İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 17, 2024, 2, 30-45

Geliş/Received:09.07.2024, **Kabul**/Accepted: 16.10.2024

Araştırma Makalesi / Research Article

Ridge-Robust-Boosting topluluk regresyon yaklaşımı

Ayşegül Han[1]

İnönü Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümü, Malatya, Türkiye
aysegullhann@gmail.com
ORCID: 0000-0002-3390-2129

Mehmet Güngör

İnönü Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Ekonometri Bölümü, Malatya, Türkiye
m.gungor@inonu.edu.tr
ORCID: 0000-0001-6869-4043

Öz

Çalışmanın amacı, regresyon analizinde karşılaşılan çoklu bağlantı ve aykırı değer sorunlarına aynı anda çözüm getirebilen bir regresyon modeli geliştirmektir. Önerilen Ridge-Robust-Boosting Topluluk Regresyon modeli, Ridge regresyonunu kullanarak çoklu bağlantıyı azaltmakta ve böylece bağımsız değişkenler arasındaki korelasyonu dengelemektedir. Ayrıca, Sağlam regresyonu kullanarak aykırı değerlere karşı dirençli olmayı hedeflemektedir. Bu sayede, nadir ancak etkili gözlemlerin tahminler üzerindeki etkisini azaltmaktadır. Ayrıca, Boosting yöntemleri kullanılarak tahmin edicinin başarımı arttırılmıştır.

Anahtar sözcükler: Çoklu doğrusal bağlantı, aykırı değer, Ridge-Robust-Boosting regresyon.

Abstract

Ridge-Robust-Boosting Ensemble Regression Approach

The aim of the study is to develop a regression model that can simultaneously solve the multicollinearity and outlier problems encountered in regression analysis. The proposed Ridge-Robust-Boosting Ensemble regression model reduces multicollinearity by using Ridge regression, thus stabilizing the correlation between independent variables. It also aims to be robust to outliers by using robust regression. This reduces the impact of rare but influential observations on the forecasts. Furthermore, the performance of the estimator is improved by using boosting methods.

Keywords: *Multicollinearity, outlier, Ridge-Robust-Boosting regression.*

[1] Bu çalışma Prof. Dr. Mehmet GÜNGÖR danışmanlığında Ayşegül HAN tarafından hazırlanan ve İnönü Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalında 2023 yılında sunulan "Çoklu Doğrusal Bağlantı ve Aykırı Değer Sorunu için Ridge-Robust-Boosting Topluluk Regresyon Yaklaşımı" başlıklı doktora tezinden türetilmiştir.

1. Giriş

Doğrusal regresyon modeli genel olarak bir bağımlı değişken ile bir dizi bağımsız değişken arasındaki ilişkiyi değerlendirmektedir. Regresyon katsayılarını tahmin etmek için yaygın olarak en küçük kareler (EKK) tahmin edicisi kullanılmaktadır. Bu tahmin edici, EKK ile en iyi tahminleri elde edebilmek için bazı varsayımların sağlanması gerekmektedir. Bu varsayımlar arasında doğrusallık, hata terimlerinin normal bir dağılıma ve homojen varyansa sahip olması, hata terimlerinin bağımsız olması, bağımsız değişkenler arasında çoklu bağlantı olmaması ve hata terimlerinin ortalamasının sıfır olması yer almaktadır.

Doğrusal regresyon modelinin güvenilirliğini etkileyebilecek iki önemli konu aykırı değerler ve çoklu bağlantı problemidir. Aykırı değerler, genel eğilimden belirgin bir şekilde saparak diğer gözlemlerden önemli ölçüde farklı olan veri noktalarıdır. Bu soruna çözüm olarak, M-tahmin edicisi gibi sağlam regresyon tahmin edicileri geliştirilmiştir. Bu tahmin ediciler, aykırı değerlere karşı daha dayanıklıdır [1]. Diğer önemli bir konu olan çoklu bağlantı ise bağımsız değişkenler arasında yüksek bir korelasyon olduğu durumunu ifade etmektedir. Bu durumda, bir bağımsız değişken diğer bağımsız değişken(ler) tarafından yaklaşık olarak ifade edilebilir. Bu durum, regresyon modelini daha karmaşık hale getirir ve bu karmaşıklık, regresyon parametre tahminlerinin varyansını artırarak en küçük kareler tahmin edicisini güvenilmez hale getirebilir [2]. Bu sorunun çözümü için Hoerl ve Kennard [3] tarafından Ridge Regresyon tahmin edicisi geliştirilmiştir. Ridge Regresyon, y yönündeki aykırı değerlerden etkilendiği için, Silvapulle [4] bu sorunu çözmek amacıyla Sağlam Ridge Regresyon tahmin edicisini tanıtmıştır. Liu [5], çoklu bağlantı etkilerini azaltmak için Liu tahmin edicisi olarak adlandırılan başka bir tahmin edici geliştirmiştir. Liu tahmin edicisi de y yönündeki aykırı değerlere karşı hassas olduğu için Arslan ve Billor [6] Sağlam Liu tahmin edicisini öne sürmüştür. Alternatif olarak, Özkale ve Kaçıranlar [7] çoklu bağlantı sorununu çözmek için β'nın iki parametreli tahmin edicisini önermişlerdir. Ancak, β'nın iki parametreli tahmin edicisinin de y yönündeki aykırı değerlere karşı hassas olduğu görülmüştür.

Khalaf ve Shukur [8], geliştirdikleri tahmin edicinin, özellikle yüksek hata varyanslarında, Hoerl ve Kennard [3] tarafından önerilen tahmin ediciden hemen hemen tüm durumlarda daha üstün olduğunu simülasyon çalışmalarıyla göstermiştir. Alkhamisi vd. [9] ise, Khalaf ve Shukur'un [8] tahmin edicisini modifiye ederek dört yeni Ridge parametre tahmin edicisi önermiş ve bu tahmin edicilerden en az birinin, EKK ve diğer tahmin edicilere kıyasla çoğu durumda daha iyi performans sergilediğini ortaya koymuştur. Al-Hassan [10], Hocking vd.'nin [11] Ridge tahmin edicisini, Alkhamisi ve Shukur'un [12] modifikasyonunu uygulayarak geliştirdiği tahmin edicinin, EKK tahmin edicisi ve diğer ilgili tahmin edicilere göre belirli koşullarda daha etkili olduğunu Monte Carlo simülasyonları ile göstermiştir. Muniz vd. [13], kareköklü dönüşüm yöntemini Khalaf ve Shukur'un [8] tahmin edicisine uygulayarak beş yeni Ridge parametre tahmin edicisi geliştirmiş ve bazı durumlarda daha iyi performans gösterdiklerini bulmuştur. Asar vd. [14], beş yeni Ridge parametre tahmin edicisini simülasyon ve gerçek veri ile altı farklı tahmin ediciyle karşılaştırmış ve önerilen tahmincilerin daha iyi performans gösterdiğini ancak gerçek veri üzerinde yeterli sonuç vermediğini belirtmiştir. Dorugade [15], Ridge parametre tahmin edicisini EKK ile tahmin edilen regresyon modelinin standart sapması olarak belirlemiş ve simülasyonlarda diğer tahmincilerden daha etkili olduğunu göstermiştir. Lukman ve Olatunji [16], Dorugade'nin [15] tahmin edicisini modifiye ederek yeni bir tahmin edici önermiş ve bu tahmin edicinin diğerlerinden daha iyi performans gösterdiğini bulmuştur. Bhat [17], yeni bir Ridge parametre tahmin edici önermiş ve önerilen tahmincinin bazı popüler tahmincilerden daha etkin ve durağan olduğunu belirtmiştir. Lattef ve Alheety [18], beş farklı Ridge parametre tahmin edicisini ve EKK yöntemini karşılaştırmış ve önerilen tahmincilerin EKK ve diğer tahmincilerden üstün olduğunu bulmuştur. Qasim vd. [19], beta regresyon modeli için yeni bir β Ridge regresyon tahmin yöntemi önermiş ve bu yöntemle bazı mevcut tahmincilerden daha iyi performans elde edilmiştir. Irandoukht [20], Ridge parametresini tahmin etmek için belirlilik katsayısının maksimize edilmesini öneren yeni bir yaklaşım geliştirmiş ve önemli tahmin gücü iyileşmeleri sağlamıştır. Khalaf [21], Hoerl ve Kennard [3] tarafından önerilen tahmin edicinin modifikasyonuna dayalı yeni bir Ridge parametre tahmin edicisi önermiş ve hata kareler ortalaması bakımından üstün olduğunu bulmuştur. Shabbir vd. [22], regresyon modelinin standart hatası ve bağımsız değişken sayısına dayalı yeni bir Ridge parametre tahmin edicisi önermiş ve bu tahmincinin EKK ve diğer tahmincilerden daha iyi olduğunu

göstermiştir. Shaheen vd. [23], üç yeni Ridge parametre tahmin edicisi önermiş ve bu tahmincilerin genellikle hata kareler ortalaması bakımından daha yüksek performans gösterdiğini belirlemiştir.

Literatür incelemesi, mevcut tahmin edicilerin regresyon analizinde istenilen güvenilirliği sağlamada yetersiz kaldıklarını göstermektedir. Bu bağlamda Han [24] bir topluluk regresyon modeli önermiştir. Bu model, Ridge Regresyon, Sağlam Regresyon ve Gradient Boosting Regresyon modellerinin avantajlarını birleştirerek, çoklu bağlantı etkilerini azaltma, aykırı değerlere karşı direnç gösterme ve regresyon parametrelerinin güvenilirliğini artırma konularında geliştirilmiş özellikler sunmaktadır. Önerilen modelin, regresyon analizindeki temel zorluklara karşı daha sağlam bir çözüm sunma potansiyeline sahip olduğu değerlendirilmektedir. Bu kapsamda çalışmanın amacı, Han [24] tarafından önerilen topluluk regresyon modelinin etkinliğini simülasyon çalışmaları ile test etmek ve bu modelin çoklu bağlantı ve aykırı değer sorunlarına karşı performansını incelemektir. Ayrıca, önerilen modelin literatürdeki diğer tahmin edicilerle karşılaştırmalı olarak üstünlüklerini ortaya koymak hedeflenmektedir.

Çalışmanın giriş bölümünde, regresyon analizindeki temel sorunlara vurgu yapılarak literatürde sıkça karşılaşılan çoklu bağlantı ve aykırı değer sorunlarına odaklanılmış ve çalışmanın amacı açıklanmıştır. Çalışmanın ikinci bölümünde regresyon analizindeki çoklu bağlantı ve aykırı değer sorunları ele alınmış, bu durumlar için önerilen regresyon modelleri incelenmiştir. Çalışmanın üçüncü bölümünde ise topluluk modeli yaklaşımı ve önerilen modelin oluşturulma süreci detaylı bir şekilde açıklanmıştır. Dördüncü bölümde simülasyon çalışmasının ayrıntılarına yer verilmiştir. Beşinci bölümde simülasyon çalışmasıyla elde edilen bulgular paylaşılmıştır. Sonuç bölümü ise genel değerlendirmeleri içermektedir.

2. Regresyon Modelleri

Bu başlık altında, doğrusal regresyon, Ridge regresyon, Sağlam regresyon ve Gradient Boosting regresyon yöntemlerinin açıklamaları sunulmaktadır.

2.1. Doğrusal regresyon

Doğrusal regresyon denklemi, bağımlı değişkeni (Y) ve bağımsız değişkenleri $(X_1, X_2, \dots, X_n)$ ilişkilendiren bir denklemle aşağıdaki gibi ifade edilmektedir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

Burada, Y bağımlı değişkeni, $X_1, X_2, \dots, X_n$ bağımsız değişkenleri, β_0 sabit terimi, $\beta_1, \beta_2, \dots, \beta_n$ ise bağımsız değişkenlerin katsayılarını ve ε ise hata terimini ifade etmektedir.

Regresyon analizi, veri bilimi alanında güçlü bir araç olmasına rağmen, doğru tahminler yapabilmesi için belirli önemli varsayımlar altında çalışmaktadır. Bu varsayımlardan çoklu bağlantı, regresyon analizinde bağımsız değişkenler arasındaki yüksek korelasyon nedeniyle ortaya çıkmaktadır. Bu durumda, hangi değişkenin gerçekten tahmin üzerinde etkili olduğunu belirlemek zor hale gelmektedir. Çoklu bağlantı, tahmin edicilerin istikrarını ve güvenilirliğini azaltabilir.

2.1. Ridge regresyon

Çoklu bağlantı sorunuyla başa çıkmak için Hoerl ve Kennard [3] Ridge regresyon yöntemini önermiştir. Bu yöntem korelasyonları kontrol etmek amacıyla L2 cezalandırma tekniği kullanmakta ve bu sayede doğru tahminler yapmayı mümkün kılmaktadır [25].

Ridge regresyon denklemini Eşitlik 2'deki gibi göstermek mümkündür [26]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \lambda\Sigma\left(\beta_i^2\right) + \varepsilon \quad (2)$$

Burada; $i = 1, 2, \ldots, n$'dir. $\lambda \geq 0$ ayar parametresidir. Ayar parametresi λ, regresyon katsayıları üzerindeki düzenleme teriminin etkisini kontrol ederek, modelin aşırı uyum eğilimini düzenlemektedir.

2.2. Sağlam regresyon

Regresyon analizinde aykırı değerler, genel veri kümesinden belirgin şekilde sapmış veya diğer gözlemlerden önemli ölçüde farklı olan veri noktalarıdır. Bu noktalar, istatistiksel analizlerin güvenilirliğini azaltabilir ve regresyon gibi modellerde yanlış tahminlere neden olabilir. Bu sorunun giderilmesi amacıyla aykırı değerlere karşı dirençli olan Sağlam Regresyon teknikleri geliştirilmiştir. Bunlar arasında En Küçük Medyan Kareler (Least Median Squares-LMS), Theil-Sen, En Küçük Mutlak Sapma (Least Absolute Deviation-LAD), M Kestiricisi ve En Az Kırpılmış Mutlak Değer (Least Trimmed Absolute-LTA) gibi çeşitli yöntemler ya da tahmin ediciler bulunmaktadır. Bu tahmin ediciler, modelin doğruluğunu artırmak için aykırı değerlere daha az ağırlık vererek daha doğru tahminler yapmayı sağlarlar.

2.2.1. LAD tahmin edicisi

LAD tahmin edicisi, gözlenen ve tahmin edilen bağımlı değişken arasındaki mutlak artıkları minimize ederek regresyon katsayılarını bulmaktadır. Bu amacı gerçekleştirmek için Eşitlik 3 kullanılmaktadır [27]:

$$\min_{\beta} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right| \tag{3}$$

Burada; $i = 1, \ldots, n$ ve $j = 0, \ldots, p-1$'dir.

2.2.2. Theil-Sen tahmin edicisi

Theil-Sen tahmin edicisi, Theil'in [28] önermiş olduğu Theil tahmin edicisi Sen [29] tarafından Theil-Sen tahmin edicisi olarak genişletilmiştir. Bu tahmin edici, veri noktaları arasındaki eğilimi belirlemek amacıyla veri noktası çiftlerinin eğimlerini hesaplayarak bu eğimlerin medyan değerini almaktadır. Bu durum matematiksel olarak Eşitlik 4 ile belirtildiği gibi gösterilmektedir [29]:

$$medyan\left(\frac{y_i - y_j}{x_i - x_j}\right) \tag{4}$$

Burada, i ve j olası veri noktası çiftlerinin tamamını içeren indekslerdir.

2.2.3. LMS tahmin edicisi

LMS tahmin edicisi, etkili gözlemlere ve aykırı değerlere karşı hassas olan klasik yöntemlere alternatif olarak Rousseeuw [30] tarafından önerilmiştir. Bu yöntem, doğrusal regresyon modelinin katsayılarını belirlemek amacıyla kullanılmaktadır. Eşitlik 5 ile LMS tahminleri elde edilmektedir [31]:

$$\min_{\beta} \operatorname*{med}_{i} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2 \tag{5}$$

Burada; β_j ise regresyon katsayısını belirtmektedir.

2.2.4. LTA tahmin edicisi

LTA tahmin edicisinin amacı, en büyük hatalara sahip olan gözlemlerin belirli bir yüzdesini dışarıda bıraktıktan sonra, geriye kalan hataların mutlak değerlerinin toplamını en aza indirmektir. LTA özellikle büyük veri kümelerinde iyi bir alternatif oluşturmaktadır. LTA tahmin edicisi için Eşitlik 6 kullanılmaktadır [32]:

$$\min_{\beta} \sum_{i=1}^{k} |r|_{i:n} \quad (6)$$

Burada, $|r|_{1:n} \leq |r|_{2:n} \leq \cdots \leq |r|_{n:n}$ kalıntıların mutlak değerlerini ve k toplamdaki mutlak kalıntı değerlerinin sayısını belirtmektedir. Kırpılacak gözlem sayısı k, genellikle toplam gözlem sayısının %10'u ya da %20'si olacak biçimde-ayarlanmaktadır.

2.2.5. M kestiricisi

M kestiricisi, aykırı değerlerin etkisini azaltmak için ağırlıklandırma fonksiyonu kullanan bir tür sağlam regresyon olup, Huber [33] tarafından sıradan en küçük karelere (OLS) bir alternatif olarak geliştirilmiştir. Bu yöntemin amacı, gözlenen ve tahmin edilen değişkenler arasındaki mutlak artıkların ağırlıklandırılmış toplamını minimize etmektir. M-regresyon tahminleri için Eşitlik 7 kullanılmaktadır [34]:

$$\min_{\beta} \{\sum_{i=1}^{n} w_i(|y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|)\} \quad (7)$$

Burada; w_i, i. gözleme atanan ağırlıktır. Her bir gözleme atanan ağırlık, aykırı gözlemlerin etkisini minimize etmek için seçilen bir ağırlıklandırma fonksiyonuna bağlı olarak belirlenir. Sıkça kullanılan bir ağırlıklandırma fonksiyonu şu şekildedir [34]:

$$w_i = \begin{cases} (1 - (r_i/c)^2)^2 & r_i \leq c \; ise \\ 0 & r_i > c \; ise \end{cases} \quad (8)$$

Burada; r_i, i. gözlem için standartlaştırılmış artık ve c tahmin edicinin sağlamlık derecesini tespit eden ayarlama parametresidir. Standartlaştırılmış artık $r_i = (y_i - \hat{y_i})/s$ olarak tanımlanmaktadır; s gözlemler arasındaki standart sapmayı göstermektedir.

2.2.6. Minimum Hacimli Elipsoid (Minimum Volume Ellipsoid-MVE)

Elipsoid, Eşitlik 9 ile ifade edilen formülle tanımlanabilir [35]:

$$(x - \mu)^T C^{-1} (x - \mu) \leq \chi^2 \quad (9)$$

Burada, x veri noktası, μ elipsoidin merkezi, C elipsoidin kovaryans matrisidir. Elipsoidin hacmi, kovaryans matrisinin determinantı ile hesaplanmaktadır [35]:

$$V = \frac{(2\pi)^{p/2}}{\Gamma(p/2)} \sqrt{\det(C)} \quad (10)$$

Burada, V elipsoidin hacmini, Γ gama fonksiyonunu, p veri setindeki özellik sayısını ve $\det(C)$ kovaryans matrisinin determinantını ifade etmektedir. MVE metodu ile elipsoidin hacminin minimize edilmesi amaçlanırken, bu hacmin veri kümesindeki çoğu veri noktasını kapsayacak şekilde olması gerekmektedir.

2.3. Gradient Boosting regresyon

Geleneksel regresyon yöntemleri ve sağlam regresyon tekniklerinin yanı sıra, veri analizi ve tahminde önemli bir rol oynayan bir diğer güçlü yöntem Gradient Boosting algoritmalarıdır. Bu algoritmalar, zayıf tahmin edicileri bir araya getirerek güçlü bir tahmin edici oluşturmaktadır. Her bir zayıf tahmin edici, modelin önceki hatalarını düzeltmeye odaklanarak genel tahmin performansını artırmaktadır. İlk olarak Breiman [36] tarafından tanıtılan bu yöntem, uygun bir kayıp fonksiyonuyla optimizasyon yöntemi şeklinde değerlendirilebileceği belirtilmiştir. Ardından, Friedman [37] tarafından bu algoritmanın daha gelişmiş bir versiyonu oluşturulmuştur. Algoritmanın öğrenme süreci, sağlam bir sınıflandırıcı belirlemek için yeni modellerin ardışık şekilde eğitilmesi olarak belirlenmiştir [38].

$S = \{x_i, y_i\}_1^N$ şeklinde bir eğitim seti verildiğinde, Gradient Boosting, kayıp fonksiyonu $L(y, F(x))$'i minimize ederek, x tahmin değişkenlerini kullanıp y bağımlı değişkenlerini bulmayı amaçlamaktadır. Eşitlik 11'de gösterildiği üzere, Gradient Boosting, ağırlıklı bir fonksiyon toplamıyla $F(x)$'in eklemeli bir yaklaşımını oluşturmaktadır:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x) \tag{11}$$

Burada; ρ_m, yeni adımın (m. adım) öğrenme oranını (learning rate) belirtmektedir. Öğrenme oranı, yeni adımın önemini ayarlamak için kullanılmaktadır. Bu değer, $[0, 1]$ aralığındadır ve yeni adımın katkısının ne kadar olacağını belirlemektedir. $h_m(x)$, m. adımda eklenen yeni tahmin fonksiyonunu temsil etmektedir. Bu fonksiyonlar topluluktaki karar ağacı modelleridir. Algoritma yaklaşımı yinelemeli şekilde Eşitlik 12'de gösterildiği gibi gerçekleştirilmektedir:

$$F_0(x) = argmin_\alpha \sum_{i=1}^N L(y_i, \alpha) \tag{12}$$

Eşitlik 13'te gösterildiği gibi ardışık temel öğrenicileri minimize etmeyi hedefler:

$$(\rho_m h_m(x)) = argmin_{\rho,h} \sum_{i=1}^N L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big) \tag{13}$$

Her h_m yeni bir eğitim seti $D = \{x_i, r_{mi}\}_{1=1}^N$ ile eğitilmektedir. Burada, ρ her bir iterasyonda ilave edilen tahmini fonksiyonun ağırlığını, r_{mi} kalıntıları belirtmektedir [39]. Kalıntılar Eşitlik 14 ile gösterildiği gibi hesaplanmaktadır:

$$r_{mi} = \left[\frac{\delta(y_i, F(x))}{\delta F(x)}\right]_{F_m(x)=F_{m-1}(x)} \tag{14}$$

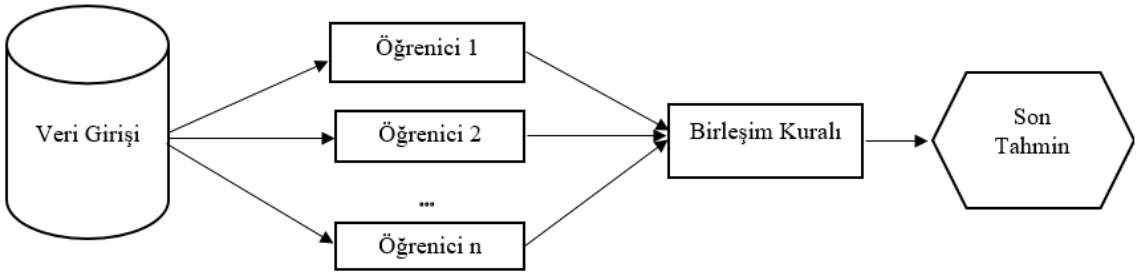
Kalıntıların hesaplanmasının ardından, ρ_m değeri bir çizgi arama optimizasyonu gerçekleştirilerek elde edilir. Yinelemeli görev uygun şekilde düzenlenmezse bu algoritma aşırı uyum sağlayabilir [40]. İkinci dereceden kayıp fonksiyonu gibi belirli kayıp fonksiyonları için, h_m yanlış kalıntılara mükemmel bir şekilde uyarsa, sonraki iterasyonda yanlış kalıntılar sıfır olmakta ve iterasyon erken sona ermektedir.

Bu temel regresyon modellerinin incelenmesinin ardından, literatürde sıkça rastlanan çoklu bağlantı ve aykırı değer sorunlarına karşı etkili bir çözüm sunmak amacıyla Ridge-Robust-Boosting (RRB) topluluk regresyon modeli Han [24] tarafından önerilmiştir. Ridge regresyonun düzenleme yeteneği, Sağlam regresyonun aykırı değerlere direnç gösteren yapısı ve Gradient Boosting regresyonunun tahmin performansındaki güçlü yanları, bu modelde entegre edilerek daha güçlü bir regresyon yaklaşımı oluşturulmuştur.

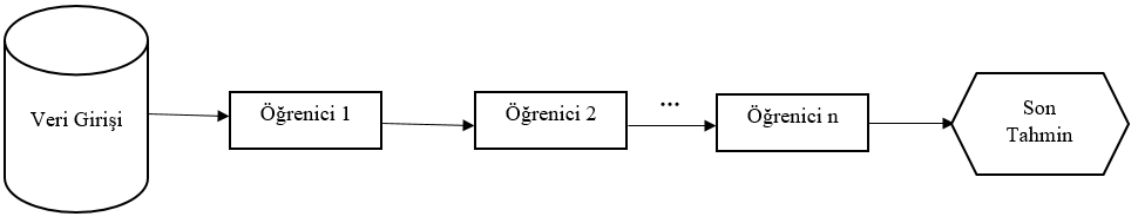
3. Topluluk Modeli Yaklaşımı

Topluluk modeli yaklaşımının temel prensibi, çeşitli öğrenme algoritmalarının verileri farklı bakış açılarıyla işlemesini sağlamak ve bu çeşitliliği birleştirerek daha etkili sonuçlar elde etmektir. Bu yaklaşımın kökeni, Dasarathy ve Sheela [41] tarafından önerilen çoklu bileşen sınıflandırıcılarına dayanmaktadır. Daha sonraki dönemde, Hansen ve Salamon [42], benzer bir yaklaşımın sınıflandırma problemlerinde tek bir sınıflandırıcıdan daha iyi performans sağladığını göstermişlerdir. Aynı zamanda, Schapire [43] boosting tekniğini geliştirerek zayıf sınıflandırıcıları güçlü bir sınıflandırıcıya dönüştürmüştür. Bu teknik, günümüzün güçlü öğrenme algoritmalarının temelini oluşturan AdaBoost, Gradient Boosting ve XGBoost gibi yöntemleri ortaya çıkarmıştır.

Topluluk modeli yaklaşımı, birden fazla makine öğrenimi algoritmasını birleştirerek daha etkili ve güçlü tahminler elde etmeyi amaçlayan bir öğrenme yaklaşımıdır. Bu teknik, farklı öğrenme algoritmalarının çeşitli bakış açılarından veriyi işlemesini ve bu farklı bakış açılarını birleştirerek daha güvenilir sonuçlar elde etmeyi hedeflemektedir. Topluluk modelleri, paralel ve sıralı topluluklar olarak sınıflandırılmaktadır. Paralel topluluklar, farklı temel sınıflandırıcıları bağımsız olarak eğitmekte ve tahminlerini bir birleştirici kullanarak birleştirmektedir [44]. Paralel topluluk algoritmaları, temel öğrenicilerin paralel olarak üretilmesini kullanarak topluluk üyeleri arasında çeşitliliği teşvik etmektedir. Buna karşılık, sıralı topluluklar temel modellere bağımsız olarak uyum sağlamaz. Bunun yerine, modeller yinelemeli olarak eğitilip her yinelemede bir önceki modelin hatalarını düzeltmeyi öğrenir. Şekil 1 ve Şekil 2'de paralel ve sıralı topluluk modelini gösteren diyagramlar bulunmaktadır [45].



Şekil 1. Paralel Topluluk Modeli Diyagramı



Şekil 2. Sıralı Topluluk Modeli Diyagramı

Bu noktadan hareketle, Han [24] tarafından önerilen RRB modeli, paralel topluluk modeli yöntemi olarak değerlendirilmektedir. RRB modeli ile Ridge regresyon, Sağlam regresyon ve Gradient Boosting regresyon yöntemleri paralel bir şekilde entegre edilerek, her bir algoritmanın güçlü yönleri birleştirilip daha güçlü ve etkili tahminler elde etmek amaçlanmıştır. Bu yaklaşım, regresyon analizindeki temel zorluklara karşı daha dirençli ve güvenilir bir çözüm sunma potansiyeli taşımaktadır.

RRB modeli, Ridge regresyonunun çoklu bağlantı sorununa karşı etkili çözümü, Sağlam regresyonunun aykırı değerlere karşı direnci ve Gradient Boosting regresyonunun karmaşık ilişkileri başarıyla modelleme yeteneğini bir araya getirmektedir. Sonuç olarak, bu model, her bir algoritmanın güçlü yönlerini kullanarak daha yüksek tahmin performansı, daha iyi genelleme yeteneği, esneklik ve güvenilirlik sağlamaktadır. Bu özellikleriyle, regresyon analizinde karşılaşılan çeşitli zorluklara karşı dirençli ve etkili bir çözüm sunma potansiyeline sahiptir.

4. Simülasyon Çalışması

RRB modelinin simülasyon aşamalarına ilişkin açıklamalar aşağıdaki gibidir:

Verilerin Oluşturulması: Monte Carlo Simülasyonunun Gibbs Algoritması ile 1000 gözlemden oluşan veri seti üretildi.

- Veri setine düşük (0.3), orta (0.6) ve yüksek (0.9) düzeyde çoklu bağlantı ve %20, %30 ve %40 oranlarında aykırı değerler eklendi.

Geman ve Geman [46] tarafından geliştirilen Gibbs Örnekleme Algoritması, çok boyutlu problemlerde sıklıkla kullanılan Markov Zinciri Monte Carlo (MCMC) algoritmasıdır. Özellikle çok boyutlu veri analizinde ve istatistiksel çıkarımlarda sıklıkla tercih edilmektedir. Algoritmanın temel prensibi, pek çok parametrenin yer aldığı karmaşık bir ortak olasılık dağılımından doğrudan örnek almak yerine, daha düşük boyutlu koşullu dağılımlardan örnek çekmektir. Bu nedenle, Gibbs Örnekleme Algoritması, diğer değişkenler sabit tutulduğunda bir değişkenin koşullu dağılımından örnek çekmeyi içermektedir.

Eğitim ve Test Serisi: Veri seti %80 eğitim ve %20 test serisi olarak bölündü.

- Eğitim serisi, modelin eğitildiği ve öğrendiği veri bölümünü temsil etmektedir. Bu bölüm, modelin içsel yapısını ve özelliklerini anlaması için ele alınmaktadır. Eğitim verileri, modelin parametrelerini tespit etmek ve ilişkileri öğrenmek amacıyla tercih edilmektedir. Model, eğitim verilerine göre ayarlanıp optimize edilmektedir.
- Test serisi ise eğitilen modelin performansını sınamak ve genelleme yeteneğini test etmek için kullanılmaktadır. Bu bölüm, modelin daha önce görmediği verileri içerir ve modelin bu verilere nasıl tepki verdiğini ölçmektedir. Test verileri, modelin gerçek verilerle ne derece iyi çalıştığını değerlendirmektedir.

Hiperparametreleri Ayarlama: Grid Search (Izgara Arama) yöntemi ile en iyi hiperparametreler tespit edilmeye çalışıldı.

Grid Search yöntemi, belirli bir aralıkta yer alan hiperparametrelerin tüm olası kombinasyonlarını test ederek, en iyi performansı veren hiperparametreleri tespit etmektedir. Bu, manuel deneme yanılma sürecini en düşük seviyeye indirerek, modelin optimize edilmiş haliyle elde edilmesini sağlamaktadır.

- Ridge Regresyon → ***alpha*** (α) ***değeri:*** Düzenleme seviyesini kontrol etmektedir ve 0'a yaklaştıkça düzenleme etkisi azalır. Böylece her özellik için bir miktar düzenleme uygulayarak aşırı uyum riskini azaltır. ***lambda*** (λ) ***değeri:*** Düzenlemenin miktarını kontrol etmektedir. Değeri ne kadar büyükse, regresyon katsayıları da o kadar kısıtlanmaktadır. Grid Search ile birden çok λ değeri denemesi, modelin aşırı uyum (overfitting) ve düşük uyum (underfitting) dengesini bulmasına olanak tanımaktadır.
- Sağlam Regresyon → ***method değeri:*** (Minimum Volume Elipsoid) Veri noktalarını kapsayan ve hacmi mümkün olduğu kadar küçük olan bir elipsoid çizimi ile aykırı değerlere dirençli model meydana getirmektedir.
- Gradient Boosting Regresyon → ***n.trees değeri:*** Gradient boosting ağaçlarının sayısını tespit etmektedir. Daha çok ağaç, modelin kompleks yapısını artırabilir, ama aynı zamanda aşırı uyuma neden olabilir. Grid aramasıyla çeşitli ağaç sayılarını test etmek, modelin en iyi performansının tespitine yardımcı olmaktadır. ***interaction.depth değeri:*** Her ağaç için en çok kaç düğüm (node) katmanına sahip olunacağını tespit etmektedir. Daha derin ağaçlar, veriyi daha detaylı olarak öğrenebilir, ama ayrıca aşırı uyuma neden olabilir. Grid aramasıyla çeşitli derinlik değerlerini sınamak önem taşımaktadır. ***shrinkage değeri:*** Her ağacın katkısını düzenlemeye destek sağlayacak bir faktörün kontrolünü sağlamaktadır. Daha düşük bir shrinkage değeri, her ağacın daha az ağırlığa sahip olacağı anlamına gelmektedir. Düşük shrinkage, daha çok ağaç kullanmanın etkisini dengelemeyi sağlamaktadır. ***n.minobsinnode değeri:*** Bir düğümde en az kaç gözlem olması gerektiğini ifade etmektedir. Düşük değerler modelin detayları öğrenmesine yardımcı olabilir, ama aynı zamanda aşırı uyuma neden olabilir. Bu nedenle optimal değeri belirlemek amacıyla, çeşitli değerleri içeren bir Grid araması yapılırken, çapraz doğrulama kullanılarak her bir değerin model performansı üzerindeki etkisi değerlendirilmiştir. Bu yöntem, modelin genelleme yeteneğini en iyi şekilde değerlendirmek ve aşırı uyumu kontrol altında tutmak için optimal değeri belirlemek amacıyla kullanışlı bir yaklaşımdır.

Modellerin Eğitimi ve Tahmin: Eğitim aşamasında ve model tahmininde çapraz doğrulama tercih edildi.

Çapraz doğrulama, makine öğrenimi modelinin performansını değerlendirip genelleme yeteneğini tahmin etmek için tercih edilmektedir. Bu yöntem, veri setini k parçaya bölerek her bir parçayı sırayla test verisi olarak kullanıp diğerlerini eğitim verisi olarak kullanmaktadır. Bu işlem, k defa tekrarlanarak modellerin güvenilir bir şekilde değerlendirilmesine olanak tanımaktadır.

Model Ağırlıklarını Elde Etme:

- İlk olarak, her katlamada (fold) Ridge Regresyon, Sağlam Regresyon ve Gradient Boosting Regresyon modellerinin ayrı ayrı hesaplanan ortalama karesel hatası (MSE) belirlenir.
- Modellerin tahminleri ile toplam karesel hata (total_mse) hesaplanır.
- Her modelin ağırlığını hesaplamak için, modelin tahmininin MSE değerinin total_mse değerine oranı alınarak bu değer 1'den çıkarılıp her model için ağırlık faktörü hesaplanmaktadır.
- Ağırlık faktörleri her katlama için ayrı ayrı bir diziye (weights) kaydedilir.
- Son olarak katlamaların hepsinin ağırlık faktörlerinin ortalaması alınarak, final_weights adlı bir vektörde bir araya getirilir. Bu, her modelin genel ağırlığını göstermektedir.

Bu adımlar, k-fold çapraz doğrulama ile elde edilen tahminlerin en iyi ağırlık kombinasyonunu belirlemeye çalışmaktadır. Böylece, her modelin katkısını dengeleyip daha iyi bir tahmin elde edilmesine yardımcı olmaktadır.

Ridge-Robust-Boosting Topluluk Regresyon Modeli Oluşturma: Oluşturulan ağırlık faktörleriyle regresyon modellerine ilişkin tahminler ağırlıkla birleştirilir. Başka bir ifadeyle her modelin tahminlerini bir araya getirmek için ağırlıklı ortalama yöntemi tercih edilmiştir.

Ağırlıklı Ortalama Yöntemi, çeşitli modellerin ya da tahminlerin performansını dengelemek ve en iyi sonucu belirlemek için tercih edilmektedir. Her modelin ya da tahminin sonucu, tespit edilen ağırlıkla çarpılarak elde edilen ağırlıklı sonuçlar bir araya getirilmektedir. Ağırlıklar, modellerin ya da tahminlerin performansına bağlı şekilde atanmaktadır. Genel olarak daha iyi performans sergileyen modeller ya da tahminler daha yüksek ağırlıklarla çarpılırken, daha zayıf performans sergileyenler daha düşük ağırlıklarla çarpılır. Bu teknik sayesinde, çeşitli modellerin ya da tahminlerin çeşitli güçlü yönleri bir araya getirilerek daha güvenilir ve genelde daha iyi sonuçlar elde edilmektedir.

Performans Değerlendirmesi: Regresyon modellerinin performansını değerlendirmek ve karşılaştırmak amacıyla kullanılan çeşitli ölçütler bulunmaktadır. Bunlar arasında en yaygın olarak kullanılanlar, MSE, RMSE, MAE ve R^2'dir.

- **MSE (Mean Squares Error-Hata Kareler Ortalaması):** Bu ölçüt, gerçek ve tahmini değerler arasındaki farkların karelerinin ortalamasını belirtmektedir. Düşük MSE değeri, daha iyi bir tahmin performansını göstermektedir.
- **RMSE (Root Mean Square Error-Hata Kareler Ortalamasının Karekökü):** MSE değerinin kareköküdür. Gerçek ve tahmini değerler arasındaki hata karelerinin ortalama değerinin karekökü alınarak hesaplanmaktadır.
- **MAE (Mean Absolute Error-Ortalama Mutlak Hata):** Gerçek ile tahmini değerler arasındaki mutlak farkların ortalamasını belirtmektedir. Kare yerine mutlak değer ile hesaplandığından büyük hataların etkisi daha dengeli olmaktadır. Düşük MAE değeri, daha iyi bir tahmin performansını göstermektedir.
- **R^2 (Belirtme Katsayısı):** Bağımsız değişkenlerin modelde bağımlı değişkendeki toplam varyansın ne kadarını açıkladığını gösteren ve 0 ile 1 arasında değer alan bir istatistiktir. 1'e yakın bir R^2 değeri, tahmin modelinin veriyi iyi açıkladığını belirtir.

Bu kapsamlı simülasyon süreci, RRB modelinin başarılı bir şekilde oluşturulmasını sağlamıştır. Elde edilen topluluk regresyon modeli, veri setindeki aykırı değerlere ve çoklu bağlantılara karşı dayanıklılığı artırarak ve Grid Search ile belirlenen hiperparametrelerle optimize edilmiş bir yapı sunarak güçlü bir performans sergilemektedir. Ayrıca, çapraz doğrulama ile eğitilen ve her bir modelin katkısını dengeleyen bu regresyon modeli, genelleme yeteneğini yüksek oranda sürdürmektedir. Sonuç olarak, performans değerlendirmesi

ölçütleri olan MSE, RMSE, MAE ve R^2 üzerinden yapılan değerlendirme, önerilen RRB modelinin güvenilir ve etkili bir tahmin aracı olduğunu doğrulamaktadır.

5. Bulgular

Doğrusal Regresyon, Ridge Regresyon, Sağlam Regresyon, Gradient Boosting Regresyon ve Ridge-Robust-Boosting Topluluk Regresyon modellerinin performanslarının kıyaslanması için gerçekleştirilen analiz sonuçları Çizelge 1'de belirtilmiştir:

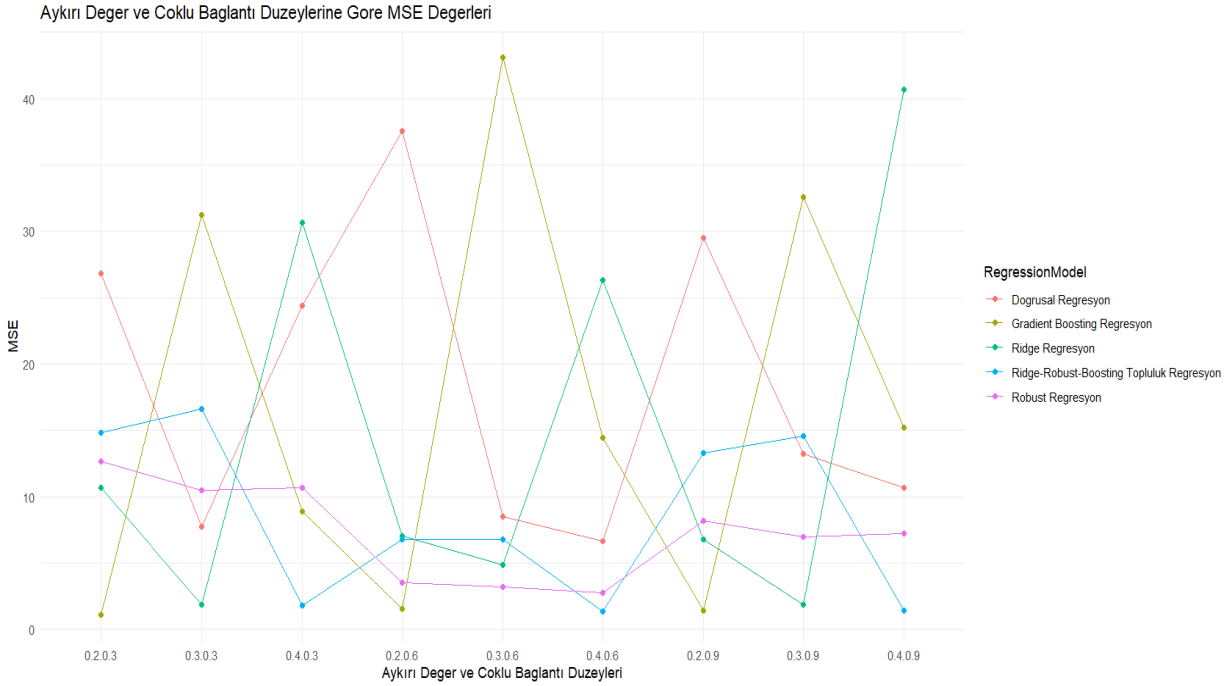
Çizelge 1. Model Performanslarının Karşılaştırılması

Çoklu bağlantı düzeyi	Aykırı değer oranı	Regresyon modeli	MSE	RMSE	MAE	R^2
0.3	0.2	Doğrusal Regresyon	26.84	5.181	4.32	0.512
		Ridge Regresyon	7.76	2.786	1.68	0.654
		Sağlam Regresyon	24.422	4.942	3.681	0.537
		Gradient Boosting Regresyon	7.049	2.655	1.585	0.685
		Ridge-Robust-Boosting Topluluk Regresyon	4.846	2.201	0.57	0.894
	0.3	Doğrusal Regresyon	26.314	5.13	3.738	0.472
		Ridge Regresyon	8.201	2.864	1.78	0.697
		Sağlam Regresyon	6.968	2.64	1.635	0.742
		Gradient Boosting Regresyon	7.261	2.695	1.651	0.731
		Ridge-Robust-Boosting Topluluk Regresyon	1.121	1.059	0.814	0.916
	0.4	Doğrusal Regresyon	31.232	5.587	3.738	0.412
		Ridge Regresyon	8.886	2.981	1.801	0.603
		Sağlam Regresyon	6.778	2.603	1.546	0.697
		Gradient Boosting Regresyon	6.811	2.61	1.548	0.696
		Ridge-Robust-Boosting Topluluk Regresyon	1.372	1.171	0.944	0.898
0.6	0.2	Doğrusal Regresyon	29.523	5.433	5.014	0.494
		Ridge Regresyon	13.253	3.64	2.368	0.608
		Sağlam Regresyon	10.655	3.264	2.086	0.685
		Gradient Boosting Regresyon	10.671	3.267	2.088	0.685
		Ridge-Robust-Boosting Topluluk Regresyon	1.862	1.365	0.899	0.864
	0.3	Doğrusal Regresyon	30.665	5.538	5	0.442
		Ridge Regresyon	3.503	1.872	0.841	0.737
		Sağlam Regresyon	3.222	1.795	0.873	0.758
		Gradient Boosting Regresyon	2.741	1.656	1.316	0.81
		Ridge-Robust-Boosting Topluluk Regresyon	1.433	1.197	0.95	0.899
	0.4	Doğrusal Regresyon	32.549	5.705	4.111	0.396
		Ridge Regresyon	15.22	3.901	2.43	0.574
		Sağlam Regresyon	14.814	3.849	2.34	0.585
		Gradient Boosting Regresyon	16.59	4.073	2.492	0.535
		Ridge-Robust-Boosting Topluluk Regresyon	1.809	1.345	0.906	0.868
0.9	0.2	Doğrusal Regresyon	37.549	6.128	4.269	0.359
		Ridge Regresyon	8.512	2.918	1.715	0.583
		Sağlam Regresyon	6.667	2.582	1.499	0.674
		Gradient Boosting Regresyon	6.792	2.606	1.509	0.668
		Ridge-Robust-Boosting Topluluk Regresyon	1.862	1.365	0.899	0.864
	0.3	Doğrusal Regresyon	40.665	6.377	4.998	0.342
		Ridge Regresyon	12.675	3.56	2.327	0.63
		Sağlam Regresyon	10.508	3.242	2.043	0.693
		Gradient Boosting Regresyon	10.677	3.268	2.057	0.688
		Ridge-Robust-Boosting Topluluk Regresyon	1.516	1.231	0.992	0.887
	0.4	Doğrusal Regresyon	43.135	6.568	5.014	0.322
		Ridge Regresyon	14.438	3.8	2.432	0.607
		Sağlam Regresyon	13.305	3.648	2.287	0.638

		Gradient Boosting Regresyon	14.547	3.814	2.356	0.604
		Ridge-Robust-Boosting Topluluk Regresyon	1.399	1.183	0.954	0.896

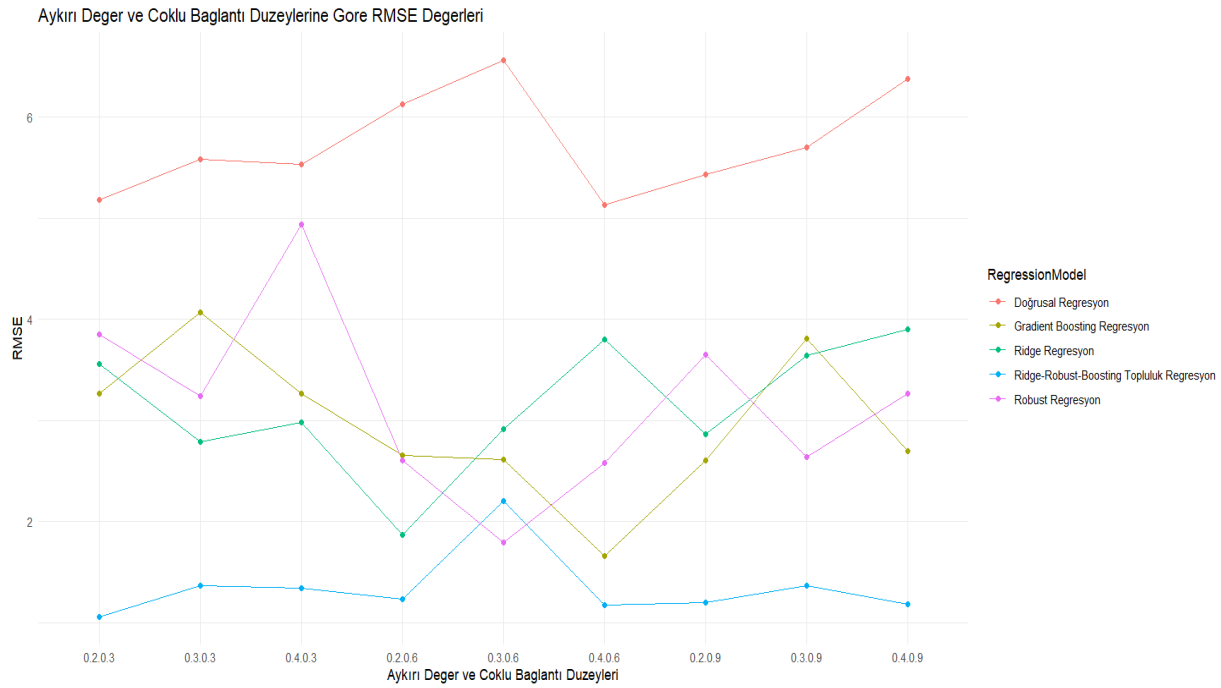
Simülasyon sonuçları incelendiğinde, ilk olarak, aykırı değer oranının arttığı durumlarda, RRB modelinin diğer modellere kıyasla daha düşük MSE, RMSE ve MAE değerlerinin olduğu gözlemlenmektedir. Aynı zamanda, R^2 değeri de diğer modellere kıyasla daha yüksek bir açıklama gücüne işaret etmektedir. Bu durum, RRB modelinin aykırı değerlere karşı dirençli olduğunu ve daha kesin tahminler gerçekleştirdiğini göstermektedir. Çoklu bağlantı düzeyi arttıkça, Ridge ve Sağlam regresyon modellerinin performansında bir azalma gözlemlenirken, Gradient Boosting Regresyon modeli bu durumdan daha az etkilenmiştir. Ancak, en etkili performansı sağlayan model yine RRB olmuştur. Model, çeşitli çoklu bağlantı durumlarında istikrarlı bir performans sergileyerek doğru tahminler yapabilme yeteneğini sürdürmüştür.

Model performanslarının karşılaştırılması amacıyla kullanılan MSE, RMSE, MAE ve R^2 değerlerine ait grafik gösterimleri Şekil 3'ten Şekil 6'ya kadar sunulmuştur:



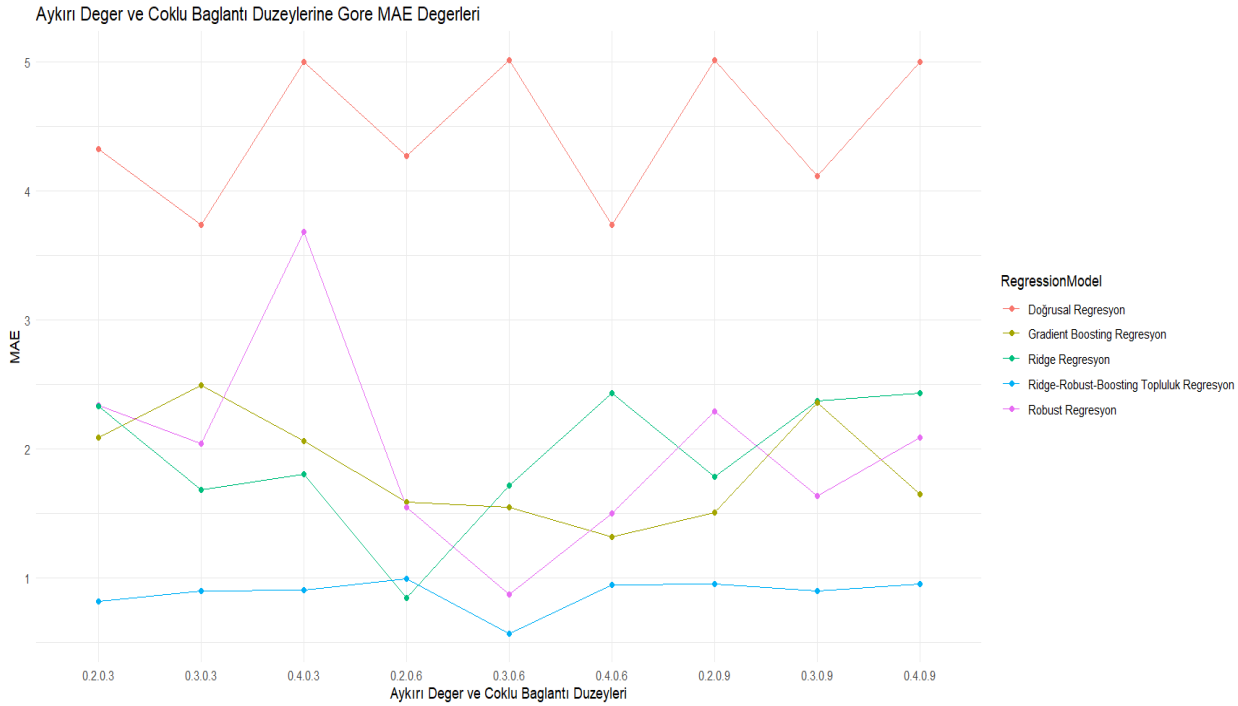
Şekil 3. Aykırı Değer ve Çoklu Bağlantı Düzeylerine Göre MSE Değerlerinin Karşılaştırılması

Şekil 3 incelendiğinde, aykırı değer ve çoklu bağlantı düzeylerinin MSE değerlerini önemli ölçüde etkilediği görülmektedir. Şekil 3'te, çoklu bağlantı ve aykırı değer oranlarının artmasıyla MSE değerlerinin de arttığı görülmektedir. Bu, çoklu bağlantı ve aykırı değerlerin modelin tahminlerini olumsuz etkilediğini göstermektedir. Regresyon modellerinin performansları karşılaştırıldığında ise RRB modelinin MSE değerlerine göre aykırı değerlere ve çoklu bağlantı sorunlarına karşı oldukça dayanıklı bir model olduğu görülmektedir.



Şekil 4. Aykırı Değer ve Çoklu Düzeylerine Göre RMSE Değerlerinin Karşılaştırılması

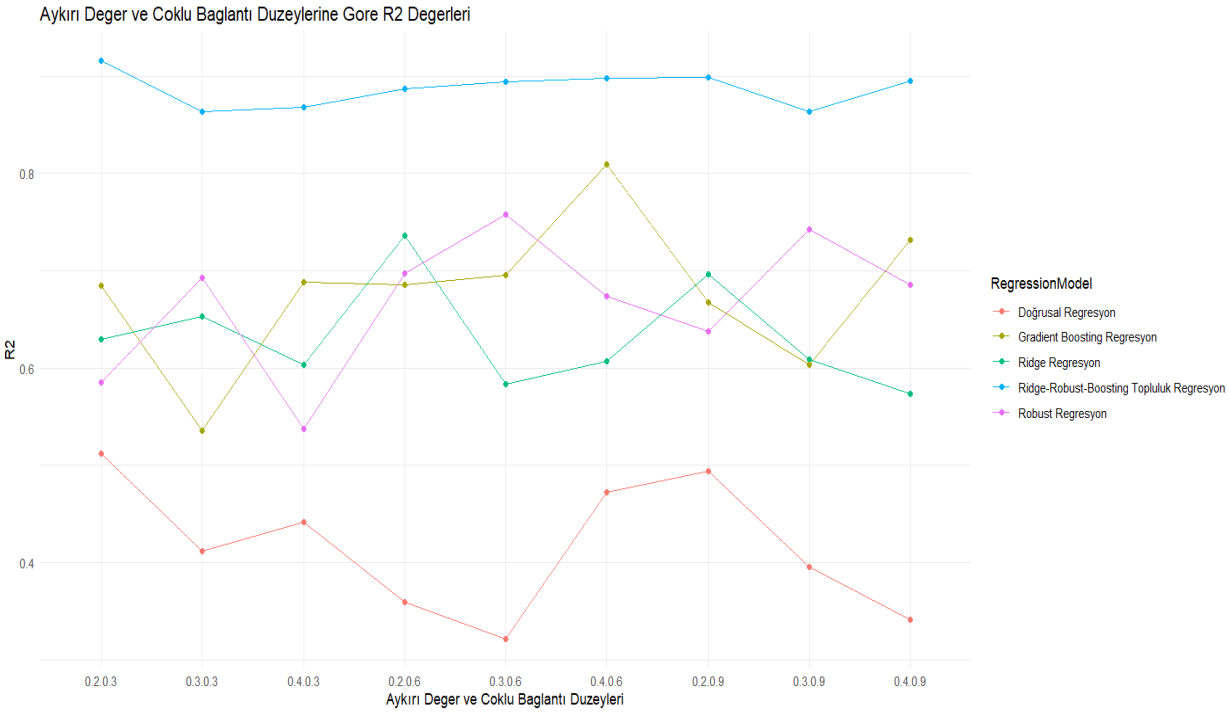
Şekil 4 incelendiğinde, MSE ile benzer şekilde aykırı değer ve çoklu bağlantı düzeyinin artmasıyla RMSE değerlerinin de arttığı görülmektedir. RRB modelinin diğer modellere kıyasla daha iyi sonuç verdiği görülmektedir. Bu model, RMSE değeri açısından diğer modellerden daha düşük değerlere sahiptir.



Şekil 5. Aykırı Değer ve Çoklu Bağlantı Düzeylerine Göre MAE Değerlerinin Karşılaştırılması

Şekil 5, aykırı değer oranı ve çoklu bağlantı düzeyine göre MAE değerlerini göstermektedir. Şekil 5'e, aykırı değer ve çoklu bağlantı düzeyinin artmasıyla MAE değerlerinin de arttığı görülmektedir. Bu, aykırı

değerler, modelin tahminlerini daha fazla bozduğu için, model performansının aykırı değerlerden olumsuz etkilendiğini göstermektedir. MAE değeri açısından da RRB modeli yine diğer modellerden daha iyi sonuç vermektedir.



Şekil 6. Aykırı Değer ve Çoklu Bağlantı Düzeylerine Göre R^2 Değerlerinin Karşılaştırılması

Şekil 6'da aykırı değer ve çoklu bağlantı düzeyinin artmasıyla R^2 değerlerinin de azaldığını görülmektedir. Aykırı değerler, modelin gerçek değerlere olan yakınlığını azalttığı için, bu durum model performansının aykırı değerlerden olumsuz etkilendiğini göstermektedir. Benzer şekilde bu durum çoklu bağlantı düzeyinin, modelin gerçek değerlere olan yakınlığını artırdığını göstermektedir. Sonuç olarak, RRB modelinin, aykırı değer ve çoklu bağlantı sorunları içeren veri setlerinde daha güvenilir ve stabil bir performans sergilediğini söylemek mümkündür. Bu model, genel olarak diğer regresyon modellerine kıyasla daha iyi sonuçlar vermiştir.

5. Sonuç

Bu çalışmada, regresyon analizinde sıkça karşılaşılan iki temel istatistiksel sorun olan çoklu bağlantı ve aykırı değer sorununu gidermek için RRB modeli önerilmiştir. Bu model ile veri setinde çoklu bağlantı ve aykırı değer sorunu olduğunda daha iyi sonuçlar elde etmek amaçlanmıştır. Bu kapsamda, Doğrusal Regresyon, Ridge Regresyon, Sağlam Regresyon, Gradient Boosting Regresyon ve önerilen Ridge-Robust-Boosting Topluluk Regresyon modellerinin performansları karşılaştırılmıştır.

Simülasyon çalışması kapsamında, 1000 gözlemden oluşan rassal veri seti Gibbs örnekleme algoritması kullanılarak üretilmiştir. Bu veri seti kullanılarak, çeşitli çoklu bağlantı seviyeleri ve aykırı değer oranlarıyla modellerin performansları ölçülmüştür. Elde edilen simülasyon sonuçları, MSE, RMSE, MAE ve R^2 ölçütleri kullanılarak değerlendirilmiştir. Analiz sonucunda, önerilen Ridge-Robust-Boosting Topluluk Regresyon modelinin diğer regresyon modellerine göre çeşitli çoklu bağlantı ve aykırı değer düzeylerinde daha üstün bir performans sergilediği görülmüştür. Bu durum, önerilen modelin, regresyon analizinde karşılaşılan hem çoklu bağlantı hem de aykırı değer sorununa daha etkili bir çözüm sunduğunu ve özellikle çoklu bağlantı ve aykırı değer sorunlarıyla başa çıkmada daha güvenilir bir seçenek olduğunu göstermektedir.

Kaynaklar

[1] P. Huber, 1981, *Robust Statistics*. Wiley, New York.

[2] D. N. Gujarati, 2004, *Basic Econometrics* (4th ed.). McGraw-Hill Companies.

[3] A. E. Hoerl, R. W. Kennard, 1970, Ridge regression biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67. https://doi.org/10.2307/1271436.

[4] M. J. Silvapulle, 1991, Robust ridge regression based on an M-estimator. *Australian Journal of Statistics*, 33(3), 319-333. https://doi.org/10.1111/j.1467-842X.1991.tb00438.x.

[5] K. Liu, 1993, A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, 22, 393-402.http://dx.doi.org/10.1080/03610929308831027.

[6] A. Arslan, N. Billor, 2000, Robust Liu estimator for regression based on an M-estimator, *Journal of Applied Statistics*, 27(1), 39-47, https://doi.org/10.1080/02664760021817.

[7] M. R. Özkale, S. Kaçıranlar, 2007, The restricted and unrestricted two-parameter estimators. *Communications in Statistics-Theory and Methods*, 36(15), 2707-2725. https://doi.org/10.1080/03610920701386877.

[8] G. Khalaf, G. Shukur, 2005. Choosing ridge parameter for regression problems, *Communications in Statistics-Theory and Methods*, 34 (5), 1177-1182. https://doi.org/10.1081/STA-200056836.

[9] M. A., Alkhamisi, G. Khalaf, G. Shukur, 2006. Some modifications for choosing ridge parameters, *Communications in Statistics-Theory and Methods*, 35(11), 2005-2020. https://doi.org/10.1080/03610920600762905.

[10] Y. M., Al-Hassan, 2010. Performance of a new ridge regression estimator, *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 9(1), 23-26. https://doi.org/10.1016/j.jaubas.2010.12.006.

[11] R. R., Hocking, F. M., Speed, M. J. Lynn, 1976. A class of biased estimators in linear regression, *Technometrics*, 18(4), 425-437. https://doi.org/10.1080/00401706.1976.10489474.

[12] M. A. Alkhamisi, G. Shukur, 2007. A Monte Carlo study of recent ridge parameters, *Communications in Statistics-Simulation and Computation*, 36(3), 535-547. https://doi.org/10.1080/03610910701208619.

[13] G. Muniz, B. M. G. Kibria, K. Mansson, G. Shukur, 2012. On developing ridge regression parameters: a graphical investigation, *Statistics and Operations Research Transactions*, 36(2), 115-138.

[14] Y., Asar, A. Karaibrahimoğlu, A. Genç, 2014. Modified ridge regression parameters: a comparative Monte Carlo study, *Hacettepe Journal of Mathematics and Statistics*, 43(5), 827-841.

[15] A. V. Dorugade, 2016. Improved ridge estimator in linear regression with multicollinearity, heteroscedastic errors and outliers, *Journal of Modern Applied Statistical Methods*, 15 (2): 362-381. https://doi.org/10.56801/10.56801/v15.i.856.

[16] A. F. Lukman, A. Olatunji, 2018. Newly proposed estimator for ridge parameter: an application to the Nigerian economy, *Pakistan Journal of Statistics*, 34(2), 91-98.

[17] S. Bhat, 2019. Performance of a weighted ridge estimator, *International Journal of Agricultural and Statistical Sciences*, 15(1), 347-354.

[18] M.N. Lattef, M. I. Alheety, 2020. Study of some kinds of ridge regression estimators in linear regression model, *Tikrit Journal of Pure Science*, 25(5), 130-142. https://doi.org/10.25130/tjps.v25i5.301.

[19] M. Qasim, K. Mansson, B. M. G. Kibria, 2021. On some beta ridge regression estimators: method, simulation and application, *Journal of Statistical Computation and Simulation*, 91(9), 1699-1712. https://doi.org/10.1080/00949655.2020.1867549.

[20] A. Irandoukht, 2021. Optimum ridge regression parameter using R-squared of prediction as a criterion for regression analysis, *Journal of Statistical Theory and Applications*, 20(2), 242- 250. https://doi.org/10.2991/jsta.d.210322.001.

[21] G. Khalaf, 2022. Improving the ordinary least squares estimator by ridge regression, *Open Access Library Journal*, 9(5), 1-8.

[22] M. Shabbir, S. Chand, F. Iqbal, 2023. A new ridge estimator for linear regression model with some challenging behavior of error term, *Communications in Statistics-Simulation and Computation*, 1-11. https://doi.org/10.1080/03610918.2023.2186874.

[23] N. Shaheen, I. Shah, A. Almohaimeed, S. Ali, H. N. Alqifari, 2023. Some modified ridge estimators for handling the multicollinearity problem, *Mathematics*, 11(11), 2522. https://doi.org/10.3390/math11112522.

[24] A. Han, 2023, Çoklu doğrusal bağlantı ve aykırı değer sorunu için Ridge-Robust-Boosting Topluluk Regresyon yaklaşımı. *Yayınlanmamış Doktora Tezi,* İnönü Üniversitesi Sosyal Bilimler Enstitüsü, Malatya, Türkiye.

[25] H. Zou, 2020, Comment: Ridge regression-still inspiring after 50 years. *Technometrics*, 62(4), 456-458. https://doi.org/10.1080/00401706.2020.180125.

[26] C. Aktaş, V. Yılmaz, 2003, Çoklu bağıntılı modellerde Liu ve Ridge regresyon kestiricilerinin karşılaştırılması. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 4(2), 189-194.

[27] Y. Li, G. R. Arce, 2004, A maximum likelihood approach to least absolute deviation regression. *EURASIP Journal on Advances in Signal Processing*, 1-8. https://doi.org/10.1155/S1110865704401139.

[28] H. Theil, 1950, A rank-invariant method of linear and polynomial regression analysis. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen*, 53, 386-392, 521-525, 1397-1412.

[29] P. K. Sen, 1968, Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324), 1379-1389. https://doi.org/10.2307/2285891.

[30] P. J. Rousseeuw, 1984, Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880. https://doi.org/10.1080/01621459.1984.10477105.

[31] L. Öztürk, 2003, Doğrusal regresyonda sağlam kestirim yöntemleri ve karşılaştırılmaları. *Yayınlanmamış Doktora Tezi*. Mimar Sinan Üniversitesi, İstanbul.

[32] H. Türkay, 2004, Doğrusal regresyon modellerinin robust (dayanıklı) yöntemlerle tahmini ve karşılaştırmalı uygulamaları. *Yayınlanmamış Doktora Tezi*, İstanbul Üniversitesi, İstanbul.

[33] P. J. Huber, 1973, Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 1, 799-821. http://dx.doi.org/10.1214/aos/1176342503.

[34] R. R. Wilcox, 2017, *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.). Academic Press.

[35] K. V. Mardia, J. T. Kent, J. M. Bibby, 1979. *Multivariate analysis*. Academic Press, London.

[36] L. Breiman, 1996, Bagging predictors. *Machine Learning*, 24, 123-140. https://doi.org/10.1007/BF00058655.

[37] J. H. Friedman, 2002, Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2.

[38] A. Natekin, A. Knoll, 2013, Gradient boosting machines, a tutorial. *Frontiers Neurorobot*, 7(21), 1-21. https://doi.org/10.3389/fnbot.2013.00021.

[39] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, 2004, Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. Association for Computing Machinery, New York, USA. https://doi.org/10.1145/1015330.1015432.

[40] J. H. Friedman, 2001, Greedy function approximation: A gradient boosting machine, *Annals Statistics*, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451.

[41] B. V. Dasarathy, B. V. Sheela, 1979, A composite classifier system design: Concepts and methodology. *IEEE Xplore*, 67(5), 708-713. Doi: 10.1109/PROC.1979.11321.

[42] L. K. Hansen, P. Salamon, 1990, Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993-1001. http://dx.doi.org/10.1109/34.58871.

[43] R. E. Schapire, 1990, The strength of weak learnability. *Machine Learning*, 5(2), 197–227. https://doi.org/10.1007/BF00116037.

[44] H. Liu, A. Gegov, M. Cocea, 2016, Ensemble learning approaches. In *Rule Based Systems for Big Data: A Machine Learning Approach*. Switzerland: Springer, 63-73. https://doi.org/10.1007/978-3-319-23696-4_6.

[45] I. D. Mienye, Y. Sun, 2022, A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*, 10, 99129-99149. https://doi.org/10.1109/access.2022.3207287.

[46] S. Geman, D. Geman, 1984, Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721-741. Doi: 10.1109/TPAMI.1984.4767596.

İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 17, 2024, 2, 46-58

Geliş/Received:22.05.2024, **Kabul**/Accepted: 24.12.2024

Araştırma Makalesi / Research Article

Incorporation of variance calculation differences in reliability predictions

Ongun Yücesan

Atılım Universitesi
Sivil Havacılık Yüksekokulu
Gölbaşı, Ankara, Türkiye
ongun.yucesan@atilim.edu.tr
ORCID: 0000-0003-2263-6803

Abstract

This paper investigates the possible incorporation of the variability of a failure durations observation series into reliability estimates considering failure time series {3, 3, 5, 6, 6.2657} and {2.5097, 4.0678, 4.9942, 5.6460, 6.1684} with very similar mean and variance. The first series of failures of a device would yield the feeling of having a failure in the duration of the third period. Later on, it would have a user experience of failing anytime. The aim of the paper is to investigate how variant conditions can be incorporated into a reliability estimate. For this purpose, distinct schemes were chosen. Initial consideration was the averaging of estimated reliability functions predicted for different deviations. Later considerations involved estimating a single standard deviation figure by averaging possible deviations. Thirdly, is the normalized version of the preceding method. The conclusions drawn from the observations indicate consideration of variability into reliability calculations impacting the estimates, which would be dependent on operational conditions, such as reset costs and safety needs.

Keywords: Mean time between failures, Reliability functions, Reliability estimates

Öz

Varyans hesaplama farklarının güvenilirlik tahminlerine işlenmesi

Makale, arızalar arası süre gözlem serisi varyansının güvenilirlik tahminlerine olası dâhil edilmesini araştırmaktadır. Çok benzer ortalama ve varyansa sahip {3, 3, 5, 6, 6.265} ve {2.5097, 4.0678, 4.9942, 5.6460, 6.1684}hatalar arası süre serileri göz önüne alındığında. Birinci serinin geldiği cihazın arızaları sanki üçüncü periyotta arıza varmış gibi bir his uyandıracaktır. İkinci serinin geldiği cihaz ise sanki her an arıza olacakmış gibi bir kullanıcı deneyimi yaşayacaktır. Bu makale, değişken koşulların bir güvenirlik tahminine nasıl dâhil edildiğini araştırmaktır. Bu amaçla farklı yöntemler seçilmiştir. İlki, farklı sapma değerleri için tahmin edilen güvenilirlik fonksiyonlarının ortalamasının alınmasıdır. Daha sonraki yöntem ise, tek bir standart sapmanın farklı değerlerin ortalaması ile hesaplanılmasıdır. Üçüncü olarak ise, ikinci metot sonucunun normalize edilmesini içermektedir. Sonuç değişkenliğin güvenilirlik hesaplamalarına dâhil edilmesinin, her iş için oluşacak koşullara, sıfırlama maliyetlerine ve güvenlik ihtiyaçlarına göre değerlendirilmesi gereğini ortaya koymaktadır.

Anahtar sözcükler: *Arızalar arasındaki ortalama süre, Güvenirlik fonksiyonları, Güvenilirlik tahminleri*

1. Introduction

For successful operation, users needs to know when a failure is most probable. Many systems, such as aircraft, consist of many electronic components. These are mostly composed of embedded systems and software. To ensure their quality, experiments on reliability are made during the development phase. A healthy duration of operations for the system is used to make reliability predictions [1].

On the other hand, there are limitations for good predictions. In many cases, the virtue of fit tests could result positively [2]. Furthermore, over fit to the available data would mean memorization rather than generalization [3].Therefore, besides basic distribution parameters, additional information about the shape of the underlying histograms is beneficial.

For instance, an overlooked issue is the deviation amount (δ_i) of each sample (x_i) from the simple mean (μ), which is $\delta_i = (x_i - \mu)$. This series represents changes in device behaviour. If the range of samples is limited, it may be simple to define a safe reset period. If they are distributed all around the available sample space, this prediction may be less reliable. The first case can be modeled with a narrow, bell-shaped normal distribution. The second case can be modeled with an exponential distribution. A uniform or a wider shaped normal distribution can as well be considered. These models all have associated errors. For the sake of clarity, comparisons will be performed based on normal distribution assuming similar error margins.

The software can have failures at accumulated or spread time indices. The series from both cases can still result in similar mean and standard deviation. Such a condition represents a hypothetical scenario, emphasizing two different observed characters. These cases could be modeled by the employment of different distributions, with different histogram shapes. However, while samples are few or results are volatile, classical and well known, simple to use distributions are more practical. The aim is to consider different methods of incorporating the deviation amount into reliability predictions by finding ways to represent these scenarios with simple distributions.

The rest of the document is organized as follows: Some bibliographic background, the data, the method of collection of the data and the processes performed on it are described in the first portion of *Section II. Material and Method*, following the brief *introduction in Section I*. to the topic. Section *III. Discussions* includes a comparison of the outcomes of the performed methods. The paper concludes with *Section IV. Conclusions* section.

2. Material and method

This section describes the methodology for observation and incorporation of the deviation amount from reliability observations. First, the techniques related to variance in existing literature are presented. The data set and how it was collected is detailed. Later on, the method of data processing is described.

2.1. Relative history of variance considerations

Variance incorporation is a topic that has been considered in more of a sense that of a developmental item which has a chance of intervention. The recent studies are more limited. The aim was at times minimizing variance RAI et al. [4], at times choosing the minimum variance among an available set of options Yiang et al. [5], or sometimes estimating a lower and upper boundary for the variance of the stochastic process Chadjiconstantinidis et al. [6]. Employment of this measurement for identifying a difference in two forms of the same information and making predictions is also possible, Khieovongphachanh et al. [7]. Deviated cases can be identified and eliminated, Shibo, et al. [8], Abu-Shawiesh [9], Joglekar et al. [10]. There are some fast methods also to estimate these parameters by incorporation of Daubeshie's Wavelet transform Ding et al. [11]. There are employment of deviation

metrics for getting around noise and clutter Wang et al. [12]. There are techniques and algorithms that take help from standard deviation to make identifications, detections and improvements to content Yousefpour et al. [13], Zhao et al.[14], Pengsen et al. [15], Prakoso et al. [16], Yang et al.[17]. Some studies consider a deviation to be more significant than others Zhang et al. [18]. Sometimes available samples are limited and prediction of standard deviation may require additional effort Luo [19].

2.2. Data set

The Data Set considered in the study is an array of times to observe a failure series. Those are composed from observations from several test case outcomes and their repetitions. These are performed during a study Yucesan et al. [1] that was already formed into an array during that paper.

All gathered data is over a peer-to-peer industrial workshop electronic Open Platform Collaboration Unified Access (OPC UA) communications. Most of the involved equipment is commercial off the shelf (COTS), while data reading software is modified to the needs from a COTS library. The communication is working in a very similar fashion or being used exactly as in many student rocket probes Matevskaet.al [20] where most of the software and tools are used intact for internal communication of this low-cost vehicle.

2.3. Preparations and methodology for analysis

The method at the beginning involves calculating statistics and probability distribution arrays, then the generation of reliability estimates that are the basis for any safe usage guideline. Whenever the mean μ and the standard deviation σ are at hand, one can generate a cumulative distribution function (CDF). The reliability function is obtained by subtracting CDF (i.e. P(X<x)) from one (1) (see Trivedi et. al. [21]) as in equation (1) and (2). The reliability function is an indicator about chance of proper functioning. Therefore when a failure is probable, before any harm, a reset, repair or a replacement could be made.

$$\mathrm{f}(\mathrm{x},\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)} \; where -\infty < x < \infty \qquad (1)$$

$$R(t) \;=\; 1 - F(\mathrm{x},\mu,\sigma) = \; 1 - \mathrm{P}(\mathrm{X}<\mathrm{x}) = 1 - \int_{-\infty}^{x} f(x,\mu,\sigma)dx \qquad (2)$$

Therefore, the probability distribution function (pdf) for variance occurrences is an as crucial component as in equation (1). Gaussian distribution is also considered as being a two parameter one. The read data consists of a bulky accumulated history of 20 counters increasing simultaneously with different step sizes. Algorithmic description of the technique for obtaining a reliability estimate is as follows:

testingMethod()

```
I=0;
while testing
  I=I+1;
  J=0;
  while system working
          Read data
          J=J+1;
  end
  X_i = J;
  Reset
end
```

return X_i

In this algorithm testing Method (), the experiment is conducted till a failure is encountered. The discrete count of repetitions is noted to form series X_i. Read data is not used. The following flow chart in Fig.1 illustrates the prediction technique. The returned value is the observation series X_i. This series is used in the reliability predictions.

Once obtained, the pdf for the variance is incorporated into the reliability estimates. It can be done in a multitude of ways. Nevertheless, three (3) methods have been considered in this study. All methods involve generating a gaussian pdf for considering the variance distribution. This distribution is obtained around the mean of the absolute differences from the "overall mean" of the observation series.

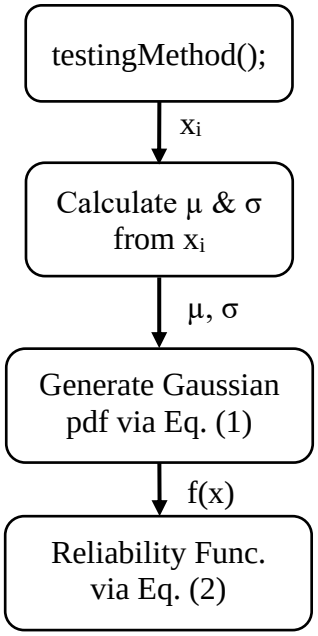


Figure 1. The method for predicting Gaussian Reliability function

Considering the method for calculation of a variance, $\delta_i = x_i - \mu$ we would obtain the aforementioned series denoted as δ_i, namely the series of the absolute differences or as *the variance calculations series*.

Methods are differentiated in how the probability of each deviant value is reflected in the estimation of reliability. For the first technique for incorporating deviance, the Gaussian pdf obtained from the variance series is reflected onto the standard reliability function by calculating the weighted average of the Normal normalized distributions for each possible variance value. Over the alternative approaches, calculating this mean value corresponds to the multiplication of the individual possible variance value's distribution in an interval with their corresponding probability of occurrence from the variance series pdf, which can be deemed as a weighted expectation as in equation (4) where $p_{dev}(x)$ is as in equation (3) representing the histogram of variance calculation series in absolute terms using Gaussian pdf of Equation (1). $\mu_{variance}$ and $\sigma_{variance}$ are mean and standard deviation obtained from the absolute values of the series δ_i.

$$p_{dev}(\delta_i) = f(|\delta_i|, \mu_{variance}, \sigma_{variance}), \qquad \delta_i \in \left[\mu_{variance} - \frac{\sigma_{variance}}{2}, \mu_{variance} + \frac{\sigma_{variance}}{2}\right] \quad (3)$$

$$\overline{R1(x)} = 1\text{-}\sum_{\forall\delta_i} F(x,\mu,\delta_i).p_{dev}(\delta_i)\,, \quad \delta_i \in \left[\mu_{variance} - \frac{\sigma_{variance}}{2}, \mu_{variance} + \frac{\sigma_{variance}}{2}\right] \quad (4)$$

The second technique comes with a single variance or standard deviation (σ) figure to generate the final reliability function as in equation (5).

$$\overline{R2(x)} = 1 - F(x,\mu,\sum_{\forall\delta_i}\delta_i.p_{dev}(\delta_i)), \quad \delta_i \in \left[\mu_{variance} - \frac{\sigma_{variance}}{2}, \mu_{variance} + \frac{\sigma_{variance}}{2}\right] \quad (5)$$

The final method is to normalize this generated estimate reliability function with its maximum to obtain a function starting from one as in equation (6).

$$R3(x) = \overline{R2(x)} \Big/ \max\left(\overline{R2(x)}\right) \quad (6)$$

The paper attempts to identify differences in reliability estimates based on the inclusion of the variance technique. The variance series mean $\mu_{variance} = 133.8776$ and the standard deviation of this series is $\sigma_{variance} = 160.2485$. The pdf generated over these parameters is as in Fig. 1. There is a small issue which is the total sum of all probabilities over the mentioned interval $\left[\mu_{variance} - \frac{\sigma_{variance}}{2}, \mu_{variance} + \frac{\sigma_{variance}}{2}\right]$ is one. The operation corresponds to slight over-normalization rather than division by the sum over the interval $[+\infty, -\infty]$. This normalization is required to satisfy any random variable's pdf conditions.

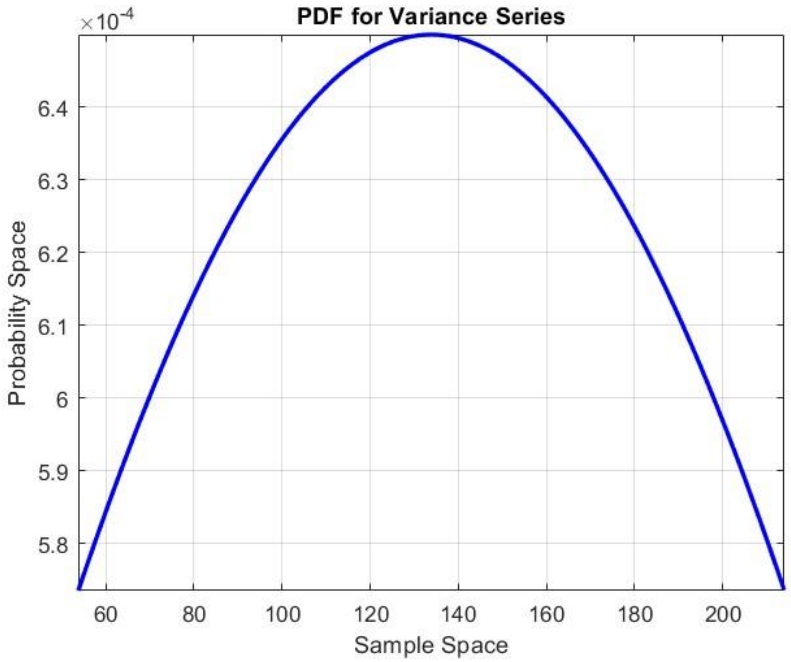


Figure 2. The Gaussian pdf estimated for the variance calculation series

The observation series employed to obtain the statistics (i.e. mean and variance). Using this information a Gaussian distribution is formalized. This approach predicts probability of occurrence. The resulting pdf for variance calculation series is visible from Fig.2.

2.4. Highest and lowest variance generated PDFs

One can see from Fig.3 that the domain of the histogram for collected data is the same interval as the variance calculation series. In this figure the highest encountered deviation observation is used as the deviation parameter for pdf, next to the minimal deviation pdf. The lighter-colored curve is with maximum deviation and is not grouped at the center. So the wings of the histogram are wider. In contrast to the low deviation, a tendency to uniform nature exists. In this case, every outcome is equally possible rather than a common time of arrival for error. Whereas with low variance pdf, most of the outcomes would have been coming from samples nearby the mean.

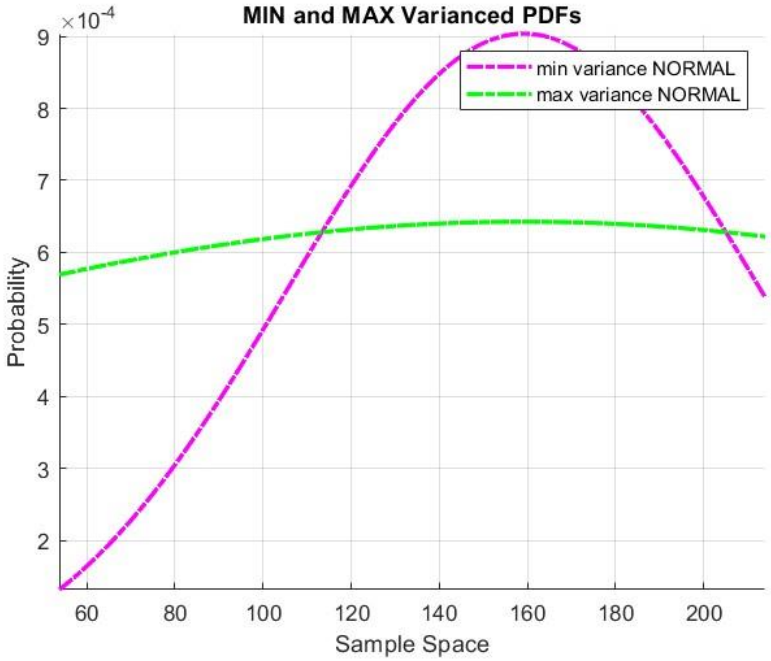


Figure 3. The Gaussian PDFs for minimum considered variance and maximum considered variance

2.5. PDFs for observations and variance calculation series

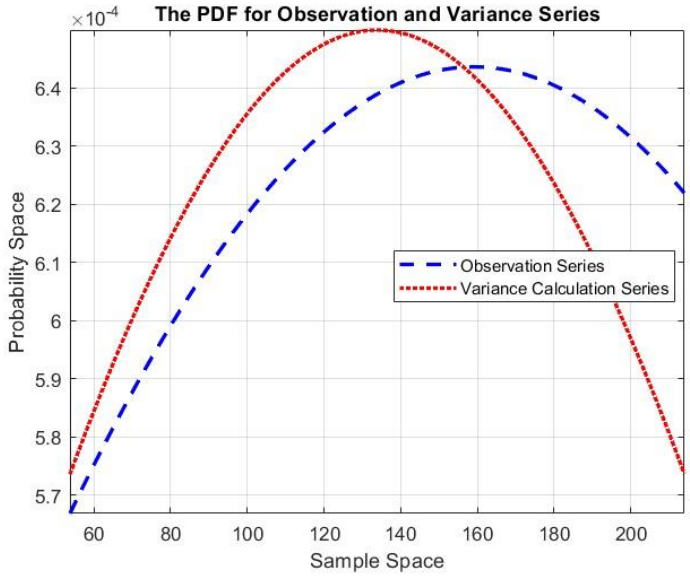


Figure 4.The PDFs Distribution predictions for originating series and variance calculation series

From Fig. 4, expresses Gaussian histograms for observed data (dashed line) and the variance calculation series δ_i . They have different means; the original series is with parameters $\mu_{original} = 159.1885$ and $\sigma_{original} = 209.1672$. Later it had, as previously mentioned,$\mu_{variance} = 133.8776$ and $\sigma_{variance} = 160.2485$ with respective order of the parameters.

2.6. Method of weighted sum of the CDFs for all sample space

This section presents the results of the first method, calculating the weighted average over histograms using each deviation amount used as second parameter for Gaussian reliability functions.

Fig. 5 shows the low-variance prediction at the highest position. It has most of the energy accumulated around the mean. In contrast, high-variance series looks like a straight line placed at the lowest position. The Weighted Average by increments of 0.1 in the interval $\left[\mu_{variance} - \frac{\sigma_{variance}}{2}, \mu_{variance} + \frac{\sigma_{variance}}{2}\right]$ is presented with evident dashed lines in the middle of both curves. It carries some of the impact of the accumulated energy of the lowest variance series and the wide wings of the highest variance series. Here the CDFs used to generate the R(x) estimates are normalized. This causes the series to start and end with similar points due to the R(x) = 1-CDF(x) operation. Their starting point is therefore very similar but different mean and deviations are the case. This figure indicates, under the conditions expressed, employing the first method gives a more curved character to reliability prediction without being as curvy as minimum deviation.

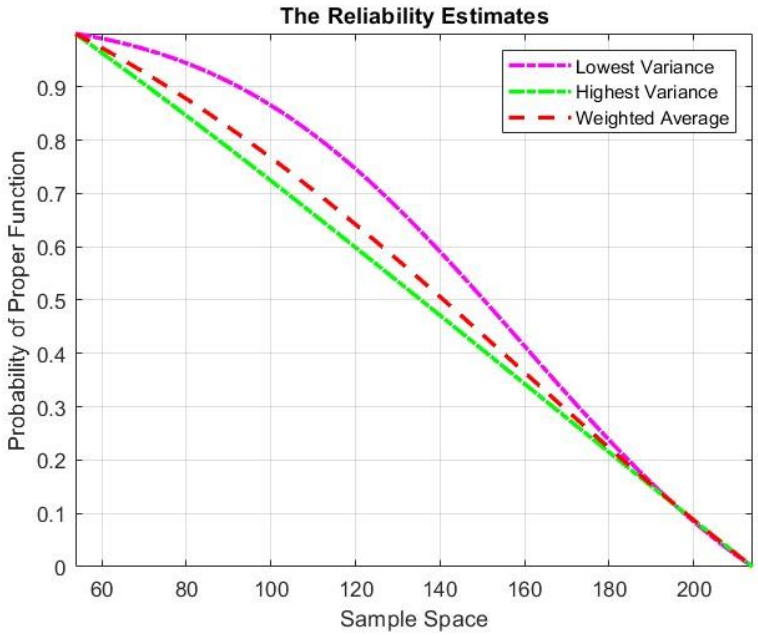


Figure 5. The Reliability Estimates of Highest, Weighted Average, Lowest variance figures from left to right

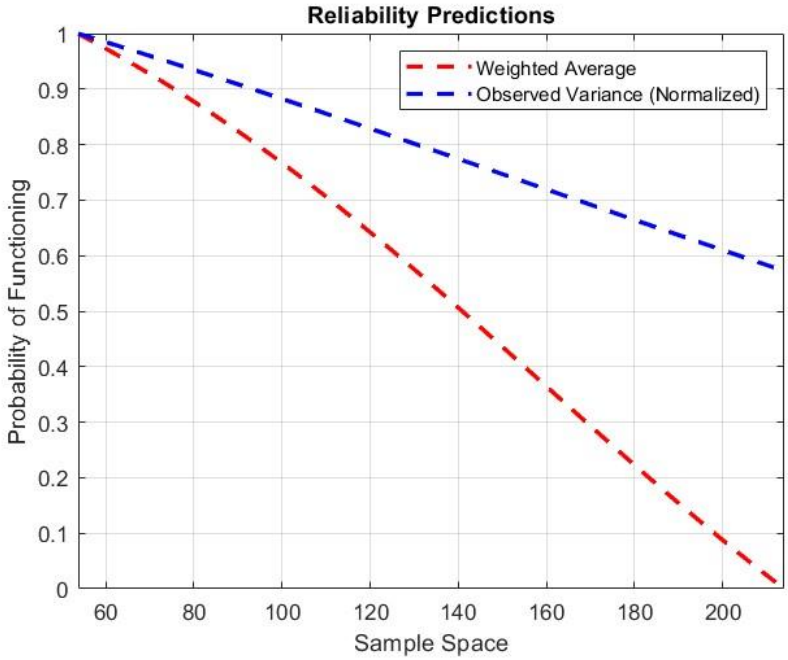


Figure 6. Normalized Observation Series Reliability Estimate and obtained Weighted Average Rel. Estimate

In Fig. 6, there are two estimates. The higher residing curve is the estimate obtained by simply using the mean and variance from the observation series normalized with its maximum value starting from one. Matlab function was used in this case. The lower residing one is the weighted average series from Fig.5.

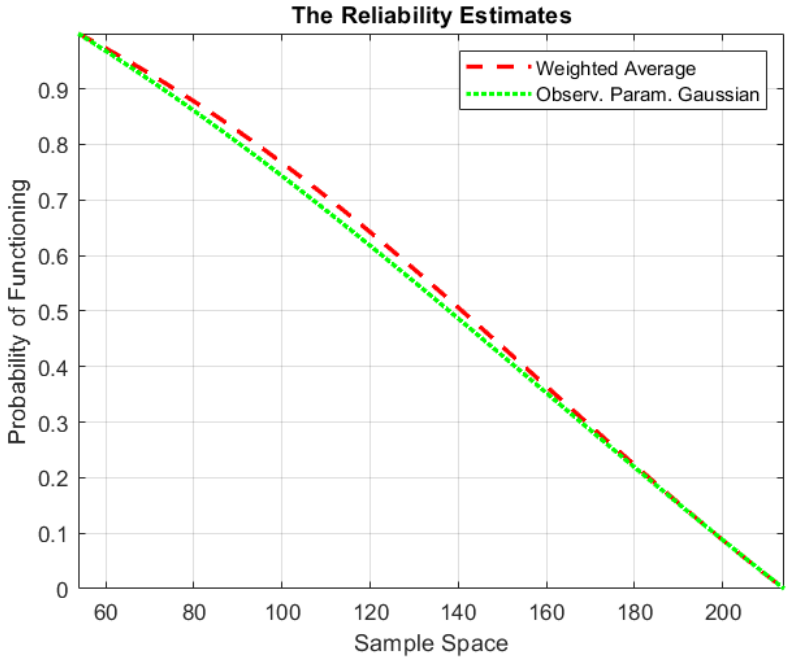


Figure 7. Observation Series Reliability Estimate and obtained Weighted Average Rel. Estimate as from same method

Fig. 7.illustrates a comparison based on Gaussian reliability estimates with classical standard deviation obtained from the original (dotted) and as the alternative weighted average (dashed) series. It presents more curvature in the weighted average case because of the contributions of low-

variance estimates. The high-variance reliability functions are flat with little impact. Therefore, the weighted average predicts the reliability higher, at risk to the customer. The harm that can take place during maintenance activities might justify such a condition.

2.7. Method of weighted averages for deviations

In this case, rather than averaging the sum of the squares and then calculating the square root for this value as in a variance calculation, 1_{st} norm of variance calculation series is divided by its series length. So this figure is not variance, not standard deviation, but a smaller or equal value. Considering an all absolute of one (1) deviations case, standard deviation and 1_{st} norm divided by length would be equal. Since this case is not the situation and the deviations are greater than one (1), and since integers are the case, the 1_{st} norm divided by length yields a smaller value. This smaller value can be regarded as an average standard deviation (STD) figure obtained and used for Gaussian pdf as the second parameter. It is in actuality the median of the variety. The first being 133.8553 later being 133.7533. If the higher median was used, they would actually be the same.

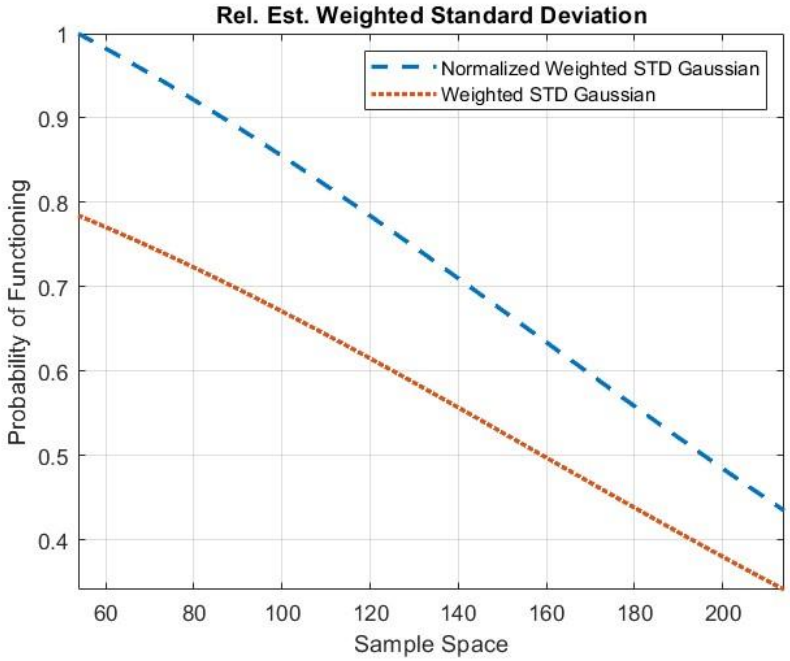


Figure 8. The average (mean) STD calculated by weighing with the probability of the individual sample

Fig.8, illustrates the single average deviation figure obtained from second and third methods. Two different graphs are visible: the dotted line represents the direct and the dashed line represents the normalized estimations.

2.8.The reliability estimate calculated with total sum normalization

In Fig. 9, during the outline of experimentation, the outcome of the first and third methods is also included. The second method is avoided in the figure also due to low start point. The exponential is seen among these reliability estimates according to its mean generated by MATLAB function. Minimum variance curve is the highest residing one, and the dashed line below that is the curve of the first method. The light coloured dotted line is the curve of the third method. The highest deviation curve is one of the lower parts of the group, with slightly darker dotted lines. This curve has second parameter (classical standard deviation) slightly less than the maximum deviation (213.95) with $\sigma_{original} = 209.1672$. The solid curve starting from 0.7 is the exponential curve reliability estimate.

It is rapidly falling in early samples; however, it is relatively constant after a while as samples progress.

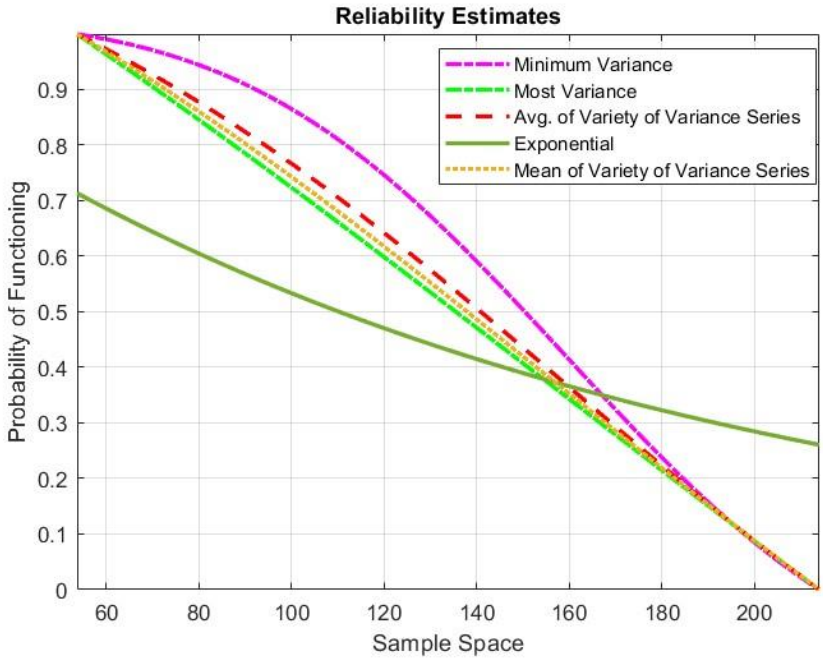


Figure 9. The Exponential along with the variety of reliability estimates

3. Discussions

The figure that drives the discussion mainly is Fig. 9, since it is the figure comparing the normalized estimates of reliability together. Exponential was not normalized in this case for the sake of relevance to common theory. Fig.10 indicates the similar not normalized version starts around 0.85 in the experiment domain starting from 60. The low deviation estimates stay around level 1 for a while. However, the weighted average case (first method) indicates a shift towards a low-variance scenario as an alternative to relatively flat estimate of the high variability. Such an understanding shows that the risk is to the customer if low variability is considered for early values. Considering the cost for repairs and maintenance risks, such an alternative can be handy at the times these factors are relevant. It may the case where the devices are at remote locations that are hard to reach.

In cases where the reset has the probability of causing additional safety hazards or failures, there would be a trade-off. A reset requires some intensive activities. These can also cause the processor to overwork, heating it to the extent that it may not function anymore or for a while at least.

Fig.10, illustrates a MATLAB generated comparison for high variance and low variance scenarios. In this figure, the “extended duration usage case” is clearly visible. In long duration, low variance calculates the reliability lower. In that situation, it enables safer usage predictions for reset periods. When the usage is extended the functioning probability is already less than 50%. If such consideration turns out to be necessary, the lesser the distributed results the safer the predictions.

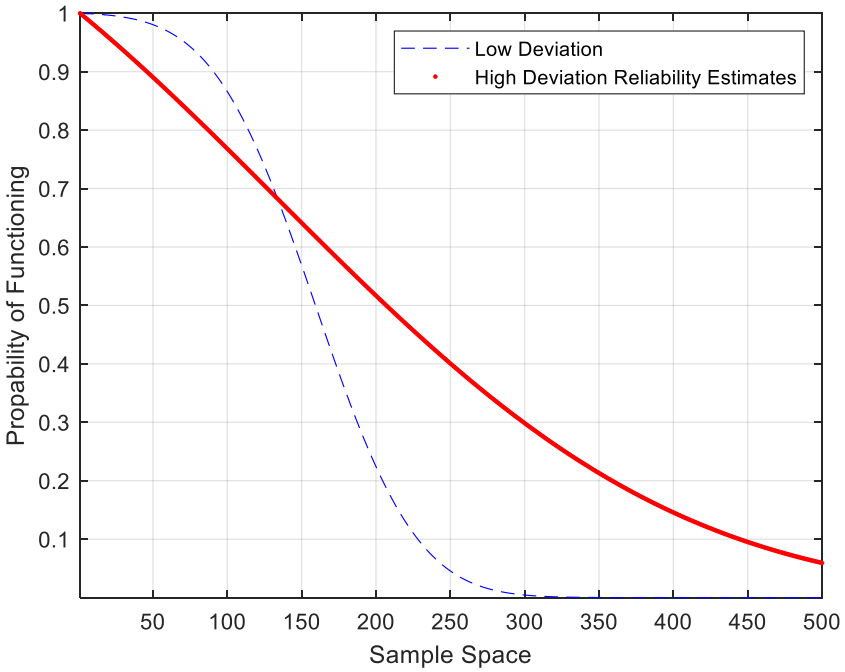


Figure 10. The Gaussian Reliability Estimates with highest and lowest deviations.

Gaussian is more useful for estimates around mean values [2] such as MTBF. It could be possible to compare other distributions with variance, but Gaussian is sufficient for a fair comparison. They were based on the impact of different deviation amounts by normalizations. It is notable that the differences are not strong enough to justify for pure reliance on any proposed method. A logical path for resolving this issue is to consider the situation with available data. The study makes comparisons under similar conditions/methods of estimations, aiming to guide such considerations. The comparison is for cases, where the purpose is to predict a safe reset period. Obviously, any product with lower variance is safer for any usage scenarios. However, a safer period to reset is using the highest variance possible.

The methods proposed may not be useful for safety critical considerations. Being ignorant of possible high-variance scenarios may lead to underestimating problems. The chance of failure is distributed over time. In such a case, if possible, improvements to system would be helpful. However, under the scenarios available, safest usage is with consideration if a high variance exists. This consideration comes at a price of making a reset cost. If it can be formulized, it can further be a good decision problem as a future work. Watching the chances of a failure taking place for a reset, using higher variety estimates might come in handy assuring the highest safety per users.

4. Conclusions

Simply averaging the reliability predictions can be handy whenever cost for a reset is significant. It is essential to take into account the operational needs and the deviant conditions for reliability estimates. For a reset period identification, using higher variety estimates is beneficial assuring the highest safety per users.

References

[1] O. Yucesan , A. Ozkil ve E. Ozbek , “A Reliability Assessment of an Industrial Communication Protocol on a Windows OS Embedded PC for an Oil Rig Control Application” in Journal of Science, Technology and Engineering Research, c. 2, volume.2, pages. 22-30, Dec. 2021, doi:10.53525/jster.971534

[2] O. Yucesan , A. ̈Ozkil ve M. E. ̈Ozbek , “Validity of Exponential Distribution for Modelling Inter-failure Arrival Times of Windows based Industrial Process Control Data Exchange” inJournal of Science, Technology and Engineering Research, c. 3,Volum. 1, pages. 1-8, june. 2022, doi:10.53525/jster.1017004

[3] L. Wang, O. Thakkar and R. Mathews, "Unintended Memorization in Large ASR Models, and How to Mitigate It," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 4655-4659, doi: 10.1109/ICASSP48485.2024.10446083.

[4] A. Rai, R. Valenzuela, B. Tuffin, G. Rubino, P. Dersin ”Approximate Zero-Variance Importance Sampling for static network reliability estimation with node failures and ap-
plication to rail systems.” Conference: 2016 Winter Simulation Conference (WSC) December 2016, 3201-3212.10.1109/WSC.2016.7822352.

[5] Y. Jiang, J. Lin, B. Cukic and T. Menzies, ”Variance Analysis in Software Fault Prediction Models,” 2009 20th International Symposium on Software Reliability Engineering, Mysuru, India,2009, pp. 99-108, doi: 10.1109/ISSRE.2009.13

[6] StathisChadjiconstantinidis, ”Some bounds for the renewalfunction and the variance of the renewal process,” in Applied Mathematics and Computation, Volume 436, 2023, 127497, ISSN 0096-3003, https://doi.org/10.1016/j.amc.2022.127497

[7] V. Khieovongphachanh, S. Kanthavong, K. Hamamoto and M. Chanthavong, "Image quality criterion of ultrasonic echo image using Standard Deviation," The 4th Joint International Conference on Information and Communication Technology, Electronic and Electrical Engineering (JICTEE), Chiang Rai, Thailand, 2014, pp. 1-3, doi: 10.1109/JICTEE.2014.6804082.

[8] S. Xin, Y. Wang and W. Lv, "Standard deviation control chart based on weighted standard deviation method," Proceedings of the 33rd Chinese Control Conference, Nanjing, China, 2014, pp. 3574-3579, doi: 10.1109/ChiCC.2014.6895533.

[9] Abu-Shawiesh, Moustafa. A control chart based on robust estimators for monitoring the process mean of a quality characteristic. International Journal of Quality & Reliability Management.(2009) 26. 480-496. 10.1108/02656710910956201.

[10]P. Joglekar, T. Katale, A. Katale, S. Deshpande, A. Nirgude and A. Chotrani, "A Novel approach for formation of Dense Clusters by Outlier Elimination and Standard Deviation," 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Bhilai, India, 2024, pp. 1-6, doi: 10.1109/ICAECT60202.2024.10468952

[11] F. Ding and T. Cao, "Application of Daubechies Wavelet Transform in the Estimation of Standard Deviation of White Noise," 2011 Second International Conference on Digital Manufacturing & Automation, Zhangjiajie, China, 2011, pp. 212-215, doi: 10.1109/ICDMA.2011.59.

[12] Y. Wang, C. Zheng and H. Peng, "Covariance Mean-to-Standard-Deviation Factor for Ultrasound Imaging," 2020 IEEE International Ultrasonics Symposium (IUS), Las Vegas, NV, USA, 2020, pp. 1-4, doi: 10.1109/IUS46767.2020.9251390.

[13] A. Yousefpour, R. Ibrahim, H. N. Abdul Hamed, U. H. Hair Zaki and K. A. Mohamed Khaidzir, "Feature subset selection using mutual standard deviation in sentiment mining," 2017 IEEE Conference on Big Data and Analytics (ICBDA), Kuching, Malaysia, 2017, pp. 13-18, doi: 10.1109/ICBDAA.2017.8284100.

[14] L. Zhao, L. Wang and D. Liu, "Hyperspectral Imagery Band Selection Based on Maximal Standard Deviation," 2015 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2015, pp. 59-62, doi: 10.1109/ISCID.2015.141.

[15] K. Pengsen and Y. Zhenming, "Image blurred region detection based on RGB color space information and local standard deviation," 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 2017, pp. 2177-2181, doi: 10.1109/IAEAC.2017.8054403.

[16] B. S. Prakoso, I. K. Timotius and I. Setyawan, "Palmprint identification for user verification based on line detection and local standard deviation," 2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering, Semarang, Indonesia, 2014, pp. 155-159, doi: 10.1109/ICITACEE.2014.7065733.

[17] Yang and Zheng Zhou, "Research and implementation of image enhancement algorithm based on local mean and standard deviation," 2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM), Kuala Lumpur, 2012, pp. 375-378, doi: 10.1109/EEESym.2012.6258668.

[18] J. Zhang, "Study on Software Project Significant Deviation Standard," in Computer Science and Software Engineering, International Conference on, null, 2008, pp. 15-18, doi: 10.1109/CSSE.2008.721.

[19] W. Luo, "A comment on "Estimating the standard deviation from extreme Gaussian values"," in IEEE Signal Processing Letters, vol. 12, no. 2, pp. 109-111, Feb. 2005, doi: 10.1109/LSP.2004.840840.

[20] Matevska, Jasminka &Noack, Enrico & Reinhold, Manuel &Diekmann, Eike. (2020). Decentralised Avionics and Software Architecture for Sounding Rocket Missions. 10.18420/SE2020_66

[21] K.S. Trivedi Probability & Statistics With Reliability, Queuing And Computer Science Applications, 2Nd Ed Wiley India Pvt.Limited, 2008, isbn 9788126518531 Conference Proceedings