Black Sea Journal of Statistics

Volume 1 | Issue 1

ISSN: 3108-4265

KAR-YAY

BLACK SEA JOURNAL OF STATISTICS

(BSJ STAT)

Black Sea Journal of Statistics (BSJ Stat) is an international a peer-reviewed journal published electronically twice a year in June and December. BSJ Stat is an international e-journal that publishes open access. The general aim of BSJ Stat is to provide a platform for statistical scientists and related researchers to publish their research findings on a global scale, to encourage their professional development and to provide new perspectives for further research. BSJ Stat publishes original research articles, reviews and technical notes produced from studies carried out in the fields of theoretical and applied statistics. BSJ Stat is published electronically and all articles are provided to the readers free of charge.

e-ISSN: 3108-4265

Phone: +90 362 408 25 15

Fax: +90 362 408 25 15

Email: bsjstat@blackseapublishers.com

Web site: http://blackseapublishers.online/index.php/stat

Sort of publication: Periodically 2 times (June and December) in a year

Publication date and place: June 15, 2025- Samsun, TÜRKİYE

Publishing kind: Electronically

OWNER

KARYAY Karadeniz Yayımcılık ve Organizasyon Ticaret Limited Company

DIRECTOR IN CHARGE

Prof. Dr. Hasan ÖNDER

EDITOR BOARDS

EDITOR IN CHIEF

Assoc. Prof. Dr. Cem TIRINK, Iğdır University, Türkiye

SECTION EDITORS*

Prof. Dr. Dariusz PIWCZYŃSKI, Bydgoszcz University of Science and Technology, POLAND

Prof. Dr. Özgür Hakan AYDOĞMUŞ, Social Sciences University of Ankara, TÜRKİYE

Prof. Dr. Taner TUNÇ, Ondokuz Mayıs University, TÜRKİYE

Prof. Dr. Adnan ÜNALAN, Nigde Ömer Halisdemir University, TÜRKİYE

Assoc. Prof. Dr. İlker ÜNAL, Çukurova University, TÜRKİYE

* The ranking is arranged alphabetically within the academic title

STATISTIC EDITOR

Prof. Dr. Mehmet TOPAL, Kastamonu University, TÜRKİYE

ENGLISH EDITOR

Assist. Prof. Dr. Betül ÖZCAN DOST, Ondokuz Mayis University, TÜRKİYE

TURKISH EDITOR

Prof. Dr. Serkan ŞEN, Ondokuz Mayis University, TÜRKİYE

REVIEWERS OF THE ISSUE*

Prof. Dr. Çiğdem TAKMA, Ege University, Department of Animal Science, Biometry, TÜRKİYE

Assoc. Prof. Dr. Aycan Mutlu YAĞANOĞLU, Atatürk University, Department of Animal Science, Biometry, TÜRKİYE

Assoc. Prof. Dr. Yasin ALTAY, Eskişehir Osmangazi University, Department of Animal Science, Biometry, TÜRKİYE

Assist. Prof. Dr. Esra YAVUZ, Şırnak University, Department of Accounting and Tax, Biometry, TÜRKİYE

Dr. Ebru ERGÜNEŞ BERKİN, Hassa District Directorate of Agriculture and Forestry, Biometry, TÜRKİYE

Dr. Hasan Alp ŞAHİN, Ondokuz Mayıs University, Hemp Research Institute, Biometry, TÜRKİYE

---

* The ranking is arranged alphabetically within the academic title

Volume 1, Issue 1 (2025)

Table of Contents

Research Articles

1. USAGE OF R SOFTWARE IN EDUCATIONAL RESEARCH IN TÜRKİYE: FREQUENCY, FUNCTIONS, AND ADVANTAGES

Hatice DİLAVER, Kâmil Fatih DİLAVER……1-6

2. PERMUTATION TESTS AND ITS BIBLIOMETRIC ANALYSIS

Hasan ÖNDER……7-12

3. CLASSIFICATION OF STUDENT ACHIEVEMENT USING DATA MINING TECHNIQUES: A COMPARATIVE STUDY

Hatice DİLAVER, Kâmil Fatih DİLAVER……13-17

4. DETERMINING THE RELATIONS OF DAILY LIVE WEIGHT GAIN OF SAANEN KIDS USING CONCORDANCE CORRELATION

Burcu KURNAZ……18-21

5. BINARY LOGISTIC REGRESSION PROCEDURE WITH AN APPLICATION

Mustafa ŞAHİN……22-26

Black Sea Journal of Statistics

Doi: 10.5281/zenodo.18234368

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 1-6 / June 2025

USAGE OF R SOFTWARE IN EDUCATIONAL RESEARCH IN TÜRKİYE: FREQUENCY, FUNCTIONS, AND ADVANTAGES

Hatice DİLAVER[1]*, Kâmil Fatih DİLAVER[2]

[1]*Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye*
[2]*Niğde Ömer Halisdemir University, Faculty of Engineering, Department of Electric and Electronics, 51200 Niğde, Türkiye*

Abstract: This study explores the use of R statistical software in educational sciences in Türkiye. Although R is a powerful, free, and open-source software widely used globally, its adoption in Türkiye, especially in social sciences, remains very limited. The study investigates the frequency of R and other statistical software in SSCI-indexed educational journals between 2010 and 2014. Results reveal that SPSS is overwhelmingly preferred, while R was used in only one instance. The advantages of R, its basic commands, assumptions testing, descriptive and inferential statistics, and non-parametric tests are also presented. The study concludes with recommendations for wider use of R in Turkish academia.

Keywords: Educational, Function, Statistical software, Frequency

***Corresponding author:** Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye
E mail: haticedilaver509@gmail.com (H. DİLAVER)
Hatice DİLAVER https://orcid.org/0000-0002-4484-5297
Kâmil Fatih DİLAVER https://orcid.org/0000-0001-7557-9238

Received: April 22, 2025
Accepted: May 23, 2025
Published: June 15, 2025

Cite as: Dilaver, H., & Dilaver, K. F. (2025). Usage of R software in educational research in Türkiye: Frequency, functions, and advantages. *Black Sea Journal of Statistics, 1*(1), 1–6.

1. Introduction

One of the most critical stages of the scientific research process is analyzing the collected data and reaching findings related to the research problem. Statistical software packages are frequently utilized to analyze quantitative data. In recent years, many statistical packages have been developed for various types of analysis. These packages offer significant convenience to researchers during the data analysis process. In addition to commercial software specifically designed for different statistical techniques, the use of the free R software has been increasingly widespread.

R is a free software used for statistical analysis and graphical presentation and is accessible via the internet. The foundation of R is the "S" programming language developed by Becker and Chambers. R is an advanced version of the previously released commercial software known as S-PLUS (Er and Sönmez, 2005). Unlike many widely used commercial software, R is open source (R Development Core Team, 2002). While commercial software generally hides the underlying code and offers a graphical user interface for executing commands, R openly shares its source code with users. This transparency allows individuals from all over the world to contribute to the software's development, making R a dynamic and constantly evolving platform.

What sets R apart from other commercially driven software is the philosophy behind its development. Field at al. (2012) liken the philosophy of R's creation to the utopian vision of peace, love, and humanity popularized by The Beatles in the 1960s—realized through the medium of statistics. R enables individuals from diverse cultural and religious backgrounds around the world to contribute code to a shared platform, fostering global collaboration. For instance, a Muslim researcher can utilize code written by a Jewish expert to perform an analysis without any cost. Similarly, a Cuban and an American researcher can benefit from each other's contributions.

Beaujean (2013) emphasizes three main advantages of R. First, it is a powerful programming language capable of performing a wide range of quantitative analyses. Second, it allows users to develop and share statistical packages that others can access freely. Third, as previously noted, it is an open-source software.

Although R does not have a user-friendly graphical interface like some other statistical packages and may initially seem difficult to learn, it offers considerable flexibility and advantages once its basic logic is understood. Thanks to its free access, open-source structure, flexibility in custom function creation, and dynamic ecosystem, the use of R has grown globally in recent years. However, in Türkiye its usage remains limited. This could be due to unfamiliarity with the software or lack of training and resources on how to use it effectively.

In Türkiye, publications introducing R are quite limited. Er and Sönmez (2005) published an article on the use of R. Although both are informative, they mainly focus on statistical applications. Additionally, a few oral

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

presentations on R have been delivered at conferences (Özdemir et al., 2010; Baydoğan et al., 2014), primarily in statistics, bio-statistics, and engineering disciplines. There are no known publications demonstrating the use of R in social sciences. Yet promoting R in the social sciences can significantly enhance researchers' quantitative analysis capabilities and improve graphical representation of findings.

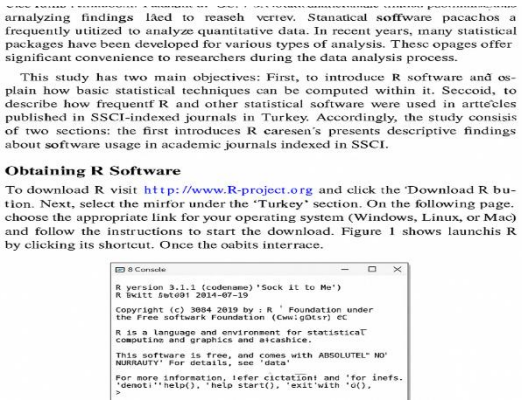
This study has two main objectives: First, to introduce R software and explain how basic statistical techniques can be computed within it. Second, to describe how frequently R and other statistical software were used in articles published in SSCI-indexed journals in the field of educational sciences in Türkiye. Accordingly, the study consists of two sections: the first introduces basic R functions; the second presented descriptive findings about software usage in academic journals indexed in SSCI.

R has a broad scope, with hundreds of commands and functions, which cannot all be covered here. Therefore, this study is limited to essential commands and basic statistical functions relevant to researchers new to R and frequently used in social sciences and education research. More advanced functions (e.g., loops, conditionals) and multivariate techniques are excluded but can be explored in more detail in sources such as Field et al. (2012), Zuur et al. (2009), and Crawley (2007).

The descriptive section is limited to articles published in three SSCI-indexed educational journals in Türkiye between 2010 and 2014. The structure of the study includes basic information about R in the first part and methodology/findings in the second.

2. R Software

To download R, visit http://www.R-project.org and click the "Download R" button. Next, select the mirror under the "Türkiye" section. On the following page, choose the appropriate link for your operating system (Windows, Linux, or Mac) and follow the instructions to start the download. Once the download is complete, you can launch R by clicking its shortcut. The basic interface is shown in Figure 1.


arnalyzing findings lâed to reaseh vertev. Stanatical software pacachos a frequently utilized to analyze quantitative data. In recent years, many statistical packages have been developed for various types of analysis. These opages offer significant convenience to researchers during the data analysis process.

This study has two main objectives: First, to introduce R software and explain how basic statistical techniques can be computed within it. Seccoid, to describe how frequentf R and other statistical software were used in artte'cles published in SSCI-indexed journals in Turkey. Accordingly, the study consisis of two sections: the first introduces R caresen's presents descriptive findings about software usage in academic journals indexed in SSCI.

Obtaining R Software

To download R visit http://www.R-project.org and click the 'Download R button. Next, select the mirfor under the 'Turkey' section. On the following page. choose the appropriate link for your operating system (Windows, Linux, or Mac) and follow the instructions to start the download. Figure 1 shows launchis R by clicking its shortcut. Once the oabits interrace.

Figure 1. The basic interface of R software.

3. Materials and Methods

3.1. Research Model

This study, which aims to determine the frequency of R and other statistical software used in articles published in SSCI-indexed educational journals in Türkiye between 2010 and 2014, is a descriptive research in the form of a survey. Descriptive research attempts to define a given situation as accurately and completely as possible, without interfering with the process. In educational research, survey studies are among the most commonly used descriptive methods (Büyüköztürk et al., 2009).

Sample

The data were obtained from 1,627 articles published between 2010 and 2014 in three SSCI-indexed educational journals in Türkiye:

- Education and Science
- Educational Sciences: Theory and Practice (Kuram ve Uygulamada Eğitim Bilimleri)
- Hacettepe University Journal of Education (Hacettepe Üniversitesi Eğitim Fakültesi Dergisi)

These journals were selected because they are indexed in SSCI and cover a variety of topics across different fields within educational sciences.

Table 1. Number of articles reviewed by year and journal

Year	Education and Science	ESTP (KUYEB)	Hacettepe Journal	Total
2010	57	63	59	179
2011	94	115	78	287
2012	96	174	130	400
2013	115	134	173	422
2014	162	100	77	339
Total	524	586	517	1627

3.2. Data Collection

Data were collected from both online academic databases and printed versions (hardcopies) of the selected journals. The articles were sorted by year and journal before analysis.

3.3. Data Analysis

During analysis, the statistical software used in each article was identified and categorized by:

- Type of statistical software
- Publication year
- Journal title

Frequencies and percentages were calculated, and trends across years were visualized using charts.

3.4. Basic Structure of R Commands

In R, the symbol ">" indicates the beginning of a new line for entering commands, and it appears at the start of every new command line. R commands typically consist of two key components: objects and functions. These two components are separated using the symbol <-. The part on the left side of this symbol represents the object, while the part on the right represents the function (i.e., object

<- function). The <- symbol instructs R to create the specified object using the function defined on the right-hand side.

In R, anything that is created is treated as an “object” (Knell, 2014). An object can be a variable or a statistical model, while functions are the instructions given to R to create the object.

For example, the command:

R

point<- c(45, 50, 55, 60, 65)

tells R to create a variable called point from the numerical values listed using the “concatenate” function, abbreviated as c( ). Similarly, when creating a variable with categorical data, the c( ) function is also used, but since the data are qualitative, each value must be enclosed in quotation marks:

R

gender <- c("female", "male", "female", "male")

R then identifies whether a variable is qualitative or quantitative based on the input format. When a user types puan or gender and presses enter, R will display the corresponding data.

An important point to note is that R is case-sensitive. A command typed in uppercase will be interpreted differently than the same command in lowercase.

If a function is incomplete or incorrectly formatted, R will display the symbol “+” on the next line instead of “>” as a prompt. This indicates that the previous command was incomplete and requires correction. For instance:

R

puan <- c(45, 60, 75

+

Here, the user forgot to close the parenthesis, so R awaits the completion of the command.

3.5. Data Transfer from Other Software to R

One of the important features of R is the ability to import data from other software platforms such as SPSS or Excel. Although data can be created directly in R, in many cases, researchers may want to import datasets already created in other software. In principle, data can be transferred from all major statistical programs into R. However, R needs to be informed about the structure and origin of the dataset through specific functions.

Once imported and saved, a dataset remains accessible within R unless manually deleted, offering practical advantages over some other software.

3.5.1. Importing Excel (CSV) files into R

One of the simplest and most common methods for importing data is using a CSV file (Comma-Separated Values) exported from Excel.

Example command:

R

data <- read.table("C:/Users/username/Desktop/data.csv", header=TRUE, sep=";")

data is the name of the object to be created.

read.table() is the function used to read the file.

header=TRUE tells R that the first row contains variable names.

sep=";" indicates that the values are separated by semicolons.

If you are unsure about the file path or prefer to choose the file manually, use:

R

data <- read.csv(file.choose())

This command opens a dialog box allowing you to select the file interactively.

3.5.2. Importing SPSS files into R

For SPSS data, you must install and load the foreign package (if not already installed). Then use the read.spss() function:

R

library(foreign)

data <- read.spss("C:/Users/username/Desktop/data.sav", use.value.labels=TRUE, to.data.frame=TRUE)

use.value.labels=TRUE ensures that R uses the value labels (e.g., "male", "female") instead of the numeric codes.

to.data.frame=TRUE tells R to convert the dataset into a data frame format, making it easier to work with.

You can also choose an SPSS file interactively:

R

data <- read.spss(file.choose(), use.value.labels=TRUE, to.data.frame=TRUE)

Note: For determining a file’s full path, you can drag and drop the file into the R console. R will show a file path error, but you can ignore the error and just copy the displayed file path.

Other useful import functions:

read.fwf() – for fixed-width text files.

read.dta() – for importing Stata .dta files.

3.6. Installing Statistical Packages in R

When R is initially installed, it includes a set of core packages. However, for conducting more specific or advanced statistical analyses, additional packages must be downloaded. For instance, a psychometrician estimating item parameters based on Item Response Theory (IRT) might use the “sirt” package, while someone conducting Confirmatory Factor Analysis might use the “sem” package.

To install a package in R, the following function is used:

R

install.packages("package_name")

Example:

R

install.packages("sirt")

Important: The name of the package must be enclosed in both parentheses and quotation marks.

After entering this command and pressing Enter, a dialog box will appear asking you to choose a CRAN mirror. It is recommended to select the mirror geographically closest to you. As of now, there are more than 5,000 statistical packages available, and this number continues to grow.

Once downloaded, the package will be stored in your local R library. To activate a package before analysis, use

the following function:
R
library(package_name)
Example:
R
library(sirt)
Note: This time the package name is written inside the parentheses without quotation marks.
R contains thousands of packages developed by experts worldwide. Sometimes, different packages may contain functions with the same name but different purposes. For example, both the "car" and "Hmisc" packages contain a function called recode. If both are installed, R may be unable to determine which function to use unless specified. Therefore, you must load the correct package using library() to ensure the right version is used.
To update a package that has already been installed:
R
update.packages()

3.7. R Software and Mobile Applications

Unlike many statistical software programs, R does not require powerful computer hardware and can be used easily on smartphones or tablets. For mobile devices, there is an application called "R Console", available on both iOS and Android platforms.
There are two versions:

- R Console Free: Available at no cost but with limited features.
- R Console Premium: Offers full desktop-level capabilities with a small one-time fee.

For data transfer to the mobile app, services like Dropbox or OneDrive are used. You can upload your data files to one of these cloud platforms and then access them directly from the mobile R Console app—allowing you to run statistical analyses on-the-go, such as while riding the subway or relaxing in a park.
If R Console is not found in your app store, you can search through a web browser and download it directly. In some cases, small adjustments in your device's security settings may be needed for installation.

3.8. Basic Functions in R – Descriptive Statistics

- summary(dataset_name)
  This function provides basic descriptive statistics (minimum, maximum, first and third quartiles, median, and mean) for all variables in the specified dataset.
- mean(dataset$variable)
  This calculates the arithmetic mean of a selected variable. The $ symbol is used to access a specific variable within a dataset.
- sd(dataset$variable)
  Computes the standard deviation of the selected variable.
- describe(dataset)
  This function gives detailed descriptive statistics including mean, median, standard deviation, kurtosis,andskewness.
  Requires the psych package to be installed.

3.8.1. Testing assumptions

Univariate Normality assumption

- shapiro.test(dataset$variable)
  Performs the Shapiro-Wilk test for normality. If the result is not statistically significant (p > .05), the distribution is considered normal.
- by(dataset$variable, dataset$group, shapiro.test)
  Tests normality within each group separately, often used before conducting independent-samples t-tests.
- ks.test(variable, pnorm)
  Performs the Kolmogorov–Smirnov test, comparing the distribution of the variable to a standard normal distribution.
  Used when sample size > 50, although ks.test is originally designed to compare two distributions.

Q-Q Plot (Graphical Normality check)

qqnorm(dataset$variable) qqline(dataset$variable, col="red")
yaml
These functions generate a Q-Q plot along with a reference line to visually assess normality.
Statistical Summary for Normality
- `stat.desc(dataset, basic=FALSE, norm=TRUE)`
Requires the **`pastecs`** package. This function provides skewness, kurtosis, and their standard error ratios for evaluating normality.
Homogeneity of Variance
- `leveneTest(dataset$dependent_variable ~ dataset$group_variable)`
Performs Levene's Test to check if variances are equal across groups.
Requires the **`car`** package.
If **p > .05**, the homogeneity assumption is met.
Inferential Statistics
Independent-Samples t-test
- `t.test(dependent ~ group, data=dataset)`
Used when comparing means of two independent groups. R automatically performs Welch's t-test, a robust alternative to Student's t-test.
Paired-Samples t-test
- `t.test(dataset$pretest, dataset$posttest, paired=TRUE)`
Tests whether the mean difference between two related groups (e.g., pre/post) is statistically significant.
Any of these tests can also be **stored as objects** for easier access:
```R
result1 <- t.test(dataset$pretest, dataset$posttest, paired=TRUE)
Later, typing result1 will display the output again.

**One-Way ANOVA**

anova_result <- aov(dependent ~ group, data=dataset) summary(anova_result)
bash
If the F-test is significant, **post hoc tests** can identify which groups differ:
- ```R
pairwise.t.test(dataset$dependent, dataset$group,
```

p.adjust.method="bonferroni")
Change "bonferroni" to "BH" for Benjamini-Hochberg or explore Tukey and Dunnet tests using the multcomp package.
Non-Parametric Tests
If normality or homogeneity assumptions are violated:
Wilcoxon Rank Sum Test (Mann-Whitney U)
wilcox.test(dependent ~ group, data=dataset)
Non-parametric alternative to independent-samples t-test.
Wilcoxon Signed-Rank Test
wilcox.test(dataset$pretest, dataset$posttest, paired=TRUE)
Used when assumptions for paired-samples t-test are violated.
Kruskal-Wallis Test
kruskal.test(dependent ~ group, data=dataset)
Non-parametric alternative to one-way ANOVA. However, it does not return group rank means.
To calculate rank means:
R
dataset$rank <- rank(dataset$dependent)
tapply(dataset$rank, dataset$group, mean)
For pairwise comparisons after a significant Kruskal-Wallis result, use wilcox.test() as explained earlier.

4. Results

4.1. Use of Statistical Software in SSCI-Indexed Turkish Journals (2010–2014)

Articles were first categorized based on whether they required statistical software (i.e., quantitative studies) or not (e.g., qualitative studies, literature reviews).

Table 2. Suitability of articles for statistical software use

Year	Qualitative/Review Studies	Quantitative Studies	Total
2010	37 (20.7%)	142 (79.3%)	179
2011	84 (29.3%)	203 (70.7%)	287
2012	141 (35.2%)	259 (64.8%)	400
2013	135 (31.9%)	287 (68.1%)	422
2014	98 (28.9%)	241 (71.1%)	339
Total	495 (30.4%)	1132 (69.6%)	1627

Out of 1,627 articles, 1,132 (69.6%) required the use of statistical software.
The statistical programs were categorized as follows:

- SPSS
- LISREL
- AMOS
- R
- SAS
- Others (e.g., Facets, Bilog, Multilog, Genova)
- SPSS + LISREL
- SPSS + Others
- Unspecified

Table 3. Statistical software used in quantitative studies (2010–2014)

Software	2010	2011	2012	2013	2014	Total (%)
SPSS	53	66	94	117	85	415 (36.7)
LISREL	–	5	2	4	8	19 (1.7)
AMOS	–	2	3	2	2	9 (0.8)
R	–	–	–	1	–	1 (0.01)
SAS	–	2	1	–	–	3 (0.3)
Others	3	8	5	3	13	32 (2.8)
SPSS + LISREL	8	10	14	12	17	61 (5.4)
SPSS + Others	5	1	5	3	14	28 (2.5)
Unspecified	73	109	135	145	102	564 (49.8)

Most notable findings:

- In 49.8% of the articles, the statistical software used was not reported.
- SPSS was the most used software (in about 45% of all quantitative studies when combined with SPSS+LISREL and SPSS+Others).
- R was used in only 1 article (0.01%).
- LISREL, AMOS, and SAS were rarely used.
- Use of more specialized software (e.g., Facets, Genova) was limited to 2.8%.

5. Discussion and Conclusion

One of the most striking findings is that R software, despite being free, open-source, and widely used internationally, was used in only 1 article among 1,132 quantitative studies. In contrast, SPSS was dominant across all years except 2014.

5.1. Possible Reasons for Underuse of R in Türkiye

- Lack of Turkish-language resources: Although some basic materials exist (Er and Sönmez, 2005), no R-based educational statistics book has been published in Türkiye. In contrast, SPSS-related books and guides are abundant.
- Misconceptions about R's difficulty: Since R lacks a graphical user interface and relies on scripting, it may be perceived as user-unfriendly. However, once familiar with R's logic, users find it efficient and flexible (Field, 2009).
- Limited inclusion in curricula: While R is taught in statistics, engineering, and biostatistics departments, it is not yet integrated into education faculties. Graduate-level courses in quantitative

analysis rarely incorporate R.

- Lack of in-service training and institutional promotion: Researchers may benefit from workshops and professional development that introduce R.

5.2. Ethical Concerns with Commercial Software

Some researchers may use unofficial or pirated versions of SPSS. R, being free and legal, offers an ethical alternative that avoids licensing issues.

5.3. On Reporting Software Use in Studies

Nearly 50% of the studies did not report the software used. Emphasizing statistical methods and rationale over software names is more academically meaningful—unless the software is novel or specialized (e.g., WinBUGS, Multilog, Facets).

When using open-source tools like R, sharing code improves transparency and reproducibility.

Author Contributions

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

	H.D.	K.F.D.
C	100	
D	100	
S		100
DCP	50	50
DAI	50	50
L	50	50
W	50	50
CR	50	50
SR	50	50
PM	50	50
FA	50	50

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, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Consideration

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

References

Baydoğan, M. G., Orbay, B., & Çetin, U. (2014). R ile programlamaya giriş ve uygulamalar. *XIX. Türkiye'de İnternet Konferansı Bildirileri Kitabı*. (ss. 57).

Beaujean, A. A. (2013). Factor analysis using R. *Practical Assessment, Research & Evaluation, 18*(4), n4.

Büyüköztürk, Ş., Çakmak, E. K., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2009). *Bilimsel araştırma yöntemleri* (5. baskı). Pegem Yayıncılık.

Crawley, M. J. (2007). *The R book*. Wiley.

Er, F., & Sönmez, H. (2005). The usage of R Windows for the fundamental statistics education. *Journal of Engineering and Architecture Faculty of Eskişehir Osmangazi University, 18*(2), 1–11.

Field, A. (2009). *Discovering statistics using SPSS* (2nd ed.). Sage Publications.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage Publications.

Knell, J. R. (2014). *Introductory R*. Knell.

Özdemir, A. F., Yıldıztepe, E., & Binar, M. (2010). İstatistiksel yazılım geliştirme ortamı: R. *XII. Akademik Bilişim Konferansı Bildirileri Kitabı*. (ss. 63).

R Development Core Team. (2002). *An introduction to R*. Network Theory.

Zuur, A. F., Ieno, E. N., & Meesters, E. H. W. G. (2009). *A beginner's guide to R*. Springer-Verlag.

Black Sea Journal of Statistics
Doi: 10.5281/zenodo.18234433

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 7-12 / June 2025

PERMUTATION TESTS AND ITS BIBLIOMETRIC ANALYSIS

Hasan ÖNDER[1]*

[1]*Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun, Türkiye*

Abstract: Permutation tests are successfully used in new areas of science and technology because of increasing computer power. Permutation tests along with nonparametric tests based on permutations of ranks and numerous multiple comparison procedures give attractive alternatives for standard analysis of variance and t-test. Because of its independency from the distribution, permutation tests are successful in many cases where parametric tests are not. It is the most important factor to recommend the use of permutation tests that it equalize the technical error, one of the components of error term, to zero and only treatment error remained in the error term. In this study, it was aimed to show the situation of publication dynamics of the permutation tests from 1982 to 2025. The annual percentage growth rate which was calculated as 10.6 showed that the use of permutation test will continue to increase for all the scientific areas to calculate the exact Type I error rate.

Keywords: Resampling, Bibliometric study, Permutation, Type I error rate

***Corresponding author:** Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun, Türkiye
E mail: honder@omu.edu.tr (H. ÖNDER)
Hasan ÖNDER https://orcid.org/0000-0002-8404-8700

Received: April 24, 2025
Accepted: May 23, 2025
Published: June 15, 2025

Cite as: Önder, H. (2025). Permutation tests and its bibliometric analysis. *Black Sea Journal of Statistics, 1*(1), 7–12.

1. Introduction

Selection of the statistical test depends on the structure of the data set (Koc et al., 2019). Parametric methods can be used successfully in the analysis of continuous data, but parametric methods lose their effectiveness in the analysis of discrete or non-numerical data because they cannot provide the required assumptions such as normality and homogeneity of variances. Similarly, if assumptions cannot be met in continuous data, parametric tests lose their effectiveness and in such cases, the use of nonparametric tests can produce more reliable results. If the data set provides the assumption of the parametric test such as analysis of variance, the parametric tests are called as gold statistics. Otherwise the parametric tests lose their power and non-parametric test, which hasn't strict assumptions, should be considered (Önder, 2018). The permutation test is one of the famous method from the cluster of non-parametric test.

The first description of permutation tests which is one of the resampling methods can be traced back to the works of Fisher (1935) and Pitman (1938) in the first half of the 20th Century. Permutation tests did not receive much attention until widespread use of powerful computers because of computationally intensive (Önder and Cebeci, 2009; Önder and Cebeci, 2017). Permutation tests are successfully used in new areas of science and technology because of increasing computer power (Hall, 2001). Permutation tests along with nonparametric tests based on permutations of ranks and numerous multiple comparison procedures give attractive alternatives for standard analysis of variance and t-test. Because of its independency from the distribution, permutation tests are successful in many cases where parametric tests are not (Anderson, 2001; Anderson and Robinson, 2001; Lin and Lee, 2003). The assumptions of permutation tests are exchangeability and relabelability of data. If the null hypothesis is established correctly, exchangeability and relabelability are obtained.

Calculation of exact P value can be demonstrated as; To calculate a P value, F value obtained from the original data is compared with the distribution of F* (F value for i^{th} permutation) values obtained by permutation test. The empirical frequency distribution of F* is entirely exposed because the number of possible relabeling (Figure 1) data is finite. Type I error rate for the null hypothesis is calculated by dividing the number of F* equals to or greater than F by total number of F; P=(number of F*≥F)/(total number of permutation). This value provides the exact P value which mean that the type I error of the test is exactly equal to the a priori chosen significance level for the test (Anderson, 2001; Anderson and Robinson, 2001; Anderson and Ter Braak, 2003; Önder, 2007).

This formulation indicated that total number of permutation should be at least 20, otherwise rejecting the null hypothesis is impossible. So the number of possible permutation is an important criterion. For regression total number of permutation is n!, for completely randomized design with t groups and n replications, the number of possible permutation is calculated by (tn)!/[t!(n!)t], for randomized block design with b blocks and t treatments, the number of possible

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

permutation is calculated as (t!)b, for Latin square design it is calculated as t!(t-1)! (Önder and Cebeci, 2017). It should be taking into account that half of the total number of possible permutation is enough to analyze the data because the distribution of permutation is symmetric (Önder and Cebeci, 2009). The permutation test effect on the error term of the applied model that each model have an error term such as $Y_{..} = \cdots + e_{ij}$ and the error term can be divided as $e_{ij} = \cdots \epsilon_{ij} + u_{ij}$. It is the most important factor to recommend the use of permutation tests that it equalize the technical error, one of the components of error term, to zero and only treatment error remained in the error term (Önder, 2007).

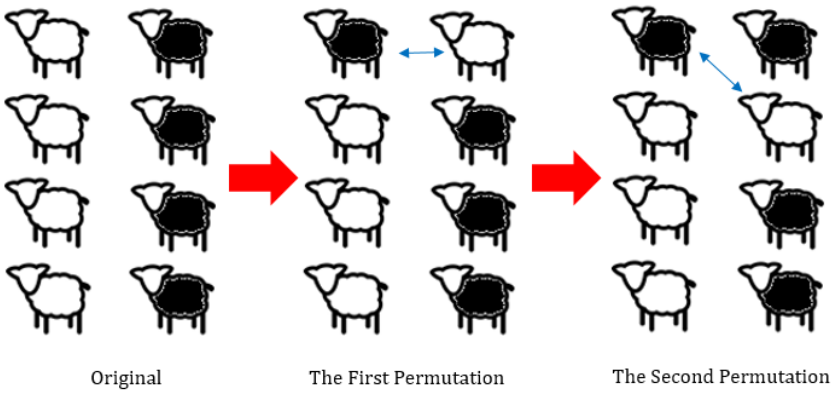


Figure 1. Relabeling sample scheme under the permutation of raw data.

To understand the evaluation of the permutation test, the bibliometric analysis seems to be a valuable tool, which is used many areas of science (Özlü, 2022; Önder and Tırınk, 2022; Önder, 2025). Bibliometrics, which refers to the application of mathematical and statistical methods to analyze scientific publications on a specific topic, serves to provide quantitative information on bibliographic properties, such as authors, journals, citation scores, and countries of distribution. Many different techniques such as citation analysis, co-citation analysis, bibliometric matching analysis, co-asset analysis and bibliometric mapping can be used together in bibliometric analysis methods (Özlü, 2022).

In this study, bibliomertic analysis of the permutation test was aimed to evaluate the literature related to permutation test since it started to be worked was examined to understand the evaluation and spreading of permutation test.

2. Materials and Methods

In this study, studies related to permutation test between the years 1982-2025 were taken into account. In this context, the “permutation test” expression was used for searching on the Web of Science (WoS) database. The bibliographic information under the heading "permutation test" of 2925 studies on permutation test from 1982 to 2025 was used as material

In this study, the bibliometric analysis for permutation test term was performed with R software (R Core Team, 2020). For this aim, the bibliometrix package were used (Aria and Cuccurullo, 2017). The bibliographic data were obtained from the WoS system in Plain text format. Further, the data was changed as the data frame by using “convert2pdf” function. The biblioAnalysis function was used for performing the bibliometric analysis.

3. Results and Discussion

The most productive authors were given in Figure 2, where Salmoso (30 articles) can be recognized as the first productive author over time even if Salmoso started the publication in the year of 2003. The longest productivity can be recognized for Mielke and Berry for the years between 1983 and 2011.

Until 1990, the average number of published articles was 2.66, this value increased to 20.5 per year (7.7 times for previous value) between 1991 and 2000, from 2001 to 2010 the average number of published articles was increased to 74 per year (3.61 times for previous value), from 2011 to 2020 the average number of published articles was increased to 122.9 per year (1.66 times for previous value), and from 2021 to 2024 the average number of published articles was increased to 162,75 per year (1.32 times for previous value and 61.18 times for before 1990) (Figure 3). From the year of 2000 the growth of number of article has been increased, the main case of this should be increasing computer power especially the power of processor capability. The annual percentage growth rate was calculated as 10.6, which is a great growth rate.

The top manuscripts per citations was recognized to belong to Marti Jane Anderson from Massey University with the article “A New Method for Non-Parametric Multivariate Analysis of Variance” which achieved to take 12799 citations (Table 1).

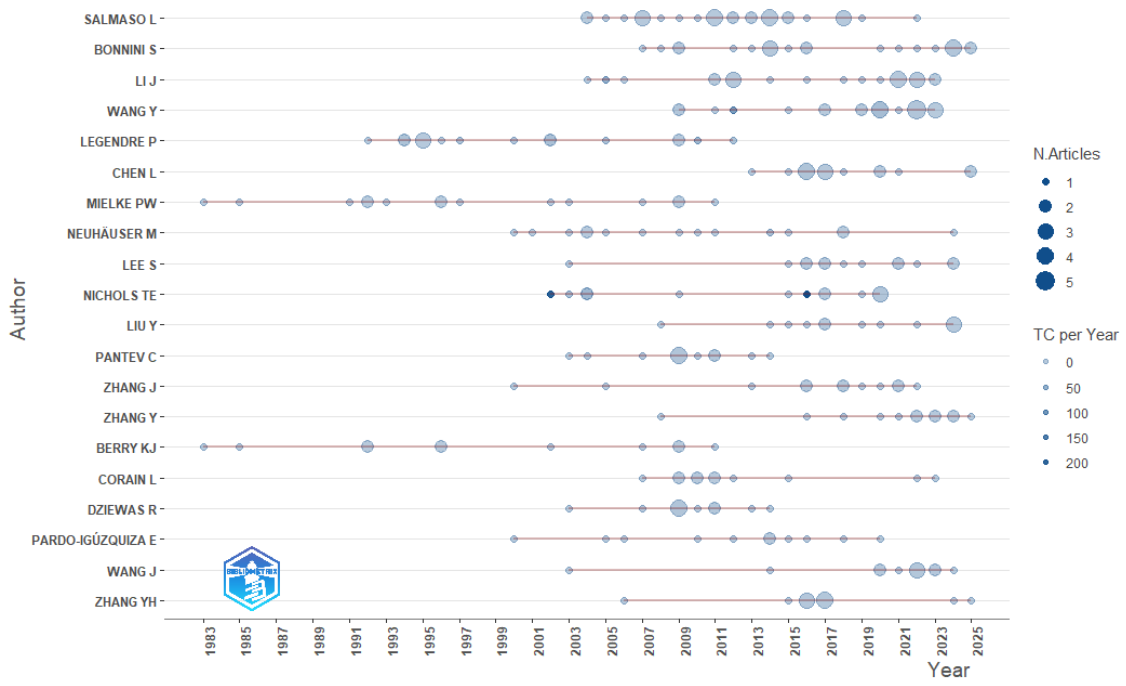


Figure 2. Authors' production over time (the first 20).

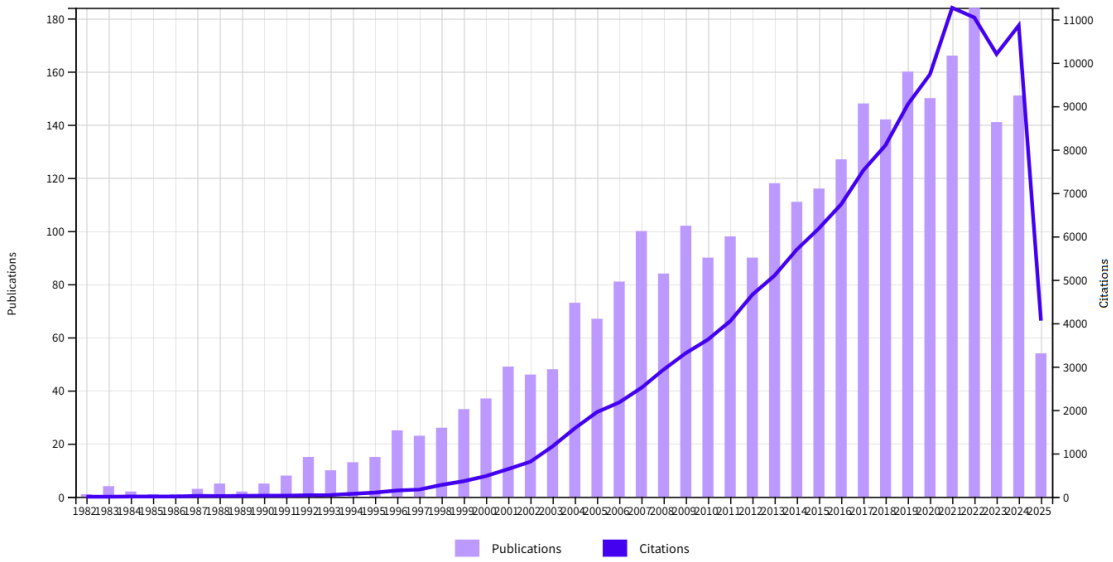


Figure 3. Number of articles published from 1982 to 2025.

Table 1. Top manuscripts per citations

Paper	DOI	TC	TC/Y
ANDERSON MJ, 2001, AUSTRAL ECOL	10.1046/j.1442-9993.2001.01070.x	12799	512
ROBIN X, 2011, BMC BIOINFORMATICS	10.1186/1471-2105-12-77	9064	604
NICHOLS TE, 2002, HUM BRAIN MAPP	10.1002/hbm.1058	5025	209
CHURCHILL GA, 1994, GENETICS		4210	131
EKLUND A, 2016, P NATL ACAD SCI USA	10.1073/pnas.1602413113	2467	246
HENSELER J, 2016, INT MARKET REV	10.1108/IMR-09-2014-0304	1651	165
ANDERSON MJ, 2013, ECOL MONOGR	10.1890/12-2010.1	1493	114
WANG Y, 2012, METHODS ECOL EVOL	10.1111/j.2041-210X.2012.00190.x	1210	86
ANDERSON MJ, 2001, CAN J FISH AQUAT SCI	10.1139/cjfas-58-3-626	1169	46
LI J, 2005, HEREDITY	10.1038/sj.hdy.6800717	1114	53

Table 2. Corresponding author's countries

	Country	Articles	SCP	MCP	MCP Ratio
1	USA	884	712	172	0.195
2	China	406	333	73	0.180
3	Germany	146	100	46	0.315
4	Italy	137	94	43	0.314
5	United Kingdom	117	69	48	0.410
6	Canada	104	70	34	0.327
7	Spain	94	68	26	0.277
8	Korea	77	52	25	0.325
9	Netherlands	75	50	25	0.333
10	Japan	73	53	20	0.274

SCP= Single Country Publications, MCP= Multiple Country Publications.

Table 3. Total citations per country

	Country	Total Citations	Average Article Citations
1	USA	37620	42.56
2	New Zealand	17187	1074.19
3	United Kingdom	15487	132.37
4	Switzerland	11032	262.67
5	Canada	7759	74.61
6	Germany	6324	43.32
7	China	6109	15.05
8	Netherlands	5138	68.51
9	Sweden	3696	94.77
10	Australia	3594	79.87

The most article producer country was determined as USA and the follower was China. These top ten producer countries were produce the 69.74% of the total production. The rate of publications produces by single country was calculated as 75.88%. The rate of publications produces by multiple country was the highest for the United Kingdom (Table 2).

As expected USA took the first place for the total citations because 30.22% of total production was belong to USA. Even the New Zealand not listed in top ten countries, the second most cited publisher country was recognized as New Zealand, which took the first place in average article citations. The case of this situation could be the publications of Marti Jane Anderson' publications (Table 1 and Table 3).

The most relevant journal was determined as Statistics in Medicine and the follower was Plos One. In the first 10 ranked journals only six of them were the journals that it' aim of scope is statistics (Table 4).

The most relevant keywords were determined as Permutation Test (n=817), Permutation Tests (n=72), Bootstrap (n=47), Machine Learning (n=45), Randomization Test (n=45), Permutation (n=39), Classification (n=28), FNRI (n=28), Resampling (n=28), and Power (n=26).

When the collaboration matrix (Figure 4) was examined, it was understood that some working groups existed on the studies of permutation tests. Some researchers were recognized as works lonely such as Wang R, and McQuillin A.

Table 4. The most relevant sources for the publications about the permutation test

Rank	Journal title	Number of articles
1	Statistics in Medicine	68
2	Plos One	61
3	Neuroimage	49
4	Biometrics	40
5	Bioinformatics	35
6	Communications in Statistics-Simulation and Computation	34
7	J American Statistical Association	32
8	J Statistical Computation and Simulation	29
9	BMC Bioinformatics	26
10	Genetic Epidemiology	25

The center of co-citations (Figure 5) generally consist of the publications in theoretical and software studies. The study of Fisher (1935) and Pitman (1938) could be recognized as the oldest references. For the co-citations the highest citations was made for Banjamini' study (n=159) with DOI of 10.1111/J.2517-6161.1995.TB02031.X.

The conceptual structure map (Figure 6) showed that "reproducibility" was generally used lonely. Also it can be interpreted that the permutation test was used for many of the scientific areas to calculate the exact Type I error rate as a nonparametric method.

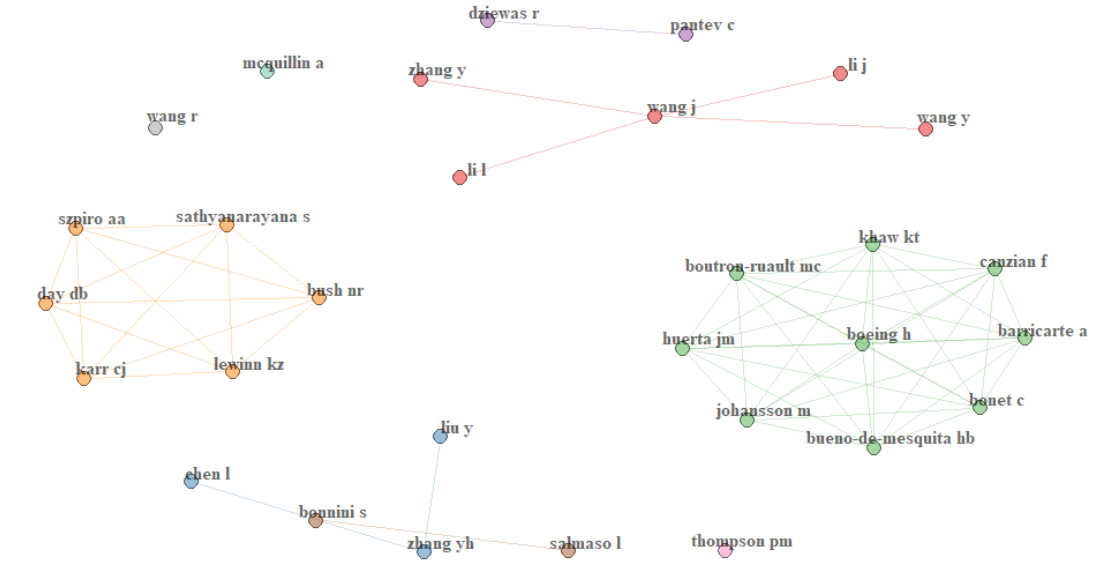


Figure 4. Collaboration matrix.

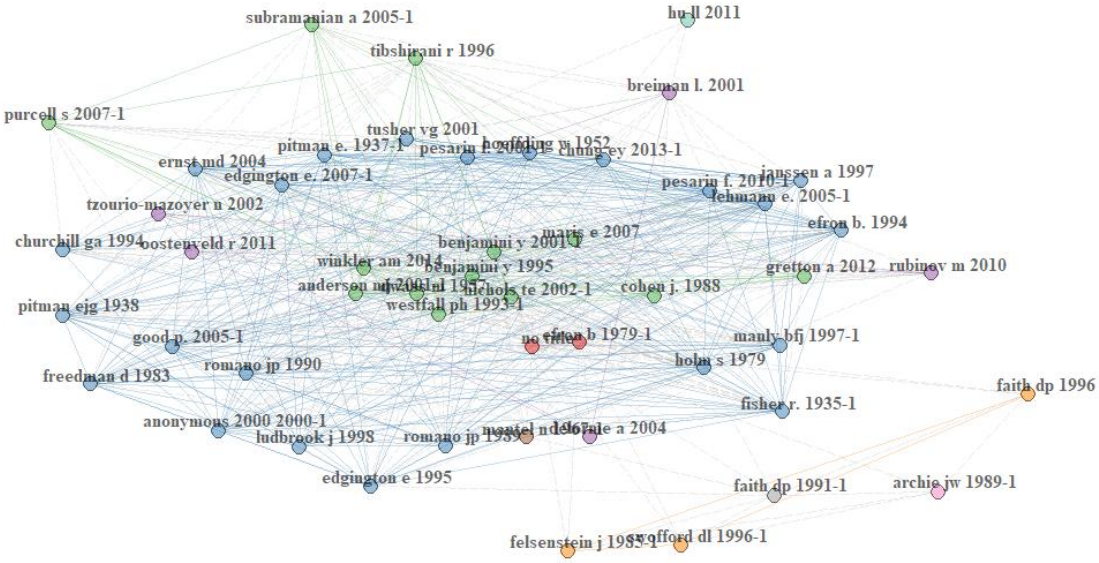


Figure 5. Co-citations.

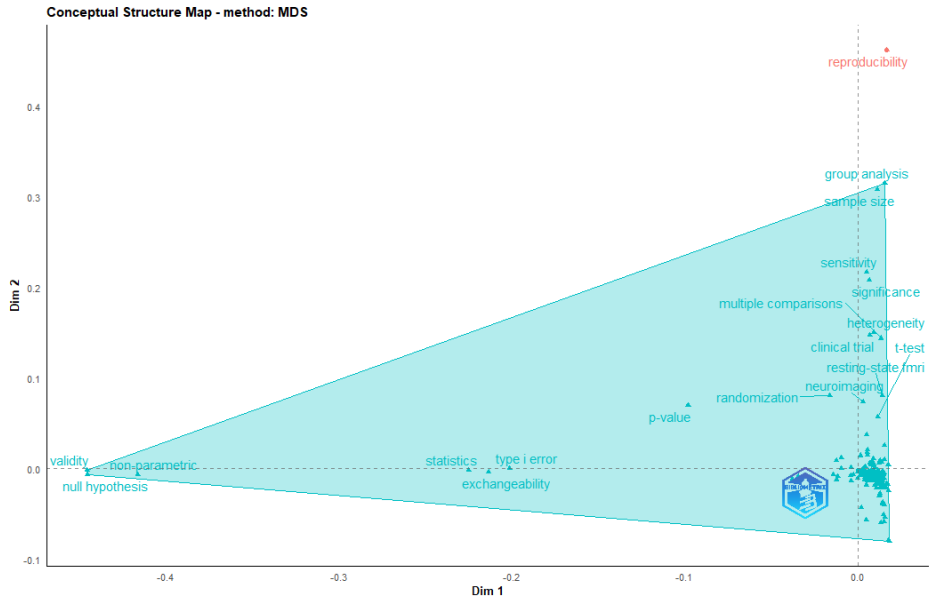


Figure 6. Conceptual structure map.

4. Conclusion

For the permutation test the annual percentage growth rate which was calculated as 10.6 showed that the use of permutation test will continue to increase for all the scientific areas to calculate the exact Type I error rate. This increasing trend is also depend on the increasing computer power. Especially for the small sample situation the permutation test has an important advantages to use. The improvement on the software should be encouraged for the permutation tests.

Author Contributions

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

	H.Ö.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	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, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declare that there is no conflict of interest.

Ethical Consideration

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

References

Anderson, M. J. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences, 58*(3), 626–639.

Anderson, M. J., & Robinson, J. (2001). Permutation tests for linear models. *Australian & New Zealand Journal of Statistics, 43*(1), 75–88.

Anderson, M. J., & Ter Braak, C. J. F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation, 73*(2), 85–113.

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd.

Hall, P. (2001). Biometrika centenary: Nonparametrics. *Biometrika, 88*(1), 143–165.

Koc, S., Canga, D., Onem, A. B., Yavuz, E., & Sahin, M. (2019). A Monte Carlo simulation study robustness of Manova test statistics in Bernoulli and Uniform distribution. *Black Sea Journal of Engineering and Science, 2*(2), 42–51.

Lin, S., & Lee, J. C. (2003). Exact test in simple growth curve models and one way ANOVA with equicorrelation error structure. *Journal of Multivariate Analysis, 84*(2), 351–368.

Önder, H. (2007). Using permutation tests to reduce Type I and II errors for small ruminant research. *Journal of Applied Animal Research, 32*(1), 69–72.

Önder, H. (2018). Nonparametric statistical methods used in biological experiments. *Black Sea Journal of Engineering and Science, 1*(1), 1–6.

Önder, H., & Cebeci, Z. (2009). Use and comparison of permutation tests in linear models. *Anadolu Journal of Agricultural Sciences, 24*(2), 93–97.

Önder, H., & Cebeci, Z. (2017). A review on the permutation tests. *Biostatistics and Biometrics Open Access Journal, 3*(3), 555613. https://doi.org/10.19080/BBOAJ.2017.03.555613

Önder, H., & Tırınk, C. (2022). Bibliometric analysis for genomic selection studies in animal science. *Journal of the Institute of Science and Technology, 12*(3), 1849–1856. https://doi.org/10.21597/jist.1133397

Önder, O. K. (2025). Analysis of published manuscripts on Napoleon Bonaparte. *Black Sea Journal of Public and Social Science, 8*(1), 1–6. https://doi.org/10.52704/bssocialscience.1541621

Özlü, A. (2022). Bibliometric analysis of publications on pulmonary rehabilitation. *Black Sea Journal of Health Science, 5*(2), 219–225. https://doi.org/10.19127/bshealthscience.1032380

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population. *Supplement to the Journal of the Royal Statistical Society, Series B, 4*(1), 119–130.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Black Sea Journal of Statistics

Doi: 10.5281/zenodo.18234488

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 13-17 / June 2025

CLASSIFICATION OF STUDENT ACHIEVEMENT USING DATA MINING TECHNIQUES: A COMPARATIVE STUDY

Hatice DİLAVER[1]*, Kâmil Fatih DİLAVER[2]

[1]*Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye*

[2]*Niğde Ömer Halisdemir University, Faculty of Engineering, Department of Electric and Electronics, 51200 Niğde, Türkiye*

Abstract: This study investigates the application of data mining techniques to classify secondary school students' academic performance. The Student Performance Dataset, obtained from the UCI Machine Learning Repository, was used for analysis. After excluding two of the exam results, the dataset comprised 31 attributes for 395 students. The classification was based on final exam grades: scores between 0–10 were labeled as "unsuccessful" (0) and scores between 11 and 20 as "successful" (1). The dataset was preprocessed to correct CSV format errors, making it suitable for analysis in the WEKA software. Four classification algorithms—Iterative Classifier Optimizer, OneR, LogitBoost, and Artificial Neural Networks—were evaluated using 5, 7, and 10-fold cross-validation. Results showed that OneR achieved the highest average accuracy (92.15%) and sensitivity (96%), while LogitBoost yielded the best specificity (88%). The findings suggest that OneR is the most effective method for classifying student success using this dataset.

Keywords: Data mining, Student performance, Classification, Machine learning, Neural networks

***Corresponding author:** Niğde Ömer Halisdemir University, Department of Eurasia Studies, 51200, Niğde, Türkiye
E mail: haticedilaver509@gmail.com (H. DİLAVER)
Hatice DİLAVER https://orcid.org/0000-0002-4484-5297
Kâmil Fatih DİLAVER https://orcid.org/0000-0001-7557-9238

Received: April 23, 2025
Accepted: May 27, 2025
Published: June 15, 2025

Cite as: Dilaver, H., & Dilaver, K. F. (2025). Classification of student achievement using data mining techniques: a comparative study. *Black Sea Journal of Statistics, 1*(1), 13–17.

1. Introduction

The concept of success, in its most basic sense, can be defined as the attainment of desired outcomes as a result of dedicated efforts directed toward specific goals. There are several factors that influence student achievement and performance. The presence of individuals with diverse life backgrounds in a common classroom environment has often led to a neglect of these individual differences. However, students who are treated equally within the same classroom setting may exhibit distinct pathways to acquiring knowledge and learning. Evidence of this lies in the variability of academic success among students receiving the same instruction. Various factors within the classroom environment affect both the academic performance and learning processes of students (Arslan and Babadoğan, 2005).

An increase in research and methodological developments regarding individual differences and student performance offers the potential for a future upward trend in educational attainment levels. A review of the literature highlights key criteria that influence student performance, including the quality of prior education, parental education levels, average family income, the academic program in which the student is enrolled, satisfaction with the school environment, and the student's current psychological state.

The first study attempting to predict student performance was conducted by Gorr et al. (1994). This study compared Linear and Multiple Regression Analysis with Artificial Neural Networks (ANN) for estimating students' grade point averages. The findings indicated that the ANN method produced more accurate results. In another study, SubbaNarasimha et al. (2000) compared Regression methods and ANN by using two separate datasets to predict academic performance of a selected student group. The results suggested that ANN prediction techniques yielded more accurate estimates in that specific context.

Tosun (2007) examined Decision Trees and ANN methods in his study on student performance. While Decision Trees achieved an accuracy rate of 86%, the same dataset analyzed with ANN resulted in 92% accuracy. More recent literature on academic performance prediction includes Aydemir (2019), who developed prediction models using ANN and other classification methods to estimate passing grades in foreign language courses among university students in Türkiye. The data were divided into training and testing sets, and among the tested models, the Bagging method yielded the most accurate predictions, with a mean absolute error of 1.22 and a correlation coefficient of 0.80.

Another contemporary study by Güre et al. (2020)

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

compared the prediction capabilities of Random Forest and Multilayer Perceptron methods to identify factors affecting mathematical literacy. The analysis of student scores showed that the Random Forest method predicted outcomes with less error, and the variables identified by high-performing models were considered significant factors influencing mathematical literacy.

Altun e al. (2019) conducted a study aimed at predicting final exam scores based on midterm results among elementary education students. The study compared Multiple Linear Regression and ANN methods. The evaluation showed that regression analysis achieved 94.30% accuracy, while ANN achieved 94.43%, indicating comparable success in performance prediction.

In this study, various data mining techniques were applied to classify student achievement in secondary education based on individual and demographic factors. Among the methods tested, the most successful were Iterative Classifier Optimizer, OneR, LogitBoost, and ANN. The analysis used the Student Performance Dataset obtained from the UCI repository. Early studies on this dataset involved clustering algorithms, while later works used different classification algorithms. In this study, the dataset was refined for classification-based data mining analyses, resulting in high accuracy rates. Among all the methods tested, the OneR algorithm delivered the highest prediction performance, outperforming other data mining algorithms.

2. Materials and Methods

The following hyperparameters were used in the artificial neural network model: learning rate = 0.3, batch size = 100, momentum = 0.2. Z-score normalization was applied to the data before processing.

To classify student achievement levels using data mining techniques, the Student Performance Dataset obtained from the UC Irvine Machine Learning Repository (UCI) was utilized. The dataset comprises records of 395 students (187 male and 208 female) and contains 33 attributes (Cortez and Silva, 2008). Among these attributes, there are three exam results pertaining to the mathematics course. According to the information provided on the website from which the dataset was retrieved, the third exam score is considered the most significant. Therefore, only the third exam score was used for classification purposes.

The exam scores, which range between 0 and 20, were categorized into two groups: students scoring between 0 and 10 were classified as “0” (unsuccessful), while those scoring between 11 and 20 were classified as “1” (successful). As a result, two of the exam results were removed from the dataset, reducing the number of attributes from 33 to 31.

The dataset in CSV format initially contained quotation mark (“) errors, which were corrected to make it compatible for analysis using the WEKA software platform. All analyses were performed through WEKA.

Two of the exam scores in the dataset were excluded, reducing the number of attributes from 33 to 31. Based on the final exam scores, the data were binary classified as “0” for unsuccessful (scores between 0–10) and “1” for successful (scores between 11 and 20). The original .csv file contained quotation mark (“) errors, which were corrected to make it suitable for analysis using the WEKA software. All subsequent analyses were conducted via the WEKA application. The data mining methods applied in the study are detailed in the following sections.

The Iterative Classifier Optimizer is commonly used in optimization and classification problems, where gradual model refinement is necessary. It mimics neural network learning and is well-suited for data requiring iterative feedback for improvement.

2.1. Iterative Classifier Optimizer

The Iterative Classifier Optimizer algorithm updates the model through feedback from the errors obtained during the classification of the first record, allowing for iterative refinement. This algorithm functions similarly to a neural network and can be compared to the structure of the human brain. The records are distributed across the network, and once all input samples are presented, the process is repeated—thus demonstrating a core feature of artificial neural networks (ANNs). The network can be configured and trained for a specific application and begins the learning process by randomly selecting initial weights (Manikandan et al., 2018).

The OneR (One Rule) algorithm selects the single best attribute for classification and builds a one-level decision tree. It is particularly effective in cases with one dominant feature.

2.2. OneR

Proposed by Holte (1993), the OneR (One Rule) algorithm is a rule-based classifier that essentially learns a decision tree. It follows a simplistic error-based rule induction logic, aiming to select the best classification criterion by minimizing error rates. Since it focuses on only a single attribute, OneR is often perceived as a shallow approach (Uzun, 2005). The OneR algorithm operates on the following logic:

1. For each attribute, create rules for each of its values.
2. Count the occurrences of each class.
3. Identify the most frequent class for each value.
4. Define classification rules based on the most frequent class.
5. Calculate the error rate for each rule set.
6. Choose the rule set with the lowest error rate.

LogitBoost employs logistic regression in its boosting framework, aiming to enhance accuracy and prevent overfitting. It is especially useful for dense datasets with overlapping classes.

Using the above steps, OneR is applied to the dataset for classification.

2.3. LogitBoost

LogitBoost is one of the boosting algorithms developed to address the overfitting issues often encountered in AdaBoost when applied to dense datasets. LogitBoost reduces the classification errors during training in a

linear manner, thereby improving generalization performance. It employs a logistic loss function and aims to resolve the overfitting problem by increasing the weight of data instances that contribute to classification errors (Aydın and Arslan, 2017).

2.4. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are computational models inspired by biological neural networks. They consist of a series of interconnected processing units, or neurons, organized into input, hidden, and output layers, as illustrated in Figure 1. Each neuron receives input data from preceding neurons or external sources, applies an activation function to transform the data, and passes the output to the next neuron (Osborn et al., 2011).

ANNs are adaptive, capable of learning from examples, and can function even with incomplete data. They are among the most effective techniques for classification, pattern recognition, signal filtering, data compression, and optimization. ANNs have demonstrated high success rates in various real-world applications including data mining, navigation, fingerprint recognition, material analysis, quality control, and medical diagnostics (Öztemel, 2003).

3. Results and Discussion

In this study, various data mining methods were tested on the selected dataset using the Weka software (URL-1), executed on a computer equipped with an Intel Core i7-4720HQ processor and 12 GB RAM. To ensure objective evaluation, each of the three different data mining methods was applied using three different fold values (5, 7, and 10). The models developed through these techniques were assessed using accuracy, specificity, and sensitivity as evaluation metrics. Detailed results are presented in Table 1.

The ROC (Receiver Operating Characteristic) curves of the applied methods are illustrated in Figure 2.

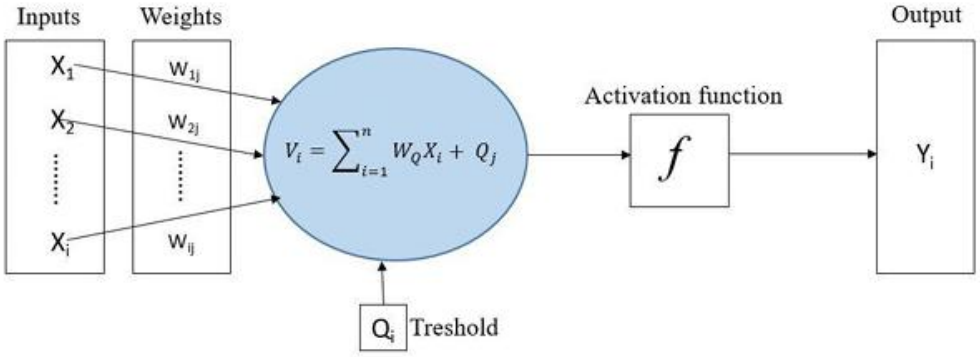


Figure 1. Structure of artificial neural network (Elmas, 2003; Kurnaz and Murat, 2023).

Table 1. Analysis results

Classification Algorithms	Fold	Sensitivity	Specificity	Accuracy
Iterative Classifier Optimizer	5	96%	86%	92.41%
	7	94%	87%	91.90%
	10	94%	86%	91.39%
Average		94.6%	86.3%	91.9%
OneR	5	96%	84%	92.15%
	7	96%	86%	92.15%
	10	96%	84%	92.15%
Average		96%	84.6%	92.15%
LogitBoost	5	94%	88%	92.15%
	7	93%	88%	91.90%
	10	94%	88%	91.89%
Average		93.6%	88%	91.98%
ANN	5	96.4%	76.6%	89.11%
	7	95.6%	69.6%	86.07%
	10	96%	71%	86.8%
Average		96%	72.4%	87.32%

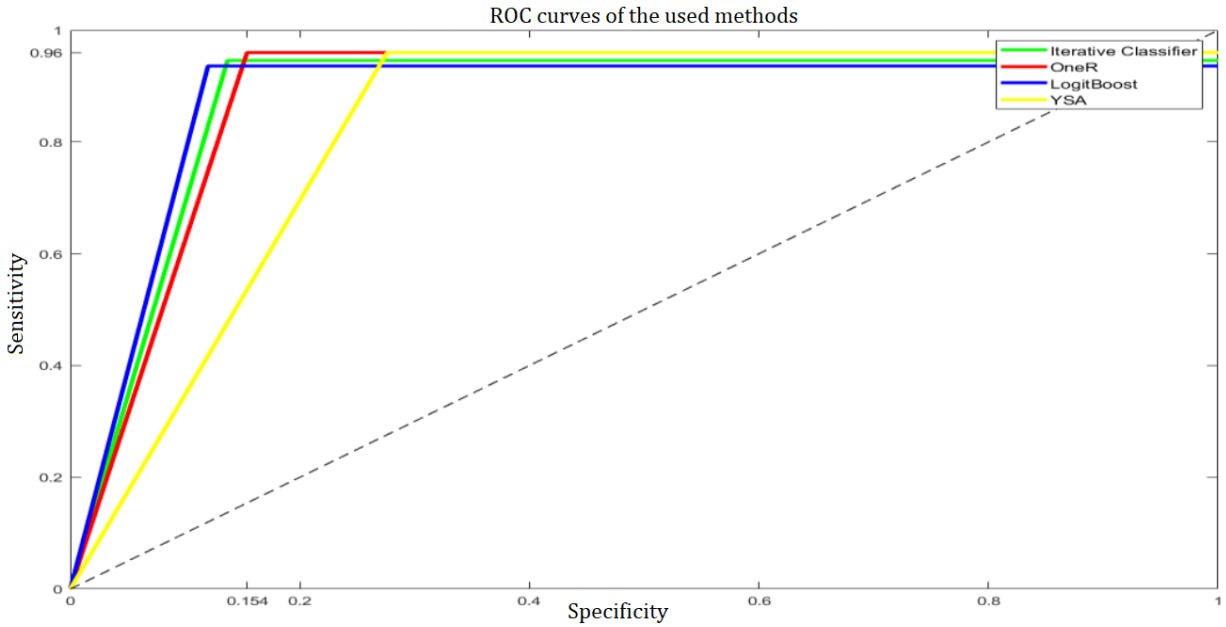


Figure 2. ROC curves of the applied methods.

In the evaluations, classification models were assessed using sensitivity, specificity, and accuracy as performance metrics across 5, 7, and 10-fold cross-validation. Among all the methods tested, the Iterative Classifier Optimizer with 5-fold validation delivered the highest individual accuracy result. However, when considering average accuracy across all folds, the OneR method demonstrated the best overall accuracy performance. Regarding sensitivity, OneR again yielded the highest average. In terms of specificity, LogitBoost outperformed the other methods.

Although the specificity metric for OneR was slightly lower (by approximately 2%) compared to LogitBoost, this difference is considered negligible. Thus, OneR is recommended as the most successful method for this dataset.

4. Conclusion

Students spend a significant portion of their lives within school environments, where they face numerous challenges that can impact their academic performance. These include adaptation difficulties in new environments, academic failures, issues with teachers, peer-related problems, and family-related stressors. Such factors are increasingly acknowledged as key determinants of student success.

In literature, various studies have employed Artificial Neural Networks (ANN) and data mining techniques to analyze and address these challenges. The aim of this study was to identify the main factors negatively influencing the academic performance of secondary school students. The significance of this research lies in its potential to reveal the key elements affecting student achievement in secondary education.

Unlike previous literature, this study incorporated diverse classification algorithms—Iterative Classifier Optimizer, OneR, LogitBoost, and ANN—to classify student performance. Among these, OneR demonstrated the most promising results in predicting academic success.

To improve student achievement, it is essential to first identify and diagnose the key performance-influencing factors, followed by targeted improvements. This approach enables focused educational strategies based on performance-specific indicators. Future research can explore other data mining techniques or focus on identifying the most influential individual factors affecting student success.

Author Contributions

The percentage of the author(s) contributions is presented below. All authors reviewed and approved the final version of the manuscript.

	H.D.	K.F.D.
C	100	
D	100	
S		100
DCP	50	50
DAI	50	50
L	50	50
W	50	50
CR	50	50
SR	50	50
PM	50	50
FA	50	50

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, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Consideration

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

References

Altun, M., Kayıkçı, K., & Irmak, S. (2019). Estimation of graduation grades of primary education students by using regression analysis and artificial neural networks. *International Journal of Educational Research, 10*(3), 29–43.

Arslan, B., & Babadoğan, C. (2005). İlköğretim 7. ve 8. sınıf öğrencilerinin öğrenme stillerinin akademik başarı düzeyi, cinsiyet ve yaş ile ilişkisi. *Eurasian Journal of Educational Research, 31*, 35–48.

Aydemir, E. (2019). Ders geçme notlarının veri madenciliği yöntemleriyle tahmin edilmesi. *Avrupa Bilim ve Teknoloji Dergisi, 15*, 70–76.

Aydın, F., & Arslan, Z. (2017). Yapay öğrenme yöntemleri ve dalgacık dönüşümü kullanılarak nörodejeneratif hastalıkların teşhisi. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 32*(3), 745–754.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. İçinde *Proceedings of the 5th Annual Future Business Technology Conference* (ss. 5–12).

Elmas, Ç. (2003). *Yapay sinir ağları teori, mimari, eğitim, uygulama* (1. baskı). Seçkin Yayıncılık.

Gorr, W. L., Nagin, D., & Szczypula, J. (1994). Comparative study of artificial neural network and statistical models for predicting student grade point averages. *International Journal of Forecasting, 10*(1), 17–34.

Güre, Ö. B., Kayri, M., & Erdoğan, F. (2020). PISA 2015 matematik okuryazarlığını etkileyen faktörlerin eğitsel veri madenciliği ile çözümlenmesi. *Eğitim Bilim, 45*, 251–270.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*(1), 63–90.

Kurnaz, G., & Murat, N. (2023). Determination of harness production time and defective product formation risk factors with artificial neural network. *Black Sea Journal of Engineering and Science, 6*(4), 325–329. https://doi.org/10.34248/bsengineering.1296187

Manikandan, G., Aravind, V., & Anitha, B. (2018). A survey to identify an efficient classification algorithm for heart disease prediction. *International Journal of Pure and Applied Mathematics, 119*(2), 13337–13345.

Osborn, J., Francisco Javier De C. J., Guzman, D., Butterley, T., Myers, R., Guesalaga, A., & Laine, J. (2011). Using artificial neural networks for open-loop tomography. *Optics Express, 20*(3), 2420–2432. https://doi.org/10.1364/OE.20.002420

Öztemel, E. (2003). *Yapay sinir ağları*. Papatya Yayıncılık.

SubbaNarasimha, P. N., Arinze, B., & Anandarajan, M. (2000). The predictive accuracy of artificial neural network and multiple regression in the case of skewed data: Exploration of some issues. *Expert Systems with Applications, 19*(2), 117–123.

Tosun, S. (2007). *Artificial neural networks, decision tree comparison in classification analysis: An application on students' success* [Yayımlanmamış yüksek lisans tezi]. İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü.

Uzun, Y. (2005). *Machine learning algorithms and learning rules with fuzzy logic on medical data* [Yayımlanmamış yüksek lisans tezi]. Selçuk Üniversitesi Fen Bilimleri Enstitüsü.

Black Sea Journal of Statistics
Doi: 10.5281/zenodo.18234538

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 18-21 / June 2025

DETERMINING THE RELATIONS OF DAILY LIVE WEIGHT GAIN OF SAANEN KIDS USING CONCORDANCE CORRELATION

Burcu KURNAZ[1]*

[1]*Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun Türkiye*

Abstract: In this study, the Concordance Correlation Coefficient (CCC) method was used to evaluate the consistency between the daily live weight gain (DLWG) data obtained in the six-month period from birth. CCC is a powerful analysis tool in terms of determining both the strength of the relationship between repeated measurements and how close the measurements are to each other. As a result of the analysis, it was seen that the CCC values were low especially between the first month and the following months, but these values increased significantly from the third month onwards. The findings obtained revealed that the CCC method was effective in evaluating the consistency of the weight gains that changed over time. Therefore, CCC can be used as a reliable statistical tool in the analysis of growth dynamics in animal science studies.

Keywords: Concordance correlation, Weight, Saanen, Relation, Autocorrelation

***Corresponding author:** Ondokuz Mayıs University, Faculty of Agriculture, Department of Animal Science, 55139, Samsun Türkiye
E mail: burcu2039@hotmail.com (B. KURNAZ)
Burcu KURNAZ https://orcid.org/0000-0001-5613-6992

Received: April 23, 2025
Accepted: May 27, 2025
Published: June 15, 2025

Cite as: Kurnaz, B. (2025). Determining the relations of daily live weight gain of Saanen kids using concordance correlation. *Black Sea Journal of Statistics, 1*(1), 18–21.

1. Introduction

Various correlation analyses are used to evaluate the relationships between variables on data obtained in the field of animal science (Kurdal and Önder, 2020). Correlation is a statistical method that measures how and in which direction two variables are related to each other. This analysis helps to understand how an increase or decrease in one variable causes a change in the other variable. However, the method to be used in correlation analysis depends on the structure of the data and the provision of certain statistical assumptions. For example; while the Pearson correlation coefficient is preferred in the presence of continuous and normally distributed variables and independent observations; the Spearman correlation coefficient can be applied to ordinal or continuous/discrete data where at least one of the variables is not normally distributed (Kurdal and Önder, 2020).

The concordance correlation coefficient can potentially be an excellent tool in many types of goodness-of-fit evaluations, by simply examining how well the observed outcomes concord with the hypothesized values (Lawrance and Lin, 1992). In general, these measures do not address both precision and accuracy as does the concordance correlation coefficient; however, the equivalency and similarities of the intraclass correlation coefficient to the concordance correlation coefficient under certain scenarios has been discussed. The CCC measures how far the fitted linear relationship of two variables deviates from the concordance line (accuracy) and how far each observation deviates from the fitted line (precision) (Crawford et al., 2007).

Concordance Correlation Coefficient (CCC) allows the evaluation of the level of relationship and equivalence together, especially in repeated measurements, by taking into account not only the directional relationship between variables but also the degree to which the measurements are close to each other and of similar magnitude. The dataset used in this study was analyzed with the Concordance Correlation Coefficient (CCC) method using the R software *epiR* package in order to evaluate the level of agreement between repeated measurements obtained from the same individuals over a period of six months.

2. Materials and Methods

The continuous data in this study was the data set consisting of daily live weight gain from birth to the 6th month of life of 75 Saanen kids used in the study conducted by Önder and Abacı (2015). Analyses were performed using the R software version 4.4.3 (R Core Team, 2025). The codes used in the analysis was given:

```
install.packages("readxl")
install.packages("epiR")
library(readxl)
library(epiR)
file.choose()
data <-
read_excel("C:/Users/Username/Desktop/data.xlsx")
ccc1 <- epi.ccc(data$variable1, data$variable2)
```


This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

```
print(ccc1)
```

2.1. Concordance Correlation Coefficient

The CCC measures agreement between two methods or time points by measuring the variation of their linear relationship from the 45° line through the origin. Therefore, this coefficient is not only measuring how far each observation deviates from the line fit to the data (precision), but also how far this line deviates from the 45° line through the origin (accuracy).

Lin characterizes the degree of concordance between two variables X and Y by the expected value of the squared difference, and defines the CCC as given in Equation 1:

$$\rho_c = 1 - \frac{E[(X-Y)^2]}{E_{indep}[(X-Y)^2]} = \frac{2\sigma_{XY}}{\sigma_{XX} + \sigma_{YY} + (\mu_X - \mu_Y)^2} \quad (1)$$

where μ_X = E(X), μ_Y = E(Y), σ_{XX} = var(X), σ_{YY} = var(Y), and σ_{XY}= cov(X, Y). This coefficient is related to the Pearson correlation coefficient ρ in that when $\mu_X = \mu_Y$ and $\sigma_{XX} = \sigma_{YY}$, then $\rho_c = \rho$ (Lawrance and Lin, 1989). Where μ_X and σ_{XX} represent the mean and variance for the first rater, μ_Y and σ_{YY} represent the mean and variance for the second rater, and σ_{XX} is the covariance for the first and second rater (Barnhart et al., 2002).

3. Results and Discussion

According to the results obtained from the analysis, the change in the average live weight values of individuals for six months from birth is shown. A regular and nearly linear increase is observed in the weight values from birth. There is a clear increasing trend between the monthly measurements, and it is seen that there is approximately a similar amount of live weight gain each month. In addition, the error bars (standard error) of the data points show a slight widening over time, which shows that the variation between individuals increases slightly in the following months. In general, the graph reveals that growth progresses continuously and steadily (Figure 1).

According to Figure 2, the daily live weight gain (DLWG) values calculated at the end of each month from birth are shown. The daily weight gain was at its highest level in the first month, and a gradual decrease was observed in the following months. The slowdown in weight gain became apparent especially from the 3rd month onwards. This decrease coincides with the period when the kids are weaned and reflects the effect of weaning on growth performance. Although the daily gain values follow a more stable course after the 4th month, it is seen that the average gain level is lower compared to the previous months from the 3rd month onwards. The error bars show that this decreasing trend can be evaluated together with the variation arising from the differences between individuals.

According to the results obtained, it provides important clues about how the growth process changes over time and which periods are more similar to each other (Table 1). It is seen that the CCC values especially between M1-B (first month - birth) and the following months are quite low. The coefficient of concordance between M1-B and M6-B is only 0.178, which shows that the weight gain in the first month after birth does not establish a consistent and predictable relationship with the other periods up to the sixth month. Similarly, it was determined that the values between M1-B and M4-B (0.288) and M1-B and M5-B (0.223) are also low. However, the values after M2-B, especially between M3-B and M4-B (0.902), M4-B and M5-B (0.962) and M5-B and M6-B (0.960), were determined to be quite high. This finding shows that following weaning, weight gain values reach a more stable, predictable and close structure.

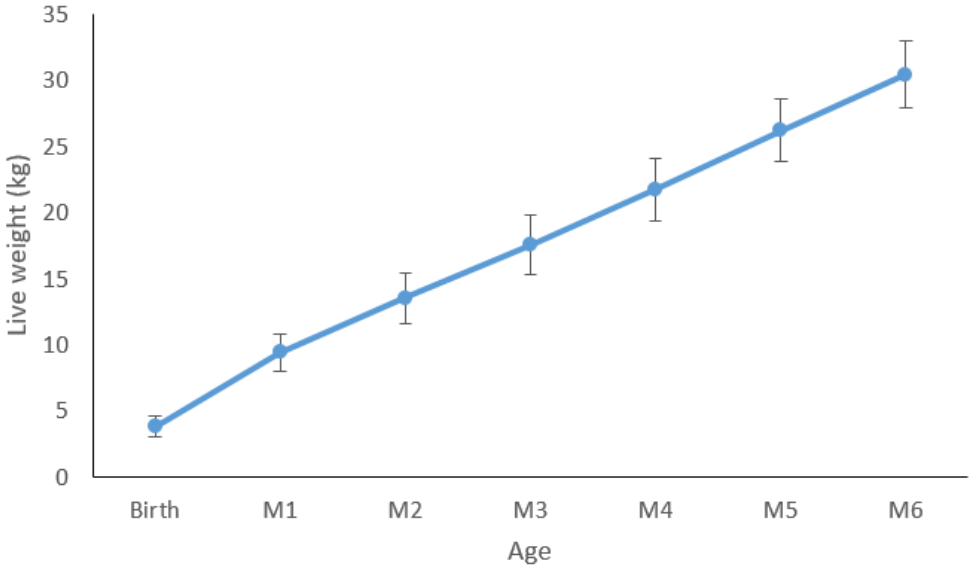


Figure 1. Live weight according to the age. M1= 1st month, M2= 2nd month, M3= 3rd month, M4= 4th month, M5= 5th month, M6= 6th month.

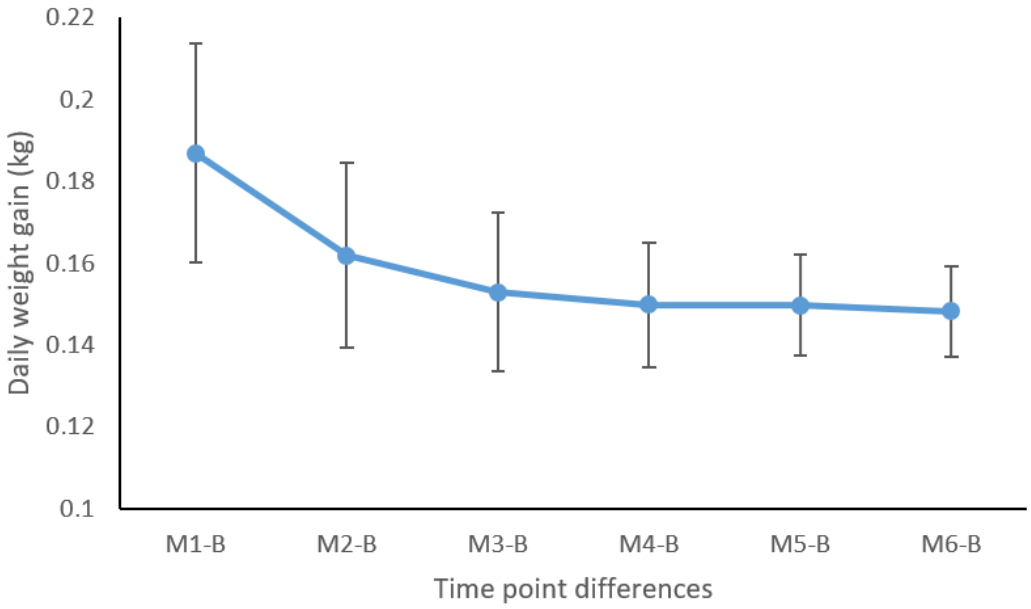


Figure 2. Absolute daily weight gain. M1-B= from birth to the 1st month, M2-B= from birth to the 2nd month, M3-B= from birth to the 3rd month, M4-B= from birth to the 4th month, M5-B= from birth to the 5th month, M6-B= from birth to the 6th month.

Table 1. Concordance Correlation Coefficients between time points for absolute daily weight gain

	M2-B	M3-B	M4-B	M5-B	M6-B
M1-B	0.610	0.420	0.287	0.223	0.178
M2-B		0.883	0.717	0.600	0.493
M3-B			0.902	0.792	0.676
M4-B				0.962	0.892
M5-B					0.960

According to the obtained table, the early period (first 1-2 months) shows a process in which individual differences are more dominant; however, after the 3rd month, environmental conditions stabilize and growth begins to follow a more standard course.

4. Conclusion

In this study, the Concordance Correlation Coefficient (CCC) method was used to evaluate the agreement between 6-month daily live weight gain data obtained from the same individuals. CCC is a powerful method in that it shows not only the relationship between measurements but also how close they are to each other.

The analysis results showed that the compliance was low, especially between the first month and the following months; however, this compliance increased significantly from the third month onwards. This situation shows that the compliance between the monthly live weight gain values was low in the first months of the growth process, but over time these values became closer to each other and the development began to follow a more stable course.

CCC is therefore quite useful for assessing the consistency of growth data over time in animal science studies. It provides the opportunity to evaluate both the accuracy and precision of the live weight gains recorded by individuals over certain time periods, especially in cases where repeated measurements are taken. In addition, the fact that CCC evaluates both the directional relationship (correlation) and the closeness of values to each other (concordance) makes it a more comprehensive method than traditional correlation analyses. In future studies, more comprehensive analyses can be performed by separating the data set by gender, working with larger sample groups, and integrating the effects of different genotypes or environmental conditions into the model. This method is recommended as an effective analysis tool for comparing developmental periods in similar longitudinal studies to be conducted in the future.

Author Contributions

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

	B.K.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	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, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declare that there is no conflict of interest.

Ethical Consideration

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

References

Barnhart, H. X., Haber, M., & Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics, 58*(4), 1020–1027.

Crawford, S. B., Kosinski, A. S., Lin, H. M., Williamson, J. M., & Barnhart, H. X. (2007). Computer programs for the concordance correlation coefficient. *Computer Methods and Programs in Biomedicine, 88*(1), 62–74.

Kurdal, N., & Önder, H. (2020). Estimating the nonparametric confidence interval for correlation coefficient on animal data. İçinde *The IV. International Congress on Domestic Animal Breeding, Genetics and Husbandry Bildiri Kitabı*. (ss. 97).

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*(1), 255–268.

Lawrence, I., & Lin, K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics, 48*(2), 599–604.

Önder, H., & Abacı, H. S. (2015). Path analysis for body measurements on body weight of Saanen kids. *Kafkas Üniversitesi Veteriner Fakültesi Dergisi, 21*(3), 351–354. https://doi.org/10.9775/kvfd.2014.12500

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Black Sea Journal of Statistics

Doi: 10.5281/zenodo.18234609

Open Access Journal
e-ISSN: 3108-4265

Research Article
Volume 1 - Issue 1: 22-26 / June 2025

BINARY LOGISTIC REGRESSION PROCEDURE WITH AN APPLICATION

Mustafa ŞAHİN[1]*

[1]*Kahramanmaraş Sütçü İmam University, Faculty of Agriculture, Department of Agricultural Biotechnology, 46100, Kahramanmaraş, Türkiye*

Abstract: Binary logistic regression is widely used in health, social, and science sciences, particularly in studies involving extensive categorical independent variables. Considering that binary logistic regression calculates the probability of a specific event occurring or not occurring relative to the opposite event using one or more independent variables, its widespread use becomes even clearer. Another reason for its widespread use is that regression equations based on the least squares method cannot be used in data sets with numerous categorical variables affecting a two-level dependent variable. Another advantage is its ability to make future predictions and identify risk factors and impact magnitudes that influence the occurrence of events. However, in practice, many errors can occur, especially during model construction. Determining the individual effects of variables, joint effects, interaction effects, and how independent variables enter the model are highly sensitive processes. This study will examine the operations performed on independent variables in binary logistic regression until the final model is formed, using a numerical example. Thus, an important resource will be created for researchers to obtain the correct models with the correct process flow in binary logistic regression.

Keywords: Logistic regression, Procedure, Independent variable, Binary, Discrete data

***Corresponding author:** Kahramanmaraş Sütçü İmam University, Faculty of Agriculture, Department of Agricultural Biotechnology, 46100, Kahramanmaraş, Türkiye
E mail: ms66@ksu.edu.tr (M. ŞAHİN)
Mustafa ŞAHİN https://orcid.org/0000-0003-3622-4543

Received: March 28, 2025
Accepted: May 06, 2025
Published: June 15, 2025

Cite as: Şahin, M. (2025). Binary logistic regression procedure with an application. *Black Sea Journal of Statistics, 1*(1), 22–26.

1. Introduction

Binary logistic regression (BNR) is a statistical modeling method used when the dependent variable consists of two categories. This model analyzes the relationship between one or more independent variables and the dependent variable to estimate the probabilities of observed outcomes. The aim of binary logistic regression is to estimate the probability that an observation belongs to a particular class. In this respect, the logistic regression model stands out as an effective tool for supporting decision-making processes and making probability-based predictions in a wide variety of disciplines. For example, in the field of healthcare (Bircan, 2004; Dasgupta and De, 2007; Şahin and Efe, 2018) (presence-absence of disease), education (Eratlı Şirin and Şahin, 2020) (success-failure), marketing (Maharani, 2021) (whether the customer will purchase the product), and animal husbandry (Adekunle, 2021; Gök and Tolun, 2023). Production being below or above a certain limit) has found a wide range of use. On the other hand, it is also used in many fields such as credit risk analysis, evaluation of insurance applications (Yin et al., 2020, Bakrie et al., 2023), estimation of crime rates, meteorology (Ruiz and Villa, 2008; Bertolin and De Santis, 2022), attitude and behavior analysis in social sciences (Kariyam, 2020; Okeke et al., 2020).

Above all, the effectiveness of binary logistic regression analysis, which has found such widespread application, in relevant fields depends on the analysis being conducted using the correct statistical steps. For example, how independent variables with more than two discrete levels are entered into the model, or how these variables are transformed into design variables, is crucial. Similarly, whether continuous variables are entered into the model as continuous variables or transformed into discrete variables, or at what stage are interacting variables introduced into the model and how are their significance verified, is crucial. Consequently, the selection of variables and their evaluation are crucial.

The process flow in binary logistic regression consists of various steps, from data preparation to model interpretation. First, the dependent variable is checked for binary nature. Dependent variables that are not binary are converted to binary, taking into account the researcher's experience or limitations outlined in the literature. Furthermore, the model is constructed after the appropriateness of the independent variables is assessed. Following model construction, regression coefficients are interpreted, significance tests are conducted, and the overall validity of the model is assessed using various statistical criteria.

In this study, after explaining the basic concepts of the binary logistic regression model, the step-by-step process flow will be detailed on a sample data set.

This work is licensed (CC BY-NC 4.0) under Creative Commons Attribution 4.0 International License

2. Materials and Methods

2.1. Materials

This study examined the stages of evaluating dependent variables in binary logistic regression, creating the model, and interpreting it using a numerical example. For this purpose, calf birth weight, live weight, birth type, calf sex, calf season, and feeding method were considered for 87 individuals of Jersey and Holstein dairy cattle. Calf birth weight (BW) was the dependent variable, while cattle breed (CB), live weight (LW), lactation order (LO), birth type (BT), calf gender (CG), birth season (BS), and feeding method (FM) were considered as independent variables. The coding of the variables is provided in Table 1.

2.2. Methods

Simple logistic regression analysis of candidate variables: In this initial stage of logistic regression, simple logistic regression analysis is performed on each variable individually. Three different tests can be used to determine whether variables are included in the model, i.e., to test the significance of the coefficients: the likelihood ratio test (G-test), the Wald test, and the score test. The G test will be used in this study. If the contribution of a candidate variable to the model is found to be insignificant by the G test, that variable is excluded from the model. The G statistic is in the form of given in Equation 1.

Table 1. Variables and coding types

	Variables							
	BW	LW	CB	LO	BT	CG	BS	FM
Coding type	0 if 30<	kg	Holstein=0	1,2,3,4	Normal=0	Female=0	Spring=0	Free=0
	1 if 30≥		Jersey=1		Other=1	Male=1	Summer=1	Controlled=1
							Autumn=2	
							Winter=3	

$$G = 2\left\{\sum_{i=1}^{n}[y_i \log(\hat{P}_i) + (1-y_i)\log(1-\hat{P}_i)] - [n_1 \log(n_1) + n_0 \log(n_0) - n\log(n)]\right\} \quad (1)$$

($\hat{P}_i = \hat{P}(x_i)$, $n_1=\Sigma y_i$ and $n_0= \Sigma(1-y_i)$)

$$\ell(\mathbf{x_i}) = \beta_0 + \sum_{k=1}^{p}\beta_k x_{ik} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots.. + \beta_p x_{ip} \quad (2)$$

Continuous variables are directly entered into the model as continuous variables when testing their individual contributions. On the other hand, variables with a level number of r>2 should be defined as design variables when checking their individual contributions. In simple logistic regression, all candidate variables whose contribution is found to be insignificant are excluded from the model, while all variables whose contribution is found to be significant are included in the multiple analysis.

2.2. Multiple Analysis

Discrete and continuous variables whose individual contributions are found to be significant are subjected to multiple analysis. In multiple analysis, the contribution of variables is again checked with the G test. A variable whose contribution is significant in simple logistic regression may be found to be insignificant in multiple analysis. In this case, the variable will be excluded from the model. A logit model for p independent variables, $x=(x_1, x_2, \ldots, x_p)$, can be written as given in Equation 2:

2.3. Deciding on How to Enter Continuous Variables into the Model

At this stage, it is necessary to decide how to enter the continuous variable whose contribution is found to be significant in simple logistic regression and multiple regression into the model. For this purpose, the continuous variable is divided into quartile groups. The first quartile group is taken as the reference group, and the design variables are defined. The continuous variable is then removed from the model, and multiple logistic regression analysis is performed using the defined design variable instead. The odds ratios for the estimated coefficients for the design variables are examined. If the odds ratios show a linear increase or decrease, the relevant independent variable is concluded to be linear with the logit method, and the variable is entered into the model as a continuous variable. If there is no linear increase or decrease, the variable is determined to be non-linear with the logit method, and the continuous variable remains in the model as a design variable. At this point, depending on the researcher's preference, the continuous variable can be made discrete and included in the model instead of the design variable. For this purpose, the continuous variable can be made discrete using a biologically significant or critical threshold. The contribution of the discrete variable to the model must be checked. Continuity checks are performed for all continuous variables in the model.

2.4. Creating the Main Effects Model

A main effects model is created using discrete variables whose contribution is deemed significant and continuous variables whose continuity is checked and their inclusion in the model is determined. Multiple logistic regression analysis is then performed by individually adding

interaction terms deemed likely or significant to the main effects model. The contribution of the added interaction terms to the main effects model is again tested using the G test. The interaction terms deemed significant are also added to the model to create the final model.

2.5. Interpreting Coefficients in Logistic Regression

In logistic regression, "odds" and "odds ratio" are used to interpret coefficients. The probability that y = 1 for any group of X (X = 0) is called "odds." "odds" is the logit without the natural logarithm, and for example, the odds values for X = 1 and X = 0 can be expressed as given in Equations 3 and 4, respectively.

$$\frac{P(x=1)}{1-P(x=1)} \tag{3}$$

$$\frac{P(x=0)}{1-P(x=0)} \tag{4}$$

The odds ratio will be denoted by Ω. The odds ratio is the ratio of the odds value calculated for x=1 to the odds value calculated for x=0. Therefore, the odds ratio can be written as given in Equation 5.

$$\Omega(1,0)=\frac{P(1)/[1-P(1)]}{P(0)/[1-P(0)]} \tag{5}$$

Accordingly, if the independent variable in logistic regression is binary and coded as 0.1, the odds ratio can be expressed as $\Omega = e^{\beta_1}$.

2.6. Determining the Goodness of Fit of the Model

The C statistic can be used to test the goodness of fit of the main effects model. From the main effects model, the logit estimate ($\hat{\ell}(x_i)$) is obtained for each subject, respectively. Then, using the $P(\mathbf{x_i})=\frac{e^{\ell(\mathbf{x_i})}}{1+e^{\ell(\mathbf{x_i})}}$ equation, the P($\mathbf{x_i}$) and 1- P($\mathbf{x_i}$) values are obtained for each subject, respectively. They are sorted from smallest to largest according to P($\mathbf{x_i}$). Ten-fold risk groups T are then created, and their observed and expected frequencies are obtained. After obtaining the observed and expected frequencies for each ten-fold risk group, crosstabs are obtained, with the independent variable in the row and the ten-fold risk groups in the column.

The $\hat{\boldsymbol{C}}$ test statistic is then used to test the goodness of fit using expected and observed frequencies. The $\hat{\boldsymbol{C}}$ test statistic, which shows the χ^2 distribution with t-2 degrees of freedom, is as given in Equation 6.

$$\hat{C}=\sum_{m=1}^{t}\left[\frac{(g_{1m}-b_{1m})^2}{b_{1m}}+\frac{(g_{0m}-b_{0m})^2}{b_{0m}}\right] \tag{6}$$

3. Results and Discussion

The results of simple logistic regression analysis of variables thought to be related to birth weight such as cattle breed (CB), live weight (LW), lactation order (LO), birth type (BT), calf gender (CG), birth season (BS) and feeding method (FM) are given in Table 2.

The lactation order variable with more than two levels was defined as a design variable (DLO) in the model (1st and 2nd lactation=0, 2nd and 3rd lactation=1) and included in the main effects model together with breed, live weight and sex, which were found to be significant.

As shown in Table 3, the live weight variable, which is included as a continuous variable in the main effects model, should be checked to see if it can be included as a continuous variable. For this purpose, the live weight variable was converted into a design variable using quartile partitioning (LW1, LW2, LW3), and the results are presented in Table 4.

Table 2. Simple logistic regression analysis of variables thought to be associated with birth weight

Parameter	$\hat{\beta}$	$SE(\widehat{\beta})$	$\widehat{\Omega}$	95% Confidence Limits	Chi-Square	-2 Log L	G	P
Intercept	1.209	0.254	-	-	22.511	93.810	-	-
CB	3.337	1.055	28.148	355-222.90	9.993	71.774	22.036	0.0016*
LW	0.0185	0.00493	1.019	1.009-1.029	14.017	64.50	29.31	0.0002*
LO	1534	0.390	4.639	2.158-9.972	15.441	68.196	25.614	0.0001*
BT	-0.518	0.513	0.595	0.218-1.519	1.205	92.792	1.018	0.312
CG	2.609	0.682	14.316	3.758-54.53	21.195	72.615	21.195	0.0001*
BS	-0.087	0.224	0.916	0.59-1.424	0.150	93.659	0.151	0.6977
FM	0.889	0.529	2.433	0.862-6.872	2.937	90.873	2.937	0.1093

* Statistically significant at α=0.10 error level.

Table 3. Multiple logistic regression analysis results of the main effects model

Parameter	$\hat{\beta}$	$SE(\widehat{\beta})$	Chi-Square	P
Intercept	-7.2223	2.6487	7.4351	0.0064*
CB	2.0715	0.8335	6.1771	0.0129*
LW	0.0131	0.00561	5.4769	0.0193*
DLO	-2.8991	1.4795	3.8395	0.0501*
CG	4.8539	1.6057	9.1382	0.0025*

* Statistically significant at α=0.10 error level.

Table 4. Multiple logistic regression analysis for the live weight variable divided into quartiles and transformed into a design variable

Parameter	$\hat{\beta}$	$SE(\widehat{\beta})$	Chi-Square	P
Intercept	-2.2379	0.9951	5.06	0.0245
CB	1.7861	0.8751	4.17	0.0413
LW1	1.5166	1.1436	1.76	0.1848
LW2	1.5615	1.1069	1.99	0.1583
LW3	3.6140	1.4186	6.49	0.0108*
DLO	-2.7252	1.5571	3.06	0.0801*
CG	4.8828	1.6510	8.75	0.0031*

* Statistically significant at α=0.10 error level.

An examination of Table 4 reveals that the odds ratios for live weight, as a result of the quartile analysis, are lowest in Group 2, higher in Group 3, and highest in Group 4. Since the estimated coefficients do not follow an increasing trend (4.55, 4.76, 37.11), it is concluded that the variable is linear with the logit. Therefore, the live weight variable will be entered into the model as a continuous variable.

After obtaining the main effects model, interactions between the variables in the model are examined. In this step, the contribution of each interaction term to the main effects model is examined. Therefore, the -2 log likelihood value, likelihood ratio test statistic (G), degrees of freedom, and P-value for each possible interaction term are presented in Table 5.

Table 5. Values obtained by adding ınteraction terms one by one to the main effects model

Parameter	-2 Log L	DF	G	P
Main Effects Model	43.80			
CB*LW	42.02	1	1.78	0.256
CB*DLO	41.13	1	2.67	0.951
CB*CG	41.31	1	2.49	0.953
LW*DLO	41.61	1	2.19	0.174
LW*CG	43.41	1	0.39	0.551
DLO*CG	43.73	1	0.07	0.966

* Statistically significant at α=0.10 error level.

Table 6. Observed and expected frequencies for ten risk groups

	1	2	3	...	9	10	∑
y=1	0	1	1	…	2	1	18
	0.114	0.783	2.615	…	2.921	1.048	18
y=0	6	5	7	…	7	5	69
	5.915	4.745	6.135	…	5.987	4.589	69

An examination of Table 5 reveals that all potential interaction effects added to the main effects model are statistically insignificant (P>0.10). At this stage, the final model can be tested for goodness of fit. Some of the observed and expected frequencies of the resulting ten-risk groups are presented in Table 6. Using these values, the C statistic can be calculated for the goodness of fit test of the main effects model.

From Table 6, the Y value can be calculated as given in Equation 7.

$$\hat{C} = \sum_{m=1}^{t}\left[\frac{(g_{1m}-b_{1m})^2}{b_{1m}}+\frac{(g_{0m}-b_{0m})^2}{b_{0m}}\right]=5.253 \qquad (7)$$

Here, since $\hat{C} = 5{,}253 < X^2_{8,0.05} = 15.507$, it can be said that the final model fits the data well.

As a result, the logistic regression model estimate for any subject from the main effects model can be written as given in Equation 8.

$$\hat{\ell}(CB, LW, DLO, CG) = \hat{\beta}_0 + \hat{\beta}_1(CB) + \hat{\beta}_2(LW) + \hat{\beta}_3(DLO) + \hat{\beta}_4(CG)= -7.2223+2.0715(CB)+0.0131(LW)-2.8991(DLO)+4.8539(CG) \qquad (8)$$

The main effects model does not include birth type, calving season, and feeding method variables, and therefore, these variables have no statistically significant effect on calf birth weight. The main effects model shows that cattle breed (CB), live weight (LW), lactation order (DLO), and calf gender (CG) have an effect on calf birth weight. The odds ratios for these effects are as follows; for cattle breed (CB); $\widehat{\Omega}$ =exp(2.0715) =7.936, live weight (LW); $\widehat{\Omega}$ =exp(0.0131) =1.031, for lactation order (DLO); $\widehat{\Omega}$ =exp(-2.8991) =0.055, and for calf sex (CG); $\widehat{\Omega}$ =exp(4.8539) =128.22. In this case, it can be said that calf

gender is the most effective variable, while lactation order has the least effect.

4. Conclusion

In binary logistic regression, variable selection and how they are included in the model play a critical role in the model's accuracy, interpretability, and generalizability. Including unnecessary or irrelevant variables can lead to problems such as multicollinearity and overfitting, thereby reducing model performance.

In this study, the linearity of the live weight variable, as a continuous independent variable, was tested using the logit method. It was concluded that it was linear using the logit method, and it was decided to include it as a continuous variable in the model. Furthermore, the lactation order variable, which has more than two levels, was also defined as a design variable and included in the model. The study demonstrated that without checking continuity for the continuous variable and defining a design variable for the independent variable with more than two levels, very different coefficient estimates, and therefore different odds ratios, would have been obtained.

In conclusion, it can be said that in binary logistic regression analysis, determining the factors affecting the dependent variable with a statistically sound process and the method of evaluating the variables are extremely important.

Author Contributions

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

	M.Ş.
C	100
D	100
S	100
DCP	100
DAI	100
L	100
W	100
CR	100
SR	100
PM	100
FA	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, PM= project management, FA= funding acquisition.

Conflict of Interest

The author declare that there is no conflict of interest.

Ethical Consideration

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

References

Adekunle, F. O. (2021). A binary logistic regression model for prediction of feed conversion ratio of *Clarias gariepinus* from feed composition data. *Marine Science and Technology Bulletin, 10*(2), 134–141.

Bakrie, P. D., Aulia, R. N., & Taufiqillah, R. (2023). Customer churn prediction for life insurance using binary logistic regression. *Economics Review and Journal, 3*(3), 45–58.

Bertolin, C., & De Santis, F. (2022). Prediction of snowmelt days using binary logistic regression in the Umbria-Marche Apennines (Central Italy). *Water, 14*(9), 1495.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2*(1), 185–208.

Dasgupta, S., & De, U. K. (2007). Binary logistic regression models for short term prediction of premonsoon convective developments over Kolkata (India). *International Journal of Climatology, 27*(6), 831–836.

Eratlı Şirin, Y., & Şahin, M. (2020). Investigation of factors affecting the achievement of university students with logistic regression analysis: School of Physical Education and Sport example. *SAGE Open, 10*(3), 215–225.

Gök, İ., & Tolun, T. (2023). A study on body weight and carcass characteristics and sex in broilers. *Türk Doğa ve Fen Dergisi, 12*(4), 129–133.

Kariyam, F. H. (2020). Application of binary logistic regression in modelling women's participation in improving the welfare of fishermen families. İçinde *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)* (ss. 179–185).

Maharani, O. C. (2021). A binary logistic regression analysis of factors influencing repeat purchase for snack product X. *Jurnal Teknik Industri Terintegrasi, 7*(4), 359–366.

Okeke, J. U., Okeke, E. N., & Dakhin, Y. V. (2020). Binary logistic models of home ownership in Wukari Nigeria. *Open Journal of Statistics, 10*(1), 50–59.

Ruiz, A., & Villa, N. (2008). *Storms prediction: Logistic regression vs random forest for unbalanced data* (arXiv Preprint, arXiv: 0804.0650). Cornell University.

Şahin, M., & Efe, E. (2018). Determining the factors affecting birth weight by using logistic regression method. *Black Sea Journal of Health Science, 1*(2), 22–27.

Yin, S., Dey, D. K., Valdez, E. A., Gan, G., & Vadiveloo, J. (2020). *Skewed link regression models for imbalanced binary response with applications to life insurance* (arXiv Preprint, arXiv: 2007.15172). Cornell University.