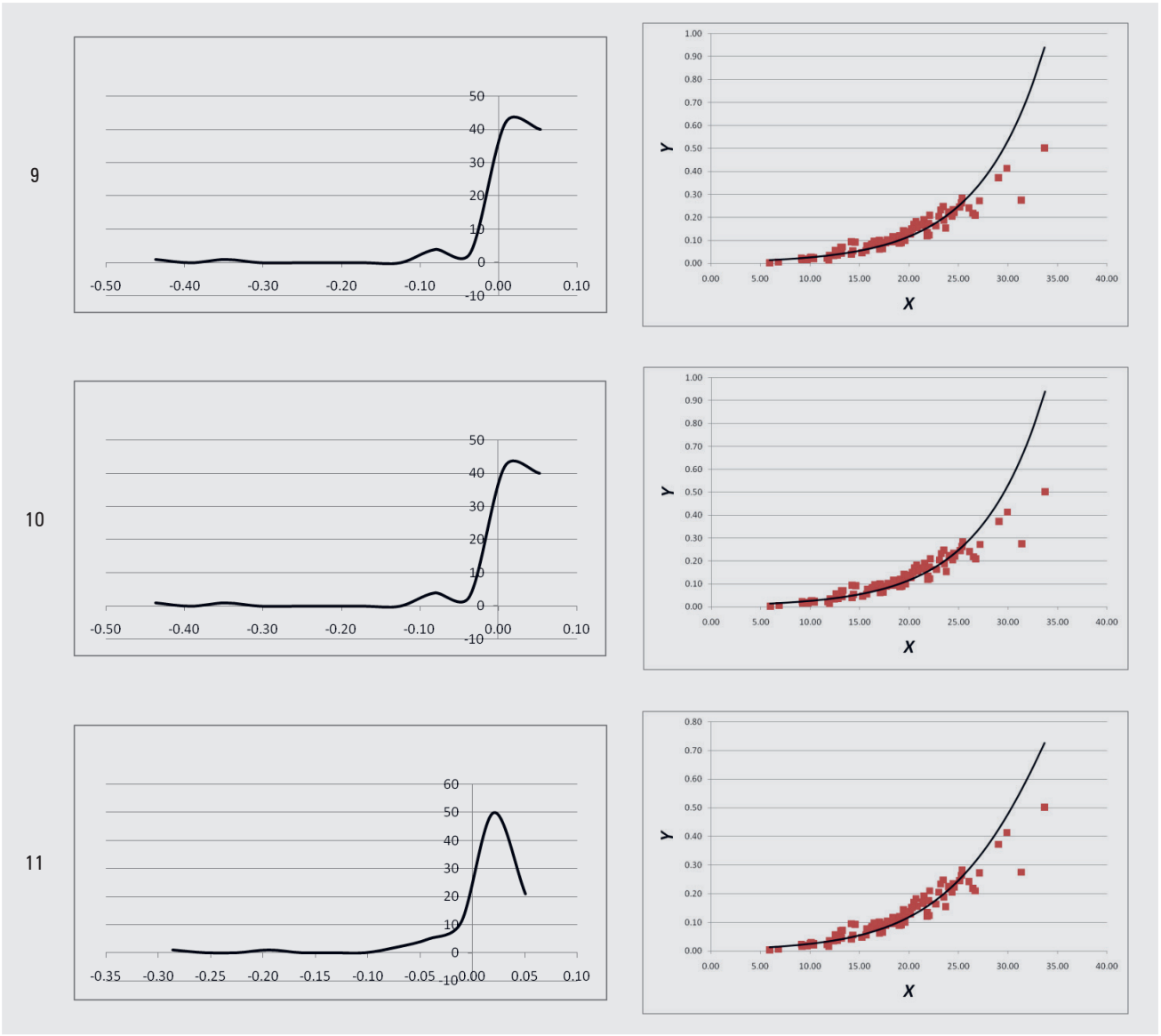

40
35
30
25
20
15
10
5
0
-5
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.10


0.60
0.50
0.40
0.30
0.20
0.10
0.00
Y
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
X

8


40
35
30
25
20
15
10
5
0
-0.10
-0.05
0.00
0.05
0.10
0.15
0.20
0.25


0.60
0.50
0.40
0.30
0.20
0.10
0.00
Y
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
X



Regarding the statistical criteria for the comparison of the 11 regression models, the coefficient of determination R^2 is calculated in cell J2, the Standard Error of the Estimate (SEE) in cell K2, and the Root of the Mean Squared Error (RMSE) in cell L2 in the corresponding spreadsheet for each of the 11 regression models.

The regression assumptions listed in Table 2 and the statistical comparison criteria listed in Table 3 are both visible in their aggregated forms, in the "comparison" spreadsheet. At this point, the researcher has the opportunity to choose a particular regression model that will serve their needs in the most efficient way. When a regression assumption is not met, the font in these cells becomes red as well. Especially for the linearity test, when at least one regression coefficient is not statistically significantly different from zero, the "violated" message is displayed in red font. In the event that all regression coefficients are statistically significantly different from zero, the message "OK" is displayed.

Regarding the independence test, cells C3:C13 are conditionally highlighted. When a value is between 1.5 and 2.5, the cell is highlighted in dark green (optimum value), for a value between 1 and 1.5 or 2.5 and 3, the corresponding cell is highlighted in light green (acceptable value), and for a value >3 or <1, the corresponding cell is highlighted in yellow (the assumption of independence is not met).

The cells where the three statistical comparison criteria are calculated (i.e., the cells H3:J13) are highlighted on a graduated scale, from dark green (best values) to yellow (worst values).

For the data by Softa (2023), the results of the "comparison" spreadsheet are given in Table 5. Based on these results, one might decide (subjectively, not necessarily) that the best fitted model is the power (no 7) model:

$$V = 0.000046 \cdot DBH^{2.6513}$$

Table 5. Results of the "comparison" spreadsheet
Tablica 5. Rezultati proračunske tablice usporedbe ("comparison")

Model	Regression Assumptions					Model	Comparison Criteria		
	Linearity	Indepen-dence	Homosce-dasticity	Residuals' zero mean	Residuals' normality		R^2	SEE	RMSE
1	OK	1.5438	0.0005	1.0000	0.0000	1	0.8577	0.0350	0.0346
2	OK	1.5459	0.1528	1.0000	0.0000	2	0.7273	0.0484	0.0479
3	OK	1.5640	0.8899	1.0000	0.0000	3	0.5438	0.0626	0.0619
4	violated	1.6663	0.0000	1.0000	0.0000	4	0.9197	0.0263	0.0260
5	violated	1.6733	0.0000	1.0000	0.0000	5	0.9198	0.0263	0.0260
6	OK	2.0161	0.0000	0.3513	0.0000	6	0.8174	0.0396	0.0685
7	OK	1.8419	0.0000	0.9686	0.0000	7	0.9172	0.0267	0.0273
8	violated	1.3253	0.0000	0.0467	0.0000	8	0.8664	0.0339	0.0403
9	OK	2.0161	0.0000	0.3513	0.0000	9	0.8174	0.0396	0.0685
10	OK	2.0161	0.0000	0.3513	0.0000	10	0.8174	0.0396	0.0685
11	OK	1.9999	0.0000	0.5945	0.0000	11	0.8695	0.0335	0.0469

because the linearity, independence and residuals' zero mean assumptions are not violated, and the comparison criteria have satisfactory values, compared with those of the other 10 models.

DISCUSSION AND CONCLUSION
RASPRAVA I ZAKLJUČAK

LineFit.xls is an all-in-one template whose outputs represent the behavior of eleven regression models, allowing any researcher to assess the fit of these models to their Y-X data. It is important to remember, however, that regression, as a parametric statistical process, may still perform pretty well even if some of the assumptions behind it are violated (Kitikidou et al., 2012; Kitikidou et al., 2013). Before rejecting a regression model because the assumptions are not met and choosing a non-parametric analysis, one should critically consider that non-parametric analyses employ rankings of values rather than the values themselves and hence cannot provide usable, quantitative estimations (Altman, 2009).

The LineFit.xls MS Excel template is available as a downloadable file on this journal's website to offer the scientific community an option to test and evaluate the most popular Y-X regression models with an all-in-one modeling tool.

REFERENCES
LITERATURA

- Altman, D., 2009: Parametric v non-parametric methods for data analysis,Brit. Med. J., 338:a3167.
- Archontoulis, S.V., F.E., Miguez, 2015: Nonlinear regression models and applications in Agricultural Research, Agron. J., 107(2): 786-798.
- Draper, N.R., H. Smith, 2019: Applied regression analysis, Wiley India Private Limited, Delhi, India.
- Durbin, J., G.S., Watson, 1950: Testing for serial correlation in least squares regression. I. Biometrika, 37(3/4): 409-428.
- Durbin, J., G.S., Watson, 1951: Testing for serial correlation in least squares regression. II. Biometrika, 38(1/2): 159-178.
- Ezequiel, M., K.A., Fox, 1959:Methods of correlation and regression analysis: Linear and curvilinear, John Wiley and Sons, New York, USA.
- Field, A.P., 2009: Discovering statistics using SPSS, SAGE, California, USA.
- Freese, F., 1964: Linear regression methods for Forest Research, U.S. Dept. of Agriculture, Forest Service, Forest Products Laboratory.
- Jarque, C.M., 2011: Jarque-Bera Test,in: Lovric, M. (ed) International Encyclopedia of Statistical Science. Springer, Berlin, Germany, pp 701-702.
- Kitikidou, K., 2005: Applied statistics using the SPSS statistical package,Greek, Tziola publications, Thessaloniki.
- Kitikidou, K., E., Milios, L., Iliadis, M., Kaymakis, 2012: Combination of M-estimators and neural network model to analyze inside/outside bark tree diameters,IFIP Adv. Inf. Comm. Te., 381:11-18.
- Kitikidou, K., E. Milios, L. Iliadis, M., Kaymakis, 2013: Pilot neural modeling of the inside bark tree diameter. A comparative study with robust regression, Eng. Intell. Syst., (2/3): 125-131.
- Koenker, R., G., Bassett, 1982: Robust tests for heteroscedasticity based on regression quantiles, Econometrica, 50(1): 43-61.
- Mathews, J.H., 1987: Numerical methods for computer science, engineering, and Mathematics, Prentice-Hall, New Jersey, USA.

- Softa, E. 2023: Dendrometrical characteristics of *Robinia pseudoacacia* in degraded areas of Xanthi region. MSc thesis, Democritus University of Thrace, Greece.
- SPSS, Inc., 2007: SPSS statistics base 17.0 user's guide, Chicago, USA.
- Student, 1908: The probable error of a mean, Biometrika, 6(1): 1-25.
- Wackerly, D.D., W., Mendenhall, R.L., Scheaffer, 2012:Mathematical statistics with applications, Brooks/Cole, California, USA.

SAŽETAK

Ključni izazov u bilo kojem području koje proučava okoliš (šumarstvo, poljoprivreda itd.), koji stavlja naglasak na analizu podataka u svrhu donošenja odluka i rješavanja problema, procjena je zavisne okolišne varijable (Y) kroz nezavisnu varijablu (X). U ovom radu predstavljen je predložak programa Microsoft Excel za procjenu jedanaest popularnih regresijskih modela Y-X. Svaki istraživač moći će koristiti LineFit.xls kao alat za modeliranje za procjenu jedanaest regresijskih modela i odabir modela koji najbolje odgovara njihovim podacima provođenjem testova na svim regresijskim pretpostavkama i uspoređivanjem modela koristeći najčešće kriterije usporedbe. Budući da je Microsoft Excel program široke primjene i jednostavan je za korištenje, omogućuje jednostavno ažuriranje, proširenje i personalizaciju testova kako bi se zadovoljile pojedinačne potrebe.

KLJUČNE RIJEČI: prilagodba podataka, modeliranje podataka, usporedba prikladnosti, MS Office softver