biology

MDPI

Article

Similarity-Based Predictive Models: Sensitivity Analysis and a Biological Application with Multi-Attributes

Jeniffer D. Sanchez [1], Leandro C. Rêgo [1,2], Raydonal Ospina [2,3], Víctor Leiva [4,*], Christophe Chesneau [5] and Cecilia Castro [6]

[1] Department of Statistics and Applied Mathematics, Universidade Federal do Ceara, Fortaleza 60020-181, Brazil; jjduartes@dema.ufc.br (J.D.S.), leandro@dema.ufc.br (L.C.R.)
[2] Department of Statistics, Universidade Federal de Pernambuco, Recife 50670-901, Brazil; raydonal@de.ufpe.br
[3] Department of Statistics, IME, Universidade Federal da Bahia, Salvador 40170-110, Brazil
[4] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso 2362807, Chile
[5] Department of Mathematics, Université de Caen, 14032 Caen, France; christophe.chesneau@unicaen.fr
[6] Centre of Mathematics, Universidade do Minho, 4710-057 Braga, Portugal; cecilia@math.uminho.pt
* Correspondence: victor.leiva@pucv.cl or victorleivasanchez@gmail.com

Simple Summary: In this study, we perform a sensitivity analysis in similarity-based predictive models using computational simulations and two distinct methodologies, while focusing on a biological application. We utilize a linear regression model as a reference point. We gauge sensitivity by calculating the coefficient of variation of the parameter estimators from three different models. Our findings show that the first approach outperforms the second one when dealing with categorical variables. Moreover, this first approach offers the advantage of being more parsimonious due to a smaller number of parameters.

Abstract: Predictive models based on empirical similarity are instrumental in biology and data science, where the premise is to measure the likeness of one observation with others in the same dataset. Biological datasets often encompass data that can be categorized. When using empirical similarity-based predictive models, two strategies for handling categorical covariates exist. The first strategy retains categorical covariates in their original form, applying distance measures and allocating weights to each covariate. In contrast, the second strategy creates binary variables, representing each variable level independently, and computes similarity measures solely through the Euclidean distance. This study performs a sensitivity analysis of these two strategies using computational simulations, and applies the results to a biological context. We use a linear regression model as a reference point, and consider two methods for estimating the model parameters, alongside exponential and fractional inverse similarity functions. The sensitivity is evaluated by determining the coefficient of variation of the parameter estimators across the three models as a measure of relative variability. Our results suggest that the first strategy excels over the second one in effectively dealing with categorical variables, and offers greater parsimony due to the use of fewer parameters.

Keywords: biological data; coefficient of variation; data science; distance measures; estimation methods; Monte Carlo simulation; predictive modeling; similarity functions

Citation: Sanchez, J.D.; Rêgo, L.C.; Ospina, R.; Leiva, V.; Chesneau, C.; Castro, C. Similarity-Based Predictive Models: Sensitivity Analysis and a Biological Application with Multi-Attributes. *Biology* **2023**, *12*, 959. https://doi.org/10.3390/biology12070959

Academic Editor: Andrés Moya and Jacques Demongeot

Received: 13 May 2023
Revised: 20 June 2023
Accepted: 27 June 2023
Published: 4 July 2023

1. Introduction

The empirical similarity prediction method does not assume a specific functional form relating the response variable to the covariates. Instead, it estimates the response variable value based on a weighted average of past response variable values, where the weights depend on the similarity of the covariate values. To apply empirical similarity in practice, a similarity measure and an estimation method are necessary.

Similarity measures are functions of distance that decrease as the distance decreases. They equal one when the distance is zero and converge to zero as the distance approaches infinity. The literature commonly considers two similarity measures: exponential inverse (EX) and fractional inverse (FR). These measures incorporate weighted distances, where the weights represent the relative importance of each covariate or level of a categorical variable. Estimating these weights from the data requires two methods: ordinary least squares (OLS) and maximum likelihood (ML).

The concept of empirical similarity has been axiomatized as a means to replicate human reasoning or natural behavior [1,2]. In [3], the identification, consistency, and distribution problems of the ML estimator for similarity models' parameters were analyzed.

Categorical data, which includes multi-attribute records, are a crucial type of biological observations as they involve separable data and qualitative characteristics. Categorical data classify samples into mutually exclusive categories, often by counting the number of objects that fall into each qualitative class [4–6]. When dealing with categorical covariates, the empirical similarity literature within biological data describes two predictive approaches. The first approach, denoted as M1 and proposed in [7], maintains the categorical variables in their original formats. It assigns equal importance (weight) to all levels of the variable. The second approach, denoted as M2 and proposed in [8], encodes the categorical variables into binary variables, treating each category as a separate variable. In this case, different weights (influences) can be associated with each category of the same variable.

In predictive models, linear regression is a well-known and often used method. However, when dealing with categorical covariates, its utility can sometimes be limited. While linear regression provides a simple and interpretable model, it may not always capture the complexities of categorical covariates effectively. Therefore, alternative methods, such as empirical similarity models, may provide more nuanced and accurate predictions when dealing with such data types. Still, in our study, linear regression is utilized as a benchmark to provide a familiar frame of reference to readers and to aid comparison.

To the best of our knowledge, no previous studies have examined the sensitivity analysis of a specific class of similarity models concerning the accuracy of predicted values and the sensitivity of parameter estimators for the M1 and M2 methods. The choice of similarity and distance measures has been subjective in previous research. Thus, this study seeks to fill such a gap by performing a sensitivity analysis. Our study provides value by identifying which method yields the most robust predictions and parameter estimates under different scenarios.

Our main objective is to select similarity and distance measures that yield lower prediction errors and parameter estimators with reduced variability. The sensitivity of these models to environmental variations is simulated by splitting the data into training and test sets and calculating the coefficient of variation (CV) over multiple repetitions. The CV is a dimensionless and standardized measure of dispersion relative to the average of a dataset [9,10]. Given the different scales of the weights in the models, the CV is a suitable measure in our context.

To demonstrate the practical utility of our analysis, we employ a dataset on tooth length growth in Guinea pigs [11]. This dataset, involving different dosage levels and delivery methods of vitamin C, illustrates the models' potential applications in biological research. The structure of the article is as follows: Section 2 provides a theoretical overview of empirical similarity and linear regression models. Section 3 describes the tooth length growth dataset used for simulating the sensitivity analysis. The methodology employed in the simulation study is detailed in Section 4. The results of the sensitivity analysis are presented in Section 5. In Section 6, our conclusion states a comparative analysis of M1, M2, and linear regression models, illustrating their competitive performances as gauged by the CV. We highlight the M1 method for its exceptional parsimony. The insights drawn from our research have potential to inform and guide researchers in selecting appropriate similarity and distance measures. Such informed selections can subsequently ensure predictions with enhanced accuracy and robustness in their parameter estimates.

2. Theoretical Background

In this study, we first introduce the linear regression model [12], as it serves like a benchmark for our detailed exploration of the performance of different predictive models under various scenarios. Consider a sample of size n with for the response variable, denoted as $Y_1, \ldots, Y_n$, which can be formulated as:

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{w} + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \tag{1}$$

where $\boldsymbol{x}_i^\top = (x_{1i}, \ldots, x_{mi})$ represents a $1 \times m$ vector of observed covariates, $\boldsymbol{w} = (w_1, \ldots, w_m)^\top$ is an $m \times 1$ vector of weights for the regression model (fixed effects), and ε_i denotes the model random error, with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. It is assumed that $(\varepsilon_1, \ldots, \varepsilon_n)^\top$ are independent and identically distributed.

Let $\boldsymbol{X}$ be an $n \times m$ matrix with rank m, where each row represents $\boldsymbol{x}_i^\top$ (note that $\boldsymbol{X}$ is the known incidence matrix relating observations to fixed effects). Hence, $Y_i \sim \mathrm{N}(\boldsymbol{x}_i^\top \boldsymbol{w}, \sigma^2)$, and the formulation stated in (1) represents a linear regression model [12,13].

The OLS estimator of $\boldsymbol{w}$, which coincides with the ML estimator in this case, is given by:

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \tag{2}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$. To predict a new observation y_t with features $\boldsymbol{x}_t^\top = (x_{1t}, \ldots, x_{mt})$, based on the model described in (1) and the estimate derived in (2), we use:

$$\widehat{y}_t = \widehat{y}(\boldsymbol{x}_t) = \boldsymbol{x}_t^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}. \tag{3}$$

It is important to note that, assuming normality of errors, the variance of $\widehat{y}_t$ in (3) can be calculated as:

$$\sigma^2 \boldsymbol{x}_t^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_t.$$

Now, we delve into the similarity model, considering the observations $(x_{1i}, \ldots, x_{mi}, y_i)$, where y_i represents the value of the random variable Y_i for $i \in \{1, \ldots, n\}$. We have a new vector of covariate values $\boldsymbol{x}_{n+1} = (x_{1(n+1)}, \ldots, x_{m(n+1)})$, and want to predict the future value of Y_{n+1} as a weighted mean of the past values y_i. The weights depend on the similarity between the past features x_i and the present value x_{n+1} [1]. The similarity is measured by a function s: $\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_+$. We provide a detailed explanation on how variations in the similarity function and other parameters impact the model's performance.

Based on this concept, we give insights into the similarity model proposed in [8] and specify it as:

$$Y_t = \frac{\sum_{i \neq t} s(\boldsymbol{x}_i, \boldsymbol{x}_t) y_i}{\sum_{i \neq t} s(\boldsymbol{x}_i, \boldsymbol{x}_t)} + \varepsilon_t, \quad 1 < t \leq n, \tag{4}$$

where the error term ε_i represents a non-observable variable that accounts for the inherent uncertainty of the phenomenon under study, and $s(\boldsymbol{x}_i, \boldsymbol{x}_t)$ is a similarity measure between $\boldsymbol{x}_i$ and $\boldsymbol{x}_t$. Notably, the error term, for $1 < t \leq n$, is uniquely defined as $\varepsilon_1 = \sqrt{n}(\overline{Y}_n - \alpha)$, where $\overline{Y}_n = (1/n) \sum_{i=1}^n Y_i$, and $\alpha = \mathbb{E}(Y_t)$. This special error term, ε_1 namely, is used to incorporate the inherent variability in the data that is not captured by the similarity measure $s(\boldsymbol{x}_i, \boldsymbol{x}_t)$. Such additional variability, as mentioned, might be due to the inherent uncertainty of the studied phenomenon or possible measurement errors. Moreover, ε_1 acts as a form of regularization, helping to avoid overfitting to the similarity model. This is particularly important for complex and high-dimensional models, where overfitting can be a relevant issue. Therefore, the specific need for ε_1, with $1 < t \leq n$, arises to capture the additional variability in the data not addressed by the similarity measure and to provide regularization, avoiding overfitting.

In [8], parametric estimation of the similarity function, s say, was conducted. The estimation is considered parametric because ε_i is assumed to follow a well-defined distribution with unknown parameters. The assumption is that the similarity function s is the same for all subjects generating Y_t with $t \leq n$. Two estimation methods are considered: OLS and ML.

Particularly, in [8], the study was focused on similarity functions that depend on a weighted Euclidean distance (WED). The square of the WED between two vectors $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\boldsymbol{x}' = (x'_1, \ldots, x'_m) \in \mathbb{R}^m$ is defined as:

$$d_{\boldsymbol{w}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j \leq m} w_j (x_j - x'_j)^2, \tag{5}$$

where $\boldsymbol{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m_+$, as mentioned, represents a weight vector.

The function defined in (5) allows for different variables to have distinct influences on the distance measure, permitting adjustments for covariates with different scales. The weights in this function do not need to add up to one, providing flexibility in the modeling process. In the present study, we recall two specific similarity functions, denoted as EX and FR, are considered. These functions are defined as:

$$s_{\boldsymbol{w}}^{\mathrm{EX}} = \mathrm{e}^{-d_{\boldsymbol{w}}}, \quad s_{\boldsymbol{w}}^{\mathrm{FR}} = \frac{1}{1 + d_{\boldsymbol{w}}}, \tag{6}$$

where $s_{\boldsymbol{w}}^{\mathrm{EX}}$ represents the exponential similarity function, and $s_{\boldsymbol{w}}^{\mathrm{FR}}$ represents the fractional similarity function. These functions are derived from the WED, $d_{\boldsymbol{w}}$ namely.

By incorporating these similarity functions into the model specified in (4), we obtain the parametric version of the empirical similarity model, which was estimated in [8] using the ML method. The ML estimation procedure is described in more detail in [7,8].

Utilizing the estimated values $\widehat{\boldsymbol{w}}$ obtained from the expression defined in (2) and the expressions given in (4), we can calculate the predicted value for a new observation $\boldsymbol{x}_t$ using:

$$\widehat{Y}_t = \frac{\sum_{i \neq t} \widehat{s}(\boldsymbol{x}_i, \boldsymbol{x}_t) y_i}{\sum_{i \neq t} \widehat{s}(\boldsymbol{x}_i, \boldsymbol{x}_t)}, \tag{7}$$

where $\widehat{s}$ represents the similarity function evaluated at $\widehat{\boldsymbol{w}}$.

In the case of handling categorical covariates, the distance measure defined in (5) is not suitable, particularly when there is no ordinal categories available. In such cases, a codification approach was proposed in [8], which involves transforming categorical variables into binary variables. This approach, referred to as M2, utilizes the WED stated in (5) to measure similarity. However, the method proposed in [8] has certain drawbacks. First, it may lead to a large number of parameters if a categorical variable has a high number of levels. Second, since different levels of the same categorical variable are treated as independent variables, they might be associated with significantly different weights, making the interpretation of the model more challenging. To address these issues, an alternative approach called M1 was introduced in [7] to handle categorical variables.

In the M1 approach, categorical variables are kept in their original format, and a weighted binary distance (WBD) is employed to measure similarity between vectors $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\boldsymbol{x}' = (x'_1, \ldots, x'_m) \in \mathbb{R}^m$. The WBD is defined as:

$$d_w(\boldsymbol{x}, \boldsymbol{x}') = \sum_{l \leq m} w_l \mathbb{1}_l(\boldsymbol{x}_l, \boldsymbol{x}'_l), \tag{8}$$

where $\mathbb{1}_l(\boldsymbol{x}_l, \boldsymbol{x}'_l)$ is an indicator function given by:

$$\mathbb{1}_l(\boldsymbol{x}_l, \boldsymbol{x}'_l) = \begin{cases} 0, & \text{if } \boldsymbol{x}_l = \boldsymbol{x}'_l, \\ 1, & \text{if } \boldsymbol{x}_l \neq \boldsymbol{x}'_l. \end{cases}$$

Thus, the WBD stated in (8) sums the weights associated with covariates that have different observed values. Consequently, the predicted value for the response variable related to a given set of features is obtained as the weighted mean of the other observed values of this variable, where observations with more features in common in relation to the given set have a higher weight.

Here, we explore the use of the weighted Minkowski distance (WMD) to handle dichotomous covariates, considering ordinal categories. The WMD of order γ between two vectors $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$ and $\boldsymbol{x}' = (x'_1, \ldots, x'_m) \in \mathbb{R}^m$ is defined as:

$$d_{\boldsymbol{w}}^{\mathrm{WMD}}(\boldsymbol{x}, \boldsymbol{x}') = \Big(\sum_{l=1}^{m} w_l |\boldsymbol{x}_l - \boldsymbol{x}'_l|^{\gamma} \Big)^{1/\gamma}. \tag{9}$$

Since the WMD stated in (9) introduces an additional parameter γ, we also introduce another parameter δ in the WBD and WED to provide more flexibility in explaining the observed data. The WBD and the WED are then stated as:

$$d_{\boldsymbol{w},\delta}^{\mathrm{WBD}}(\boldsymbol{x}, \boldsymbol{x}') = \Big(\sum_{l \le m} w_l I_l (x_l - x'_l) \Big)^{\delta}, \quad d_{\boldsymbol{w},\delta}^{\mathrm{WED}}(\boldsymbol{x}, \boldsymbol{x}') = \Big(\sum_{l \le m} w_l (x_l - x'_l)^2 \Big)^{\delta}. \tag{10}$$

Note that, as δ increases, the distances defined in (10) also increase. However, for $\delta = 1$, we obtain the standard distance measures, $d_{\boldsymbol{w},1}^{\mathrm{WBD}} = d_{\boldsymbol{w}}^{\mathrm{WBD}}$. The approach that handles categorical covariates without codification is M1. It is important to emphasize that for M2, where all covariates are binary, the Minkowski, binary, and Euclidean distances coincide.

3. Biological Application

The biological dataset used in this study investigates the effect of vitamin C on the tooth growth of Guinea pigs. Scientifically known as *Cavia porcellus*, Guinea pigs are rodents belonging to the *Caviidae* family and the *Cavia* genus [14]. The dataset consists of 60 observations, where the response variable (Y) is the length of the Guinea pig tooth measured in micrometers (µm), and the covariates are as follows:

- Vitamin C dose (X_1): This covariate is measured in milligrams (mg) and has three levels: 0.5 mg, 1.0 mg, and 2.0 mg. The vitamin C dose variable is ordinal.
- Food supplemental type (X_2): This covariate has two categories: ascorbic acid (VC) and orange juice (OJ). These categories are represented as 0 and 1, respectively. The food supplemental type variable is also ordinal.

To conduct an exploratory data analysis, violin plots are created to visualize the tooth length distribution based on the vitamin C dose and food supplemental type. Figure 1 shows the violin plots, where each plot represents the distribution of tooth length for a specific combination of the two covariates. The plots reveal that the vitamin C dose has an impact on tooth growth, showing a similar trend in both food supplemental types. However, there are differences in the central tendency and variability measures between the two types. A violin plot combines the features of a box plot and a kernel density plot, providing information about the data distribution and density peaks.

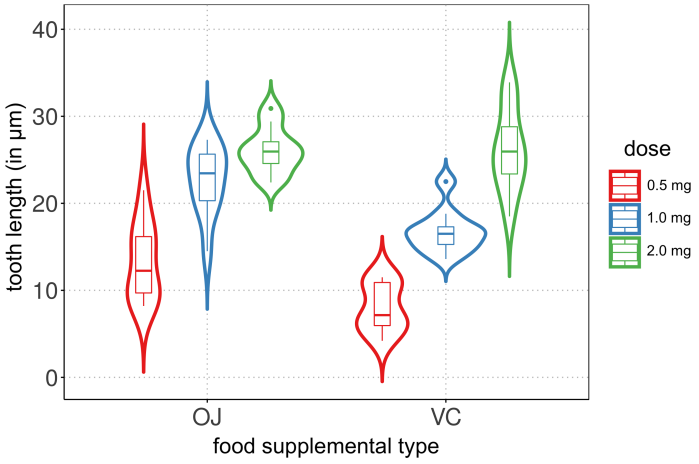

Figure 1. Violin plots of tooth length (in µm) for listed vitamin C dose (in mg) and food supplement.

4. Setup to Evaluate Sensitivity

In the Monte Carlo simulation, a sensitivity analysis was conducted to assess the variability of the parameter estimators and the predicted values of the response variable. The simulation consisted of 30 iterations, where each iteration involved a randomly generated training dataset comprising 70% of the total data, and a test dataset comprising the remaining 30%.

For each training dataset, numerical computations were performed to obtain parameter estimates using the empirical similarity methods. To initiate the estimation process, five initial parameter values were considered. The specific values of these initial parameters are not provided in the given text and should be defined based on the methodology and requirements of the empirical similarity methods used in the study as:

- For each fixed value of vitamin C dose (0.5, 1.0 and 2.0), the mean of the response variable is calculated. Let us denote these means by $\overline{y}_{11}$, $\overline{y}_{12}$ and $\overline{y}_{13}$, respectively.
- For each fixed value of the supplemental type (VC and OJ), the mean of the response variable is also computed and denoted by $\overline{y}_{21}$ and $\overline{y}_{22}$, respectively.
- The five initial parameter values (w_1^0, w_2^0) for M1 are: $(\overline{y}_{11}, \overline{y}_{21})$, $(\overline{y}_{11}, \overline{y}_{22})$, $(\overline{y}_{12}, \overline{y}_{21})$, $(\overline{y}_{12}, \overline{y}_{22})$, and $(\overline{y}_{13}, \overline{y}_{21})$.
- The five initial parameter values $(w_1^0, w_2^0, w_3^0, w_4^0)$ for M2 are: $(\overline{y}_{11}, \overline{y}_{11}, \overline{y}_{11}, \overline{y}_{21})$, $(\overline{y}_{11}, \overline{y}_{11}, \overline{y}_{11}, \overline{y}_{22})$, $(\overline{y}_{12}, \overline{y}_{12}, \overline{y}_{12}, \overline{y}_{21})$, $(\overline{y}_{12}, \overline{y}_{12}, \overline{y}_{12}, \overline{y}_{22})$, and $(\overline{y}_{13}, \overline{y}_{13}, \overline{y}_{13}, \overline{y}_{21})$.

For each of the initial parameter values, the model is estimated for the training data and the mean square error (MSE) for the prediction in the test data is calculated. The predicted response value and the estimated weights for the case with a minimal MSE are chosen.

We investigated the variability by modifying the following aspects:

- Models: M1 [7], M2 [8], and the linear regression [12].
- Estimation methods: ML and OLS.
- Similarity functions: EX and FR.
- Distance measures: WBD, WED, and WMD for M1; and WED for M2. We test the values $\{1/4, 1/2, 1, 2, 4\}$ for the parameter γ of the WMD, and the values $\{1, 2, 4\}$ for the parameter δ in the modified WBD and WED. Values of δ less than one are also tested, but they do not provide convergence in the estimation algorithm.

To quantify the variability of the parameter estimators and the predicted response variable, the empirical CV and MSE are calculated based on the 30 iterations of the Monte Carlo simulation.

For M1:

- The parameter estimators correspond to w_1 (intercept) and w_2 (associated with an increase in dose of 1.0 mg).
- The empirical CV of the parameter estimators can be calculated as the ratio of the sample standard deviation to the sample mean of the parameter estimates w_1 and w_2 across the 30 iterations.
- The MSE of the predicted response variable may be computed as the average squared difference between the predicted response variable values and the true values across the 30 iterations.

For M2:

- The parameter estimators correspond to w_1 (intercept), w_2 (associated with dose of 0.5 mg), w_3 (associated with dose of 1.0 mg), and w_4 (associated with dose of 2.0 mg and supplemental type).
- The empirical CV of the parameter estimators can be obtained as the ratio of the sample standard deviation to the sample mean of the parameter estimates w_1, w_2, w_3, and w_4 across the 30 iterations.
- The MSE of the predicted response variable can be determined as the average squared difference between the predicted response variable values and the true values across the 30 iterations.

5. Simulation Results

The simulations were carried out on a computer equipped with an Intel© Core™, i7-5500UK CPU 4 gigahertz, 16 gigabyte RAM, System Type 64 bit operating system Linux, using the R language, a software environment for statistical computing and graphics, in its version 3.5.2 [15]. Codes are available upon request from the authors.

Based on our simulation results, it is clear that the empirical similarity models (M1 and M2) and the linear regression model show comparable performance in terms of the mean MSE of the predicted values for the length of the Guinea pig tooth, a testament to the robustness of our analysis.

While we recognize that an increased sample size might result in a broader distribution of the results, the consistent findings among different models under our current conditions attest to the reliability of our work. Furthermore, our chosen sample size reflects a practical balance between computational complexity and statistical validity, a key consideration in all real-world application.

The mean MSE and standard deviation of the linear regression model, being comparable to those of M1 and M2, serve as a strong benchmark in our analysis. Furthermore, from Figure 2, we find no significant statistical difference among the MSE of the response variable predictions for all tested models, in addition to corroborating the robustness of our chosen models. Such robustness underlines the adaptability of these models to different scenarios and conditions. It serves as a valuable insight for making informed decisions on model selection, considering factors such as model complexity, interpretability, and specific objectives of the analysis. Notably, M1 stands out as the most parsimonious model, requiring only two parameters. This parsimony enhances its applicability, particularly when dealing with a large number of categorical covariates or when these covariates have numerous levels.

In summary, the empirical similarity models (M1 and M2) and the linear regression model demonstrate competitive performance in terms of the variability of the predicted values for the length of the Guinea pig tooth. These insights from the simulation and data analysis can guide the anticipation of the models' performance under different conditions and make adjustments to the research design accordingly. Our findings contribute to the current literature on empirical similarity prediction models, and we consider further research with larger Monte Carlo simulations and other comparison strategies in the future.

Figures 3 and 4 display the CVs of the parameter estimators for model M1 when the WBD and WED are considered, respectively. The plots illustrate the impact of different similarity measures and estimation methods on the variability of the parameter estimators. The results highlight the influence of the exponential inverse similarity and the choice of the ML method, particularly when $\delta = 4$, in reducing the CVs for the parameter estimates.

Figures 3 and 4 provides valuable insights, summarized as follows:

- The OLS method, when used with the fractional inverse similarity, exhibits high variability in the estimates for the parameter w_1.
- Increasing the value of the parameter δ results in parameter estimates with lower variability.
- The exponential inverse similarity generally produces parameter estimates with less variability compared to the fractional inverse similarity.
- The ML method generally provides estimates with lower variability for the parameter w_1 compared to the OLS method.
- The minimum CVs obtained for $\widehat{w}_1$ and $\widehat{w}_2$ are 0.01 and 0.05, respectively, when the ML method is utilized with the exponential inverse similarity and $\delta = 4$.

By combining these insights, we gain a comprehensive understanding of the variability in the parameter estimators for model M1 with different similarity measures and distances. Figure 5 illustrates the CVs of the parameter estimators for model M1 when the WMD is employed.

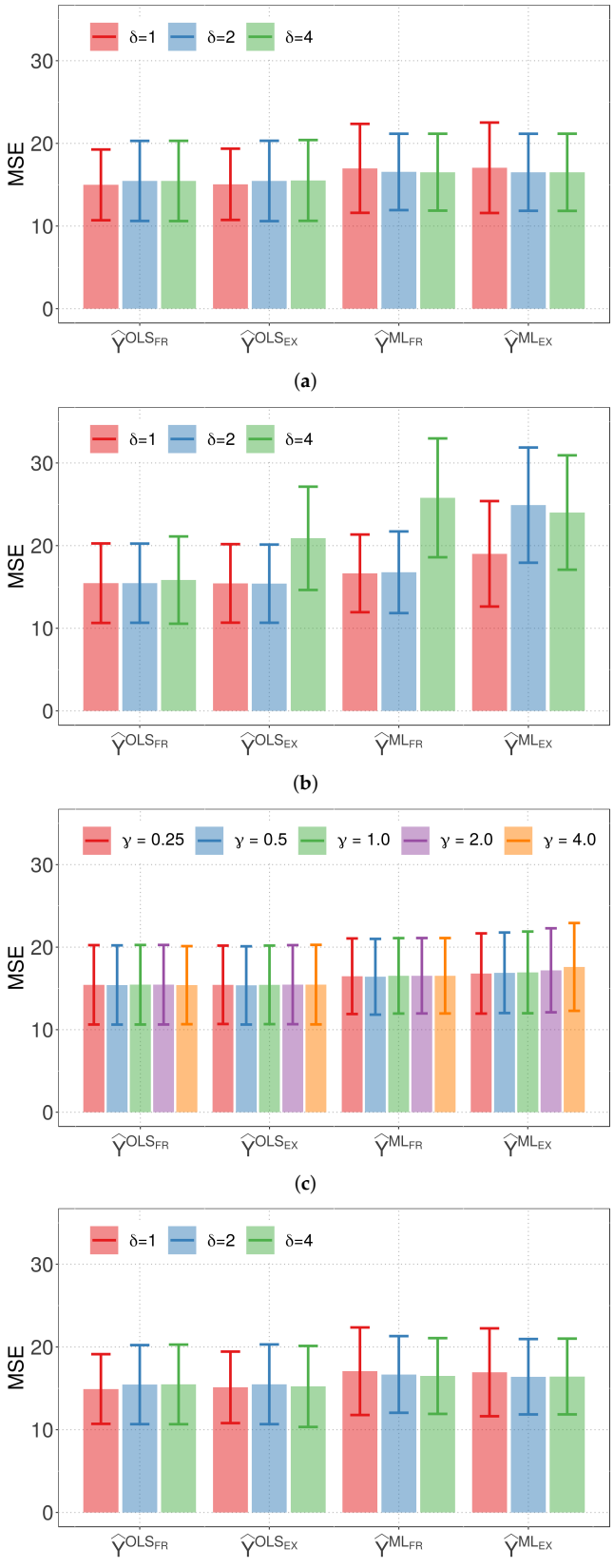

Figure 2. Plots of average MSE of the indicated response prediction and parameter (with error bars) for: (**a**) model M1 and binary distance; (**b**) Model M1 and Euclidean distance; (**c**) Model M1 and Minkowski distance; and (**d**) model M2 and Euclidean distance.

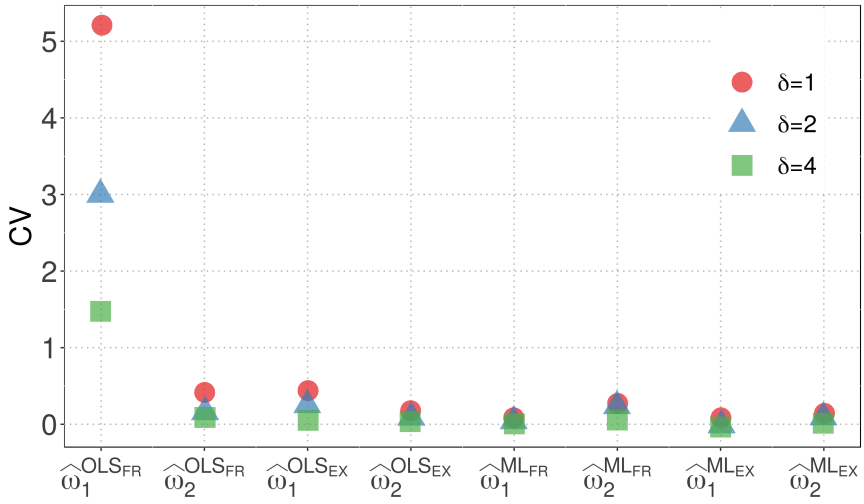

Figure 3. CVs of parameter estimators for model M1 with WBD.

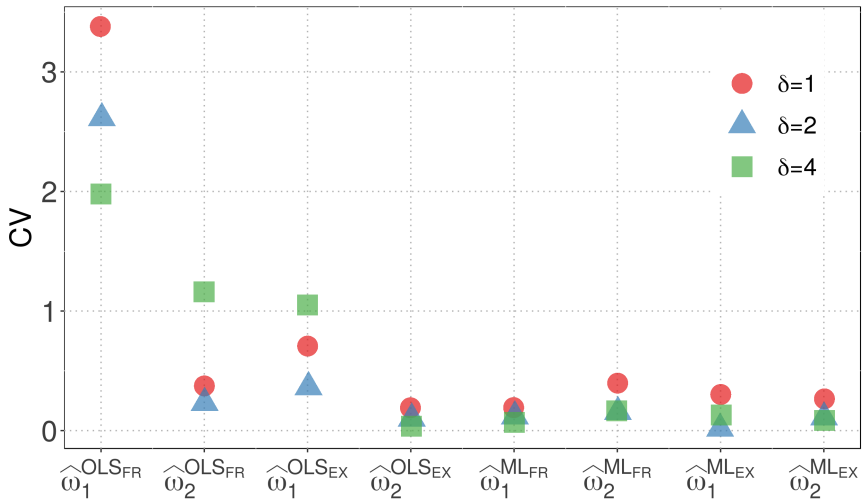

Figure 4. CVs of parameter estimators for model M1 with WED.

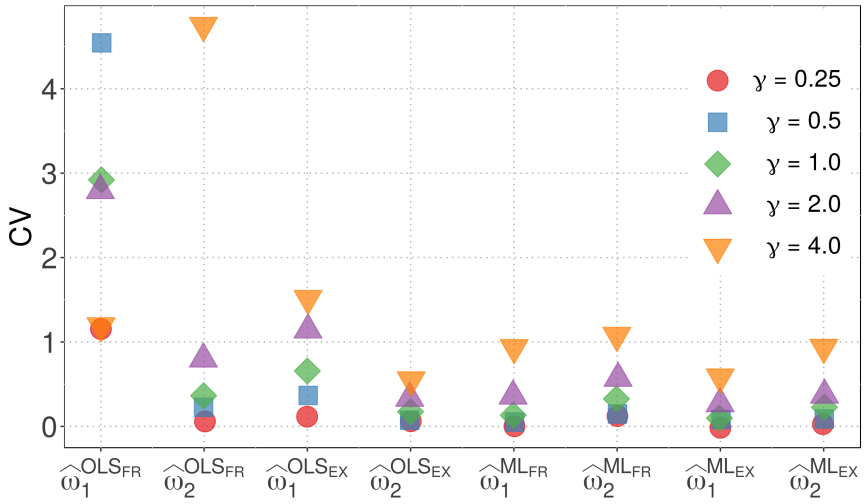

Figure 5. CVs of the parameter estimators for model M1with WMD.

The following observations can be made from Figure 5:

- The OLS method, in conjunction with the FR similarity, yields parameter estimates with notably high variability.
- With the exception of the parameter $\widehat{w}_1$ estimated using the OLS method and FR similarity, the variability of parameter estimates generally increases with higher values of γ.
- The EX similarity measure, except for the case of $\widehat{w}_1$ estimated using the OLS method and $\gamma = 4$, results in parameter estimates with less variability than the FR similarity.
- In all other cases, parameter estimates obtained using the ML method exhibit less variability compared to the corresponding estimates from the OLS method.
- The combination of the ML method with the EX similarity and $\gamma = 1/4$ yields the lowest variability, as evidenced by the sum of the CVs of $\widehat{w}_1$ and $\widehat{w}_2$, which are 0.03 and 0.07, respectively.

Figure 5 provides valuable insights into the variability of parameter estimators for model M1 with different similarity measures and the Minkowski distance. It is evident that the choice of similarity measure and estimation method can significantly impact the variability of the parameter estimators. Furthermore, the ML method, particularly when used with the EX similarity and appropriate parameter values, demonstrates superior performance in terms of reduced variability.

By considering the results from both models, we can compare the variability of the parameter estimates for M1 and M2 with the regression model. Model M2, which has the same number of parameters as the regression model, allows for a more straightforward visual comparison. In terms of the EX similarity, except for the case where $\delta = 4$, where the estimate of the weight w_2 has high variability, the estimates for model M2 exhibit competitive variability with the regression model in other cases. Figures 6 and 7 present a visualization of the CVs of the parameter estimates for model M2 and the regression model using the ML method, respectively. The figures highlight the variability of the estimates and show the competitive performance of model M2, with the EX similarity yielding consistent results except for the case where $\delta = 4$.

From Figures 6 and 7, the following observations can be stated:

- The OLS method, when used with the FR similarity, yields parameter estimates for w_1, w_2, and w_3 with high variability. Additionally, when the EX similarity is employed, the estimates for w_2 exhibit increased variability.
- Estimates of parameter w_2 are consistently zero when the ML method is utilized with $\delta = 1$ and $\delta = 2$, indicating that the variable dose of 1.0 mg has no influence on the response variable estimation.
- Among the tested scenarios, using $\delta = 4$ results in estimates with the least variability in 13 out of 16 cases.
- The EX similarity measure consistently provides parameter estimates with variability at least as low as, if not lower than, those obtained with the FR similarity in 20 out of 24 cases.
- When considering the sum of the CVs of $\widehat{w}_1$, $\widehat{w}_2$, $\widehat{w}_3$, and $\widehat{w}_4$, the combination of the ML method with the EX similarity and $\delta = 2$ yields the lowest variability. In this case, the CVs of $\widehat{w}_1$, $\widehat{w}_2$, $\widehat{w}_3$, and $\widehat{w}_4$ are 0.07, 0.00, 0.08, and 0.11, respectively.

When comparing the M1, M2, and regression models, M2 stands out as the most suitable for visual comparison due to its equal number of parameters. Based on minimal CV, we selected model M2 due to its best fit. Figures 6 and 7 also compare the regression and M2 models using the ML method, consistently showing superior results for model M2. While the estimate of weight w_2 exhibits higher variability when $\delta = 4$ in the EX similarity case, the variability of the estimates for model M2 remains competitive with that of the regression model in the other cases.

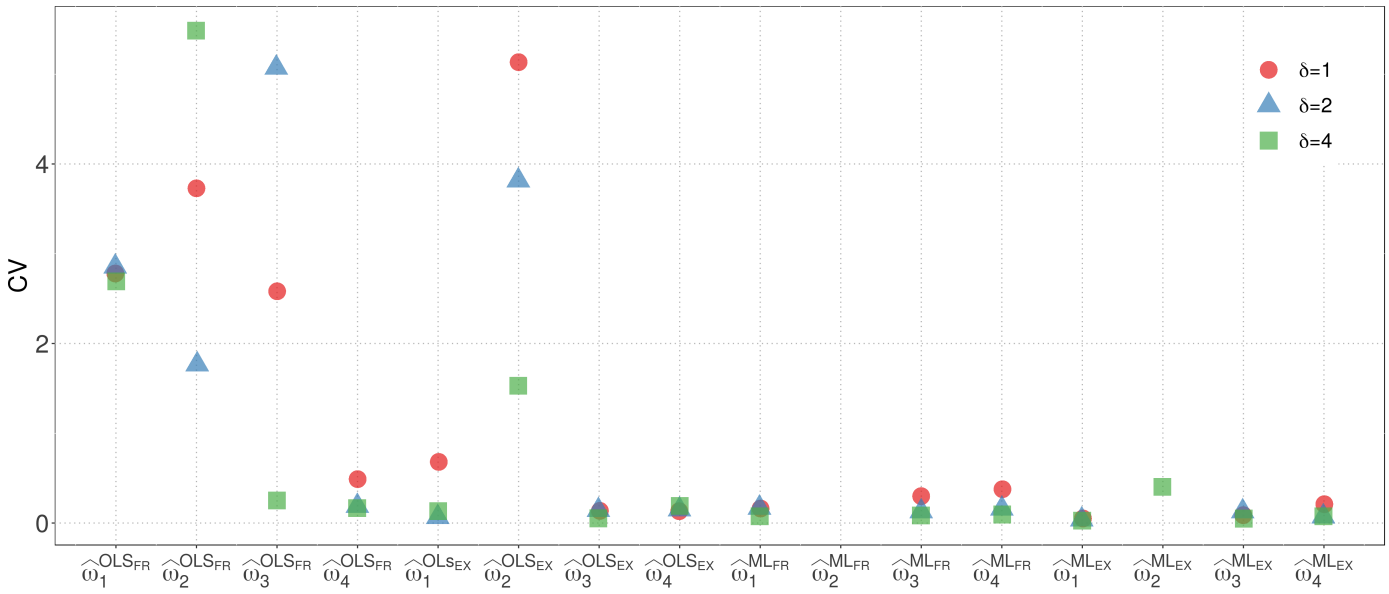

Figure 6. CVs of the parameter estimators for model M2.

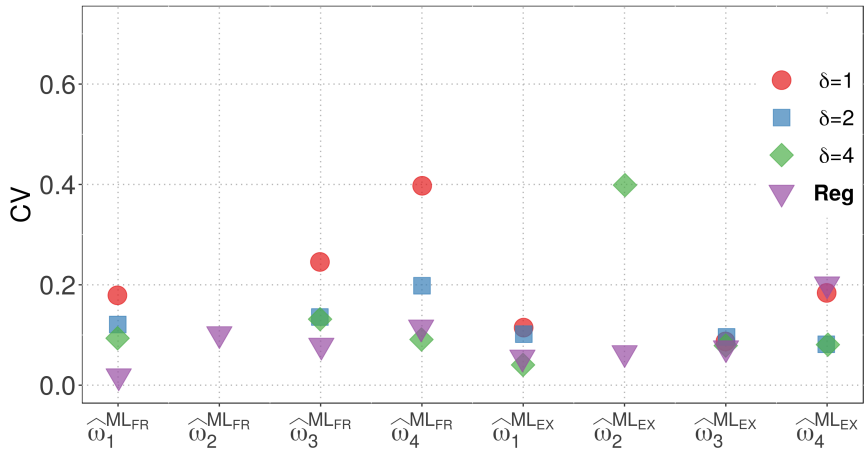

Figure 7. CVs of the parameter estimators for the regression model using the ML method.

6. Conclusions

This study aimed to evaluate the performance of empirical similarity models in a biological application, specifically in the context of predicting the length of Guinea pig teeth. Two empirical similarity models, M1 and M2, were compared against a linear regression model, serving as a benchmark. On the one hand, M1 preserved the original format of categorical covariates and utilized general distance measures with a single weight assigned to each covariate. On the other hand, M2 constructed binary variables for each level of the categorical covariates and employed similarity measures based solely on the Euclidean distance. For both M1 and M2, parameter estimation was conducted using ordinary least squares and maximum likelihood methods. It was observed that the maximum likelihood method consistently provided parameter estimates with low variability across both models, emphasizing the robustness of our approach.

In terms of the mean square error of the predicted response values, all models demonstrated competitive performance. Interestingly, M1 emerged as the most parsimonious model with only two parameters. The mean square error of the predicted values for the empirical similarity models did not exhibit dependency on the estimation method, similarity function, or distance measure. Different similarity functions were also explored for both models, including the exponential inverse and fractional inverse similarity functions.

The results indicated that the exponential inverse similarity function yielded less variability in most scenarios. For M1, three different distance functions were tested: weighted binary, Euclidean, and Minkowski distances. In the case of the binary and Euclidean distances, introducing an exponential parameter ($\delta \in \{1, 2, 4\}$) reduced variability. For the Minkowski distance, smaller values of the parameter γ resulted in better performance. However, it is important to note that values of δ greater than four or values of γ less than $1/4$ may lead to convergence issues in the estimation algorithms. In our analysis, when we employed the maximum likelihood estimation and the exponential inverse similarity function, we observed that the coefficients of variation for the parameter estimates were similar across the M1, M2, and linear regression models. This suggests that such models are comparable in terms of sensitivity. Nonetheless, M1 stands out due to its simplicity, requiring only two parameters. This simplicity can be particularly advantageous when dealing with a large number of categorical covariates or when these covariates have numerous levels.

To address potential overfitting in the M2 model, regularization techniques could be introduced [16]. However, to ensure a fair comparison between the models, we did not introduce this penalty in the empirical similarity framework in the current study. Nevertheless, we recognize the value of incorporating regularization methods in future investigations to explore their impact on the performance and generalization ability of the empirical similarity models.

It is important to acknowledge that the performance of these models can be influenced by several factors that warrant further investigation. Among these factors are the total sample size, the distribution of samples across different covariate categories, and the number of simulations performed. Although our findings contribute significantly to the existing literature on empirical similarity prediction models and methods, we deem it crucial to conduct expanded research to explore and understand the potential impact of these factors. Furthermore, the utility of simulation and data analysis is evident in our study. They serve as strategic tools, providing a clear understanding of how the models perform under different scenarios. Such insights permit informed adjustments to the models, thereby enhancing their functionality. The comprehensive analysis of model performance, encompassing different scenarios, similarity functions, and distance measures, facilitates effective decisions regarding model selection, taking into account specific objectives of the analysis, complexity, and interpretability.

In sum, our work contributes significantly to the burgeoning field of empirical similarity prediction models and methods. Our findings, which provide robust models capable of handling a variety of scenarios, serve as a foundation for future research, particularly for further exploration of the impact of sample size, number of simulations, and distribution of covariates on the performance of these models, as mentioned.

Author Contributions: Data curation: J.D.S., L.C.R. and R.O. Formal analysis: J.D.S., L.C.R., R.O., V.L., C.C. (Christophe Chesneau) and C.C. (Cecilia Castro) Investigation, J.D.S., V.L., C.C. (Christophe Chesneau) and C.C. (Cecilia Castro) Methodology: J.D.S., L.C.R., R.O., V.L., C.C. (Christophe Chesneau) and C.C. (Cecilia Castro) Writing—original draft: J.D.S., L.C.R., R.O. and C.C. (Christophe Chesneau). Writing—review and editing: V.L. and C.C. (Cecilia Castro) All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the National Council for Scientific and Technological Development (CNPq) through the grant 303192/2022-4 (R.O.) and 308980/2021-2 (L.C.R.); by the Comissão de Aperfeiçoamento de Pessoal do Nível Superior (CAPES), from the Brazilian government; by FONDECYT, grant number 1200525 (V.L.), from the National Agency for Research and Development (ANID) of the Chilean government under the Ministry of Science and Technology, Knowledge, and Innovation; and by Portuguese funds through the CMAT-Research Centre of Mathematics, University of Minho—within projects UIDB/00013/2020 and UIDP/00013/2020 (C.C. (Cecilia Castro)).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and codes are available upon request.

Acknowledgments: The authors would like to thank the Editors and four reviewers for their constructive comments which led to improvement in the presentation of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gilboa, I.; Lieberman, O.; Schmeidler, D. Empirical similarity. *Rev. Econ. Stat.* **2006**, *88*, 433–444. [CrossRef]
2. Raza, B.; Kumar, Y.J.; Malik, A.K.; Anjum, A.; Faheem, M. Performance prediction and adaptation for dataset management system workload using case-based reasoning approach. *Inf. Syst.* **2018**, *76*, 46–58. [CrossRef]
3. Lieberman, O. Asymptotic theory for empirical similarity models. *Econom. Theory* **2010**, *4*, 1032–1059. [CrossRef]
4. Xu, B.; Feng, X.; Burdine, R.D. Categorical data analysis in experimental biology. *Dev. Biol.* **2010**, *348*, 3–11. [CrossRef] [PubMed]
5. Mayya, S.S.; Monteiro, A.D.; Ganapathy, S. Types of biological variables. *J. Thorac. Dis.* **2017**, *9*, 1730. [CrossRef] [PubMed]
6. Larrabee, B.; Scott, H.M.; Bello, N.M. Ordinary least squares regression of ordered categorical data: Inferential implications for practice. *J. Agric. Biol. Environ. Stat.* **2014**, *19*, 373–386. [CrossRef]
7. Sanchez, J.D.; Rêgo, L.C.; Ospina, R. Prediction by empirical similarity via categorical regressors. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 641–652. [CrossRef]
8. Gayer, G.; Gilboa, I.; Lieberman, O. Rule-based and case-based reasoning in housing prices. *B.E. J. Theor. Econ.* **2007**, *7*, 10. [CrossRef]
9. Riquelme, M.; Leiva, V.; Galea, M.; Sanhueza, A. Influence diagnostics on the coefficient of variation of elliptically contoured distributions. *J. Appl. Stat.* **2011**, *38*, 513–532. [CrossRef]
10. Ospina, R.; Marmolejo-Ramos, F. Performance of some estimators of relative variability. *Front. Appl. Math. Stat.* **2019**, *5*, 43. [CrossRef]
11. De Miguel, C.; Saniotis, A.; Cieslik, A.; Henneberg, M. Comparative study of brain size ontogeny: Marsupials and placental mammals. *Biology* **2022**, *11*, 900. [CrossRef]
12. Bucchi, A.; Del Bove, A.; López-Lázaro, S.; Quevedo-Díaz, F.; Fonseca, G.M. Digital reconstructions using linear regression: How well can it estimate missing shape data from small damaged areas? *Biology* **2022**, *11*, 1741. [CrossRef]
13. Judge, G.G.; Griffiths, W.E.; Hill, C.; Lee, T.C. *The Theory and Practice of Econometrics*; Wiley: New York, NY, USA, 1985
14. Crampton, E. The growth of the odontoblasts of the incisor tooth as a criterion of the vitamin C intake of the Guinea pig. *J. Nutr.* **1947**, *33*, 491–504. [CrossRef]
15. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021. Available online: https://www.R-project.org/ (accessed on 19 June 2023).
16. Tutz, G.; Gertheiss, J. Regularized regression for categorical data. *Stat. Model.* **2016**, *16*, 161–200. [CrossRef]