T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ZOOTEKNİ ANABİLİM DALI

FARKLI VERİ YAPILARINDA UZAKLIK TEMELLİ REGRESYON MODELLERİNİN İNCELENMESİ

Yüksek Lisans Tezi

Burcu KURNAZ

Danışman

Prof. Dr. Hasan ÖNDER

SAMSUN
2023

TEZ KABUL VE ONAYI

Burcu KURNAZ tarafından, **Prof. Dr. Hasan ÖNDER** danışmanlığında hazırlanan **"FARKLI VERİ YAPILARINDA UZAKLIK TEMELLİ REGRESYON MODELLERİNİN İNCELENMESİ"** başlıklı bu çalışma, jürimiz tarafından 8.8.2023 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı **Üniversitesi** **Ana Bilim/Ana Sanat Dalı**	**Sonuç**
Başkan	Prof. Dr. Hasan ÖNDER Ondokuz Mayıs Üniversitesi Zootekni Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Doç. Dr. Hüseyin ERDEM Ondokuz Mayıs Üniversitesi Zootekni Ana Bilim Dalı	☒ Kabul ☐ Ret
Üye	Dr. Öğr. Üyesi Esra YAVUZ Şırnak Üniversitesi Muhasebe ve Vergi Bölümü	☒ Kabul ☐ Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

Prof. Dr. Ahmet TABAK
Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Yüksek Lisans tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

Etik Kurul Gerekli mi?

Evet ☐

Hayır ☒

16/06/2023
Burcu KURNAZ

TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı: FARKLI VERİ YAPILARINDA UZAKLIK TEMELLİ REGRESYON MODELLERİNİN İNCELENMESİ

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 19.06.2023 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 22

Tek kaynak oranı : % 4 çıkmıştır.

19/06/2023
Prof. Dr. Hasan ÖNDER

ÖZET

FARKLI VERİ YAPILARINDA UZAKLIK TEMELLİ REGRESYON MODELLERİNİN İNCELENMESİ

Burcu KURNAZ
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
Zootekni Anabilim Dalı
Yüksek Lisans, Ağustos/2023
Danışman: Prof. Dr. Hasan ÖNDER

Basit ve çoklu doğrusal regresyon analizi sonucunda elde edilecek olan regresyon modeline ait parametre kestirimlerinin güvenilir olabilmesi için modelle ilgili bazı varsayımların sağlanabilmesi gereklidir. Parametre tahmin yöntemlerinde varsayımların sağlanamadığı durumlar için çözüm olarak geliştirilen az sayıdaki modellerden biri Uzaklık Temelli Regresyon yöntemleridir. Bu yöntemlerin amacı kategorik veya gerçek değerli ve kategorik açıklayıcı değişkenlerin bir karışımı dahil olmak üzere, ölçüm değer tahmin edicileri ile problemleri doğru bir şekilde ele almaktır. Uzaklık temelli regresyon, karışık tipte açıklayıcı değişkenler kullanıldığında doğrusal regresyon modellerinde parametre tahmini için alternatif bir yöntemdir. Uzaklık temelli regresyon klasik doğrusal regresyona benzer, ancak açıklayıcı değişkenler ham değerler yerine uzaklık ölçülerine göre ölçülmektedir. Bu çalışmada, Euclidean, Gower ve Manhattan uzaklık ölçülerinin Binom, Normal, t, Ki-Kare ve Poisson dağılışlarına ait üretilmiş, örnek büyüklükleri 10, 25, 50, 100, 250 ve 500 olan veri setleri ve kesikli ve sürekli dağılış gösteren gerçek veri setleri (10, 50 ve 100 örnek büyüklüğünde) üzerinde etkisi ile Doğrusal Regresyon yönteminden elde edilen sonuçlara göre karşılaştırma yaparak belirlenmesi amaçlanmıştır. Analizi gerçekleştirmek için R paketi olan "dbstats", "cluster" ve “tidyverse” kullanılmıştır. Sonuç olarak, Poisson dağılışına sahip verilerde özellikle küçük örnek büyüklüklerinde ($n<50$) Manhattan uzaklığının kullanılmasının başarısız sonuçlar üretebileceği belirlenmiştir. Örnek büyüklüklerine göre farklı dağılışlar içerisinde Gower ve Euclidean uzaklıkları arasında kayda değer farklılık olmamasına rağmen bazı dağılışlarda Euclidean uzaklık ölçüsü kullanımının dalgalanmaya sebep olan sonuçlar ürettiği belirlenmiştir. Ancak, Gower uzaklığı daha sabit bir yapıya sahip olması nedeniyle daha uygun bir seçim olarak önerilebileceği anlaşılmıştır. En Küçük Kareler tahmin yönteminin uygulanabilirliği için bu çalışmada da bahsedilen gerekli olan varsayımların sağlanamadığı durumlarda Uzaklık Temelli Regresyon yöntemlerinin kullanılması önerilebilir.

Anahtar Sözcükler: Uzaklık ölçüleri, Regresyon, dbstats paketi, R yazılımı

ABSTRACT

EXAMINATION OF DISTANCE BASED REGRESSION METHODS FOR DIFFERENT DATA STRUCTURES

Burcu KURNAZ
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Animal Science
Master, August/2023
Supervisor: Prof. Dr. Hasan ÖNDER

In order to the parameter estimations of the regression model to be obtained as a result of simple and multiple linear regression analysis to be reliable, some assumptions about the model must be provided. One of the few models developed as a solution for situations where assumptions cannot be provide in parameter estimation methods is Distance Based Regression methods. The purpose of these methods is to properly address problems with measure value estimators, including categorical or a mix of real-valued and categorical explanatory variables. Distance-based regression is an alternative method for parameter estimation in linear regression models when mixed-type explanatory variables are used. Distance-based regression is similar to classical linear regression, except that explanatory variables are measured by distance measures rather than raw values. In this study, datasets with sample sizes of 10, 25, 50, 100, 250 and 500 produced for Binomial, Normal, t, Chi-square and Poisson distributions of Euclidean, Gower and Manhattan distance measures and real data with discrete and continuous distribution. It was aimed to determine the effect on the data sets (10, 50 and 100 sample sizes) by comparing the results obtained from the Linear Regression method. R packages "dbstats", "cluster" and "tidyverse" were used to perform the analysis. As a result, it has been determined that the use of Manhattan distance in data with Poisson distribution may produce unsuccessful results, especially in small sample sizes ($n<50$). Although there is no significant difference between Gower and Euclidean distances in different distributions according to sample sizes, it has been determined that the use of Euclidean distance measure in some distributions produces results that cause fluctuation. However, it has been understood that the Gower distance can be recommended as a more suitable choice since it has a more stable structure. For the applicability of the Least Square Estimation method, it may be recommended to use Distance Based Regression methods in cases where the necessary assumptions mentioned in this study cannot be met.

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

ÖN SÖZ VE TEŞEKKÜR

Akademik hayatıma başladığım ilk günden bugüne engin bilgi ve öğretileriyle beni geliştiren, her adımımda tecrübeleriyle desteklerini esirgemeyen danışman hocam sayın Prof. Dr. Hasan ÖNDER'e, bu süreçte bir telefon kadar yakınımda olan heyecanıma, stresime ortak olup beni sakinleştiren kıymetli arkadaşım Habibe KESKİN'e, Yüksek Lisans eğitimimde yardımlarını esirgemeyen iş arkadaşlarıma,

Bugünlere gelmemdeki en büyük mimarlarım canım annem, canım babam ve eğitim alanındaki uzmanlıklarıyla bana yol gösteren abim ve ablalarıma sonsuz sevgi ve teşekkürler.

Burcu KURNAZ

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

AIC	: Akaike Bilgi Kriteri
BIC	: Bayesian Bilgi Kriteri
EKK	: En Küçük Kareler Yöntemi
GCV	: Genelleştirilmiş Çapraz Geçerlilik
GKT	: Genel Kareler Toplamı
HKT	: Hata Kareler Toplamı
PSNR	: Tepe-Sinyal Gürültü Oranı
RKT	: Regresyon Kareler Toplamı
RMSE	: Hata Kareler Ortalamasının Karekökü

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

Gelişen bilim dünyasında araştırmacılar herhangi bir bilimsel alandan elde edilen verileri bir sonuç oluşturarak tablolar ya da grafikler ile göstermek istediklerinde birtakım yöntemler, yazılımlar ve modern bilgisayarlar kullanmaktadır. Birçok bilimsel araştırmada kullanılan veriler üzerinde karşılaştırma yapmak veya ilişki kurmak istenildiğinde istatistiksel analizler uygulanmaktadır. Regresyon analiz yöntemleri istatistiğin ilişkiler ile ilgili kısmını oluşturmakta ve bir sebep-sonuç ilişkisini incelemektedir. Bir veya birden fazla sebebin sonucu hangi yönde ve ne ölçüde etkilediğini inceleyen bir ilişki yöntemidir ve bu ilişkiler sebepten sonuca tek yönlüdür (Kurnaz ve Önder, 2021).

Regresyon analizinin genel kullanım amaçları (Dutter ve Huber, 1981):

- Erken zamanda ölçülebilen değerlerden ileriki zamandaki ölçüm değerlerini,
- Ölçümü kolay olan bir özelliğe ait veri setinden, ölçümü zor olan bir özelliğin değerlerini,
- Maliyeti düşük olan ölçüm değerlerinden yüksek maliyetli ölçüm değerlerini tahmin etmek şeklinde sıralanabilir.

Herhangi bir regresyon analizinde amaç, mevcut verileri kullanarak doğru ve güvenilir bir tahmin denklemi elde etmektir. Bu, bir yanıt değişkeni (Y) ile açıklayıcı değişken(ler) (Xi) arasında istatistiksel bir ilişki olup olmadığına ilişkin en önemli ve yaygın sorulardan biridir. Bu soruyu cevaplamanın bir seçeneği, ilişkisini modellemek için regresyon analizi kullanmaktır. Çeşitli regresyon analizi türleri vardır. Regresyon modelinin türü, yanıt değişkeninin (Y) dağılımının şekline bağlıdır; sürekli ve yaklaşık olarak normal ise doğrusal regresyon modeli kullanılır; eğer ikili ise, lojistik regresyon kullanılır; Poisson veya Multinomial ise log-lineer analiz kullanılmaktadır; sansürlü vakaların (hayatta kalma tipi) mevcudiyetinde olay-zaman verileri varsa, modelleme yöntemi olarak Cox regresyonunu kullanılır. Tahmin modelini kullanarak sonucu (Y) açıklayıcı değişkenlerin (Xi) değerlerine dayalı olarak tahmin etmek amaçlanır. Bu yöntemler, aynı modelde birden fazla değişkenin (ortak değişkenler ve faktörler) etkisinin değerlendirilebilmesini sağlar (Draper ve Smith, 1998; Rosner, 2015).

Doğrusal regresyon analizi, belirlenmek istenen değişkenden daha kolay, düşük

maliyetli veya daha erken saptanabilen değişken(ler)den yola çıkarak istenen yanıt değişkeni tahmin eden bir model oluşturmaktır (Alpar, 2010).

Doğrusal regresyon analizi basit doğrusal regresyon ve çoklu regresyon analizi olarak iki başlık altında incelenmektedir.

Basit doğrusal regresyon analizi, yanıt değişkeni ile tek bir açıklayıcı değişken arasındaki doğrusal ilişkiyi açıklar. Eğer tek bir yanıt değişkeni ve birden fazla açıklayıcı değişken arasındaki doğrusal bir ilişki tanımlanmak istenirse, ilişki çoklu doğrusal regresyon analizi ile incelenir (Okur, 2009; Weisberg, 2005).

Gerek basit gerekse çoklu doğrusal regresyon analizi sonucunda elde edilecek olan regresyon modeline ait parametre kestirimlerinin güvenilir olabilmesi için modelle ilgili bazı varsayımların sağlanabilmesi gereklidir. Basit doğrusal regresyon analizinde elde edilen regresyon denkleminin tahmin amaçlı kullanılabilmesi için; hata terimlerinin ($\epsilon_i = Y_i - \hat{Y}$) şansa bağlı normal dağılım göstermesi, hataların beklenen değerinin ortalamasının 0 ve varyansının homojen olup σ^2'ye eşit olması, hataların bağımsız olması [Cov(ϵ_i, ϵ_j)]=0, hata terimleri ile açıklayıcı değişken(ler) arasında korelasyon bulunmaması gibi bazı varsayımların sağlanması gerekmektedir (Alma ve Vupa, 2008).

Çoklu doğrusal regresyonda, basit doğrusal regresyondaki varsayımlara ilaveten açıklayıcı değişkenlerin birbirinden bağımsız olması varsayımının da sağlanması gerekmektedir (Vural, 2007). Açıklayıcı değişkenler arasındaki basit doğrusal korelasyon katsayılarının sıfır veya sıfıra çok yakın olması şartı şeklinde de açıklanabilen bu varsayım, istatistikte "Çoklu doğrusal bağlantı" bulunmaması olarak ifade edilmektedir (Orhunbilge, 2017). Çoklu bağlantı durumunda En Küçük Kareler (EKK) kestirim yöntemi gücünü kaybetmektedir (Vural, 2007).

Bu nedenle açıklayıcı değişkenler seçilirken, bu değişkenlerin yanıt değişkeni ile basit doğrusal korelasyon katsayılarının yüksek (1'e yakın), birbirleri arasındaki basit doğrusal korelasyon katsayılarının düşük (0'a yakın) olmasına dikkat edilmesi önerilmektedir (Damodar, 2001). Bu varsayımların sağlanamadığı durumlarda parametre kestirim yöntemlerinin değiştirilmesi önerilmektedir (Arı ve Önder, 2012).

Parametre tahmin yöntemlerinde yukarıda bahsi geçen durum için çözüm olarak geliştirilen az sayıdaki modellerden biri Uzaklık Temelli Regresyon yöntemleridir. Bu yöntemlerin amacı kategorik veya gerçek değerli ve kategorik açıklayıcı değişkenlerin

bir karışımı dâhil olmak üzere, gerçek olmayan değer tahmin edicileri ile problemleri doğru bir şekilde ele almaktır (Arenas ve Cuadras, 2002).

İstatistik ve veri analizinde, geometrik kavram bireyler veya popülasyonlar arasındaki uzaklık antropoloji, biyoloji, genetik, psikoloji, dilbilim ve diğerleri gibi alanlarda uygulanmıştır. Mesafe kavramı bir hipotez testi ve parametre tahmininde diğer uygulamalar arasında kullanışlı bir araçtır. Ayrıca uygunluk analizi veya çok boyutlu ölçekleme gibi bazı istatistiksel tekniklerde uzaklık kavramı temel bir araçtır (Cuadras,1988).

Çeşitli çok değişkenli yaklaşımlar, bağlantıdaki ilişkileri değerlendirebilmektedir (Varoquaux ve Craddock, 2013) ve bazı faktörler araştırmacıları çok değişkenli mesafe matrisi regresyonunu (MMR) incelemeye yöneltmiştir (Anderson, 2001; McArdle ve Anderson, 2001; Schork vd., 2008; Shehzad vd., 2014). Bunlar şunları içerir:

- Bir seferde birden fazla açıklayıcı değişkeni inceleme yeteneği (yani ortak değişkenler dahil edilebilir),
- Kategorik ve/veya sürekli değişkenler için uygulanabilirlik,
- Parametreye özgü veya analitik karar verme için minimum gereksinimler (örneğin, bir kullanıcının yalnızca mesafe ölçüsünü seçmesi gerekir)
- Regresyon benzeri analitik yapı nedeniyle yorumlanabilirlik kolaylığı.

Uzaklık temelli regresyon modeli ekoloji, genomik, genetik, insan mikrobiyomu ve nöroloji gibi çeşitli alanlarda ve çok değişkenli sonuç regresyon analizlerinde birçok uygulamaya sahiptir. Uzaklık temelli regresyon çok değişkenli varyans analizi, Çok değişkenli regresyon, Kanonik korelasyon analizi gibi geleneksel yöntemlerle kullanılmakta ve çoklu sonuçlar genellikle ilişkilendirilmektedir. Gözlemler arasında benzemezlik ölçümlerine dayalı çok değişkenli ekolojik veri analizi için Mcardle ve Anderson (2001) tarafından önerilen parametrik olmayan bir yaklaşım olan uzaklık temelli regresyon iyi bir alternatiftir. Uzaklık temelli regresyon modeli bazı farklılık ölçümleri açısından farklı gruplar arasındaki farklılıkları belirlemek için pseudo F test istatistiği kullanılmaktadır (Li vd., 2019).

Son zamanlarda, uzaklık temelli regresyon modeli birçok alanda başarıyla uygulanmıştır. Genomikte Xu vd. (2015) bir sürücü mutasyon varlığında uzaklık

temelli regresyon modeliyle ilişkili kümelere göre genleri sıralamış ve önemli geni seçmiştir. Nörobilimde, Shehzad vd. (2014), uzaklık temelli bir regresyon modeli ile beyin fenotipleriyle ilişkili vokselleri (Bilgisayarlı tomografiden alınan görüntülerdeki vücut bölümlerindeki hacim birimi) tespit etmiştir. İnsan mikrobiyom araştırmasında Chen vd. (2012) mikrobiyom bileşimini etkileyen faktörleri regresyon temelli bir yaklaşımla belirlemiştir. Tüm bu uygulamalarda, pseudo F testinin anlamlılık değerleri uzaklık temelli regresyon modelinden türetilen istatistik, bu amaç için üstün olduğu kanıtlanmış permütasyon prosedürü ile sayısal olarak hesaplanmıştır (Li vd., 2019).

Elde edilen bilgilere göre pseudo F test istatistiğinin teorik dağılım özellikleri mevcut literatürde tartışılmamıştır. Bu nedenle, spektral ayrıştırma ve matris normal varsayımına dayanarak, uzaklık ölçüleri Öklid veya Mahalanobis uzaklığı olduğunda pseudo F test istatistiğinin asimptotik özellikleri incelenmektedir. Öklid uzaklık ölçüsü için, pseudo F test istatistiğinin iki Ki-kare tipi dağılışın bölünmesiyle oluşturulan rastgele bir değişkenle aynı dağılıma sahip olduğu bilinmektedir. Böyle bir yaklaşım uzaklık temelli regresyon modellerinin uygulama aralığını daha yüksek boyutlara genişlettiği için kullanışlı olmakta ve pseudo F test istatistiği üzerinde daha fazla sonuç çıkarmaya olanak sağlamaktadır (Li vd., 2019).

Tavuk verisi üzerinde Önder ve Mercan (2020) biyoloji ve çevre bilimlerinde iki uzaklık matrisini veya daha genel olarak iki benzerlik veya yakınlık matrisini ilişkilendiren herhangi bir analizi içerebilen Mantel testi uygulayarak Bray Curtis uzaklık ölçüsü ve Nei’s genetik uzaklık ölçüsünü karşılaştırmışlardır. Bu çalışmanın sonucunda genetik farklılaşmalar ve popülasyonlar arasındaki coğrafi uzaklıklar arasındaki ilişkiyi hesaplamak için Mantel testinde, tavuk çeşitliliği verileri üzerinde büyük bir güvenilirlikle Nei'nin genetik uzaklığı yerine Bray Curtis uzaklığının kullanılabileceğini göstermişlerdir.

Bayram ve Nabiyev (2020), kamuflaj görüntüleri arka plan dokusu ile yakın özellikler gösterdiğinden, görüntü arka planından kamuflajlı nesneyi segmentlere ayırmak ve tespit etmek için çalışmalarında Öklid ve Mahalanobis uzaklık hesaplamaları kullanılarak K-means yöntemi kullanılarak kamuflaj görüntüleri üzerinde görüntü bölütleme uygulamışlardır. Bu çalışmadan elde edilen sonuca göre Öklid uzaklığı hesabı kullanılarak K-means yöntemi ile düşük RMSE değerleri elde edilirken, Mahalanobis uzaklığı hesabı kullanılarak daha düşük PSNR değerleri elde edilmiştir. Deneysel sonuçlarda; Öklid uzaklık hesaplamalı K-means yöntemi,

Mahalanobis uzaklık hesaplamalı K-means yöntemine göre daha başarılı olduğunu bildirmişlerdir.

Doğan (2003) çalışmasında, iki farklı ırkta doğum tipi ve cinsiyet faktörlerine göre elde edilen dört farklı durumun her biri, bir kombinasyon olarak dikkate alarak ve bu iki faktörün hangi kombinasyonlarındaki büyümenin birbirine benzediğinin belirlenmesi amacıyla çok boyutlu ölçekleme yöntemini kullanmıştır. Bu yöntemde uzaklıklar matrislerinden yararlanarak Öklid uzaklık ölçüsünden yararlanılmıştır. Yapılan çalışma sonucunda büyüme üzerine etkili olduğu düşünülen faktörlerden yalnızca doğum tipi ve cinsiyetin dikkate alınarak elde edilen sonuçlara göre Morkaraman ırkı kuzularda büyümede doğum tipinin ön plana çıktığı, Akkaraman ırkı kuzularda ise cinsiyetin daha etkili olduğu sonucu elde edilmiştir.

Bu çalışmada, Euclidean, Gower ve Manhattan uzaklık ölçülerinin Binom, Normal, t, Ki-Kare ve Poisson dağılışlarına ait üretilmiş, örnek büyüklükleri 10, 25 50, 100, 250 ve 500 olan veri setleri ve kesikli ve sürekli dağılış gösteren gerçek veri setleri (10, 50 ve 100 örnek büyüklüğünde) üzerinde etkisi ile Doğrusal Regresyon yönteminden elde edilen sonuçlara göre karşılaştırma yaparak belirlenmesi amaçlanmıştır.

2. MATERYAL VE YÖNTEM

2.1. Materyal

Bu çalışmada örnek büyüklüğü 10, 25, 50, 100, 250 ve 500 olan Binom, Poisson, Ki-Kare, Normal ve t dağılışlarından oluşan veri setleri analiz edilmiştir. Analizler R programı 4.2.2 versiyonu kullanılarak gerçekleştirilmiştir. Benzetim çalışmasında 10000 tekrar kullanılmıştır. Çalışmada kullanılan sürekli veri daha önce Önder ve Abacı'nın (2015) yaptığı bir çalışmada kullanılan Saanen oğlaklarına aittir. Altıncı ay canlı ağırlığı sonuç değişkeni olarak kullanılırken, altıncı aya ait vücut uzunluğu ve göğüs derinliği açıklayıcı değişken olarak alınmıştır. Çalışmada kullanılan kesikli veri ise daha önce Aerts vd. (2022) çalışmasında kullanılan Polonya Holstein Friesian sığırlarına aittir. Bu örnekte süt yağ oranı sonuç değişkeni olarak belirlenirken, günlük sağım sayısı ve mevsim açıklayıcı değişken olarak belirlenmiştir. Elde edilen sonuçların değerlendirilmesinde AIC, BIC, GCV değerlerinin ortalama, standart sapma ve hata hesaplamaları kullanılmıştır.

Çeşitli istatistiksel analizler için kullanılabilecek uygun yazılımlardan biri de R yazılımıdır. R, istatistiksel hesaplamalar ve grafikler için açık kaynak kodlu bir programdır. R programlarının güçlü yanlarından biri iyi tasarlanmış matematiksel semboller de dâhil edilmek üzere yayın kalitesinde grafikler ve gerektiğinde formüller üretebilmektedir. Grafik tasarım seçenekleri için varsayılanlara büyük ölçüde özen gösterilmesinin yanı sıra kullanıcı tüm kontrole sahip olmaktadır. Windows'un yanı sıra çeşitli UNIX platformları ve benzer bir sistem olan Linux gibi geniş bir yelpazede de çalışmaktadır (Yan ve Su, 2009).

2.2. Yöntem

Doğrusal regresyon, modelin regresyon parametrelerinde doğrusal olmasını gerektirmektedir. Regresyon analizi bir ya da birden fazla yanıt değişkeni (bağımlı değişken, açıklanan değişken, tahmin edilen değişken olarak da adlandırılır) ile tahmin ediciler (bağımsız değişken, açıklayıcı değişken olarak da adlandırılan x_1, x_2, …, x_p şeklinde belirtilir) arasındaki ilişkiyi inceleyen bir yöntemdir (Yan ve Su, 2009). Doğrusal regresyon yönteminin güvenilir sonuçlar üretebilmesi için varsayımlara ihtiyaç duyulmaktadır. Bu varsayımlar hata terimlerinin ($\epsilon_i = Y_i - \hat{Y}$) şansa bağlı normal dağılım göstermesi, hataların beklenen değerinin ortalamasının 0 ve varyansının homojen olup σ^2'ye eşit olması, hataların bağımsız olması [Cov(ϵ_i,

$\epsilon_j)]=0$, hata terimleri ile açıklayıcı değişken(ler) arasında korelasyon bulunmaması gibi bazı varsayımların sağlanması gerekmektedir (Alma ve Vupa, 2008). Açıklayıcı değişkenler arasındaki Pearson korelasyon katsayılarının sıfırdan uzak 1'e yakın olması durumunda "Çoklu doğrusal bağlantı" sorununun varlığından söz edilebilir (Orhunbilge, 2017). Çoklu bağlantı durumunda En Küçük Kareler (EKK) kestirim yöntemi gücünü kaybetmektedir (Vural, 2007).

Genel olarak doğrusal bir regresyon modeli; $Y = \boldsymbol{X\beta} + \varepsilon$ olarak tanımlanır. Burada; Y; (nx1) boyutlu yanıt değişkeni vektörü, X; (nxp) boyutlu bilinen katsayı matrisi (tasarım matrisi), β; (nx1) boyutlu bilinmeyen parametre vektörü (katsayılar vektörü), ε; (nx1) boyutlu kalıntılar (hata) vektörü olup, ortalaması sıfır ($E(\varepsilon)=0$) ve varyansı ($var(\varepsilon)=\sigma^2 I$) sabittir (Atkinson vd., 2000).

Doğrusal regresyon modeli matris yaklaşımıyla incelendiğinde;

$Yi = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$ (i = 1, 2, …, n) gibi çok açıklayıcı değişkene sahip bir model, aşağıdaki gibi bir denklem modelini göstermektedir.

$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + u_1$

$Y_2 = \beta_0 + \beta_1 X_{12i} + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + u_2$

……

$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots + \beta_k X_{kn} + u_n$

Bu modelin matris gösteriminde ifadesi ise aşağıdaki gibidir;

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \text{ veya } \boldsymbol{Y=X\,\beta+u}$$

Burada;

$\mathbf{Y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ n×1 boyutlu bağımlı değişken gözlemleri vektörü,

$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}$ n×k boyutlu açıklayıcı değişken verileri matrisi,

$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$ k×1 boyutlu katsayılar vektörü ve

$\mathbf{u}=\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$ n×1 boyutlu hata terimleri vektörüdür.

Doğrusal bir regresyon modelinin EKK tahmin edicisi;

$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$

şeklinde tanımlanabilmektedir (URL 1).

Doğrusal regresyon modelinde Genel Kareler Toplamının (GKT) matris gösterimi; GKT= $\boldsymbol{Y'Y}$, Regresyon Kareler Toplamının (RKT) matris gösterimi; RKT=$\widehat{\boldsymbol{\beta}}\boldsymbol{'X'Y}$, Hata Kareler Toplamının (HKT) gösterimi ise HKT= GKT-RKT şeklinde ifade edilmektedir.

Uzaklık Temelli Regresyon Modeli istatistik ve veri analizinde, bireyler veya popülasyonlar arasındaki geometrik uzaklık kavramı biyoloji, genetik, psikoloji gibi alanlarda uygulanmıştır.

Uzaklık Temelli Regresyon modeli, herhangi bir sayıda açıklayıcı matris üzerinde bir yanıt matrisinin çoklu regresyonunu içerir; burada her matris, n nesnenin (örnek birimler) tüm ikili kombinasyonları arasındaki mesafeleri veya benzerlikleri (ekolojik, uzamsal veya diğer nitelikler açısından) içerir; istatistiksel anlamlılık testleri permütasyon ile gerçekleştirilir. Yöntem, analiz edilebilecek veri türleri (sayılar, var-yok, sürekli, kategorik) ve yanıt eğrilerinin şekilleri açısından esneklik göstermektedir (Lichstein, 2007).

T⊆Y olmak üzere karşılık gelen çıktılar ile T=$\{\boldsymbol{t}_k\}_{k=1}^{K}$ ile referans girdi noktalarının (R⊆X ile R=$\{\boldsymbol{m}_k\}_{k=1}^{K}$) seçimi için $i = 1, \dots, N$ girdi noktaları $\boldsymbol{x}_i$ ve k'ncı referans noktası $\boldsymbol{m}_k$ arasında k'ncı sütunu uzaklıkları $d(\boldsymbol{x}_i, \boldsymbol{m}_k)$ içerecek şekilde $\boldsymbol{D}_x \epsilon \mathbb{R}^{N\times K}$ tanımlanır. Benzer şekilde $\Delta_y \epsilon \mathbb{R}^{N\times K}$ ifadesi y_i'deki N adet çıktı noktası ve k'ıncı referans noktasındaki t_k çıktısı arasındaki $\delta(\boldsymbol{y}_i, \boldsymbol{t}_k)$ uzaklıklarını içerecek şekilde tanımlanır. Girdi uzaklık matrisi D_x ve çıktı uzaklık matrisi Δ_y arasındaki ilişkiler çok değişkenli regresyon modeli için;

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

D_x matrisinin sütunları K giriş vektörlerine, Δ_y matrisinin sütunları K yanıt vektörlerine, N satırları gözlemlere karşılık gelmektedir. N×K boyutlu E matrisinin sütunları, K kalıntılarına karşılık gelmektedir.

Girdi ve çıktı uzaklık matrisleri arasındaki eşlemenin her yanıt için doğrusal bir yapıya sahip olduğu varsayıldığında, regresyon modeli Denklem (1)'de gösterildiği şekildedir:

$\Delta_y = D_x B + E.$ (1)

K×K regresyon matrisi B'nin sütunları, K yanıtlarının katsayılarına karşılık gelmektedir. Matris B, kayıp fonksiyonu olarak çok değişkenli hata kareler toplamının en aza indirilmesi yoluyla verilerden tahmin edilebilir:

$RSS(B) = tr((\Delta_y - D_x B)' (\Delta_y - D_x B))$.

Normal koşullar altında, Denklem (1)'deki örnek büyüklüğünün değişken sayısından daha fazla olduğu durumlarda, problem aşırı belirlenir (iterasyon sayısı çok fazla artabilir) ve genellikle çözümü yoktur. Bu, seçilen referans noktalarının sayısının mevcut nokta sayısından (yani K<N) daha küçük olduğu duruma karşılık gelmektedir. Bu durumda, B'nin olağan En Küçük Kareler tahmini tarafından sağlanan yaklaşık çözüm kullanılmalıdır,

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

Denklem (1)'de örnek büyüklüğü değişken sayısına eşitse (yani K=N çünkü tüm öğrenme noktaları aynı zamanda referans noktalarıdır), o zaman problem benzersiz bir şekilde belirlenir ve D_x matrisi tam ranklı ise tek bir çözümü vardır. Bu durumda,

$\hat{B} = D_x^{-1} \Delta_y$.

Denklem (1)'de örnek büyüklüğü değişken sayısından daha az olduğu durum açıkça bilindiği şekilde genellikle sonsuz sayıda çözüme sahip, yeterince belirlenmemiş bir çözümsüzlüğe yol açar (de Souza vd., 2015).

Herhangi bir doğrusal modeldeki (MANOVA, MANCOVA veya çok değişkenli regresyon için) herhangi bir terimle ilişkili kareler toplamı doğrudan bir uzaklık matrisinden hesaplanabilir.

Bunun nedeni, herhangi bir merkezileştirilmiş veri matrisi $\boldsymbol{Y}_{(n\times p)}$ için (p değişken ve n örnek için) klasik çok değişkenli istatistikte kullanılan iç çarpım matrisi Y′Y ile hesaplanabildiği gibi dış çarpım matrisi olan YY′ ile de hesaplanabilir. Ek olarak, bir dış çarpım matrisi herhangi bir $(n \times n)$ uzaklık matrisinden elde edilebilir (Gower, 1966), böylece analizin Bray-Curtis gibi semimetrik ölçümler dâhil olmak üzere seçilen bir uzaklık ölçüsüne dayalı olmasına izin verir.

$X_{(n\times m)}$, m parametre sayısı ile bir model (diğer bir deyişle tasarım veya regresyon) matrisi olsun. Klasik çok değişkenli istatistik için (pxp) toplam kareler toplamı iç çarpım matrisi Y′Y'nin parçalanması ile elde edilmektedir. Genel kareler toplamı (S_T), bu matristeki $tr(Y'Y)$ ile sembolize edeceğimiz köşegen elemanlarının izi veya toplamıdır (her değişken için karelerin toplamı). Parçalama işlemi $Y=X\beta+\epsilon$ doğrusal modeline göre yapılabilir. Burada β modeldeki parametreleri matrisi ve ϵ hata matrisini göstermektedir. β için en küçük kareler çözümü B=$(X'X)^{-1}X'Y$ dir. Tahmin değerleri matrisi $\hat{Y}=X$B =HY yazılabilir ki burada; H idempotent tahmin (hat) matrisidir ve $X(X'X)^{-1}X'$ olarak gösterilebilir. Hata (kalıntılar) matrisi R=Y-$\hat{Y}$=(I-H)Y olarak gösterilebilir. Regresyon kareler toplamı tr($\hat{Y}'\hat{Y}$)ve hata kareler toplamı tr(R′R) olmak üzere model parametrelerinin etkisinin olup olmadığının hipotezini test etmek için uygulanan bir istatistik olan pseudo F istatistiği aşağıdaki gibi hesaplanır:

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

Tek bir değişken olması durumunda pseudo F testi Fisher F testi ile aynı olarak hesaplanır. Parametrik olmayan test durumunda I. Tip hata olasılığı P=P(F*>=F) burada F* ünitenin permütasyonu ile hesaplanan değerdir. Serbestlik dereceleri olan (m-1) ve (n-m) permütasyon testi için gerekli değildir ve sabit olarak alınır (Anderson, 2001).

Bir benzerlik matrisi hesaplandıktan sonra, bu matris hakkındaki hipotezleri test eden bir regresyon analizine tabi tutulur. Çiftleri tarafından sergilenen benzerlik seviyesinde değişimin olup olmadığı o matrise yansıyan bireyler diğerleriyle bu bireylerin sahip olduğu özellikler aracılığıyla açıklanabilir (örn. belirli bir fenotip veya belirli bir özelliğin daha yüksek veya daha düşük değerlerine sahip kantitatif fenotip) (Schork ve Wessel, 2006).

2.2.1. Euclidean Uzaklık Ölçüsü

Birimler arasındaki uzaklıkları hesaplamak için en yaygın kullanılan uzaklık ölçüsü Öklid uzaklığıdır. Öklid uzaklığı iki birim arasındaki benzerliği ölçmede en yaygın kullanılan uzaklık ölçüsü olup iki birim arasına çizilecek bir düz doğrunun uzunluğunu temel alır (Ünlükaplan, 2008).

Sezgisel olarak, model, yanıtların vektörünü, daha sonra bu uzayda sanal, gözlemlenmeyen koordinatlarla değiştirilen, gözlemlenen öngörücülerden metrik çok boyutlu ölçeklendirme ile elde edilen bir Öklid uzayını yansıtır. Bu modelin temel

özellikleri, özel bir durum olarak görünen en küçük kareler ile olağan regresyonun bir uzantısı olması ve nitel veya karma açıklayıcı değişkenlere uygulanabilmesidir (Borg ve Groenen, 2005).

Öklid uzaklık ölçüsü kullanılarak iki birim arasındaki uzaklık n birim sayısı ve p değişken sayısı olmak üzere; i,j = 1,2,3,…,n, i. ve j. birimin birbirine olan uzaklığı

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

formülü ile hesaplanır (Dinler, 2014).

Öklid uzaklık ölçüsü kullanıldığında bu yöntemin klasik doğrusal regresyon modeliyle uyumlu olduğu kanıtlanmıştır (Arenas ve Cuadras, 2002).

2.2.2. Manhattan Uzaklık Ölçüsü

Nesneler arasındaki uzaklıkların boyutlara göre toplamıdır. İki boyutlu uzayda iki nesne arasındaki uzaklığın ölçümünde içerisinde gösterilen üçgenin hipotenüsü Öklid uzaklığını göstermektedir. Bu üçgenin hipotenüsünün dışındaki kenarlarının uzunluklarının toplamı Manhattan City Block uzaklığını vermektedir. Daha çok kesikli nicel verilere sahip değişkenler için kullanılması önerilmektedir. Aşağıdaki eşitlikle hesaplanmaktadır (Alpar, 2013):

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

Manhattan City Block uzaklığı, aykırı değerlere karşı daha az hassas olan bir uzaklık ölçüsüdür (Timm, 2002).

2.2.3. Gower Uzaklık Ölçüsü

Gower uzaklığı 1971 yılında Gower tarafından önerilmiştir. Gower uzaklığının en temel özelliği, hem kategorik hem de sürekli verilerin bulunduğu veri setinde kullanılabilmesidir. Gower uzaklığı standardize edilmiş veriler kullanılarak hesaplanır. Gower uzaklığı sadece sürekli veriler kullanıldığı zaman ayrı bir formül ile hesaplanmaktadır. Hem kategorik hem de sürekli verilerin bulunduğu veri seti için kullanılan uzaklık Gower genel benzerlik ölçüsü olarak adlandırılır (URL 2). Gower kategorik değişkenler için genel benzerlik ölçüsünü,

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}}$$

biçiminde ifade etmiştir. Burada S_{ijk} , *k.* değişken değerine göre *i.* ve *j.* gözlemler arasındaki benzerlik ölçüsüdür. W_{ijk} ise *i.* ve *j.* gözlem *k.* değişkene göre karşılaştırmasında değişken değeri bulunmuyorsa 0 diğer durumlarda 1 değerini almaktadır. Gower (1971), verideki sürekli değişkenler için benzerlik ölçüsünü;

$$S_{ij} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$$

biçiminde tanımlamıştır. Burada R_k, i. ve j. gözlemin k. değişken değerlerinin değişim aralığı (range) olarak tanımlanmaktadır (Servi, 2009).

2.2.4. Karşılaştırma Ölçütleri

Model seçim kriterlerinde yaygın olarak kullanılan AIC Akaike'nin Bilgi Kriterleri anlamına gelmektedir. Akaike'nin Bilgi Kriterleri 1973'te Hirotsugu Akaike tarafından geliştirilmiştir. AIC, tahmin edilen herhangi bir istatistiksel modelin uyum iyiliğinin bir göstergesi olarak adlandırılabilmektedir. Akaike'nin Bilgi Kriterleri asimptotik olarak çapraz geçerliliğe eşdeğer olmaktadır (URL 3).

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

eşitliği ile bulunmaktadır. Burada k sabit terim dâhil parametre sayısı ve n gözlem sayısını, L=olabilirliği göstermektedir (Ucal, 2006).

AIC'nin bazı özellikleri söyle sıralanabilir:

• En düşük AIC değerini üreten model her zaman model karşılaştırma amacıyla tercih edilmelidir.

• AIC çalışılan örnek büyüklüğünden farklı örnek büyüklükleri için yapılacak tahminler için geçerlidir (Ucal, 2006).

Bayes bilgi kriteri (BIC) veya Schwarz bilgi kriteri (ayrıca SIC, SBC, SBIC), sonlu bir model seti arasından model seçimi için bir kriterdir; genellikle daha düşük BIC'li modeller tercih edilmektedir. Kısmen olasılık fonksiyonuna dayanır ve Akaike bilgi kriteri (AIC) ile yakından ilişkilidir. Modelleri uydururken, parametreler ekleyerek maksimum olasılığı artırmak mümkündür, ancak bunu yapmak aşırı uydurmaya neden olabilir. Hem BIC hem de AIC, modeldeki parametre sayısı için bir ceza terimi getirerek bu sorunu çözmeye çalışır; ceza terimi, 7'den büyük örneklem boyutları için BIC'de AIC'den daha büyüktür (McQuarrie ve Tsai, 1998) ve aşağıda gösterildiği şekilde hesaplanabilir.

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

BIC eşitliğin sağ tarafındaki örnek büyüklüğüne bağlı olan ikinci kısım itibariyle AIC'den farklılık gösterir. Fakat AIC ve BIC arasındaki yüzeysel benzerliğe rağmen, daha sonraları Bayes yapısı içinde farklılıklar gösterdiği belirlenmiştir (Raftery, 1995; Wasserman, 2000).

Craven ve Wahba tarafından 1979 yılında geliştirilen Genelleştirilmiş Çapraz Geçerlilik (GCV) ölçütü en uygun modeli seçme kriterlerinden biri olmaktadır (Adıgüzel, 2021). GCV kriteri hataların minimizasyonuna dayandığı gibi modelin karmaşıklığını da dikkate almaktadır (Yıldız, 2022).

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

Eşitlikte C(M) geçerli temel fonksiyonlar için model karmaşıklığını cezalandıran fonksiyonu, y_i bağımlı değişkenin gözlem değerlerini ve $f_M(x_i)$ tahmin değerlerini, N ise gözlem sayısını ifade etmektedir (Chen vd., 2012).

Analizde kullanılan örnek kod dizisi;

```
library("dbstats")
library("cluster")
library(tidyverse)
simsay=10000
sampsize=25
sonuc=matrix(nrow=simsay, ncol=10)
for(i in 1:simsay) {
        Y=rnorm(sampsize,0,1)


        X1=rchisq(sampsize,5)
        X2=rchisq(sampsize,5)
        Model1 <- dbglm(formula = Y ~ X1 + X2, family = gaussian(), method = "GCV", full.search = TRUE, metric = "euclidean", weights = NULL, range.eff.rank = c(1, 2))
        Model2 <- dbglm(formula = Y ~ X1 + X2, family = gaussian(), method = "GCV", full.search = TRUE, metric = "gower", weights = NULL, range.eff.rank = c(1, 2))
```

```
        Model3 <- dbglm(formula = Y ~ X1 + X2, family = gaussian(), method = "GCV", full.search = TRUE, metric = "manhattan", weights = NULL, range.eff.rank = c(1, 2))
        glm1 <- glm(Y ~ X1 + X2, family = gaussian())
 sonuc[i,1]=summary(Model1$aic.model)[4][1]
 sonuc[i,2]=summary(Model1$bic.model)[4][1]
 sonuc[i,3]=summary(Model1$gcv.model)[4][1]
 sonuc[i,4]=summary(Model2$aic.model)[4][1]
 sonuc[i,5]=summary(Model2$bic.model)[4][1]
 sonuc[i,6]=summary(Model2$gcv.model)[4][1]
 sonuc[i,7]=summary(Model3$aic.model)[4][1]
 sonuc[i,8]=summary(Model3$bic.model)[4][1]
 sonuc[i,9]=summary(Model3$gcv.model)[4][1]
 sonuc[i,10]=summary(glm1$aic)[4][1]
}
sink("D:/deneme.txt")
sonuc
sink()
```

3. BULGULAR

Dağılışın ve Uzaklık ölçülerinin AIC, BIC ve GCV üzerindeki etkilerini incelemek amacıyla faktöriyel deneme desenine göre analizler yapılmış olup örnek büyüklüğü kovaryet olarak kullanılmıştır. Dağılış × Uzaklık ölçüsü interaksiyonu önemsiz bulunduğu (P>0,05) için sadece ana etkiler sunulmuştur. Kovaryet olarak kullanılan örnek büyüklüğünün beklenildiği şekilde istatistiksel olarak anlamlı olduğu belirlenmiş (P<0,01) bu nedenle tablolarda marjinal ortalamalar ve standart hata değerleri verilmiştir. Elde edilen bulgulara göre dağılışın AIC, BIC ve GCV üzerinde istatistiksel olarak anlamlı etkisi olduğu (P<0,01) uzaklık ölçülerinin ise sadece AIC değeri üzerinde etkisinin olduğu (P<0,01) belirlenmiştir (Tablo 3.1. ve Tablo 3.2.).

Tablo 3.1. Dağılışın AIC, BIC ve GCV üzerindeki etkileri

Dağılış	AIC	BIC	GCV
Binom	156,841 ± 0,027^{bc}	159,956 ± 0,031^{b}	0,984 ± 0,001^{b}
Ki-Kare	156,870 ± 0,027^{b}	159,978 ± 0,031^{b}	0,981 ± 0,001^{bc}
Normal	156,763 ± 0,027^{c}	159,876 ± 0,031^{b}	0,979 ± 0,001^{c}
Poisson	157,002 ± 0,029^{a}	160,421 ± 0,034^{a}	0,989 ± 0,001^{a}
t	156,812 ± 0,027^{bc}	159,915 ± 0,031^{b}	0,983 ± 0,001^{bc}
P	<0,001	<0,001	<0,001

Ko-varyet olan Örnek büyüklüğü 54,3092 olarak belirlenmiştir.
^{a,b}: Aynı sütunda bulunan farklı harfler istatistiksel farklılığı (P<0,05) göstermektedir.

Tablo 3.2. Uzaklık ölçülerinin AIC, BIC ve GCV üzerindeki etkileri

Uzaklık	AIC	BIC	GCV
Euclidean	156,557 ± 0,024^{b}	160,025 ± 0,025	0,983 ± 0,001
Gower	156,555 ± 0,024^{b}	160,022 ± 0,025	0,983 ± 0,001
Manhattan	156,576 ± 0,024^{b}	160,040 ± 0,025	0,984 ± 0,001
LR	157,743 ± 0,024^{a}		
P	<0,001	0,862	0,402

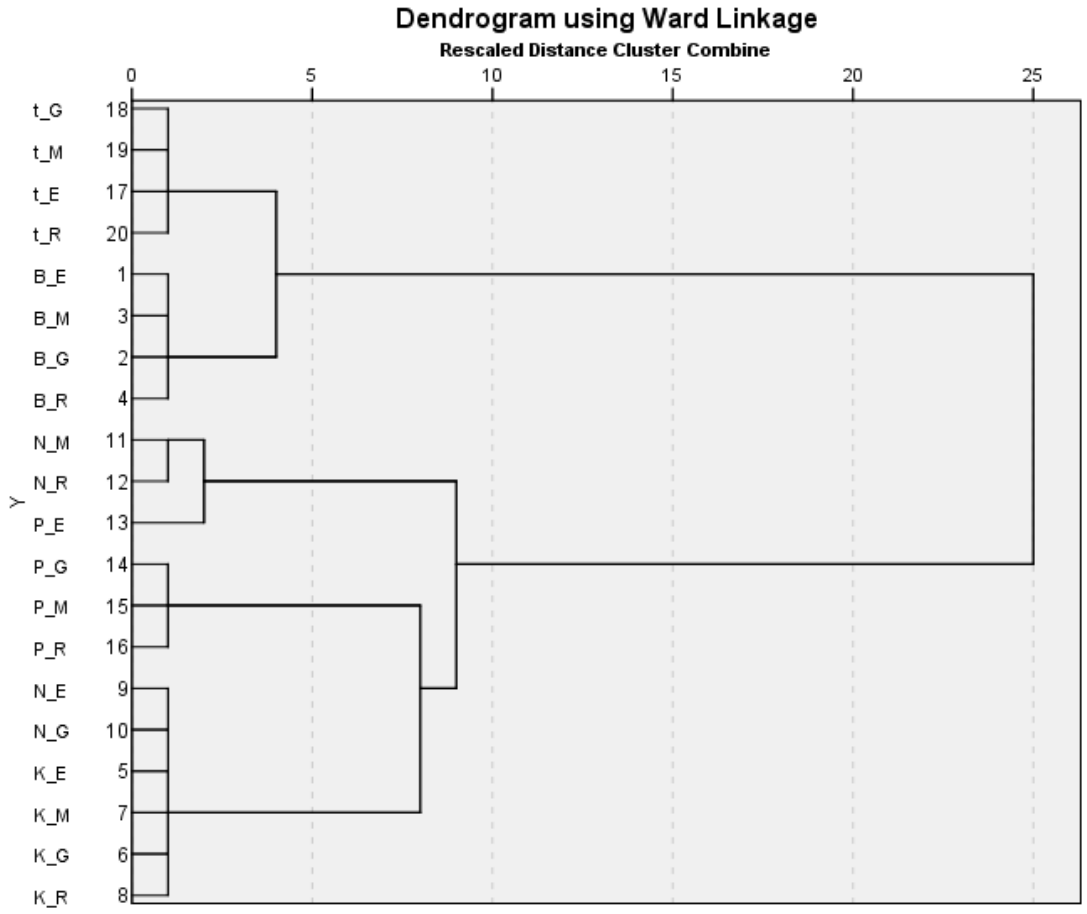
Ko-varyet olan Örnek büyüklüğü 54,3092 olarak belirlenmiştir.
a,b: Aynı sütunda bulunan farklı harfler istatistiksel farklılığı (P<0,05) göstermektedir.

En düşük AIC değeri Normal dağılışa sahip verilerden elde edilmesine rağmen Normal, Binom ve t dağılışından elde edilen sonuçlar arasında AIC değeri bakımından farklılık bulunmamaktadır. En yüksek AIC değerleri ise Poisson dağılışa sahip verilerden elde edilmiş ve diğer dağılışlarla arasında önemli farklılık belirlenmiştir. Ki-kare dağılışından elde edilen AIC değerlerinin Normal ve Poisson dağılışlarından elde edilen AIC değerlerine göre farklı olduğu ancak Binom ve t dağılışından elde

edilen sonuçlar arasında AIC değeri bakımından farklılık bulunmadığı belirlenmiştir. BIC değerleri üzerinde dağılışın etkisi değerlendirildiğinde sadece Poisson dağılışından elde edilen BIC değerlerinin diğer dağılışlardan elde dilenlere göre önemli bir farklılıkla yüksek olduğu belirlenmiştir. Diğer dağılışlardan elde dilen BIC değerlerinin ise benzer olduğu gözlemlenmiştir. Dağılışların GCV değerleri üzerindeki etkisi değerlendirildiğinde elde edilen sonuçların AIC üzerinde olan etkisiyle benzer olduğu görülmüştür (Tablo 3.1).

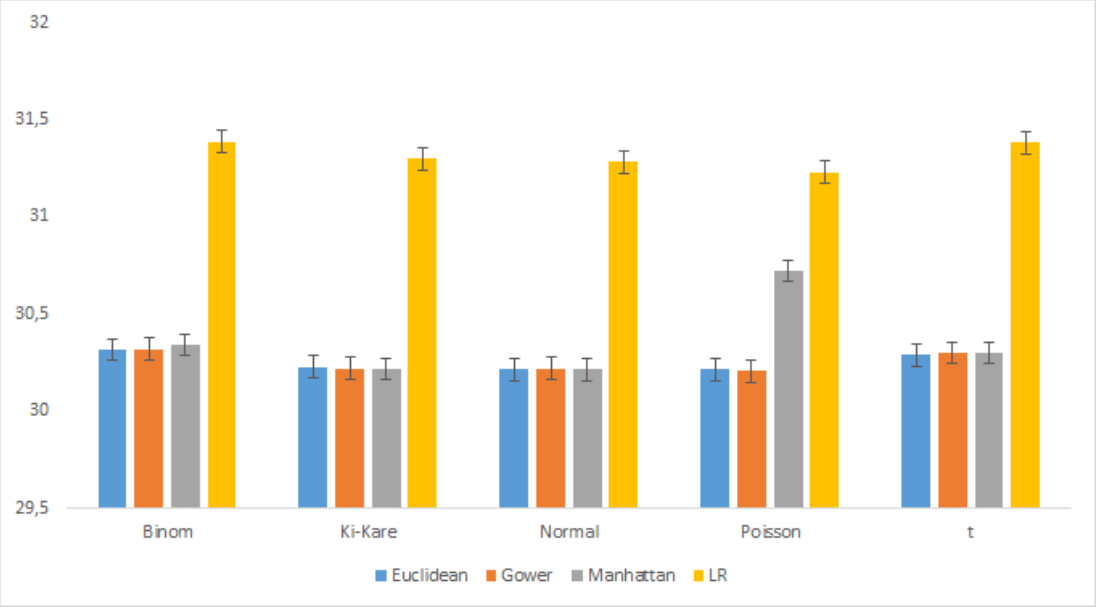
Tablo 3.2. incelendiğinde AIC değeri üzerindeki farklılığın doğrusal regresyon en küçük kareler yönteminden elde edilen değerlerden kaynaklandığı ve Euclidean, Gower ve Manhattan uzaklık ölçüleri arasında farklılık olmadığı anlaşılmaktadır.

Dağılış × Uzaklık ölçüsü kombinasyonu için AIC, BIC ve GCV ölçümleri birlikte değerlendirilerek Ward yöntemi ve Karesel Euclidean uzaklığı kullanılarak çizilen hiyerarşik kümeleme dendogramı Şekil 3.1'de verilmiştir.



Şekil 3.1. Dağılış × Uzaklık ölçüsü kombinasyonu için AIC, BIC ve GCV ölçümleri birlikte değerlendirilerek çizilen hiyerarşik kümeleme dendogramı.

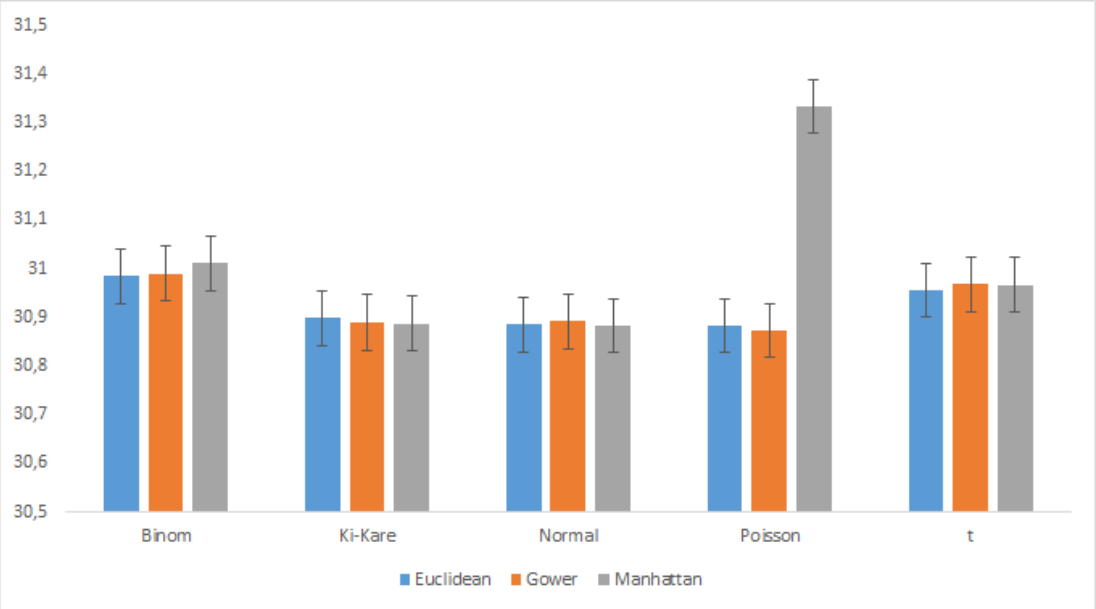
Elde edilen sonuçlara göre toplam 20 kombinasyonun beş küme içerisinde incelenebileceği belirlenmiştir. Buna göre; t dağılışında kullanılan tüm uzaklık ölçülerin ve EKK çözümün ayrı bir küme oluşturduğu (küme 1), Binom dağılışında kullanılan tüm uzaklık ölçülerin ve EKK çözümün ayrı bir küme oluşturduğu (küme 2), Normal dağılışta Manhattan ve EKK çözümü ile Poisson dağılışında Euclidean uzaklığının ayrı bir küme oluşturduğu (küme 3), Poison dağılışında Manhattan, Gower ve EKK çözümünün ayrı bir küme oluşturduğu (küme 4), Ki-kare dağılışında tüm yöntemlerin ve Normal dağılışta Euclidean ve Gower çözümlerinin ise son kümeyi oluşturduğu (Küme 5) belirlenmiştir. İlk iki kümenin diğer kümelere daha uzak olduğu anlaşılmaktadır. Binom ve t dağılışlarının kendine özgü kümeler oluşturduğu ancak diğer dağılışların karışık kümeler oluşturduğu belirlenmiştir. Son üç küme incelendiğinde; normal dağılışın hem Poisson hem de Ki-kare dağılışının farklı çözümleri ile birlikte değerlendirilebileceği ancak Ki-kare ve Poison dağılışlarının aynı kümede yer almadığı anlaşılmaktadır.



Şekil 3.2. n=10 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

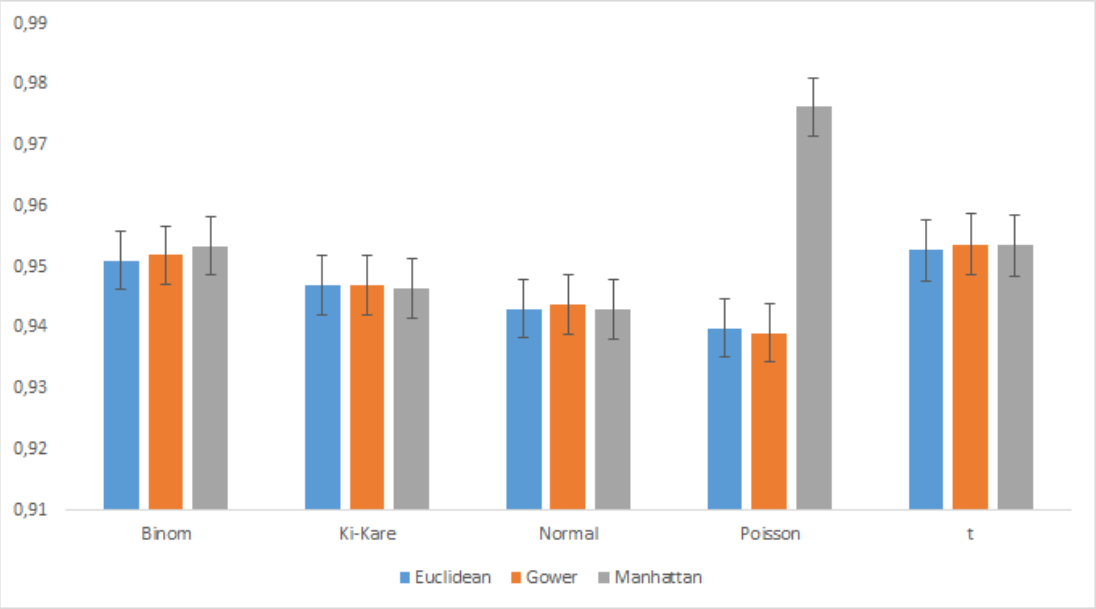
En küçük örnek büyüklüğü olan n=10 için AIC değerleri incelendiğinde tüm dağılışlar için doğrusal regresyondan elde edilen AIC değerlerinin diğerlerine göre önemli ölçüde ($P<0,05$) yüksek olduğu belirlenmiştir. Poisson dağılışı için Manhattan uzaklık ölçüsünün Euclidean ve Gower uzaklık ölçülerine göre daha yüksek AIC değeri ürettiği belirlenmiştir. Euclidean ve Gower uzaklık ölçüleri arasında farklılık olmadığı gözlemlenmiştir. Poisson hariç diğer tüm dağılışlarda uzaklık ölçüleri

arasında AIC değeri bakımından farklılık olmadığı belirlenmiştir.



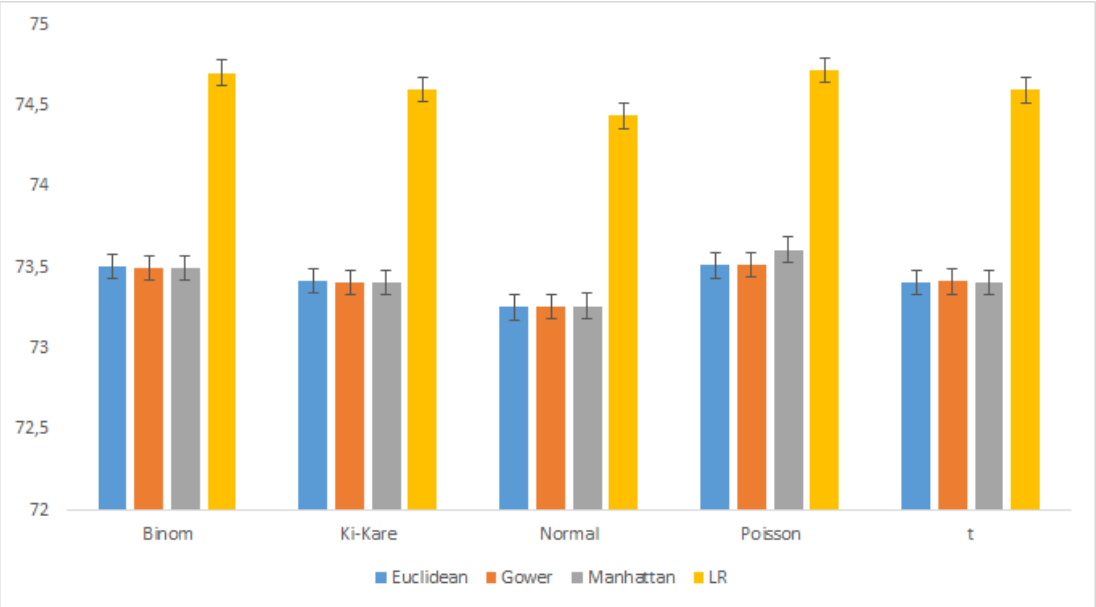
Şekil 3.3 n=10 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

Poisson dağılışı için n=10 örnek büyüklüğünde Manhattan uzaklığının en yüksek BIC değerini ürettiği belirlenmiştir. Diğer dağılışlar içinde uzaklık ölçüleri arasından önemli bir farklılık olmadığı (P>0,05) anlaşılmaktadır.



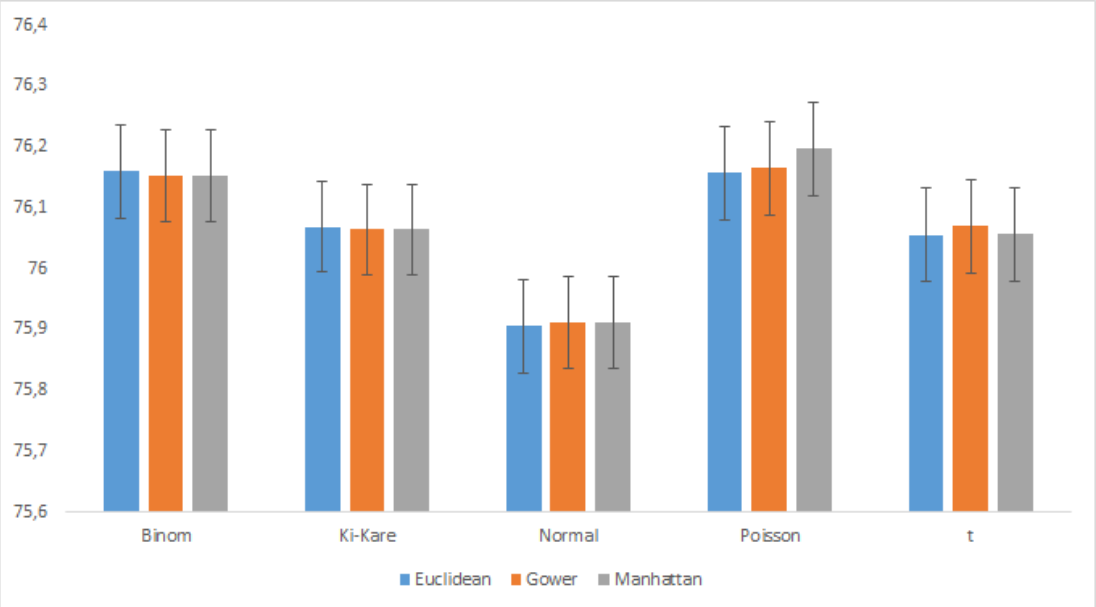
Şekil 3.4 n=10 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

Poisson dağılışı için n=10 örnek büyüklüğünde Manhattan uzaklığının en yüksek GCV değerini ürettiği belirlenmiştir. Diğer dağılışlar içinde uzaklık ölçüleri arasından önemli bir farklılık olmadığı (P>0,05) anlaşılmaktadır.



Şekil 3.5. n=25 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

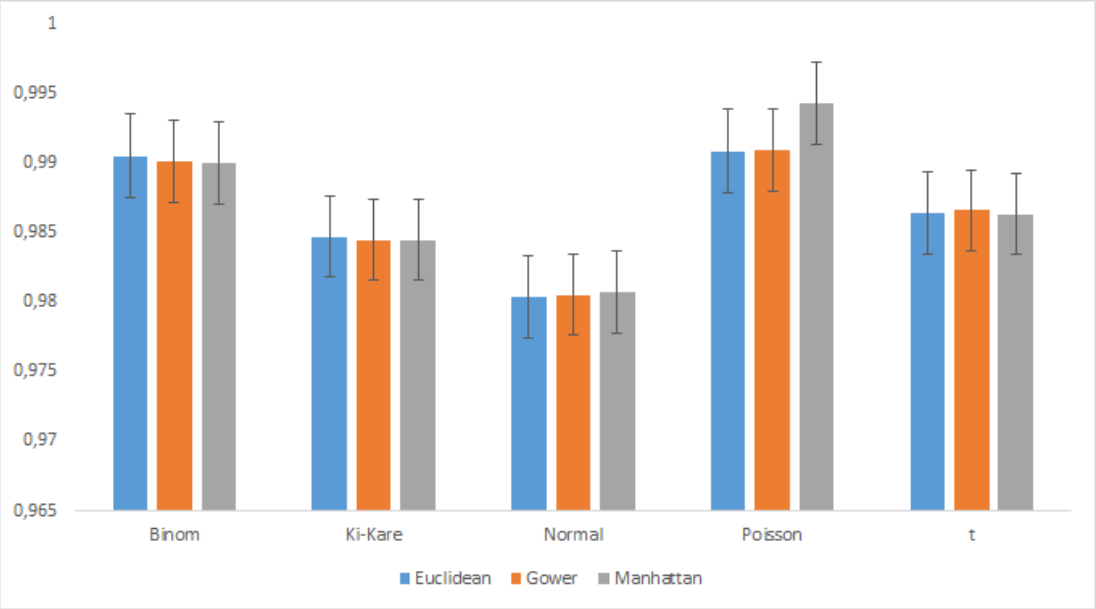
Örnek büyüklüğü n=25 için doğrusal regresyondan elde edilen AIC değerinin uzaklık ölçülerinden elde edilen değerden daha yüksek olduğu gözlemlenmiştir. Dağılışlar ve uzaklık ölçüleri bakımından önemli bir farklılık olmadığı (P>0,05) anlaşılmıştır.



Şekil 3.6. n=25 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

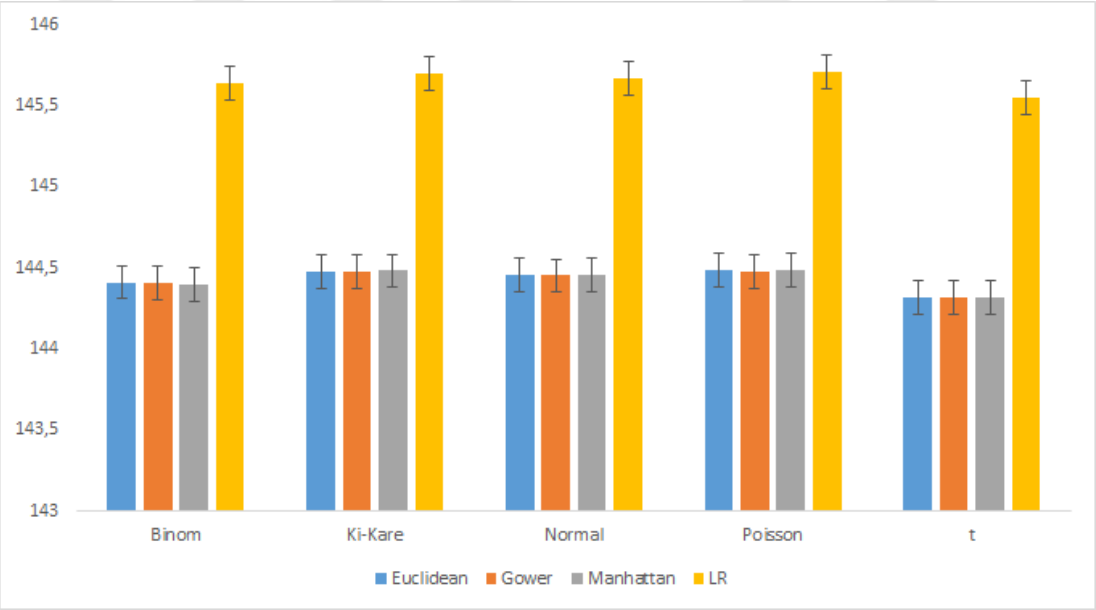
Normal dağılış için n=25 olduğunda Euclidean, Gower ve Manhattan uzaklık ölçüleri için en düşük BIC değeri ürettiği belirlenmiştir. Normal dağılışa en yakın t ve Ki-Kare dağılışından elde edilen BIC değerleri olduğu belirlenmiştir. Uzaklık ölçüleri

arasında anlamlı bir farklılık olmadığı (P>0,05) gözlemlenmiştir.



Şekil 3.7. n=25 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

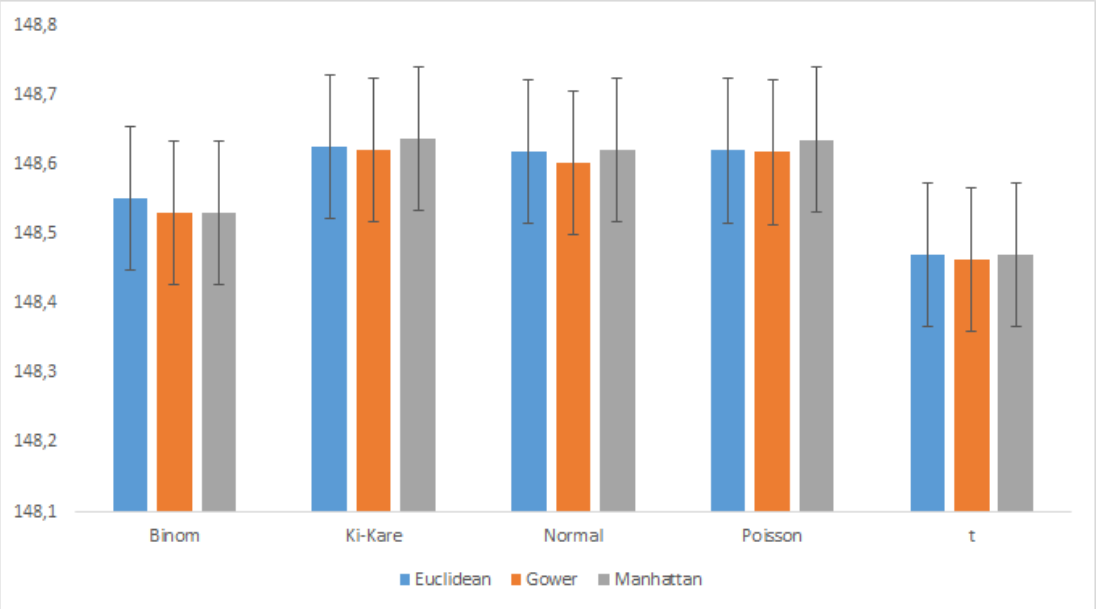
GCV değerleri için n=25 örnek büyüklüğüne bakıldığında dağılışlar arasında en düşük değeri üreten Normal dağılış olduğu belirlenmiştir. GCV değerlerine bakılarak uzaklık ölçüleri arasında önemli bir farklılık olmadığı (P>0,05) anlaşılmaktadır.



Şekil 3.8. n=50 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

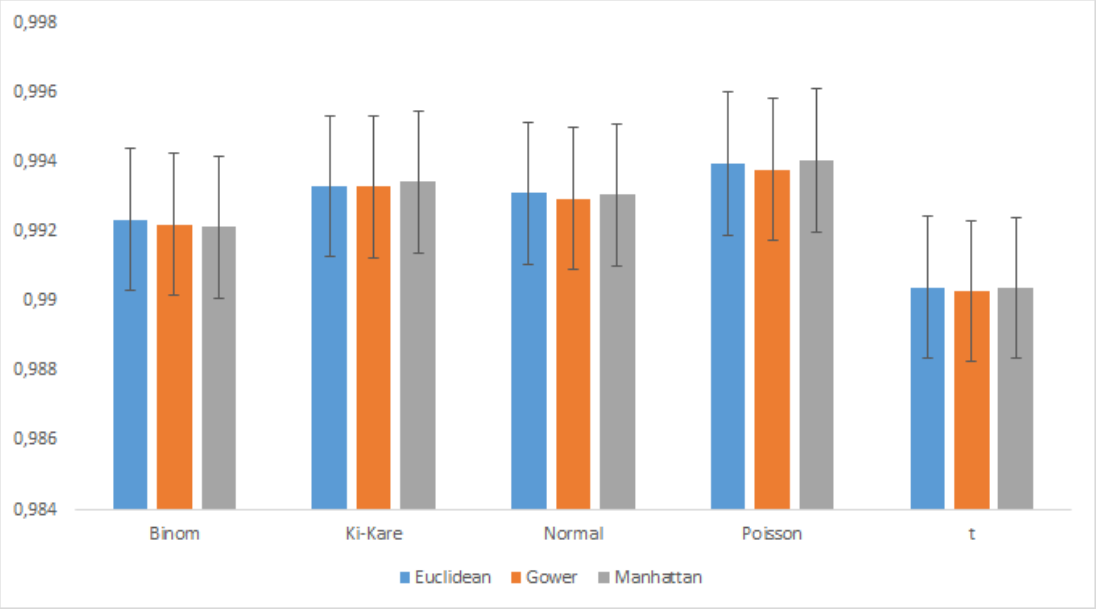
Örnek büyüklüğü n=50 olduğunda AIC değerleri için dağılışlar ve uzaklık ölçüleri arasında anlamlı bir farklılık olmadığı (P>0,05) anlaşılmıştır. Doğrusal regresyondan elde edilen AIC değerlerinin uzaklık ölçülerine göre daha yüksek bir

değer olduğu belirlenmiştir.



Şekil 3.9. n=50 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

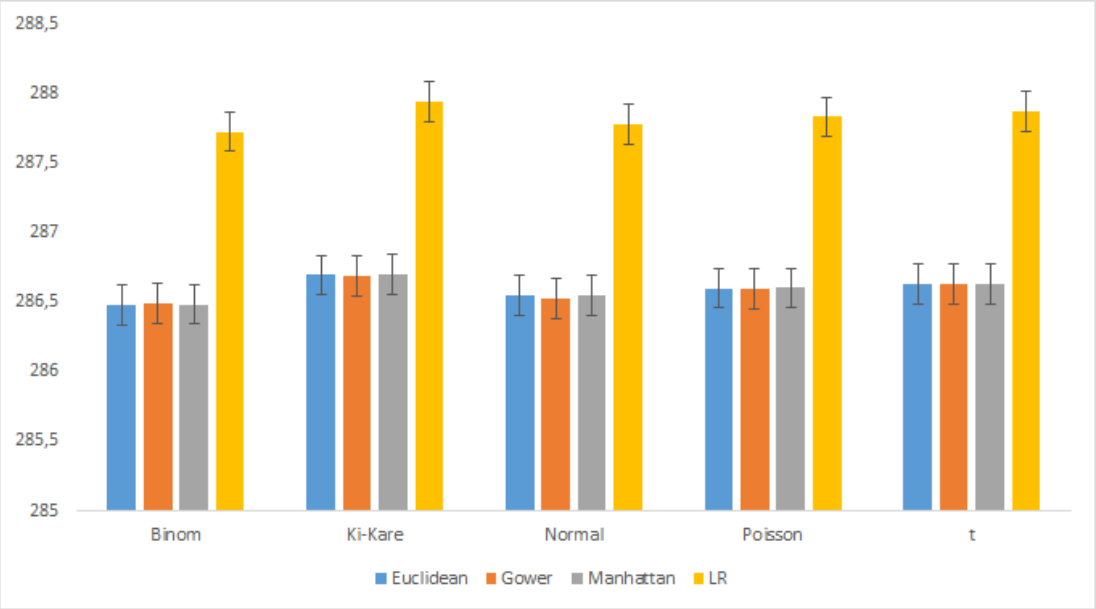
t dağılışının n=50 örnek büyüklüğü için diğer dağılışlara göre uzaklık ölçülerinden elde edilen en düşük BIC değeri ürettiği belirlenmiştir. Uzaklık ölçüleri olan Euclidean, Gower, Manhattan arasında önemli bir farklılık olmadığı (P>0,05) anlaşılmaktadır.



Şekil 3.10. n=50 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

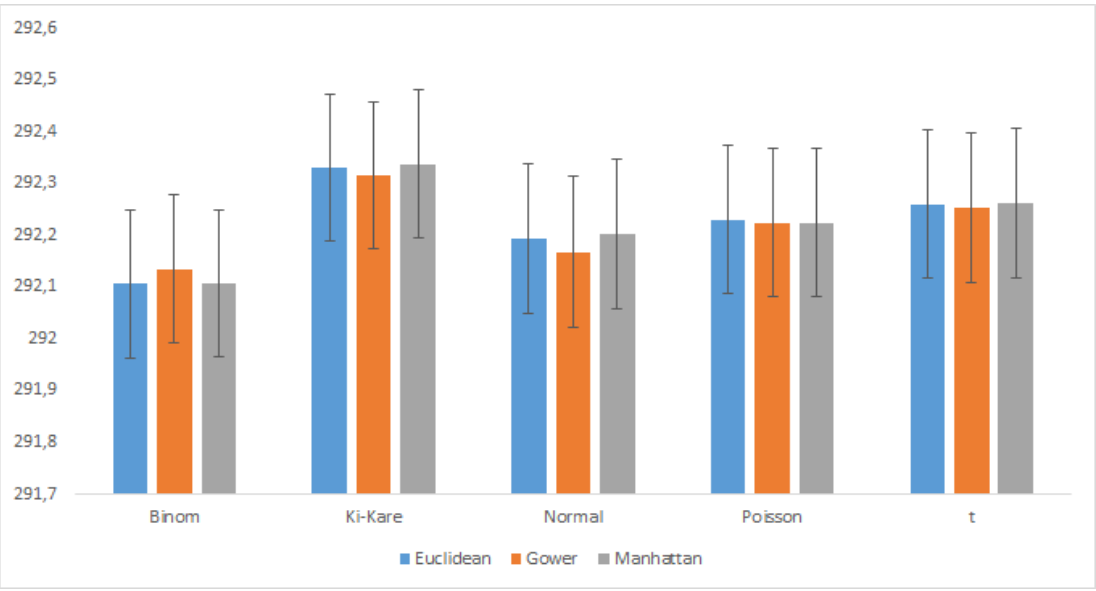
GCV değerlerine bakılarak uzaklık ölçüleri arasında anlamlı bir farklılık olmadığı (P>0,05) belirlenmiştir. Örnek büyüklüğü n=50 için t dağılışı diğer

dağılışlara göre en düşük GCV değerini üretmiş olduğu gözlemlenmiştir.



Şekil 3.11. n=100 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

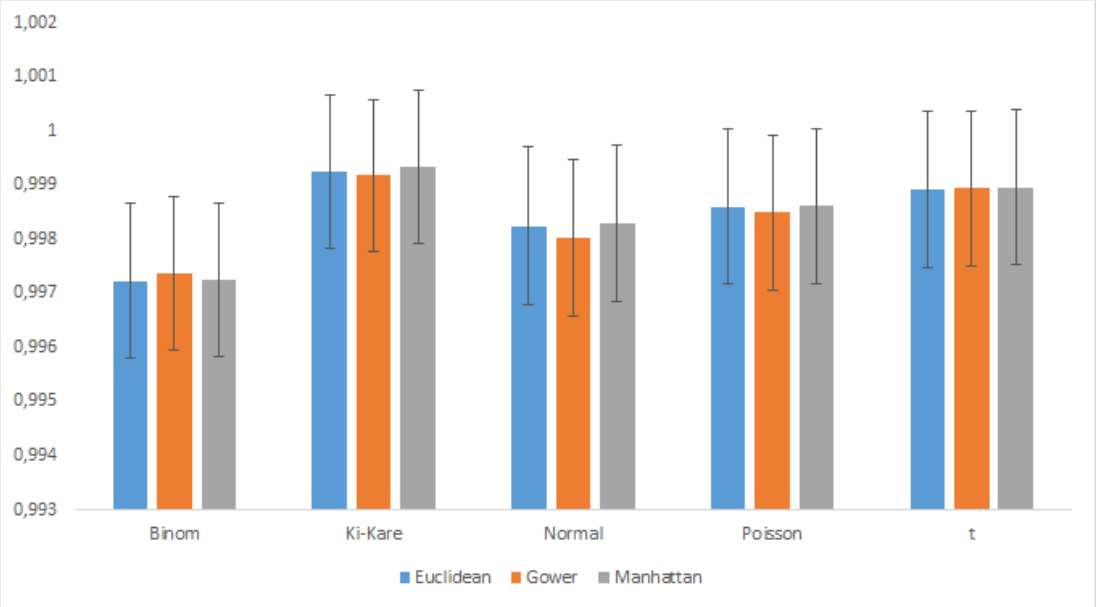
Örnek büyüklüğü n=100 olduğunda doğrusal regresyondan elde edilen AIC değerleri uzaklık ölçülerinden üretilen değerlerden daha yüksek olduğu anlaşılmaktadır. Euclidean, Gower ve Manhattan uzaklık ölçüleri arasında ve bu uzaklık ölçülerinin uygulandığı dağılışlar arasında anlamlı bir farklılık olmadığı (P>0,05) belirlenmiştir.



Şekil 3.12. n=100 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

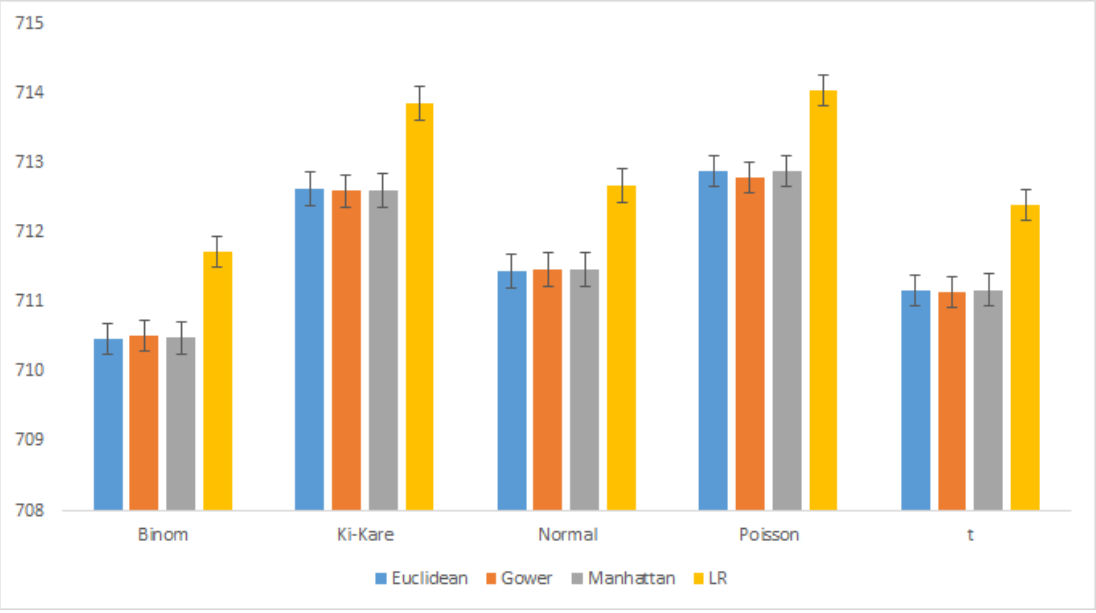
Elde edilen sonuçlara göre Binom dağılışından üretilen BIC değerinin diğer dağılışlara göre en düşük değer olduğu belirlenmiştir. Örnek büyüklüğü n=100 için

uzaklık ölçüleri arasında BIC değerleri arasından anlamlı bir farklılık olmadığı (P>0,05) anlaşılmıştır.



Şekil 3.13. n=100 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

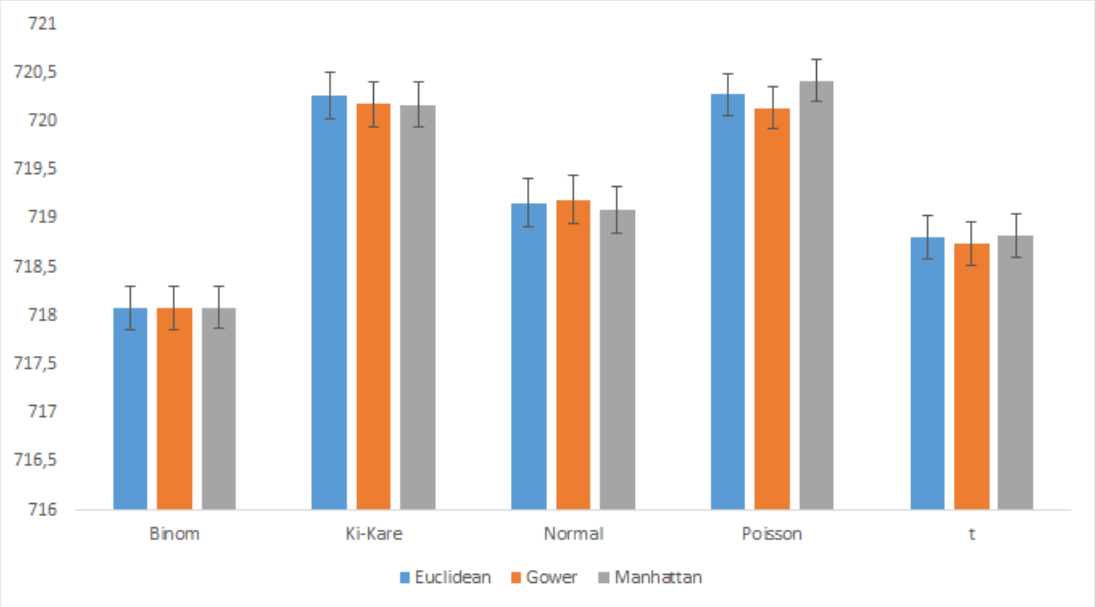
Örnek büyüklüğü n=100 için dağılışlar arasında en küçük GCV değerini üreten dağılışın Binom dağılışına ait olduğu belirlenmiştir. Euclidean, Gower ve Manhattan uzaklık ölçüleri arasında anlamlı bir farklılık olmadığı (P>0,05) belirlenmiştir.



Şekil 3.14. n=250 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

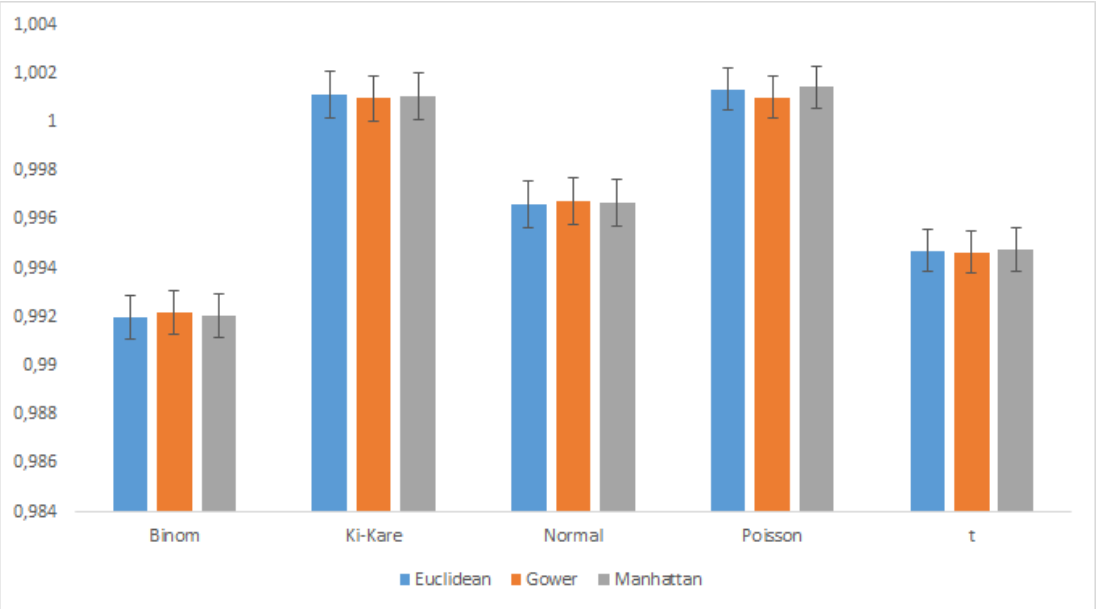
Örnek büyüklüğü n=250'ye yükseldiğinde tüm dağılışlar için doğrusal regresyondan elde edilen AIC değerleri ile uzaklık ölçülerinden elde edilen değerler

arasındaki farkın azaldığı anlaşılmaktadır. Yine de en yüksek AIC değerlerinin doğrusal regresyondan üretilen değerler olduğu belirlenmiştir. Binom dağılışı için uzaklık ölçülerinden elde edilen AIC değerlerinin diğer dağılışlara göre daha küçük olduğu belirlenmiştir. Uzaklık ölçüleri arasında önemli ölçüde bir farklılık olmadığı (P>0,05) anlaşılmıştır.



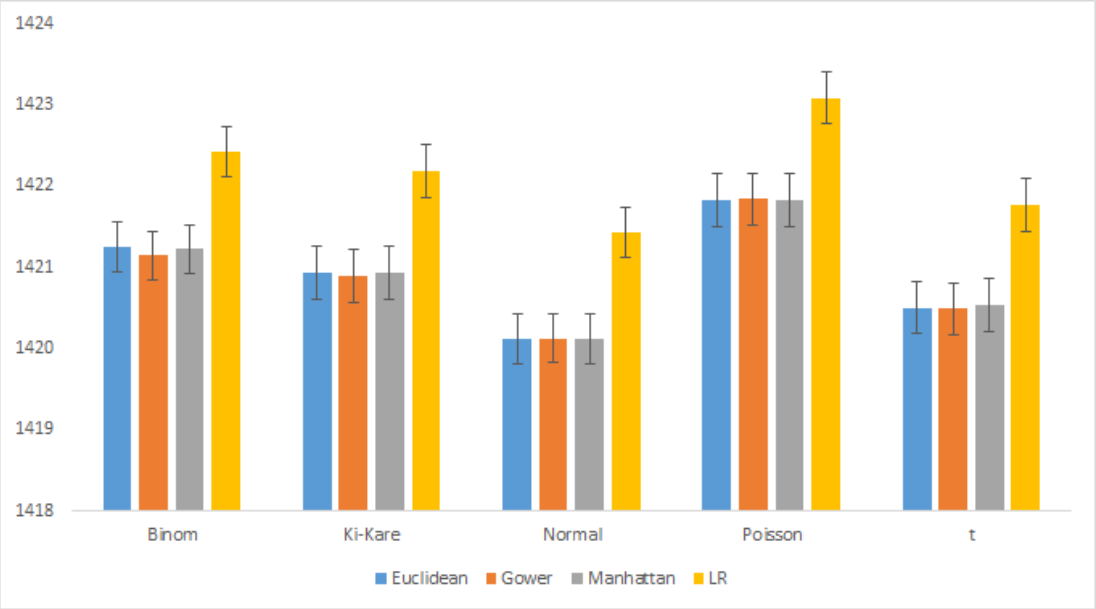
Şekil 3.15. n=250 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

Dağılışlar arasında n=250 örnek büyüklüğü olduğunda tüm uzaklık ölçüleri için en küçük BIC değerlerini üreten Binom dağılışı olduğu belirlenmiştir. Tüm dağılışlarda uzaklık ölçüleri bakımından anlamlı bir farklılık olmadığı (P>0,05) gözlemlenmiştir.



Şekil 3.16. n=250 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

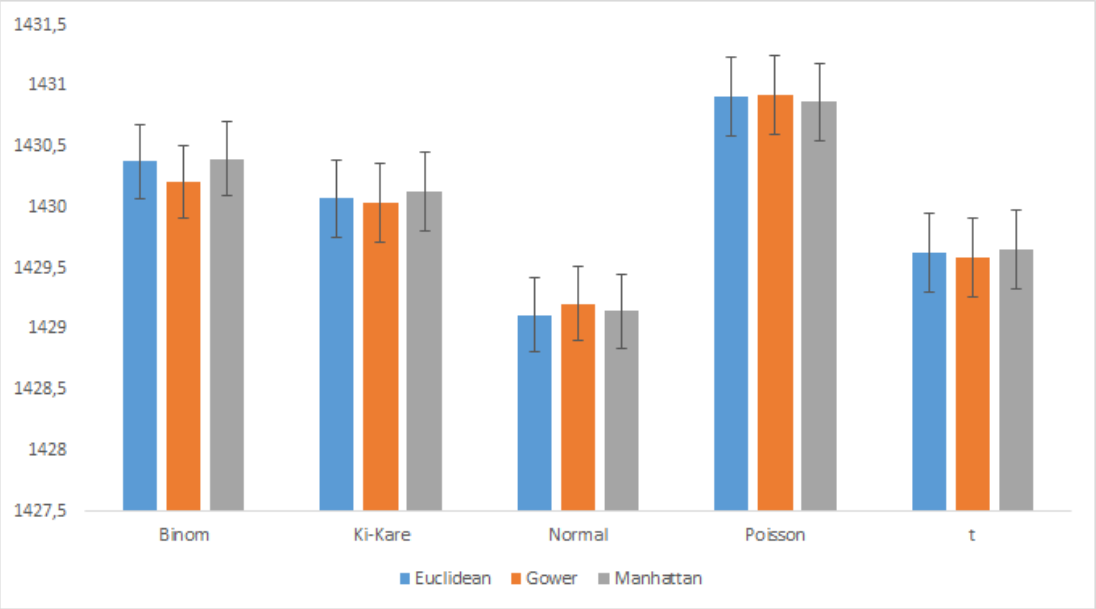
Örnek büyüklüğü n=250 için dağılışlara göre uzaklık ölçülerinden üretilen GCV değerleri arasından en küçük değere Binom dağılışının sahip olduğu anlaşılmıştır. Euclidean, Gower ve Manhattan uzaklık ölçüleri arasında önemli ölçüde bir farklılık olmadığı (P>0,05) belirlenmiştir.



Şekil 3.17. n=500 için dağılış ve uzaklık ölçülerine göre AIC değerleri.

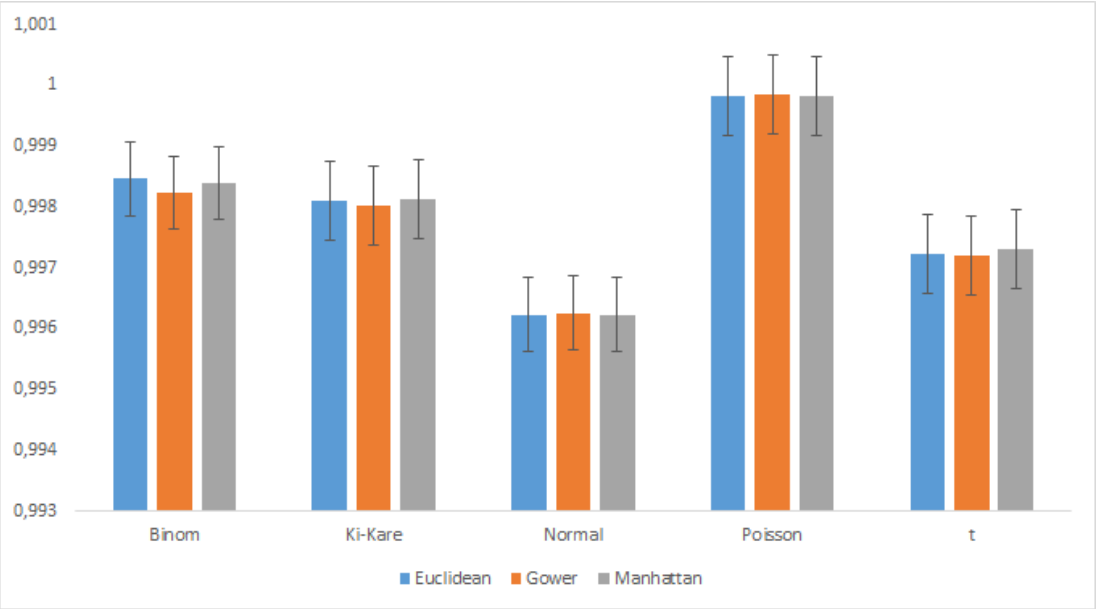
Örnek büyüklüğü n=500 için en yüksek AIC değerlerinin doğrusal regresyondan elde edilen değerler olduğu belirlenmiştir. Dağılışlar arasında en küçük AIC değeri üreten dağılışın Normal dağılış olduğu anlaşılmıştır. Uygulanan uzaklık

ölçüleri arasında anlamlı bir farklılık olmadığı (P>0,05) belirlenmiştir.



Şekil 3.18 n=500 için dağılış ve uzaklık ölçülerine göre BIC değerleri.

Normal dağılışın n=500 örnek büyüklüğü için diğer dağılışlardan üretilen BIC değerlerine göre daha küçük değer ürettiği belirlenmiştir. Elde edilen BIC değerlerine bakılarak uzaklık ölçüleri arasından önemli bir farklılık olmadığı (P>0,05) anlaşılmaktadır.



Şekil 3.19 n=500 için dağılış ve uzaklık ölçülerine göre GCV değerleri.

GCV değerlerine bakıldığında en küçük değeri üreten dağılışın Normal dağılışa ait olduğu anlaşılmaktadır. Örnek büyüklüğü n=500 olduğunda uzaklık ölçüleri

arasında anlamlı bir farklılık olmadığı (P>0,05) belirlenmiştir.

Uzaklık ölçülerinin örnek büyüklüğü 10, 50 ve 100 için sürekli ve kesikli dağılışa ait gerçek veri yapısı üzerine etkileri Tablo 3.3. ve Tablo 3.4.'te verilmiştir.

Tablo 3.3. Uzaklık ölçülerinin sürekli veri üzerine etkisi

Bilgi Kriterleri	Uzaklık ölçüleri	n=10	n=50	n=100
AIC	Euclidean	41,72	208,47	338,74
	Gower	41,72	209,33	340,35
	Manhattan	41,77	208,77	340,10
	EKK	43,44	208,50	338,70
BIC	Euclidean	42,32	214,21	345,96
	Gower	42,32	213,16	347,57
	Manhattan	42,38	214,50	347,32
GCV	Euclidean	2,57	3,50	3,47
	Gower	2,57	3,56	3,54
	Manhattan	2,59	3,52	3,53

Analiz sonucunda elde edilen bulgulara göre n=10 için EKK yönteminin uzaklık ölçülerine göre daha yüksek AIC değeri ürettiği belirlenmiştir. Uzaklık ölçüleri arasında Euclidean ve Gower aynı ve en düşük bilgi kriteri değerlerine sahip olduğu anlaşılmaktadır. Örnek büyüklüğü n=50 olduğunda EKK yönteminden elde edilen AIC değeri en düşük Euclidean ölçüsüne en yakın değeri ürettiği gözlemlenmiştir. Sürekli dağılış için örnek büyüklüğü n=50 için Gower uzaklık ölçüsü en büyük AIC değerine sahip olduğu görülmektedir. Örnek büyüklüğü n=100 için en küçük AIC değerinin doğrusal regresyondan elde edilen değerin olduğu belirlenmiştir. Bu değere en yakın AIC değeri Euclidean uzaklık ölçüsüne ait olduğu anlaşılmaktadır. Elde edilen sonuçlara göre EKK tahmin yönteminin örnek büyüklüğü küçük olan sürekli dağılışlı gerçek verilerde uzaklık temelli regresyondan elde edilen seçim kriterlerine göre güvenilir sonuçlar üretmediği belirlenmiştir. Örnek büyüklüğü arttığında doğrusal regresyon yöntemleri varsayımları sağlandığı durumlarda sürekli veriler için uzaklık ölçüsü olan Euclidean ile kayda değer bir farklılık olmadığı anlaşılmaktadır. Gower uzaklık ölçüsünün örnek büyüklüğü n=50 için en küçük BIC değerine sahip olduğu gözlemlenmiştir.

Tablo 3.4. Uzaklık ölçülerinin kesikli veri üzerine etkisi

Bilgi Kriterleri	Uzaklık ölçüleri	n=10	n=50	n=100
AIC	Euclidean	20,31	48,03	147,00
	Gower	20,44	48,11	146,63
	Manhattan	20,38	50,21	147,77
	EKK	22,27	48,41	148,60
BIC	Euclidean	21,10	51,81	152,21
	Gower	21,24	51,89	151,84
	Manhattan	21,18	53,99	152,98
GCV	Euclidean	0,26	0,14	0,24
	Gower	0,26	0,14	0,24
	Manhattan	0,26	0,15	0,25

Kesikli dağılışa ait veri setlerinde örnek büyüklüğü n=10 için AIC değerlerine bakılarak en küçük değeri üreten Euclidean uzaklık ölçüsünden elde edilmiştir. EKK yönteminin en yüksek Akaike Bilgi Kriteri değerine sahip olduğu gözlemlenmiştir. BIC değerlerine bakıldığında en küçük değeri Euclidean uzaklık ölçüsünün üretmiş olduğu anlaşılmaktadır. Tüm uzaklık ölçülerinden elde edilen GCV değerleri arasında bir farklılık olmadığı belirlenmiştir. Örnek büyüklüğü n=50 olduğunda en yüksek AIC değerini Manhattan uzaklık ölçüsünün ürettiği belirlenmiştir. Doğrusal regresyondan elde edilen kriter değeri en küçük kriter değerine sahip olan Euclidean ölçüsüne yakın değer aldığı gözlemlenmiştir. Bunun nedeni doğrusal regresyon için varsayımların sağlanmış olması olabilir. Kesikli dağılış için örnek büyüklüğü n=100 için en küçük AIC değerini üreten kesikli verilerde güvenilir sonuçlar oluşturan Gower uzaklık ölçüsü olduğu belirlenmiştir. BIC değerleri örnek büyüklükleri 10 ve 50 için en küçük bilgi kriteri üreten değerin Euclidean uzaklık ölçüsüne aitken, örnek büyüklüğü 100 olduğunda en küçük değere sahip uzaklık ölçüsü Gower olduğu belirlenmiştir. Tüm örnek büyüklükleri için GCV değerleri bakımından önemli ölçüde bir farklılık olmadığı anlaşılmaktadır.

4. SONUÇ

Uzaklık Temelli Regresyon yöntemlerinin dağılışlar üzerindeki etkisi değerlendirildiğinde en düşük bilgi kriterleri değerini Normal dağılış yapılı açıklayıcı değişkenlerden oluşan veri setinde olduğu görülmektedir ki bu sonuç teorik olarak bir beklentidir. Uzaklık ölçülerinden elde edilen AIC, BIC ve GCV değerlerine bakıldığında Gower uzaklık ölçüsü diğer uzaklık ölçülerine göre en düşük bilgi kriterleri değerlerine sahip olduğundan bu uzaklık ölçüsünün model seçimi için kullanılması önerilebilir.

Elde edilen bulgular değerlendirildiğinde, Poisson dağılışına sahip verilerde özellikle küçük örnek büyüklüklerinde ($n<50$) Manhattan uzaklığının kullanılmasının başarısız sonuçlar üretebileceği söylenebilir. Örnek büyüklüklerine göre farklı dağılışlar içerisinde Gower ve Euclidean uzaklıkları arasında kayda değer farklılık olmamasına rağmen bazı dağılışlarda Euclidean uzaklık ölçüsü kullanımının dalgalanmaya sebep olan sonuçlar ürettiği belirlenmiştir. Ancak, Gower uzaklığı daha sabit bir yapıya sahip olması nedeniyle daha uygun bir seçim olarak öne sürülebilir. Bunun nedeni Gower uzaklığının standardize edilmiş veriler kullanılarak hesaplanması olabilir.

Model seçiminde bilgi kriterleri değerlerine bakıldığında veri setlerine ait tüm dağılışlar ve örnek büyüklükleri için en yüksek AIC değeri üreten yöntem doğrusal regresyon için EKK tahmin yöntemi olduğu belirlenmiştir. Bilgi kriterleri değerleri model seçimi amacıyla değerlendirildiğinde en küçük değeri üreten modelin seçilmesi tahmin başarısını artırmaktadır. Dolayısıyla doğrusal regresyon analiz yöntemlerinden elde edilen modelin güvenilir olmayacağı söylenebilir. EKK tahmin yönteminin uygulanabilirliği için bu çalışmada da bahsedilen gerekli olan varsayımların sağlanamadığı durumlarda Uzaklık Temelli Regresyon yöntemlerinin kullanılması önerilebilir.

Gelecekte yapılacak çalışmalarda Bray-Curtis, Orloci's Chord, Chi-square, Canberra ve Hellinger gibi diğer uzaklık ölçülerin değerlendirilmesi ve/veya farklı dağılışa sahip açıklayıcı değişken kombinasyonlarının incelenmesinin konu açısından yararlı olabileceği değerlendirilmiştir.

KAYNAKLAR

Adıgüzel, M. B. (2021). *Çok değişkenli uyarlanabilir regresyon eğrilerinde alternatif bilgi kriterleri ile model seçimi*. Basılmamış Doktora Tezi. Ondokuz Mayıs Üniversitesi Lisansüstü Eğitim Enstitüsü İstatistik Anabilim Dalı, 86, Samsun.

Aerts, J., M. Kolenda, D. Piwczynski, B. Sitkowska and H. Önder. (2022). Forecasting milking efficiency of dairy cows milked in an automatic milking system using the decision tree technique. *Animals*. 12. 1040.

Alma, G. Ö. ve Ö. Vupa. (2008). Regresyon analizinde kullanılan en küçük kareler ve en küçük medyan kareler yöntemlerinin karşılaştırılması. *SDÜ Fen Edebiyat Fakültesi Fen Dergisi*, 3 (2). 2019-229.

Alpar, R. (2010). *Basit doğrusal regresyon çözümlemesi: Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik*. Ankara: Detay Yayıncılık.

Alpar, R. (2013). *Uygulamalı çok değişkenli istatistiksel yöntemler*. Ankara: Detay Yayıncılık.

Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*. 26. 32-46.

Arenas, C. and C. M. Cuadras. (2002). Recent statistical methods based on distances. *Contributions to Science*. 2(2). 183-191.

Arı, A. ve H. Önder. (2013). Farklı veri yapılarında kullanılabilecek regresyon yöntemleri. *Anadolu Tarım Bilimleri Dergisi*. 28(3) 168-174.

Atkinson, A. C., M. Riani, and M. Riani. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Bayram, E., ve V. Nabiyev. (2020, Ekim). "Image segmentation by using K-means clustering algorithm in Euclidean and Mahalanobis distance calculation in camouflage images". *28th Signal Processing and Communications Applications Conference (SIU)*. Gaziantep.

Borg, I. and P. J. Groenen. (2005). *Modern multidimensional scaling: Theory and applications*. Berlin: Springer Science & Business Media.

Chen, J., K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, H. Li. (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*. 28(16). 2106-2113.

Cuadras CM. (1988) Statistical distances. *Estadística Española*, 30. 295-378.

Damodar, N. 2001. *Temel ekonometri*. İstanbul: Literatür Yayıncılık.

de Souza Jr A. H., F. Corona, G. A. Barreto, Y. Miche, and A. Lendase. (2015). Minimal learning machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing*. 164. 34-44.

Dinler, M. (2014). *Kümeleme analizi yöntemlerinin hayvancılık verilerinde karşılaştırılmalı olarak incelenmesi*. Basılmamış Yüksek Lisans Tezi. Bingöl Üniversitesi Fen bilimleri Enstitüsü Zootekni Anabilim Dalı, 89, Bingöl.

Doğan, İ. (2003). Kuzularda büyümenin çok boyutlu ölçekleme yöntemi ile değerlendirilmesi. *Uludag Univ. J. Fac. Vet. Med*., 22. 33-37.

Draper, N. R. and H. Smith. (1998). *Applied regression analysis*. New York: John Wiley & Sons.

Dutter, R. and Hubber P. J. (1981). Numerical methods for the nonlinear robust regression problem. *Journal of Statistical Computation and Simulation*. 13 (2). 79-113.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. 53(3-4). 325-338.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*. 857-871.

Kurnaz B, Önder H. (2021, June). "Distance Based Regression Models". *II. International Applied Statistics Conference (UYIK-2021).* Tokat.

Li, J., W. Zhang, S. Zhang and Q. Li. (2019). A theoretic study of a distance-based regression model. *Sci China Math*, 62. 979-998.

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, 188. 117-131.

McArdle, B. H. and M. J. Anderson. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82 (1). 290-297.

McQuarrie, A. D. and C. L. Tsai. (1998). *Regression and time series model selection*. London: World Scientific Publication Co Pte. Ltd.

Okur, S. (2009). *Parametrik ve parametrik olmayan doğrusal regresyon analiz yöntemlerinin karşılaştırılmalı olarak incelenmesi*. Basılmamış Yüksek Lisans Tezi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü Zootekni Anabilim Dalı, 62, Adana.

Orhunbilge, N. (2017). *Uygulamalı regresyon ve korelasyon analizi*. Ankara: Nobel Yayıncılık.

Önder, H. ve Abacı, S. H. (2015). Path analysis for body measurements on body weight of saanen kids. *Kafkas Univ Vet Fak Derg*., 21 (3). 351-354.

Önder, H. ve Mercan L. (2020). Comparison of Bray Curtis and Nei's genetic distance on Mantel test for chicken diversity data. *Black Sea Journal of Engineering and Science*, 3 (3). 76-80.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25. 111-163.

Rosner, B. (2015). *Fundamentals of biostatistics*. Boston: Harvard University.

Schork, N. J., J. Wessel, and N. Malo. (2008). DNA sequence-based phenotypic association analysis. *Advances in genetics*, 60. 195-217.

Servi, T. (2009). *Çok değişkenli karma dağılım modeline dayalı kümeleme analizi*. Basılmamış Doktora Tezi. Çukurova Üniversitesi Fen bilimleri Enstitüsü, İstatistik Anabilim Dalı, 266, Adana.

Shehzad, Z., C. Kelly, P. T. Reiss, R. C. Craddock, J. W. Emerson, K. McMahon, M. P. Milham. (2014). A multivariate distance-based analytic framework for connectome-wide association studies. *Neuroimage*. 93. 74-94.

Timm, N.H. (2002). *Applied Multivariate Analysis*. New York: Splinger-Verlag.

Ucal, M. Ş. (2006). Ekonometrik Model Seçim Kriterleri Üzerine Kisa Bir İnceleme. *CÜ İktisadi ve İdari Bilimler Fakültesi*, 7 (2). 41-57.

URL 1: https://acikders.ankara.edu.tr/pluginfile.php/130799/mod_resource/content/0/6-%20Matris.pdf (Erişim: 12.05.2023).

URL 3: https://avys.omu.edu.tr/storage/app/public/kamilal/108861/HPTZ.HF10.pdf (Erişim: 16.05.2023).

URL2: http://emredunder.blogspot.com/2011/06/kumeleme-analizinde-kullanlan-baz.html (Erişim: 18.06.2021).

Ünlükaplan, Y. (2008). *Çok değişkenli istatistiksel yöntemlerin peyzaj ekolojisi araştırmalarında kullanımı.* Basılmamış Doktora Tezi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü Peyzaj Mimarlığı Ana Bilim Dalı, 156, Adana.

Varoquaux, G., and R. C. Craddock. (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, 80. 405-415.

Vural, A. (2007). *Aykırı değerlerin regresyon modellerine etkileri ve sağlam kestiriciler*. Basılmamış Yüksek Lisans Tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı, 73, İstanbul.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44 (1). 92-107.

Weisberg, S. (2005). *Applied linear regression*. New Jersey: John Wiley & Sons.

Wessel, J. and N. J. Schork. (2006). Generalized genomic distance–based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79 (5). 792-806.

Xu, Y., X. Guo, J. Sun, Z. Zhao. (2015). Snowball: resampling combined with distance-based regression to discover transcriptional consequences of a driver mutation. *Bioinformatics*. 31(1). 84-93.

Yan, X. and X. Su. (2009). *Linear regression analysis: theory and computing*. New York: World Scientific.

Yıldız, M. (2022). Dolar ve Euro kurları üzerinde etkili faktörlerin iki bağımlı değişkenli MARS modeli ile belirlenmesi. *Kastamonu Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 24 (1). 6-29.

ÖZ GEÇMİŞ

Burcu Kurnaz, Samsun Bafra Kızılırmak Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Fen-Edebiyat Fakültesi, İstatistik bölümünden 2018 yılında mezun oldu. 2020 yılında OMÜ LEE Zootekni Anabilim Dalı Yüksek Lisans programına girdi. Orta derecede İngilizce bilmektedir (16.06.2023).

İletişim Bilgileri

ORCID ID : 0000-0001-5613-6992

Yayınlar:

1. Kurnaz B, Önder H, Piwczynski D, Kolenda M, Sitkowska B. (2021). Determination of the Best Model to Predict Milk Dry Matter in High Milk Yielding Dairy Cattle. *Acta Sci. Pol. Zootechnica.* 20 (3). 41–44. DOI:10.21005/asp.2021.20.3.05. (Uluslararası dergi)

2. Kurnaz B, Yüksel HM, Önder H, Tırınk C. (2022). 3-D Classification of agricultural areas of Turkey using mammalian livestock existence. *BSJ Agri.* 5 (3). 311-313. DOI: 10.47115/bsagriculture.1116612. (TR Dizin)

3. Kurnaz B, Önder H. (2021, June). "Distance Based Regression Models". *II. International Applied Statistics Conference (UYIK-2021).* Tokat

4. Kurnaz B, Önder H, Piwczynski D, Kolenda M, Sitkowska B. (2021, October). "Determination of the Best Model to Predict Milk Dry Matter in High Milk Yielding Dairy Cattle". *Acta Scientiarum Polonorum Zootechnica.* October 14, 2021, Szczecin.

5. Kurnaz B, Önder H. (2022, October). "Estimating the Significance of Pearson Correlation Coefficient by Permutation Test". *VI. International Congress on Animal Breeding, Genetics and Husbandry (ICABGEH-22).* Samsun.

6. Kurnaz B, Önder H. (2022), June). "Comparison of Liu, Ridge and Least Square Estimators on Relative Feed Value Estimation". *III. International Applied Statistics Conference (UYIK-2022).* Skopje.

7. Kolenda M, Sitkowska B, Önder H, Piwczynski D, Kurnaz B, Şen U. (2022, September). "Single step genomic prediction of milk yield in Polish Holstein Friesian dairy cattle". *EcoSET Closing Conference.* Bydgoszcz.

8. Sitkowska B, Kolenda M, Piwczynski D, Önder, H, Kurnaz B, Şen U. (2022, September). "Polymorphism that change frequencies of genotypes affecting milk yield". *EcoSET Closing Conference.* Bydgoszcz.

Yurtdışı Deneyimleri

1. EcoSET projesi kapsamında 16-30 Ekim 2021 tarihleri aralığında Bydgoszcz University of Science and Technology, Faculty of Animal Breeding and Biology