Exploring the Impact of Diverse Biological Factors on Metabolic Rates in a Wide Spectrum of Living Organisms through Linear Regression Analysis

By: Makayla Avendano

Introduction

Assessing metabolic rate is a focal point when evaluating living organisms. At a superficial level, metabolic rate is defined as the energy expended by a living organism to sustain life and basic bodily functions.

Problem

Even though the scientific understanding of metabolic rate is relatively advanced, there is still significant opportunity for knowledge advancement and scientific exploration. Building a linear regression model will help assess the relationship between various factors and metabolic rates while allowing for further ecological and environmental understanding. “Assessing the relationship between energy flux and the quantity of biomass it sustains offers the potential to understand the biological “carrying capacity” for ecosystems on Earth and beyond. Our work supports this understanding by quantifying the energy–biomass relationship for the global biosphere and an environmentally diverse range of its components, and by exploring the factors—including the impact of humanity—that affect that relationship.” (Hoehler et al., 2023)

Hypothesis

Hypothesis: At least one of the studied variables has a significant influence on the metabolic rate of the included living organisms.
Null Hypothesis: There is no significant influence of any of the studied variables on the metabolic rate of the included living organisms.

Data Analysis Process

Due to the nature of this project’s topic, data will not need to be collected and will instead reference a completed data set. This data set was retrieved from the Data is Plural newsletter that highlighted the data set created and published within the environmental sciences journal Proceedings of the National Academy of Sciences. The data set that will be used is a database that was compiled from multiple different published sources. These sources then compiled data to include 10,000+ rows of various living organisms.

Data Cleaning

Data cleaning is a necessary step to make sure we are addressing potential data integrity threats, the unnecessary skewing of data, and the potential inflation or deflation of values. The focus for data cleaning was on duplicates, missing values, and outliers.

Duplicates: No duplicates were found within the data, so therefore duplicates did not need to be addressed.
Missing values: There were 5 numerical variables with missing values. The missing values were treated using imputation and the median values of each variable.
Outliers: There were also 5 numerical variables with outliers. Outliers were treated with the retain method. All outliers found were within reason and were not a preexisting error, therefore all outliers were kept and not treated.

Data Analysis

The analysis technique used to address metabolic rates within living organisms is multiple linear regression. Multiple linear regression is, “a statistical tool used to model the relationship between a continuous response (dependent) variable and one or more continuous and/or categorical explanatory (independent) variables.” (Middleton, 2022) Variance Inflation Factor (VIF) and backward stepwise elimination were also used to reduce and improve the analysis model. VIF is used to address and reduce multicollinearity within the data. Backward stepwise elimination is a method to reduce the model based on variables p-values.

Outline of the findings

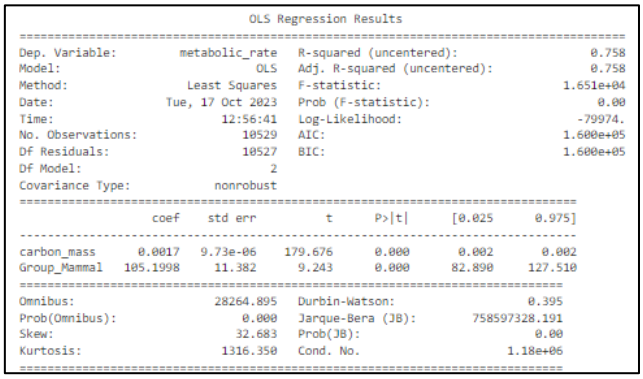
Model Reduction

Variable elimination is a vital step in multiple linear regression and the creation of a strong model. To begin, the model utilized 31 variables. In the final reduced model, 2 variables were remaining. Variance Inflation Factor (VIF) was utilized to reduce multicollinearity which gives values of correlation between variables in a regression model. Due to low multicollinearity being an assumption of linear regression, it’s important to address this issue using VIF. After performing VIF, variables are removed based on the highest value. Values were removed until a VIF value of 10 was reached which resulted in 16 variables being removed.

The linear regression model was then created, and a backward stepwise elimination method was used to reduce the model further. Backward stepwise elimination is defined as a wrapper method that, “feed the features (variables) for your model and based on the model performance you add/remove the features.” (Middleton, 2022) This was performed by evaluating the initial model variables and their corresponding p-values. At the end of the evaluation, I removed a total of 15 variables. With this technique, we were able to reduce the model and focus on those variables that have more influence on the dependent variable and get rid of those variables that do not.

The residual standard error (RSE) was calculated for both the initial and reduced model (shown to the right). The residual standard error (RSE) is, “used to measure how well a regression model fits a dataset. In simple terms, it measures the standard deviation of the residuals in a regression model.” (Middleton, 2022) The initial model RSE was 481.74 and the reduced model RSE was 481.41. Since a lower RSE is better, it appears that the model got slightly better after reduction even though the reduction is extremely small.

Final Model



OLS Regression Results

Dep. Variable:	metabolic_rate	R-squared (uncentered):	0.758
Model:	OLS	Adj. R-squared (uncentered):	0.758
Method:	Least Squares	F-statistic:	1.651e+04
Date:	Tue, 17 Oct 2023	Prob (F-statistic):	0.00
Time:	12:56:41	Log-Likelihood:	-79974.
No. Observations:	10529	AIC:	1.600e+05
Df Residuals:	10527	BIC:	1.600e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
carbon_mass	0.0017	9.73e-06	179.676	0.000	0.002	0.002
Group_Mammal	105.1998	11.382	9.243	0.000	82.890	127.510

Omnibus:	28264.895	Durbin-Watson:	0.395
Prob(Omnibus):	0.000	Jarque-Bera (JB):	758597328.191
Skew:	32.683	Prob(JB):	0.00
Kurtosis:	1316.350	Cond. No.	1.18e+06

Through the performed analysis and resulting model, it’s evident that the variables carbon mass and the mammal group influence the metabolic rate of various living organisms. This conclusion can be reached due to the statistical significance of our model. To determine statistical model significance, I focused on the F-statistic and the p-value. The p-value or Prob(F-statistic) in the summary of the reduced model indicated a value of 0.0. This value is less than the significance level of 0.05 and therefore leads to the

conclusion that we reject the null hypothesis and determine that there is a relationship that exists within this model. (Straw, 2023) As mentioned above, another evaluation metric to look at would be the F-statistic which was 16,510. Utilizing the F-table of critical values, I discovered that the F-statistic (16,510) was greater than the critical value leading to the conclusion that the model is statistically significant. (Straw, 2023)

Coefficients

These coefficients indicate the effect that each variable has on metabolic rate. When a living organism is characterized as being a part of the mammal group, this causes an increase in metabolic rate by approximately 105. Similarly, a one-unit increase in carbon mass is associated with a 0.017 increase in metabolic rate.

Variable	Coefficient Value
Mammal Group	105.1998
Carbon Mass	0.017

Limitations

Throughout the performed analysis, there are a few notable limitations. A few stem from the data cleaning phase of the analysis: imputation for missing values and the retention of outliers. These data cleaning procedures allow for unnecessary influence causing an added skewness to the data as well as potentially causing bias. The remaining limitations stem from the actual analysis techniques. Multiple linear regression is highly sensitive to outliers leading to a potentially inaccurate regression line. Variance Inflation Factor (VIF) is used to eliminate multicollinearity, but it may not be completely successful and may leave small amounts of multicollinearity within the model. Lastly, backward stepwise elimination could cause the potential elimination of important variables that have casual effects on metabolic rate.

Summary of Proposed Actions

Based on the results, a recommended course of action would be a more exploratory approach on the most influential variables. Digging deeper into carbon mass, mammals, and metabolic rate would allow for a more thorough understanding of this relationship. Performing exploratory analysis, creating visualizations to identify trends, and looking at different statistical breakdowns can reveal a more in-depth interpretation of the influence of these factors on metabolic rate.

Future studies could also be beneficial focusing on exploratory analysis and a statistical ANOVA test looking at metabolic rates and group of living organism significance. Another future study includes a closer look at endotherms and metabolic rate. Temperature was not a focus of this study and looking at temperature effects could help with further information on metabolic plasticity.

Benefits

Taking the information directly from the analysis, we can also benefit from understanding that carbon mass is more impactful than dry and wet mass. This could lead to further studies that focus solely on carbon mass rather than dry and wet mass. Lastly, mammals increase their metabolic rate by

approximately 105. This information can be used to address metabolic plasticity in mammals and help with understanding environmental changes in the future.

As mentioned previously, although the scientific understanding of metabolic rate is relatively advanced, there is still significant opportunity for knowledge advancement and scientific exploration. This analysis holds substantial potential for scientists and biologists, serving as a valuable foundation for advancing knowledge and enhancing understanding in their fields. Areas of advancement and exploration include a comprehensive understanding of the flow of energy throughout an ecosystem, the carrying capacity of organisms, and the energy-to-biomass relationship within the biosphere, with fluctuations in environments, and we can look to advance our knowledge on metabolic plasticity and the responses for different living organisms.

Sources

Hoehler, T. M., Mankel, D. J., Girguis, P. R., McCollom, T. M., Kiang, N. Y., & Jørgensen, B. B. (2023). The metabolic rate of the biosphere and its components. Proceedings of the National Academy of Sciences of the United States of America, 120(25), e2303764120. https://doi.org/10.1073/pnas.2303764120

Middleton, K. (2022). D208 – Webinar: Getting Started with D208 Part I [Powerpoint Slides]. https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=15e09c73-c5aa-439d-852f-af47001b8970.

Straw, E. (2023) Dr. Straw's Tips for Success in D208. Student Facing Resources. https://westerngovernorsuniversity.sharepoint.com/sites/DataScienceTeam/Shared%20Documents/Forms/AllItems.aspx?csf=1&web=1&e=9ccodm&cid=1dfa8779%2D624c%2D4a36%2D80d2%2D62325caa29f6&FolderCTID=0x01200022092E63FD85A64A8ABFB4F5AEA4839A&id=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD208%2FStudent%20Facing%20Resources%2FDr%2E%20Straw%20tips%20for%20success%20in%20D208%2Epdf&parent=%2Fsites%2FDataScienceTeam%2FShared%20Documents%2FGraduate%20Team%2FD208%2FStudent%20Facing%20Resources