



## EXAMINATION OF DISTANCE BASED REGRESSION METHODS FOR DIFFERENT DATA STRUCTURES IN ANIMAL SCIENCE

Burcu KURNAZ<sup>1\*</sup>, Hasan ÖNDER<sup>1</sup>


<sup>1</sup>Ondokuz Mayıs University, Agricultural Faculty, Department of Animal Science, 55139, Samsun, Türkiye


**Abstract:** Distance-based regression is an alternative method for parameter estimation in linear regression models when mixed-type explanatory variables are used. Distance-based regression is similar to classical linear regression, except that explanatory variables are measured by distance measures rather than raw values. In this study, datasets with sample sizes of 10, 25, 50, 100, 250 and 500 produced for Binomial, Normal, t, Chi-square and Poisson distributions of Euclidean, Gower and Manhattan distance measures and real data with discrete and continuous distribution that body weight at sixth months was used as outcome variable, body length and chest depth at sixth months of Saanen kids were used as explanatory variables as continuous data. Milk fat ratio was determined as the response variable, while the number of milking per day and the season of Polish Holstein Friesian cattle were determined as the explanatory variables as discrete data. It was aimed to determine the effect on the data sets (10, 50 and 100 sample sizes) by comparing the results obtained from the Linear Regression method. R packages "dbstats", "cluster" and "tidyverse" were used to perform the analysis. As a result, it has been determined that the use of Manhattan distance in data with Poisson distribution may produce unsuccessful results, especially in small sample sizes ( $n < 50$ ). Although there is no significant difference between Gower and Euclidean distances in different distributions according to sample sizes, it has been determined that the use of Euclidean distance measure in some distributions produces results that cause fluctuation. However, it has been understood that the Gower distance can be recommended as a more suitable choice since it has a more stable structure. For the applicability of the Least Square Estimation method, it may be recommended to use Distance Based Regression methods in cases where the necessary assumptions mentioned in this study cannot be met.

**Keywords:** Distance measures, Package dbstats, Regression, R software

\*Sorumlu yazar (Corresponding author): Ondokuz Mayıs University, Agricultural Faculty, Department of Animal Science, 55139, Samsun, Türkiye

E mail: burcu2039@hotmail.com (B. KURNAZ)

Burcu KURNAZ  <https://orcid.org/0000-0001-5613-6992>

Hasan ÖNDER  <https://orcid.org/0000-0002-8404-8700>

Received: December 11, 2024

Accepted: January 16, 2025

Published: March 15, 2025

**Cite as:** Kurnaz B, Önder H. 2025. Examination of distance based regression methods for different data structures in animal science. BSJ Eng Sci, 8(2): 354-362.

### 1. Introduction

In any relationship analysis, the goal is to obtain an accurate and reliable prediction equation using the available data. This is one of the most important and common questions about whether there is a statistical relationship between a response variable ( $Y$ ) and the explanatory variable(s) ( $X_i$ ). One option to answer this question is to use regression analysis to model its relationship. There are several types of regression analysis. The type of regression model depends on the shape of the distribution of the response variable ( $Y$ ) (Ari and Önder, 2013; Kurnaz and Önder, 2021; Kurnaz et al., 2021).

Linear regression analysis is a method of creating a model that predicts the desired response variable based on the variable(s) that can be detected more easily, at lower cost, or earlier than the variable to be determined. Simple linear regression analysis explains the linear relationship between the response variable and a single explanatory variable. If a linear relationship between a single response variable and more than one explanatory variable is desired, the relationship is examined by

multiple linear regression analysis (Weisberg, 2005; Okur, 2009; Alpar, 2010). In order for the parameter estimations of the regression model, which will be obtained as a result of both simple and multiple linear regression analysis, to be reliable, some assumptions about the model must be provided. In order to use the regression equation obtained in the simple linear regression analysis for estimation; The error terms ( $\epsilon_i = Y_i - \hat{Y}$ ) show normal distribution due to chance, the mean of the expected value of the errors is 0 and the variance is homogeneous and equal to  $\sigma^2$ , the errors are independent [ $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ ], with error terms it is necessary to provide some assumptions such as the absence of correlation between the explanatory variable(s) (Alma and Vupa, 2008). In multiple linear regression, in addition to the assumptions in simple linear regression, the assumption that the explanatory variables are independent from each other should also be provided (Vural, 2007). This assumption, which can also be explained as the condition that the simple linear correlation coefficients between the explanatory variables are zero or very close to zero, is expressed as



the absence of "multicollinearity" in statistics (Orhunbilge, 2017). In case of multiple connections, Least Squares estimation method loses its power (Vural, 2007). For the biological researches, some discrete or count variables such as sex, genotype, level of crossbreed (F1, G1, etc.) used as explanatory variables that break the assumptions of linear regression. In cases where these assumptions cannot be met, it is recommended to change the parameter estimation methods (Ari and Önder, 2013).

One of the few models developed as a solution for the above-mentioned situation in parameter estimation methods is Distance Based Regression methods. The purpose of these methods is to correctly address problems with non-true value estimators, including categorical or real-valued and categorical explanatory variables (Arenas and Cuadras, 2002). In statistics and data analysis, the geometric concept of distance between individuals or populations has been applied in fields such as anthropology, biology, genetics, psychology, linguistics, and others. The concept of distance is a useful tool for hypothesis testing and parameter estimation among other applications. In addition, the concept of distance is a basic tool in some statistical techniques such as fitness analysis or multidimensional scaling (Cuadras, 1988). Various multivariate approaches can evaluate relationships in connectivity (Varoquaux and Craddock, 2013), some factors have led researchers to examine multivariate distance matrix regression (MMR) (Anderson, 2001; McArdle and Anderson, 2001; Schork et al., 2008; Shehzad et al., 2014). These include:

- Ability to examine more than one explanatory variable at a time (i.e. covariates can be included),
- Applicability for categorical and/or continuous variables,
- Ease of interpretability due to the regression-like analytical structure.

Recently, the distance-based regression model has been successfully applied in many fields. In genomics, Xu et al. (2015) sequenced genes according to clusters associated with a distance-based regression model in the presence of a driver mutation and selected the important gene. In neuroscience, Shehzad et al. (2014) identified voxels (units of volume in body parts in images from computed tomography) associated with brain phenotypes with a distance-based regression model. In human microbiome research, Chen et al. (2012) determined the factors affecting the composition of the microbiome with a regression-based approach. In all these applications, the statistical significance derived from the distance-based regression model of the pseudo F test was calculated numerically with the permutation procedure, which proved to be superior for this purpose (Li et al., 2019).

In this study, datasets with sample sizes of 10, 25, 50, 100, 250 and 500 produced belonging to Binomial, Normal, t, Chi-square and Poisson distributions of Euclidean, Gower and Manhattan distance measures and real datasets showing discrete and continuous

distribution (10, 50 and 100 sample sizes) were examined. It is aimed to determine the effect of linear regression by comparing the results obtained from the linear regression method.

## 2. Materials an Methods

In this study, data sets consisting of Binomial, Poisson, Chi-Square, Normal, t distributions and linear regression (LR) (no distance measure) with sample sizes of 10, 25, 50, 100, 250 and 500 were analyzed. Analyzes were performed using the R software version 4.2.2. 10000 repetitions were used in the simulation study. The continuous data used in the study belong to the Saanen kids used in a study by Önder and Abacı (2015). While body weight at 6 months was used as outcome variable, body length and chest depth at 6 months were used as explanatory variables. The discrete data used in the study were previously reported by Aerts et al. (2022) belong to the Polish Holstein Friesian cattle used in the study. In this example, milk fat ratio was determined as the response variable, while the number of milking per day and the season were determined as the explanatory variables. Mean, standard deviation and error calculations of AIC, BIC, GCV values were used in the evaluation of the obtained results.

Generally a linear regression model; It is defined as  $Y = X\beta + \epsilon$ . Here;  $Y$ ; ( $n \times 1$ ) dimensional response variable vector,  $X$ ; ( $n \times p$ ) dimensional known coefficient matrix (design matrix),  $\beta$ ; ( $n \times 1$ ) dimensional unknown parameter vector (vector of coefficients),  $\epsilon$ ; It is an ( $n \times 1$ ) dimensional residuals (error) vector, with a mean of zero ( $E(\epsilon)=0$ ) and a constant variance ( $\text{var}(\epsilon)=\sigma^2 I$ ) (Atkinson and Riani, 2000).

Least Square estimator (LSE) of a linear regression model (equation 1) can be defined as (Anon, 2023a);

$$\hat{\beta}=(X'X)^{-1}X'Y \quad (1)$$

here,  $X$  is  $n \times k$  dimensional explanatory variable data matrix (design matrix),  $\hat{\beta}$  is  $k \times 1$  dimensional vector of coefficients and  $Y$  is  $n \times 1$  dimensional vector of dependent variable observations.

Matrix representation of General Sum of Squares (GSS) in linear regression model;  $GSS= Y'Y$ , matrix representation of Regression Sum of Squares (RSS);  $RSS=\hat{\beta}'X'Y$ , mean of Squares of Error (ESS) is expressed as  $ESS= GSS-RSS$ .

The Distance Based Regression model involves multiple regression of a response matrix over any number of explanatory matrices; wherein each matrix contains distances or similarities (in terms of ecological, spatial or other attributes) between all binary combinations of  $n$  objects (sample units); Statistical significance tests are performed by permutation. The method shows flexibility in terms of the types of data that can be analyzed (numbers, absent, continuous, categorical) and the shapes of the response curves (Lichstein, 2007).

For a selection of reference input points  $R=\{m_k\}_{k=1}^K$  with  $R \subseteq X$  and corresponding outputs  $T=\{t_k\}_{k=1}^K$  with  $T \subseteq Y$ ,

define  $D_x \in \mathbb{R}^{N \times K}$  in such a way that its  $k$ th column contains the distances  $d(x_i, m_k)$  between the  $i = 1, \dots, N$  input points  $x_i$  and the  $k$ th reference point  $m_k$ . Analogously, define  $\Delta_y \in \mathbb{R}^{N \times K}$  in such a way that its  $k$ th column contains the distances  $\delta(y_i, t_k)$  between the  $N$  output points  $y_i$  and the output  $t_k$  of the  $k$ th reference point. The mapping  $g$  between the input distance matrix  $D_x$  and the corresponding output distance matrix  $\Delta_y$  can be reconstructed using the multiresponse regression model (equation 2).

$$\Delta_y = g(D_x) + E. \quad (2)$$

The columns of the matrix  $D_x$  correspond to the  $K$  input vectors and the columns of the matrix  $\Delta_y$  correspond to the  $K$  response vectors, the  $N$  rows correspond to the observations. The columns of the  $N \times K$  matrix  $E$  correspond to the  $K$  residuals. Assuming that mapping  $g$  between input and output distance matrices has a linear structure for each response, the regression model has the form (equation 3).

$$\Delta_y = D_x B + E \quad (3)$$

The columns of the  $K \times K$  regression matrix  $B$  correspond to the coefficients for the  $K$  responses. The matrix  $B$  can be estimated from data through a minimization of the multivariate residual sum of squares as loss function (equation 4):

$$RSS(B) = \text{tr}((\Delta_y - D_x B)' (\Delta_y - D_x B)) \quad (4)$$

Under the normal conditions where the number of equations in equation 3 is larger than the number of unknowns, the problem is overdetermined and, usually, with no solution. This corresponds to the case where the number of selected reference points is smaller than the number of available points (i.e.  $K < N$ ). In this case, we must rely on the approximate solution provided by the usual least squares estimate of  $B$  (equation 5),

$$\hat{B} = (D_x' D_x)^{-1} D_x' \Delta_y. \quad (5)$$

The problem is uniquely determined if the number of equations equals the number of unknowns (i.e.  $K = N$  because all the learning points are also reference points) in equation 3. It has a single solution if the matrix  $D_x$  is full-rank (equation 6). So,

$$\hat{B} = D_x^{-1} \Delta_y. \quad (6)$$

Clearly less interesting is the case where in equation 3 the number of equations is smaller than the number of unknowns (i.e. for  $K > N$ , corresponding to the situation where, after selecting the reference points, only a smaller number of learning points is used). This case usually leads to an underdetermined problem with infinitely many solutions (de Souza et al., 2015).

The sum of squares associated with any term in any linear model can be calculated directly from a distance matrix. The reason for this is that for any centralized data matrix  $Y_{(n \times p)}$  (for  $p$  variables and  $n$  samples), it can be calculated with the inner product matrix  $Y'Y$  used in

classical multivariate statistics, as well as with the outer product matrix  $YY'$ . In addition, an outer product matrix can be obtained from any  $(n \times n)$  distance matrix (Gower, 1966), thus allowing the analysis to be based on a chosen distance measure, including semimetric measures such as Bray-Curtis.

Let  $X_{(n \times m)}$  be a model (i.e. design or regression) matrix with  $m$  number of parameters. For classical multivariate statistics (p x p), the total sum of squares is obtained by breaking down the inner product matrix  $Y'Y$ . The total sum of squares (SST) is the trace or sum of the diagonal elements in this matrix (sum of squares for each variable), which we will symbolize by  $\text{tr}(Y'Y)$ . The fragmentation process can be done according to the linear model  $Y = X\beta + \epsilon$ . Here  $\beta$  denotes the parameters matrix in the model and  $\epsilon$  denotes the error matrix. The least squares solution for  $\beta$  is  $\beta = (X'X)^{-1}X'Y$ . The prediction values matrix can be written as  $\hat{Y} = X\hat{\beta} = HY$  where;  $H$  is the idempotent prediction (hat) matrix and can be represented as  $X(X'X)^{-1}X'$ . The error (residuals) matrix can be represented as  $R = Y - \hat{Y} = (I - H)Y$ . The pseudo F statistic, which is a statistic applied to test the hypothesis of whether there is an effect of the model parameters, the regression sum of squares  $\text{tr}(\hat{Y}'\hat{Y})$  and the error sum of squares  $\text{tr}(R'R)$ , is calculated as follows (equation 7):

$$F = \frac{\text{tr}(\hat{Y}'\hat{Y})/(m-1)}{\text{tr}(R'R)/(n-m)}. \quad (7)$$

In case of a single variable, the pseudo F test is calculated the same as the Fisher F test. In the case of non-parametric testing, the probability of Type I error is  $P = P(F^* \geq F)$  where  $F^*$  is the value calculated by permutation of the unit. The degrees of freedom  $(m-1)$  and  $(n-m)$  are not required for the permutation test and are taken as constants (Anderson, 2001).

Once a similarity matrix is calculated, it is subjected to a regression analysis that tests hypotheses about this matrix. Whether there is a change in the level of similarity exhibited by pairs of individuals reflected in that matrix can be explained through the characteristics that these individuals have with others (e.g. a certain phenotype or a quantitative phenotype with higher or lower values of a certain feature) (Wessel and Schork, 2006).

## 2.1. Euclidean Distance Measure

Euclidean distance is the most commonly used distance measure to measure the similarity between two units and is based on the length of a straight line to be drawn between two units (Ünlükaplan, 2008). Using the Euclidean distance measure, the distance between two units is as follows:  $n$  is the number of units and  $p$  is the number of variables;  $i, j = 1, 2, 3, \dots, n$ ,  $i$  and  $j$  distance of unit from each other can be calculated as (equation 8):

$$(d_{i,j}) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2}. \quad (8)$$

This method has been proven to be compatible with the

classical linear regression model when the Euclidean distance measure is used (Arenas and Cuadras, 2002).

## 2.2. Manhattan Distance Measure

It is the sum of the distances between objects according to their dimensions. In measuring the distance between two objects in two-dimensional space, the hypotenuse of the triangle shown inside shows the Euclidean distance. The sum of the lengths of the sides of this triangle outside the hypotenuse gives the distance to Manhattan City Block. It is recommended to be used mostly for variables with discrete quantitative data. It is calculated as follows (equation 9) (Alpar, 2013):

$$d_{i,j} = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (9)$$

Manhattan City Block distance is a distance measure that is less sensitive to outliers (Timm, 2002).

## 2.3. Gower Distance Measure

The most basic feature of Gower distance is that it can be used in data sets that contain both categorical and continuous data. Gower distance is calculated using standardized data. Gower distance is calculated with a separate formula only when continuous data is used. The distance used for a data set containing both categorical and continuous data is called the Gower general similarity measure (URL2). Gower expressed the general similarity measure for categorical variables in the form (equation 10):

$$S_{ij} = \frac{\sum_{k=1}^p W_{ijk} S_{ijk}}{\sum_{k=1}^p W_{ijk}} \quad (10)$$

here  $S_{ijk}$ , k. according to variable value i. and j. It is a measure of similarity between observations.  $W_{ijk}$  is i. and j. observation k. When comparing by variable, it takes the value 0 if there is no variable value, and 1 in other cases. For continuous variables in the data, Gower (1971) defined the similarity measure as (equation 11):

$$S_{ij} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k} \quad (11)$$

here  $R_k$  is defined as the range of change of k. variable values of i. and j. observations (Servi, 2009).

## 2.4. Comparison Criteria

Akaike Information Criterion (AIC) can be called an indicator of the goodness of fit of any estimated statistical model. Akaike Information Criteria are asymptotically equivalent to cross-validation (URL3). It can be calculated with the equation (12):

$$AIC = -2\log(L) + 2k. \quad (12)$$

here, k is the number of parameters including the constant term, n is the number of observations, and L is likelihood (Ucal, 2006).

The Bayesian information criterion (BIC) is based in part on the likelihood function and is closely related to the Akaike information criterion (AIC). When fitting models, it is possible to increase the maximum likelihood by

adding parameters, but doing so can lead to overfitting. Both BIC and AIC try to solve this problem by introducing a penalty term for the number of parameters in the model; The penalty term is larger in BIC than AIC for sample sizes greater than 7 (McQuarrie and Tsai, 1998) and can be calculated as shown below (equation 13):

$$BIC = -2\log(L) + k \log(n) \quad (13)$$

BIC differs from AIC in that the second part on the right-hand side of the equation depends on the sample size. However, despite the superficial similarity between AIC and BIC, it was later determined that they differ within the Bayesian structure (Raftery, 1995; Wasserman, 2000).

The Generalized Cross Validation (GCV) criterion developed by Craven and Wahba in 1979 is one of the criteria for selecting the most appropriate model (Adıgüzel, 2021). The GCV criterion is based on the minimization of errors and also takes into account the complexity of the model (equation 14) (Yıldız, 2022).

$$GCV(M) = \frac{1}{N} \sum_{i=1}^N \frac{[y_i - f_M(x_i)]^2}{\left[1 - \frac{C(M)}{N}\right]^2} \quad (14)$$

In the equation,  $C(M)$  is the function that penalizes the model complexity for valid basis functions,  $y_i$  is the observation values of the dependent variable and  $f_M(x_i)$  is the prediction values, and  $N$  is the number of observations (Chen et al., 2012).

All statistical evaluation was performed using R software (R Core Team, 2022). The sample code sequence used in the analysis is given below;

```
library("dbstats")
library("cluster")
library(tidyverse)
simulnumber=10000
sampsiz=25
results=matrix(nrow= simulnumber, ncol=10)
for(i in 1: simulnumber) {
  Y=rnorm(sampsiz,0,1)
  X1=rchisq(sampsiz,5)
  X2=rchisq(sampsiz,5)
  Model1 <- dbglm(formula = Y ~ X1 + X2, family
= gaussian(), method = "GCV", full.search = TRUE, metric
= "euclidean", weights = NULL, range.eff.rank = c(1, 2))
  Model2 <- dbglm(formula = Y ~ X1 + X2, family
= gaussian(), method = "GCV", full.search = TRUE, metric
= "gower", weights = NULL, range.eff.rank = c(1, 2))
  Model3 <- dbglm(formula = Y ~ X1 + X2, family
= gaussian(), method = "GCV", full.search = TRUE, metric
= "manhattan", weights = NULL, range.eff.rank = c(1, 2))
  glm1 <- glm(Y ~ X1 + X2, family = gaussian())
  results[i,1]=summary(Model1$aic.model)[4][1]
  results[i,2]=summary(Model1$bic.model)[4][1]
  results[i,3]=summary(Model1$gcv.model)[4][1]
  results[i,4]=summary(Model2$aic.model)[4][1]
  results[i,5]=summary(Model2$bic.model)[4][1]
  results[i,6]=summary(Model2$gcv.model)[4][1]
```



```
results[i,7]=summary(Model3$aic.model)[4][1]
results[i,8]=summary(Model3$bic.model)[4][1]
results[i,9]=summary(Model3$gcv.model)[4][1]
results[i,10]=summary(glm1$aic)[4][1]
}
sink("D:/result_matrix.txt")
results
sink()
```

### 3. Results and Discussion

In order to examine the effects of distributions and distance measures on AIC, BIC and GCV, analyzes were made according to the factorial experimental design and the sample size was used as a covariate. Since the distribution  $\times$  distance interaction was found to be insignificant ( $P>0.05$ ), only main effects are presented. The sample size used as a covariate was determined to be statistically significant as expected ( $P<0.01$ ), therefore marginal means and standard error values are given in the tables. According to the findings, it was determined that the distribution had a statistically significant effect on AIC, BIC and GCV ( $P<0.01$ ), while distance measures had an effect only on the AIC value ( $P<0.01$ ) as seen in table 1 and 2.

Although the lowest AIC value is obtained from data with normal distribution, there is no difference in terms of AIC value between the results obtained from Normal, Binomial and t distribution. The highest AIC values were obtained from data with Poisson distribution and a significant difference was determined between them and other distributions. It was determined that the AIC values obtained from the Chi-Square distribution were different from the AIC values obtained from the Normal and Poisson distributions, but there was no difference in terms of AIC values between the results obtained from the Binomial and t distribution. When the effect of

distribution on BIC values was evaluated, it was determined that the BIC values obtained only from the Poisson distribution were significantly higher than those obtained from other distributions. It was observed that the BIC values obtained from other distributions were similar. When the effect of distributions on GCV values was evaluated, it was understood that the results obtained were similar to the effect on AIC (Table 1).

It is understood that the difference in the AIC value arises from the values obtained from the linear regression least squares method and that there is no difference between the Euclidean, Gower and Manhattan distance measures (Table 2). Boj et al. (2002) mentioned that using distance measures when the existence of non-normal variables were more reliable than classic linear regression, which supports our results.

For the combination of distribution  $\times$  distance measure, AIC, BIC and GCV measurements are evaluated together and the hierarchical clustering dendrogram drawn using the Ward method and Square Euclidean distance is given in figure 1.

According to the results obtained, it was determined that a total of 20 combinations could be examined in five clusters. According to this; All distance measures and the LSE solution used in the t distribution form a separate cluster (cluster 1), All distance measures and the LSE solution used in the binomial distribution form a separate cluster (cluster 2), Manhattan and the LSE solution in the normal distribution and Euclidean distance in the Poisson distribution form a separate cluster. It was determined that Manhattan, Gower and LSE solutions formed a separate cluster in the Poisson distribution (cluster 4), and all methods in the Chi-square distribution and Euclidean and Gower solutions in the normal distribution formed the last cluster (cluster 6).

**Table 1.** Effects of distributions on AIC, BIC and GCV

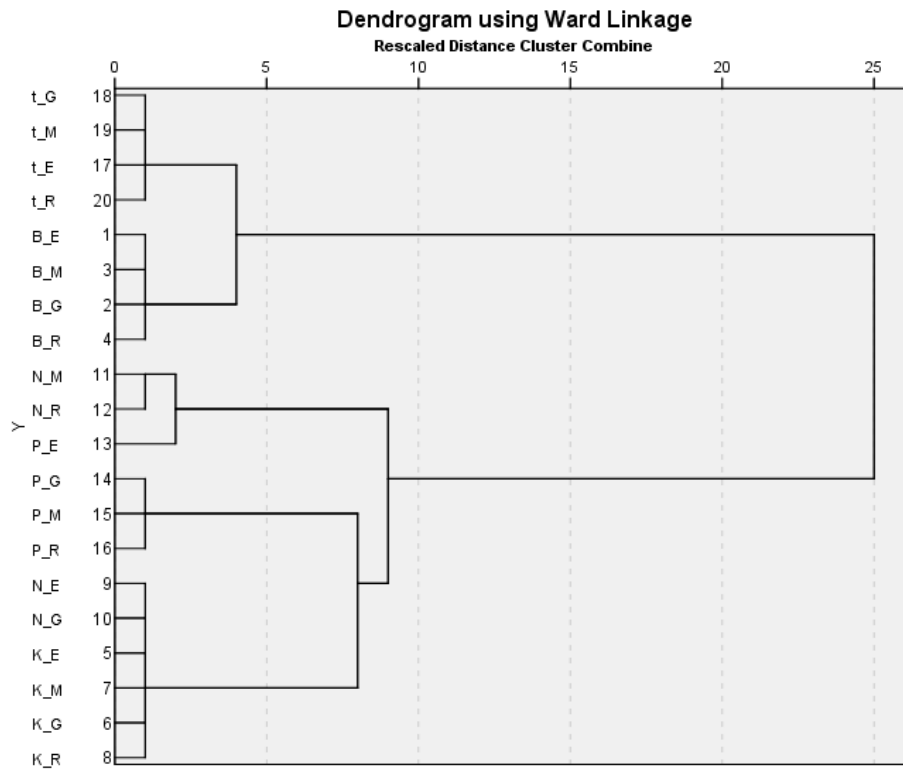
Distributions	AIC	BIC	GCV
Binom	156.841 $\pm$ 0.027 <sup>bc</sup>	159.956 $\pm$ 0.031 <sup>b</sup>	0.984 $\pm$ 0.001 <sup>b</sup>
Chi-Square	156.870 $\pm$ 0.027 <sup>b</sup>	159.978 $\pm$ 0.031 <sup>b</sup>	0.981 $\pm$ 0.001 <sup>bc</sup>
Normal	156.763 $\pm$ 0.027 <sup>c</sup>	159.876 $\pm$ 0.031 <sup>b</sup>	0.979 $\pm$ 0.001 <sup>c</sup>
Poisson	157.002 $\pm$ 0.029 <sup>a</sup>	160.421 $\pm$ 0.034 <sup>a</sup>	0.989 $\pm$ 0.001 <sup>a</sup>
t	156.812 $\pm$ 0.027 <sup>bc</sup>	159.915 $\pm$ 0.031 <sup>b</sup>	0.983 $\pm$ 0.001 <sup>bc</sup>
P	<0.001	<0.001	<0.001

The co-variate, sample size, was determined as 54.3092; <sup>a,b=</sup> different letters in the same column indicate statistical difference ( $P<0.05$ ).

**Table 2.** Effects of distance measures on AIC, BIC and GCV

Distances measure	AIC	BIC	GCV
Euclidean	156.557 $\pm$ 0.024 <sup>b</sup>	160.025 $\pm$ 0.025	0.983 $\pm$ 0.001
Gower	156.555 $\pm$ 0.024 <sup>b</sup>	160.022 $\pm$ 0.025	0.983 $\pm$ 0.001
Manhattan	156.576 $\pm$ 0.024 <sup>b</sup>	160.040 $\pm$ 0.025	0.984 $\pm$ 0.001
LR	157.743 $\pm$ 0.024 <sup>a</sup>		
P	<0.001	0.862	0.402

The co-variate, sample size, was determined as 54.3092; <sup>a,b=</sup> different letters in the same column indicate statistical difference ( $P<0.05$ ).



**Figure 1.** Hierarchical clustering dendrogram drawn by evaluating AIC, BIC and GCV measurements together for the distribution x distance measure combination. X\_X: the first letter indicates the distribution and the second letter indicates distance measure.

It is understood that the first two clusters are farther from the other clusters. It has been determined that binomial and t distributions form unique clusters, but other distributions form mixed clusters. When the last three clusters are examined; it is understood that normal distribution can be evaluated with different solutions of both Poisson and Chi-square distribution, but Chi-square and Poisson distributions are not in the same cluster. This interpretation was supported by the study of Zapala and Schork (2012) that they mentioned the choice of an appropriate distance measure may be problematic, although our experience suggests that different distance measures provide roughly the same inferences.

### 3.1. Real Data Results

The effects of distance measures on the real data structure of continuous and discrete distribution for sample sizes of 10, 50 and 100 are given in table 3 and 4. According to the findings obtained as a result of the analysis, it was determined that the LSE method produced higher AIC values than distance measurements for  $n = 10$ . It is understood that Euclidean and Gower have the same and lowest information criterion values among distance measures. Li et al. (2019) supports our results that distance-based regression with Euclidean distance was more reliable than linear regression for embryonic imprint data with the sample size of 24. Also Kim et al. (2001) mentioned that using Cook distance in local polynomial regression was more robust than standard approaches. It was observed that when the sample size was  $n=50$ , the AIC value obtained from the

LSE method produced the lowest value closest to the Euclidean measure. For continuous distribution, it is seen that the Gower distance measure has the largest AIC value for the sample size  $n = 50$ . It was determined that the smallest AIC value for the sample size  $n=100$  was the value obtained from linear regression. It is understood that the AIC value closest to this value belongs to the Euclidean distance measure. According to the results obtained, it was determined that the LSE estimation method did not produce reliable results according to the selection criteria obtained from distance-based regression in continuously distributed real data with a small sample size. It is understood that when the sample size increases, there is no significant difference with Euclidean, which is the distance measure for continuous data, when the assumptions of linear regression methods are met. It was observed that the Gower distance measure had the smallest BIC value for the sample size  $n = 50$ . Lichstein (2007) argued with supporting our results that use of Bray-Curtis distances was superior to linear regression.

In data sets belonging to discrete distribution, the AIC values for sample size  $n = 10$  were obtained from the Euclidean distance measure that produced the smallest value. It has been observed that the LSE method has the highest Akaike Information Criterion value. When we look at the BIC values, it is understood that the Euclidean distance measure produced the smallest value. It was determined that there was no difference between the GCV values obtained from all distance measurements.

**Table 3.** Effect of distance measurements on continuous data

Criteria	Distance measure	n=10	n=50	n=100
AIC	Euclidean	41.72	208.47	338.74
	Gower	41.72	209.33	340.35
	Manhattan	41.77	208.77	340.10
	LR	43.44	208.50	338.70
BIC	Euclidean	42.32	214.21	345.96
	Gower	42.32	213.16	347.57
	Manhattan	42.38	214.50	347.32
GCV	Euclidean	2.57	3.50	3.47
	Gower	2.57	3.56	3.54
	Manhattan	2.59	3.52	3.53

**Table 4.** Effect of distance measurements on discrete data

Criteria	Distance measures	n=10	n=50	n=100
AIC	Euclidean	20.31	48.03	147.00
	Gower	20.44	48.11	146.63
	Manhattan	20.38	50.21	147.77
	EKK	22.27	48.41	148.60
BIC	Euclidean	21.10	51.81	152.21
	Gower	21.24	51.89	151.84
	Manhattan	21.18	53.99	152.98
GCV	Euclidean	0.26	0.14	0.24
	Gower	0.26	0.14	0.24
	Manhattan	0.26	0.15	0.25

It was determined that the Manhattan distance measure produced the highest AIC value when the sample size was  $n=50$ . It was observed that the criterion value obtained from linear regression was close to the Euclidean measure, which has the smallest criterion value. This may be because the assumptions for linear regression were met. It has been determined that the Gower distance measure, which produces reliable results in discrete data, produces the smallest AIC value for the sample size  $n=100$  for discrete distribution. While BIC values were determined to be the Euclidean distance measure that produced the smallest information criterion for sample sizes of 10 and 50, it was determined that the Gower distance measure had the smallest value when the sample size was 100. It appears that there is no significant difference in GCV values for all sample sizes. Kurnaz and Önder (2021) supports our results that they mentioned the comparison of Euclid, Manhattan and Gower distance measures within the scope of distance based regression can give more reliable results especially existence of discrete explanatory variables. Cuadras and Arenas (1990) mentioned the use of Gower distance for mixed explanatory variables (discrete and continuous) to predict normal response variable was more powerful than linear regression, which supports our results. Ferreira Barreto et al. (2020) found that distance-based estimation for fNIRS signals and behavioral data was successful. Haron et al. (2019) also mentioned that distance-based regression is a good alternative method for estimating the unknown parameters in regression modeling when dealing with mixed-type of exploratory variables.

#### 4. Conclusion

When the effect of Distance Based Regression methods on distributions is evaluated, it is seen that the lowest information criteria value is in the data set consisting of explanatory variables with a normal distribution structure, which is a theoretical expectation. Considering the AIC, BIC and GCV values obtained from distance measures, it can be recommended to use this distance measure for model selection since the Gower distance measure has the lowest information criteria values compared to other distance measures.

When the findings are evaluated, it can be said that using Manhattan distance in data with Poisson distribution, especially in small sample sizes ( $n < 50$ ), may produce unsuccessful results. Although there is no significant difference between Gower and Euclidean distances in different distributions depending on sample sizes, it has been determined that the use of Euclidean distance measure in some distributions produces results that cause fluctuations. However, the Gower distance can be suggested as a more appropriate choice because it has a more stable structure. This may be because the Gower distance is calculated using standardized data.

When the information criteria values in model selection were examined, it was determined that the method that produced the highest AIC value for all distributions and sample sizes of the data sets was the LSE estimation method for linear regression. When information criteria values are evaluated for model selection, choosing the model that produces the smallest value increases the prediction success. Therefore, it can be said that the model obtained from linear regression analysis methods will not be reliable. In cases where the necessary assumptions mentioned in this study cannot be met for

the applicability of the LSE estimation method, it may be recommended to use Distance Based Regression methods.

It is considered that in future studies, evaluating other distance measures such as Bray-Curtis, Orloci's Chord, Chi-square, Canberra and Hellinger and/or examining combinations of explanatory variables with different distributions may be useful for the subject.

#### Author Contributions

The percentages of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	B.K.	H.Ö.
C	50	50
D	100	
S		100
DCP	80	20
DAI	90	10
L	100	
W	80	20
CR	90	10
SR	100	

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision.

#### Conflict of Interest

The authors declared that there is no financial/commercial conflict of interest.

#### Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

#### Acknowledgements

This study is short summary of MSc thesis of first author under the supervision of the second author.

#### Data Availability

To obtain the data, please contact the corresponding author.

#### References

- Adıgüzel MB. 2021. Çok değişkenli uyarlanabilir regresyon eğrilerinde alternatif bilgi kriterleri ile model seçimi. PhD Thesis, Ondokuz Mayıs University, Graduate School of Education, Department of Statistics, Samsun, Türkiye, pp: 86.
- Aerts J, Sitkowska B, Piwczynski D, Kolenda M, Önder H. 2022. The optimal level of factors for high daily milk yield in automatic milking system. *Livestock Sci*, 264: 105035.
- Alma ÖG, Vupa Ö. 2008. Regresyon analizinde kullanılan en küçük kareler ve en küçük medyan kareler yöntemlerinin karşılaştırılması. *SDÜ Fen Ede Fak Fen Derg*, 3(2): 2019-229.
- Alpar R. 2010. Basit doğrusal regresyon çözümlemesi: Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlilik. Detay Yayıncılık, Ankara, Türkiye, pp: 672.

- Alpar R. 2013. Uygulamalı çok değişkenli istatistiksel yöntemler. Detay Yayıncılık, Ankara, Türkiye, pp: 853.
- Anderson MJ. 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecol*, 26: 32-46.
- Anonymous. 2011. Kümeleme analizinde kullanılan bazı farklılık ve benzerlik ölçülerinin incelenmesi. URL: <http://emredunder.blogspot.com/2011/06/kumeleme-analizinde-kullanilan-baz.html> (accessed date: June18, 2023).
- Anonymous. 2023a. Ankara Üniversitesi açık ders. URL: [https://acikders.ankara.edu.tr/pluginfile.php/130799/mod\\_resource/content/0/6-%20Matris.pdf](https://acikders.ankara.edu.tr/pluginfile.php/130799/mod_resource/content/0/6-%20Matris.pdf) (accessed date: May 12, 2023).
- Anonymous. 2023b. Model ve dağılım seçme URL: <https://avys.omu.edu.tr/storage/app/public/kamilal/108861/HPTZ.HF10.pdf> (accessed date: May 16, 2023).
- Arenas C, Cuadras M. 2002. Recent statistical methods based on distances. *Contrib Sci*, 2(2): 183-191.
- Arı A, Önder H. 2013. Farklı veri yapılarında kullanılabilecek regresyon yöntemleri. *Anadolu Tar Bil Derg*, 28(3): 168-174.
- Atkinson AC, Riani M. 2000. Robust diagnostic regression analysis. Springer, New York, US, pp: 328.
- Boj E, Claramunt MM, Fortiana J, Vidiella A. 2002. The use of distance-based regression and generalized linear models in the rate making process: An empirical study. *Universitat de Barcelona. Institut de Matemàtica [IMUB], Barcelona, Spain*, pp: 47.
- Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinf*, 28(16): 2106-2113.
- Cuadras CM, Arenas C. 1990. A distance based regression model for prediction with mixed data. *Commun Stat A. Theory Methods*, 19: 2261-2279.
- Cuadras CM. 1988 Statistical distances. *Estadística Española*, 30: 295-378.
- de Souza Jr AH, Corona F, Barreto GA, Miche Y, Lendase A. 2015. Minimal learning machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing*, 164: 34-44.
- Ferreira Barreto CDS, Zimeo Morais GA, Vanzella P, Sato JR. 2020. Combining the intersubject correlation analysis and the multivariate distance matrix regression to evaluate associations between fNIRS signals and behavioral data from ecological experiments. *Experl Brain Res*, 238: 2399-2408. <https://doi.org/10.1007/s00221-020-05895-8>
- Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4): 325-338.
- Gower JC. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 1971: 857-871.
- Haron NH, Ahad NA, Mahat NI. 2019. Distance-based regression for non-normal data. The 4th Innovation and Analytics Conference & Exhibition (IACE 2019), March 21-28, Kedah, Malaysia. <https://doi.org/10.1063/1.5121118>
- Kim C, Lee Y, Park BU. 2001. Cook's distance in local polynomial regression. *Stat Probab Lett*, 54: 33-40.
- Kurnaz B, Önder H, Piwczynski D, Kolenda M, Sitkowska B. 2021. Determination of the best model to predict milk dry matter in high milk yielding dairy cattle. *Acta Sci Pol Zootechnica*, 20(3): 41-44. <https://doi.org/10.21005/asp.2021.20.3.05>
- Kurnaz B, Önder H. 2021. Distance based regression models. II. International Applied Statistics Conference (UYIK-2021), June 29- July 2, Tokat, Türkiye, pp: 120-126.
- Li J, Zhang W, Zhang S, Li Q. 2019. A theoretic study of a



- distance-based regression model. *Sci China Math*, 62: 979-998.
- Lichstein JW. 2007. Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecol*, 188: 117-131.
- McArdle BH, Anderson MJ. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1): 290-297.
- McQuarrie AD, Tsai CL. 1998. Regression and time series model selection. World Scientific Publication Co Pte. Ltd., London, UK, pp:45
- Okur S. 2009. Parametrik ve parametrik olmayan doğrusal regresyon analiz yöntemlerinin karşılaştırılması olarak incelenmesi. MSc Thesis, Çukurova University, Institute of Science, Department of Animal Science, Adana, Türkiye, pp: 62.
- Orhunbilge N. 2017. Uygulamalı regresyon ve korelasyon analizi. Nobel Yayıncılık, Ankara, Türkiye, pp: 394.
- Önder H, Abacı SH. 2015. Path analysis for body measurements on body weight of Saanen kids. *Kafkas Üniv Vet Fak Derg*, 21(3): 351-354.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022, URL: <https://www.R-project.org/>.
- Raftery AE. 1995. Bayesian model selection in social research. *Sociol Methodol*, 25: 111-163.
- Schork NJ, Wessel J, Malo N. 2008. DNA sequence-based phenotypic association analysis. *Adv Genet*, 60: 195-217.
- Servi T. 2009. Çok değişkenli karma dağılım modeline dayalı kümeleme analizi. PhD Thesis, Çukurova University, Institute of Science, Department of Statistics, Adana, Türkiye, pp: 266.
- Shehzad Z, Kelly C, Reiss PT, Craddock RC, Emerson JW, McMahon K, Milham MP. 2014. A multivariate distance-based analytic framework for connectome-wide association studies. *Neuroimage*, 93: 74-94.
- Timm NH. 2002. Applied Multivariate Analysis. Springer-Verlag, New York, US, pp: 718.
- Ucal MŞ. 2006. Ekonometrik Model seçim kriterleri üzerine kısa bir inceleme. *CÜ İİBF Derg*, 7(2): 41-57.
- URL1: [https://acikders.ankara.edu.tr/pluginfile.php/130799/mod\\_resource/content/0/6-%20Matris.pdf](https://acikders.ankara.edu.tr/pluginfile.php/130799/mod_resource/content/0/6-%20Matris.pdf) (accessed date: May 12, 2023).
- URL2: <http://emredunder.blogspot.com/2011/06/kumeleme-analizinde-kullanilan-baz.html> (accessed date: June18, 2023).
- URL3: <https://avys.omu.edu.tr/storage/app/public/kamilal/108861/HPTZ.HF10.pdf> (accessed date: May 16, 2023).
- Ünlükaplan Y. 2008. Çok değişkenli istatistiksel yöntemlerin peyzaj ekolojisi araştırmalarında kullanımı. PhD Thesis, Çukurova University, Institute of Science, Department of Statistics, Adana, Türkiye, pp: 156.
- Varoquaux G, Craddock RC. 2013. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80: 405-415.
- Vural A. 2007. Aykırı değerlerin regresyon modellerine etkileri ve sağlam kestiriciler. MSc Thesis, Marmara University, Institute of Social Sciences, Department of Econometry, İstanbul, Türkiye, pp: 73.
- Wasserman L. 2000. Bayesian model selection and model averaging. *J Math Psychol*, 44(1): 92-107.
- Weisberg S. 2005. Applied linear regression. John Wiley & Sons, New Jersey, US, pp: 340.
- Wessel J, Schork NJ. 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. *Amer J Human Genet*, 79(5): 792-806.
- Xu Y, Guo X, Sun J, Zhao Z. 2015. Snowball: resampling combined with distance-based regression to discover transcriptional consequences of a driver mutation. *Bioinformatics*, 31(1): 84-93.
- Yıldız M. 2022. Dolar ve Euro kurları üzerinde etkili faktörlerin iki bağımlı değişkenli MARS modeli ile belirlenmesi. *Kastamonu Üniv İİBF Derg*, 24(1): 6-29.
- Zapala MA, Schork NJ. 2012. Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Front Genet*, 3: 190. <https://doi.org/10.3389/fgene.2012.00190>