Aktüerya Derneği

İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 15, 2022, 1, 1-18

Geliş/Received:06.06.2022, Kabul/Accepted: 25.06.2022

Araştırma Makalesi / Research Article

Ridge regresyon parametre seçimi: Türkiye'nin doğrudan yabancı yatırım örneği

Bahadır YÜZBAŞI

İnönü Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
Ekonometri Bölümü
Malatya, Türkiye
b.yzb@hotmail.com
0000-0002-6196-3201

Mustafa PALA

İnönü Üniversitesi
İktisadi ve İdari Bilimler Fakültesi
Ekonometri Bölümü
Malatya, Türkiye
mustafapalaa@gmail.com
0000 0002 5390 1190

Öz

Bu çalışmada çoklu doğrusal regresyon modelin bağımsız değişkenler arasında çoklu doğrusal bağlantı olduğu durumlarda En Küçük Kareler (EKK) yöntemine alternatif olarak kullanılan Ridge regresyon metodu için ayar parametresi seçimine yardımcı olacak bazı kriterler, Akaike Bilgi Kriteri (AIC), Bayes Bilgi Kriteri (BIC), Mallow's Cp, Çapraz Geçerlilik (CV) ve Genelleştirilmiş Çapraz Geçerlilik Ölçütü (GCV) karşılaştırılmıştır. Kullanılan model seçim kriterlerinin performansları Monte Carlo simülasyon çalışması ve ekonometrik bir veri kullanılarak hata kareler ortalaması (HKO) ve tahmin hatası (TH) kriterleri yardımıyla karşılaştırılmıştır. Nümerik çalışmalar sonucunda, çoklu doğrusal bağlantının olduğu durumlarda önerilen kriterler ile ayar parametresi seçilen Ridge regresyon yöntemlerinin daha düşük HKO ve TH değerleri ile daha üstün performans gösterdiği bulunmuştur.

Anahtar sözcükler: *Çoklu Doğrusal Bağlantı, Ridge Regresyon, LASSO.*

Abstract

Ridge regression parameter selection: Turkey's example of foreign direct investment

In this study, some criteria such as Akaike Information Criteria (AIC), Bayes Information Criteria (BIC), Mallow's Cp, Cross Validity (CV) and Generalized Cross Validity Measure (GCV) that will help the selection parameter for the Ridge regression method, which is used as an alternative to the Least Squares (Least Squares) method in cases where the multiple linear regression model has multiple linear connections between the independent variables, are compared. The performances of the model selection criteria used were compared using the Monte Carlo simulation study and econometric data, with the help of mean squares error (MSE) and prediction error (PE) criteria. As a result of the numerical studies, it was found that the Ridge regression methods, whose adjustment parameter was selected with the suggested criteria in cases of multicollinearity, showed superior performance with lower MSE and PE values.

Keywords: *Multicollinearity, Ridge Regression, LASSO.*

1. Giriş

Aşağıda verilen çoklu doğrusal regresyon modelini göz önüne alalım;

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_P X_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n, \tag{1}$$

Burada Y_i'ler bağımlı rassal değişkenler, $X_{i1}, X_{i2}, \dots, X_{ip}$'ler bağımsız açıklayıcı değişkenler, β_0 sabit regresyon katsayısı, $\beta_1, \beta_2, \dots, \beta_p$ kısmi regresyon katsayıları ve $\varepsilon_1\ \varepsilon_2, \dots, \varepsilon_n$ aynı dağılımlı özdeş rassal hata terimidir. i alt indisleri belirlenmiş olan mümkün gözlemi, n gözlem sayısını ve p ise parametre sayısını ifade eder. (1) eşitliğini matris formatında yazacak olursak:

$$Y = X\beta + \varepsilon$$

Burada $Y = (Y_1, Y_2, \dots, Y_n)'$, $X = (X_1, X_2, \dots, X_n)'$, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})'$ $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ ve $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ şeklinde ifade edilmektedir ve $(')$ kullanılan üst simge bir vektörün veya matrisin devriğini almayı gösterir.

Çoklu doğrusal regresyon analizinde her bir gözlem için bağımsız değişkenler arasında tam veya tama yakın doğrusal bir korelasyon olursa ortaya çıkan probleme çoklu doğrusal bağlantı (ÇDB) adı verilir. Bu durumda doğrusal ve yansız tahmin edicileri içinde küçük varyansa sahip olan EKK tahmin edicisi önemli oranda olumsuz etkilenmektedir [1]. Modeldeki bağımsız değişkenlerin aynı eğilime sahip olmaları, modelin değişken sayısının gözlem sayısından çok olması, kullanılan veri toplama yöntemlerinin uygun olmaması çoklu doğrusal bağlantının nedenleri arasında gösterilebilir. Bağlantı sorunun çözümü için panel veri, soruna neden olan değişkenleri modelden çıkarmak ve yanlı tahmin edicileri kullanarak bağlantı sorununu çözmek için başvurulan yollardandır.

Çoklu doğrusallık probleminin tarihi Frisch [2] çalışmasına dayanır. İlk olarak Hoerl ve Kennard [3, 4] tarafından tanımlanan Ridge regresyon bu probleme çözüm yöntemlerinden biridir. Ridge regresyonun ayar parametresinin bulmak araştırmacıların temel ilgisidir. Pek çok araştırmacı bu ayar parametresinin tahminini elde etmek farklı yöntemler önermişlerdir. Bunlardan bazıları, McDonald ve Galarneau [5], Monte Carlo simülasyonları ile k ayar Ridge parametresini belirlemek için iki analitik yöntem önermiş ve hata kareler ortalaması açısından değerlendirmişlerdir. Lawless ve Wang [6], Ridge ve diğer regresyon parametre tahmin edicileriyle yaptıkları simülasyon çalışması sonucunda, yaygın olarak kullanılan iki hata kareler ortalaması kriterine göre, iki sıradan Ridge tahmin edicisinin hem en küçük karelerden hem de diğer tahmin edicilerden önemli ölçüde daha iyi performans gösterdiğini belirtmişlerdir. Golub, Heath ve Wahba [7], iyi bir Ridge parametresi seçme yöntemi olarak genelleştirilmiş çapraz doğrulama yöntemini incelemişlerdir. Khalaf ve Shukur [8], tasarım matrisinin sütunları arasında çoklu bağlantı olduğunda, Ridge ayar parametresinin k seçimi için yeni bir yaklaşım önermiş ve hata kareler ortalaması açısından simülasyon teknikleri ile değerlendirmişlerdir. Alkhamisi, Khalaf ve Shukur [9], tasarım matrisinin sütunları arasında çoklu bağlantı olduğunda k Ridge ayar parametresini seçmek için Khalaf ve Shukur [8] tarafından önerilen tahmin edicilerin geliştirilmiş versiyonu olan tahmin edicilere dört değişiklik önermişlerdir. Alkhamisi ve Shukur [10], k Ridge ayar parametresini elde etmek için yeni bir yaklaşım önermiş ve ardından Monte Carlo simülasyonları ile değerlendirmişlerdir. Lukman ve Olatunji [11], regresyon katsayılarından bağımsız ve standart regresyon hatasının bir fonksiyonu olan bir Ridge parametresi önermişlerdir. Owolabi, Ayinde ve Alabi [12], iki parametreli bir Ridge tipi tahmin edici önermişlerdir ve istatistiksel özelliklerini teorik olarak ve Monte Carlo simülasyon çalışmaları yoluyla ortaya koymuşlardır.

Bu çalışma aşağıdaki şekilde organize edilmiştir: Ridge regresyonun ayar parametre seçim kriterleri çalışmanın ikinci bölümde ayrıntıları ile birlikte incelenecektedir. Çalışmanın üçüncü bölümünde, Monte Carlo simülasyon modeli ile seçim kriterlerinin performansları karşılaştırılmıştır. Bu karşılaştırmalara, literatürde en popüler cezalı tahmin edicilerinden biri olan En Küçük Mutlak Büzülme ve Seçim Operatörü (LASSO) tahmin edicisi de eklenmiştir. Çalışmanın dördüncü bölümünde ekonometrik bir veri seti kullanılarak, doğrudan yabancı yatırımı etkileyen faktörlerle kurulan çoklu doğrusal regresyon modeliyle tahminler yapılacaktır. Yapılan ön incelemeler sonucunda, ilgili verinin bağımsız değişkenler arasında korelasyon olması sebebiyle EKK yönteminden elde edilen tahminin varyansının olması gerektiğinden daha büyük çıkmasına neden olmaktadır. Söz konusu bu durumda çoklu doğrusal regresyon modelinin tahmininde kullanılan EKK yöntemi doğru sonuçlar verememektedir. Bağlantı sorunun çözümü için önerilen yöntemlerden yanlı regresyon tahmin edicisi olan Ridge regresyon ve modele dahil edilen yanlılık parametresi ile analiz tekrardan yapılıp sorun giderilmeye çalışılmaktadır. Çalışmanın asıl amacı da modele dahil edilen Ridge regresyon ayar parametresi olan k'nın seçimi için kullanılan kriterler de analize dahil edilerek en iyi sonucu veren kriter seçimi de yapılmıştır.

2. Ridge regresyon

ÇDB problemi durumunda EKK yönteminin kullanılması tahminlerin büyük varyansa sahip olmasına neden olmaktadır. 1970 yılında Hoerl ve Kennard çoklu doğrusal bağlantı problemi olduğunda bu problemi gidermek için Ridge tahmin ediciyi önermişlerdir. Böylelikle daha küçük varyanslı tahmin ediciler elde edilmektedir. ÇDB problemi olduğunda $X'X$ matrisi tekil değildir. Hoerl ve Kennard ilk kez 1962 yılında $X'X$ matrisine k negatif olmayan bir sayı olmak üzere, kI_p sabitini modele ekleyerek, Ridge tahmin edicisini elde etmişlerdir [3].

$X'X$ matrisinde ÇDB' dan dolayı bir veya daha fazla öz değerin küçük olacağını ve bu nedenle β ile onun EKK tahmin edicisi $\hat{\beta} = (X'X)^{-1}X'Y$ ile değerleri arasındaki uzaklığın yüksek olacağını açıklamışlardır. Bu sorunun çözümü için de en uygun tahmin edicinin Ridge tahmin edicisi olduğunu söylemişlerdir [3].

Açıklayıcı değişkenler arasında ÇDB olması durumunda Ridge regresyon yöntemi ile tahmin edilen β regresyon katsayılarının EKK yöntemiyle yapılan tahminlerden daha küçük HKO'ya sahip olduğu Hoerl ve Kennard [3] tarafından gösterilmiştir. Hoerl ve Kennard [3] tarafından önerilen Ridge regresyon tahmin edicisi aşağıdaki kayıp fonksiyondan elde edilmiştir;

$$L(\beta) = \sum_{i=1}^{n}(Y_i - \beta_0 - X_i'\beta)^2 + k\sum_{j=1}^{p}\beta_j^2 .$$

β'ya türevini alıp, sıfıra eşitlenirse;

$$\hat{\beta}_k = (X'X + kI_p)^{-1}X'Y, \tag{2}$$

elde edilir. Burada I_p ise $p \times p$ boyutunda birim matrisidir ve $k \geq 0$ ayar parametresi olup, $k = 0$ ise $\hat{\beta}_k = \hat{\beta}$, $k = \infty$ ise $\hat{\beta}_k = 0$'dır.

2.1. Ridge tahmin edicisi ile EKK tahmin edicisi arasındaki ilişki

EKK tahmin edicisinin her iki tarafı da $X'X$ ile çarpılırsa;

$$X'X\beta = X'Y$$

elde edilir.

Denklem 2'de Ridge tahmin edicisi olan denklemde $X'Y$ yerine eşiti yazılırsa;

$$\hat{\beta}_k = (X'X + kI_p)^{-1}X'X\,\hat{\beta}$$

elde edilir.

$X'X$ matrisinin tersinin tersi kendisi olduğu için;

$$\hat{\beta}_k = (X'X + kI_p)^{-1}\,[(X'X)^{-1}]^{-1}\hat{\beta}$$

şeklinde yazılabilir. Her iki matriste tekil olmadıkları için;

$$\hat{\beta}_k = \left[(X'X)^{-1}\,(X'X + kI_p)\right]^{-1}\hat{\beta}$$

şeklinde yazılabilir. Buradan da,

$$\hat{\beta}_k = [(X'X)^{-1}\,X'X + k(X'X)^{-1}]^{-1}\;\hat{\beta}$$

şeklinde ifade edilir. Yapılan bu işlemlerden sonra

$$\hat{\beta}_k = \left[I_p + k(X'X)^{-1}\right]^{-1}\hat{\beta}$$

olur. $Z = \left[I_p + k(X'X)^{-1}\right]^{-1}$ olarak tanımlanırsa;

$$\hat{\beta}_k = Z\hat{\beta}$$

şeklinde ifade edilir. Bu eşitlik Ridge tahmin edicisinin EKK tahmin edicinin bir dönüşümü olduğunu göstermektedir [13].

2.2. Ridge regresyon parametre seçim kriterleri

Ayar parametresinin seçimi yanlı regresyon tahmin edicilerinin performanslarını etkileyen en önemli etkendir. Ayar parametre tahmini için birçok kriter bulunmaktadır. Bu kriterlerden AIC, BIC, CV, GCV ve CP alt başlıklar halinde verilecektir. Bu kriterlerde kullanılan değerler aşağıdaki şekilde özetlenmiştir.

$$\hat{\varepsilon} = Y - X\hat{\beta}_k$$

$$H = X(X'X + kI_p)^{-1}X'$$

$$df = tr(H)$$

$$HKT(\text{Hata Kareler Toplamı}) = \hat{\varepsilon}'\varepsilon$$

Burada, H izdüşüm matrisi olarak bilinmektedir, df serbestlik derecesi ve $tr(\cdot)$ ifadesi bir matrisin izini göstermektedir.

2.2.1. Akaike Bilgi Kriteri

Akaike'nin bilgi kriterleri, Hirotsugu Akaike [14] tarafından geliştirilmiştir. AIC, tahmin edilen herhangi bir istatistiksel modelin uyum iyiliğinin bir göstergesi olarak adlandırılabilir. AIC genellikle yüksek boyutlu gerçekliğe sahip bilinmeyen bir model bulmaya çalışır. Bu modellerin AIC'deki gerçek modeller olmadığı anlamına gelir. AIC kriterleri asimptotik olarak çapraz geçerliliğe eşdeğerdir. AIC aşağıdaki şekilde tanımlanmaktadır.

$$AIC = n \times log(HKT) + 2 \times df\;,$$

Burada, n gözlem sayısını ve $log(\cdot)$ doğal logaritmayı göstermektedir.

2.2.2. Bayes Bilgi Kriteri

Bayesian bilgi kriterleri [15] Gideon E. Schwarz Bayesçi bilgi ölçütünü geliştirmiştir. BIC, farklı sayıda parametreye sahip bir grup parametrik model arasında bir model seçimidir. BIC serbest parametreleri daha güçlü cezalandırmaktadır. Bayes Bilgi kriterleri sadece gerçek modellerle karşılaşmaktadır. Bayes bilgi kriterlerinin tutarlı olduğu söylenebilir.

$$BIC = n \times log(HKT) + df \times log(n).$$

2.2.3. Çapraz Geçerlilik Model Seçim Kriteri

Eğer kare hatalar kullanılacaksa iyi bir tahmin edici olarak birini dışarıda bırakarak çapraz geçerlilik tahmin edicisi önerilebilir. Bu metot modellerin öngörü yeteneğine göre model seçimi yapılmasını benimsemektedir. Çapraz geçerlilik modeli;

$$CV = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{-1})^2 = \sum_{n=1}^{n} (\frac{Y_i - \hat{Y}_i}{1 - h_{ii}})^2$$

olarak elde edilir. Burada h_{ii}, H şapka matrisinin i. köşegen elemanıdır. Ridge ve LASSO regresyon parametre seçiminde K kat çapraz geçerlilik kullanılmaktadır. n sayıda gözlemi olan veri seti K sayıda eşit parçaya bölünür. K parçaya bölünen veri için $K - 1$ parça eğitim ve geriye kalan ise test seti olarak seçilir. İşlem K adım tekrarlanarak her adımda farklı test seti seçilir. Söz konusu her adım için hata kareler ortalaması elde edilir ve tüm değerlerin ortalaması ele alınarak çapraz doğrulama hata eğrisi elde edilir ve bu eğride minimum değeri veren parametre seçilir. Yapılan bu çalışmada $K = 10$ alınmıştır.

2.2.4. Genelleştirilmiş Çapraz Geçerlilik Ölçütü

$$CV = \sum_{i=1}^{n} (Y_i - \hat{Y}_i^{-1})^2 = \sum_{n=1}^{n} (\frac{Y_i - \hat{Y}_i}{1 - h_{ii}})^2$$

Yukarıdaki CV modeli üzerinde yapılan değişikliklerle elde edilen GCV ölçütü CV modelindeki h_{ii} yerine $\hat{H}$ matrisinin ortalaması alınarak elde edilen bir ölçüttür. GCV modeli ise;

$$GCV = \frac{n \times HKT}{(n - df)^2}$$

şeklindedir.

2.2.5. Mallow'un Cp Kriteri

Mallow'un model seçim kriteri aşağıdaki gibidir.

$$C_P = HKT + 2 \times df \times \hat{\sigma}^2$$

Burada; $\hat{\sigma}^2$, σ^2'nin tahmin edicisidir. Bulunan alt küme modelleri arasında C_p değerini minimum olan model uygun model olarak seçilir.

2.4. En Küçük Mutlak Büzülme ve Seçim Operatörü

Ridge tahmin edicisinin tahmin gücü oldukça yüksektir. Ancak tahmin sürecinde birtakım problemler ortaya çıkmaktadır. Ridge regresyon katsayıları daraltan bir tahmin edici olması sebebiyle daha kararlı bir yapıya sahiptir. Fakat regresyon katsayılarını sıfıra daraltamadığı için modelin yorumlanması zorlaşır. Bu sebeple modeli yorumlamak için ve tahmin doğruluğunu elde edebilmek için bazı katsayıları sıfıra daraltmak söz konusu olabilir. Bu işlemden sonra varyans küçülmekte fakat yanlı açıdan sapma meydana gelmektedir. Çıkan sonuçta tahmin doğruluğu daha güvenilir olduğundan bu tahmin edicisinin yansızlık özelliği göz ardı edilebilir [16]. 1996 yılında Tibshirani, LASSO yöntemi ile bazı katsayıları sıfıra daraltarak parametre tahminini ve model seçimini eş zamanlı yapmıştır [17]. LASSO yöntemi genel olarak aşağıdaki gibidir;

$$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 + k \sum_{j=1}^{p} |\beta_j| \right\},$$

Burada k ayar parametresidir.

3. Simülasyon Çalışması

Bu bölümde, Ridge ayar parametresinin seçilmesinde kullanılacak beş yöntemin (AIC, BIC, CV, GCV, Cp) ve LASSO'nun karşılaştırılması Monte Carlo simülasyon için çalışması ile yapılmıştır. McDonald ve Galarneau çalışması [5] takip edilerek farklı derecelerde çoklu bağlantıya sahip veri matrisi üretebilmek için, bağımsız değişkenler aşağıdaki şekilde üretilmiştir;

$$X_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho z_{i(p+1)} \qquad i = 1,2,\dots,n \quad j = 1,2,\dots,p,$$

Burada p, bağımsız değişken sayısı, n, örneklem hacmi olmak üzere ρ^2 herhangi iki açıklayıcı değişken arasındaki korelasyon katsayısı, z_{ij} ise $N(0,1)$ dağılımından olan rassal sayılardır. Amacımız, çoklu doğrusallığın derecesinin yüksek ve düşük olduğu durumlarda Ridge, LASSO ve EKK tahmin edicilerinin performanslarını karşılaştırmak ve ÇDB durumunda en iyi tahmini veren yöntemi seçmek amaçlanır. Bağımsız değişkenler arasındaki farklı korelasyon değerleri ile beş örnek için simülasyon çalışması yapılıp sonuç elde etmek amaçlanmıştır [18]. Simülasyonda kullanılacak doğrusal regresyon modeli aşağıdaki gibidir;

$$Y = X\beta + \sigma\varepsilon \qquad \varepsilon \sim N(0, I)$$

Modellerin değerlendirilmesi için simülasyonda eğitim verisi ve test verisi için n birimlik örneklemler kullanılmıştır. Simülasyon için R Studio'da *glmnet* paketi kullanılmıştır. Eğitim setinde en iyi ceza parametresinin değeri model seçim kriterleri ile belirlenip en iyi model oluşturmak amaçlanmaktadır. Ardından test seti ile modelin iyiliği test edilmektedir. Simülasyonlarda tüm değişkenleri eğitim setine dayalı olarak ortalarız. $\bar{X}_{eğitim} = (\bar{X}_{1,eğitim}, \dots, \bar{X}_{p,eğitim})$ eğitim verilerini vektörlerini belirtir. Test veri setindeki gözlem sayısı n_{test} ve $\bar{Y}_{eğitim} = (Y_{eğitim} - \bar{Y}_{eğitim})$ eğitim verilerinde ortalama yanıtları gösterir. Son olarak iki performans ölçüsü hesaplanmıştır. Bunlar, test hatası $HKO_Y = \frac{1}{n_{test}} r_{sim}' r_{sim}$ ise $r_{i,sim} = X_i\beta - \left(\bar{Y}_{eğitim} + \left(X_i + \bar{X}_{eğitim}\right)'\hat{\beta}\right)$ ve tahminin hata kareler ortalaması ise $HKO_\beta = \left|\hat{\beta} - \beta\right|^2$ olarak gösterilir [19].

Sırasıyla eğitim ve test setini göstermek için ./. gösterimi kullanılacaktır. Simülasyon çalışması için incelenen örnekler aşağıdaki gibidir.

1. Her veri seti $20/200$ gözlemden oluşmaktadır. $\beta = (3,1.5,0,0,2,0,0,0)'$ olarak $\sigma = 1,3$ ve $X \sim N(0, \Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3,0.6,0.9$

2. Her veri seti $20/200$ gözlemden oluşmaktadır. $\beta = \left(\underbrace{0.85,0.85,\ldots,0.85}_{8}\right)'$ olarak $\sigma = 1{,}3$ ve $X \sim N(0,\Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3{,}0.6{,}0.9$

3. Her veri seti $20/200$ gözlemden oluşmaktadır. $\beta = (3{,}1.5{,}0{,}0{,}0{,}0,-1,-1)'$ olarak $\sigma = 1{,}3$ ve $X \sim N(0,\Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3{,}0.6{,}0.9$

4. Her veri seti $100/200$ gözlemden ve 30 bağımsız değişkenlerden oluşmaktadır. $\beta = \left(\underbrace{2,\ldots,2}_{8},\underbrace{0,\ldots,0}_{22}\right)'$ olarak $\sigma = 1{,}3$ ve $X \sim N(0,\Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3{,}0.6{,}0.9$

5. Her veri seti $100/200$ gözlemden ve 40 bağımsız değişkenlerden oluşmaktadır. $\beta = \left(\underbrace{0,\ldots,0}_{10},\underbrace{2,\ldots,2}_{10},\underbrace{0,\ldots,0}_{10},\underbrace{2,\ldots,2}_{10}\right)'$ olarak $\sigma = 1{,}3$ ve $X \sim N(0,\Sigma)$ ise $\Sigma_{ij} = \rho^{|i-j|}$, $\rho = 0.3{,}0.6{,}0.9$

Yukarıda verilen örnek modellerini 1000 veri seti ile simülasyon yaparak araştırıyoruz. Bu örnek simülasyon çalışmaları aşağıdaki çizelgelerde verilmiştir. Bu örneklerle yapılan simülasyon çalışmasına Koşul İndisi (CI) eklenerek çoklu doğrusal bağlantının varlığı durumunda seçim parametrelerinin sonuçları karşılaştırılmıştır [20]. Sonuçları daha kolay karşılaştırmak için, Rölatif HKO (RHKO) değeri şu şekilde tanımlanmıştır: $RHKO(\hat{\beta}^*) = \frac{MSE(\hat{\beta})}{MSE(\hat{\beta}^*)}$, burada $\hat{\beta}^*$ yukarıda belirttiğimiz tahmin edicilerden bir tanesidir. Eğer, $RHKO(\hat{\beta}^*)$ değerinin 1'den büyük ise, o zaman $\hat{\beta}^*$ tahmin edicisi EKK tahmin edicisinden daha üstün performans gösterdiğini belirtir. Simülasyon sonuçlarına bakıldığında tüm örneklerde EKK, $RHKO_\beta$ ve $RHKO_Y$ göre de en kötü sonuçları vermiştir. Simülasyon çıktılarının yorumları aşağıda verilmiştir.

Çizelge 1: Örnek 1 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	HKO_β	$RHKO_\beta$	HKO_Y	$RHKO_Y$
0.3	1.886	1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.827	1.081	0.839	1.057
			RIDGE(AIC)	0.837	1.068	0.842	1.053
			RIDGE(BIC)	0.830	1.078	0.840	1.056
			RIDGE(CP)	0.827	1.081	0.840	1.056
			RIDGE(CV)	0.855	1.046	0.892	0.995
			LASSO(CV)	0.591	1.514	0.625	1.419
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	5.104	1.577	5.595	1.427
			RIDGE(AIC)	5.546	1.451	5.882	1.357
			RIDGE(BIC)	5.448	1.478	6.014	1.328
			RIDGE(CP)	5.040	1.597	5.559	1.436
			RIDGE(CV)	5.013	1.606	5.529	1.444
			LASSO(CV)	5.116	1.573	5.587	1.429
0.6	3.605	1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	1.249	1.195	0.802	1.106
			RIDGE(AIC)	1.288	1.160	0.807	1.100
			RIDGE(BIC)	1.257	1.189	0.805	1.102
			RIDGE(CP)	1.246	1.198	0.805	1.102
			RIDGE(CV)	1.250	1.195	0.872	1.017
			LASSO(CV)	0.863	1.730	0.606	1.464
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	6.397	2.102	4.954	1.612
			RIDGE(AIC)	7.596	1.770	5.405	1.477
			RIDGE(BIC)	6.466	2.079	5.143	1.553
			RIDGE(CP)	6.246	2.152	4.961	1.610

			RIDGE(CV)	5.705	2.356	4.621	1.728
			LASSO(CV)	6.609	2.034	4.971	1.606
0.9	10.880	1	EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	3.541	1.785	0.633	1.401
			RIDGE(AIC)	4.052	1.560	0.673	1.318
			RIDGE(BIC)	3.613	1.750	0.651	1.363
			RIDGE(CP)	3.486	1.814	0.633	1.401
			RIDGE(CV)	3.046	2.076	0.639	1.390
			LASSO(CV)	3.087	2.048	0.561	1.581
		3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	14.562	3.907	3.558	2.244
			RIDGE(AIC)	21.676	2.625	4.341	1.840
			RIDGE(BIC)	12.978	4.384	3.528	2.263
			RIDGE(CP)	13.739	4.141	3.518	2.270
			RIDGE(CV)	8.250	6.896	2.860	2.792
			LASSO(CV)	18.000	3.161	3.884	2.056

Örnek 1 için simülasyon sonuçlarına Çizelge 1'den bakıldığında, çoklu doğrusallığın düşük, orta ve yüksek olduğu aralıklarda $\sigma = 1$'de $RHKO_{\beta}$ ve $RHKO_{Y}$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV)'dir. $\sigma = 3$ ise en iyi sonucu veren kriter RIDGE (CV)'dir. Çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının artığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter RIDGE(AIC) en iyi sonucu veren kriter ise RIDGE(CV) olduğu görülmüştür.

Çizelge 2: Örnek 2 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	HKO_{β}	$RHKO_{\beta}$	HKO_{Y}	$RHKO_{Y}$
0.3	1.886	1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.649	1.358	0.740	1.199
			RIDGE(AIC)	0.707	1.266	0.756	1.173
			RIDGE(BIC)	0.665	1.346	0.747	1.187
			RIDGE(CP)	0.652	1.373	0.742	1.195
			RIDGE(CV)	0.618	1.447	0.708	1.253
			LASSO(CV)	0.940	0.952	0.964	0.920
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	3.193	2.521	4.227	1.889
			RIDGE(AIC)	4.003	2.011	4.780	1.671
			RIDGE(BIC)	3.243	2.482	4.603	1.735
			RIDGE(CP)	3.103	2.594	4.246	1.881
			RIDGE(CV)	3.198	2.517	4.286	1.863
			LASSO(CV)	5.098	1.579	6.174	1.293
0.6	3.605	1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	0.686	2.176	0.587	1.511
			RIDGE(AIC)	0.843	1.771	0.632	1.405
			RIDGE(BIC)	0.683	2.188	0.595	1.492
			RIDGE(CP)	0.657	2.273	0.587	1.511
			RIDGE(CV)	0.602	2.483	0.543	1.633
			LASSO(CV)	1.429	1.045	0.907	0.978
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	3.409	3.943	3.533	2.260
			RIDGE(AIC)	4.996	2.691	4.190	1.906
			RIDGE(BIC)	3.041	4.420	3.597	2.220
			RIDGE(CP)	3.171	4.240	3.537	2.258
			RIDGE(CV)	3.094	4.344	3.399	2.349
			LASSO(CV)	6.427	2.092	5.367	1.488
			EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	1.297	4.876	0.369	2.406

0.9	10.880	1	RIDGE(AIC)	2.090	3.024	0.450	1.971
			RIDGE(BIC)	1.140	5.544	0.364	2.436
			RIDGE(CP)	1.197	5.280	0.367	2.419
			RIDGE(CV)	0.574	11.011	0.287	3.092
			LASSO(CV)	3.674	1.720	0.654	1.356
		3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	8.990	6.328	2.681	2.979
			RIDGE(AIC)	16.477	3.453	3.549	2.250
			RIDGE(BIC)	6.601	8.619	2.531	3.155
			RIDGE(CP)	8.577	6.633	2.669	2.992
			RIDGE(CV)	3.449	16.498	2.049	3.897
			LASSO(CV)	17.448	3.261	3.948	2.023

Örnek 2 için simülasyon sonuçlarına Çizelge 2'den bakıldığında, çoklu doğrusallığın düşük, orta ve yüksek olduğu aralıklarda $\sigma = 1$'de $RHKO_\beta$ ve $RHKO_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter RIDGE(CV)'dir. $\sigma = 3$ ise en iyi sonucu veren kriter LASSO(CV)'dir. Fakat çoklu doğrusallığın yüksek ve çoklu doğrusal bağlantının artığı (CI değeri yüksek olduğu değerler) durumda ise EKK'ya göre en kötü sonucu veren kriter RIDGE(AIC) en iyi sonucu veren kriter ise RIDGE(CV) olduğu görülmektedir.

Çizelge 3: Örnek 3 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	HKO_β	$RHKO_\beta$	HKO_Y	$RHKO_Y$
0.3	1.886	1	EKK	0.894	1.000	0.887	1.000
			RIDGE(GCV)	0.805	1.111	0.819	1.083
			RIDGE(AIC)	0.821	1.089	0.827	1.073
			RIDGE(BIC)	0.808	1.107	0.821	1.081
			RIDGE(CP)	0.804	1.113	0.820	1.083
			RIDGE(CV)	0.775	1.154	0.816	1.087
			LASSO(CV)	0.660	1.355	0.691	1.285
		3	EKK	8.049	1.000	7.985	1.000
			RIDGE(GCV)	4.773	1.686	5.321	1.501
			RIDGE(AIC)	5.321	1.513	5.704	1.400
			RIDGE(BIC)	5.161	1.560	5.860	1.363
			RIDGE(CP)	4.695	1.714	5.288	1.510
			RIDGE(CV)	4.686	1.718	5.265	1.517
			LASSO(CV)	4.989	1.613	5.541	1.441
0.6	3.605	1	EKK	1.494	1.000	0.887	1.000
			RIDGE(GCV)	1.130	1.322	0.770	1.152
			RIDGE(AIC)	1.213	1.231	0.786	1.129
			RIDGE(BIC)	1.142	1.308	0.775	1.145
			RIDGE(CP)	1.116	1.338	0.771	1.150
			RIDGE(CV)	0.991	1.507	0.781	1.135
			LASSO(CV)	0.975	1.532	0.692	1.282
		3	EKK	13.443	1.000	7.958	1.000
			RIDGE(GCV)	5.783	2.324	4.831	1.653
			RIDGE(AIC)	7.077	1.900	5.274	1.514
			RIDGE(BIC)	5.865	2.292	5.097	1.567
			RIDGE(CP)	5.603	2.399	4.840	1.650
			RIDGE(CV)	5.222	2.574	4.548	1.756
			LASSO(CV)	6.358	2.114	5.129	1.557
		1	EKK	6.322	1.000	0.887	1.000
			RIDGE(GCV)	2.838	2.227	0.578	1.534
			RIDGE(AIC)	3.491	1.811	0.629	1.411
			RIDGE(BIC)	2.843	2.224	0.589	1.507
			RIDGE(CP)	2.730	2.315	0.577	1.537
			RIDGE(CV)	1.815	3.484	0.540	1.644

			LASSO(CV)	2.938	2.152	0.561	1.581
0.9	10.880	3	EKK	56.895	1.000	7.985	1.000
			RIDGE(GCV)	13.620	4.177	3.581	2.230
			RIDGE(AIC)	20.929	2.719	4.345	1.838
			RIDGE(BIC)	12.436	4.575	3.734	2.138
			RIDGE(CP)	12.846	4.429	3.586	2.227
			RIDGE(CV)	8.227	6.916	3.049	2.619
			LASSO(CV)	16.571	3.433	4.022	1.986

Örnek 3 için simülasyon sonuçlarına Çizelge 3'den bakıldığında, çoklu doğrusallığın düşük olduğu aralıkta $\sigma = 1$ de $RHKO_\beta$ ve $RHKO_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ ise en iyi sonucu RIDGE(CV) vermiştir. Çoklu doğrusallığın orta olduğu aralıkta $\sigma = 1$ de $RHKO_\beta$ ve $RHKO_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ ise en iyi sonucu RIDGE(CV) vermiştir. Çoklu doğrusallığın yüksek ve ÇDB artığı (CI değeri yüksek olduğu değerler) durumda ise EKK' ya göre en kötü sonucu veren kriter RIDGE(AIC) en iyi sonucu veren kriter ise RIDGE(CV) olduğu görülmektedir.

Çizelge 4: Örnek 4 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	HKO_β	$RHKO_\beta$	HKO_Y	$RHKO_Y$
0.3	2.551	1	EKK	0.530	1.000	0.457	1.000
			RIDGE(GCV)	0.498	1.065	0.439	1.042
			RIDGE(AIC)	0.504	1.052	0.440	1.038
			RIDGE(BIC)	0.493	1.075	0.444	1.029
			RIDGE(CP)	0.498	1.065	0.439	1.042
			RIDGE(CV)	0.536	0.989	0.532	0.860
			LASSO(CV)	0.220	2.409	0.224	2.037
		3	EKK	4.772	1.000	4.115	1.000
			RIDGE(GCV)	3.335	1.431	3.222	1.277
			RIDGE(AIC)	3.532	1.351	3.283	1.254
			RIDGE(BIC)	3.247	1.470	3.598	1.144
			RIDGE(CP)	3.328	1.434	3.224	1.276
			RIDGE(CV)	3.318	1.438	3.235	1.272
			LASSO(CV)	1.985	2.403	2.023	2.035
0.6	4.838	1	EKK	0.930	1.000	0.457	1.000
			RIDGE(GCV)	0.729	1.275	0.402	1.139
			RIDGE(AIC)	0.766	1.215	0.406	1.127
			RIDGE(BIC)	0.680	1.368	0.418	1.094
			RIDGE(CP)	0.728	1.277	0.402	1.138
			RIDGE(CV)	0.657	1.416	0.438	1.044
			LASSO(CV)	0.283	3.285	0.190	2.410
		3	EKK	8.369	1.000	4.115	1.000
			RIDGE(GCV)	3.556	2.353	2.538	1.621
			RIDGE(AIC)	4.051	2.066	2.622	1.569
			RIDGE(BIC)	2.838	2.949	2.882	1.428
			RIDGE(CP)	3.518	2.379	2.544	1.618
			RIDGE(CV)	3.531	2.370	2.547	1.616
			LASSO(CV)	2.546	3.287	1.708	2.409
		1	EKK	4.119	1.000	0.457	1.000
			RIDGE(GCV)	1.430	2.880	0.260	1.756
			RIDGE(AIC)	1.670	2.467	0.270	1.694
			RIDGE(BIC)	1.110	3.712	0.296	1.547
			RIDGE(CP)	1.409	2.923	0.261	1.753
			RIDGE(CV)	1.072	3.842	0.294	1.556
			LASSO(CV)	0.933	4.415	0.155	2.952
			EKK	37.067	1.000	4.115	1.000

0.9	18.898	3	RIDGE(GCV)	4.594	8.068	1.416	2.906
			RIDGE(AIC)	6.254	5.927	1.526	2.697
			RIDGE(BIC)	2.614	14.183	1.687	2.439
			RIDGE(CP)	4.511	8.217	1.427	2.883
			RIDGE(CV)	4.155	8.922	1.382	2.978
			LASSO(CV)	8.117	4.567	1.376	2.992

Örnek 4 için simülasyon sonuçlarına Çizelge 4'den bakıldığında, çoklu doğrusallığın düşük ve orta olduğu aralıkta $\sigma = 1$ ve $\sigma = 3$ de $RHKO_\beta$ ve $RHKO_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV) olduğu görülmüştür. Çoklu doğrusallığın yüksek ve ÇDB' nın yüksek oranda artığı (CI değeri yüksek olduğu değerler) durumda ise EKK' ya göre en kötü sonucu veren kriter $\sigma = 1$'de $RHKO_\beta$'ya göre RIDGE(AIC), $RHKO_Y$'ye göre RIDGE(BIC) en iyi sonucu ise LASSO(CV) vermiştir. $\sigma = 3$ ise en kötü sonucu $RHKO_\beta$ göre RIDGE (AIC), $RHKO_Y$ göre RIDGE(BIC)'dir, en iyi sonucu da $RHKO_\beta$ göre RIDGE(BIC), $RHKO_Y$ göre LASSO(CV) olduğu görülmüştür.

Çizelge 5: Örnek 5 için Simülasyon Çıktısı

ρ	CI	σ	YÖNTEM	HKO_β	$RHKO_\beta$	HKO_Y	$RHKO_Y$
0.3	2.724	1	EKK	0.825	1.000	0.710	1.000
			RIDGE(GCV)	0.787	1.048	0.688	1.033
			RIDGE(AIC)	0.797	1.035	0.691	1.028
			RIDGE(BIC)	0.783	1.053	0.691	1.027
			RIDGE(CP)	0.787	1.048	0.688	1.033
			RIDGE(CV)	0.931	0.886	0.938	0.757
			LASSO(CV)	0.514	1.604	0.488	1.454
		3	EKK	7.424	1.000	6.392	1.000
			RIDGE(GCV)	5.510	1.347	5.185	1.233
			RIDGE(AIC)	5.911	1.256	5.328	1.200
			RIDGE(BIC)	5.351	1.387	5.476	1.167
			RIDGE(CP)	5.496	1.351	5.185	1.233
			RIDGE(CV)	5.464	1.359	5.193	1.231
			LASSO(CV)	4.633	1.602	4.402	1.452
0.6	5.124	1	EKK	1.443	1.000	0.710	1.000
			RIDGE(GCV)	1.159	1.246	0.630	1.127
			RIDGE(AIC)	1.231	1.173	0.640	1.110
			RIDGE(BIC)	1.107	1.304	0.644	1.102
			RIDGE(CP)	1.156	1.249	0.630	1.127
			RIDGE(CV)	1.078	1.339	0.742	0.958
			LASSO(CV)	0.724	1.995	0.423	1.679
		3	EKK	12.991	1.000	6.392	1.000
			RIDGE(GCV)	5.736	2.265	3.980	1.606
			RIDGE(AIC)	6.879	1.889	4.180	1.529
			RIDGE(BIC)	4.804	2.704	4.390	1.456
			RIDGE(CP)	5.670	2.291	3.988	1.603
			RIDGE(CV)	5.698	2.280	3.988	1.603
			LASSO(CV)	6.543	1.985	3.825	1.671
0.9	20.497	1	EKK	6.400	1.000	0.710	1.000
			RIDGE(GCV)	2.567	2.493	0.417	1.701
			RIDGE(AIC)	3.120	2.051	0.444	1.599
			RIDGE(BIC)	2.210	2.896	0.464	1.530
			RIDGE(CP)	2.540	2.519	0.418	1.698
			RIDGE(CV)	2.222	2.880	0.567	1.252
			LASSO(CV)	2.927	2.187	0.435	1.633
			EKK	57.596	1.000	6.392	1.000
			RIDGE(GCV)	7.998	7.202	2.139	2.988
			RIDGE(AIC)	11.856	4.858	2.394	2.670

3	RIDGE(BIC)	5.423	10.621	2.565	2.492
	RIDGE(CP)	7.805	7.380	2.152	2.970
	RIDGE(CV)	7.341	7.845	2.080	3.073
	LASSO(CV)	22.891	2.516	3.135	2.039

Örnek 5 için simülasyon sonuçlarına Çizelge 5'den bakıldığında, çoklu doğrusallığın düşük olduğu aralıkta $\sigma = 1$ ve $\sigma = 3$ de $RHKO_\beta$ ve $RHKO_Y$ değerlerinde EKK'ya göre en iyi sonucu veren kriter LASSO(CV) olduğu görülmüştür. Çoklu doğrusallığın orta olduğu aralıkta $\sigma = 1$ en iyi sonucu veren kriter LASSO(CV), $\sigma = 3$ de $RMSE_\beta$ göre en iyi sonuç RIDGE(CV) ve $RMSE_Y$ değerlerinde ise LASSO(CV) en iyi sonucu vermiştir. Çoklu doğrusallığın ve ÇDB' nın yüksek olduğu değerde $\sigma = 1$ için en iyi sonuç $RHKO_\beta$ göre RIDGE(BIC), $RHKO_Y$ göre ise RIDGE(GCV)'dir. $\sigma = 3$ için en kötü sonucu $RHKO_\beta$ ve $RHKO_Y$ değerlerinde LASSO(CV) vermiştir. En iyi sonucu ise $RHKO_\beta$ göre RIDGE(BIC), $RHKO_Y$ göre de RIDGE(CV) vermiştir.

Simülasyon sonuçlarına bakıldığında ÇDB'nın yüksek olduğu (CI yüksek olduğu değer) durumda Ridge regresyon ayar parametresi seçim kriterleri EKK tahmininden daha iyi sonuçlar verdiği görülmüştür. ÇDB' nın en yüksek olduğu durumda EKK'dan sonra en kötü sonucu LASSO(CV) vermiştir. En iyi sonuçları ise RIDGE(BIC) ve RIDGE(CV) vermiştir.

4. Uygulamanın amacı ve modeli

Bu çalışmanın amacı, Doğrudan Yabancı Yatırımı (DYY) etkiyen faktörlerle kurulan modelle yapılan analiz sonucunda, meydan gelen ÇDB problemini çözmek için yanlı tahmin edici Ridge regresyon analizi yapmaktır. Yapılan Ridge regresyon analizi ile birlikte Ridge ayar parametre tahmin edicilerinin hangisinin daha iyi sonuç vereceği incelenmiştir.

Çalışmada kullanılan veriler Dünya Bankasından (World Bank) https://databank.worldbank.org/ (18.10.2021) elde edilmiştir. Elde edilen veri seti, Türkiye'nin 1974-2019 yıllarına ait doğrudan yabancı yatırımlarını etkileyen faktörlerin verileridir. 46 gözlem ve 8 değişkenle kurulan çoklu doğrusal regresyon modeli aşağıdaki gibidir [21];

$\log(DYY)_i = \beta_0 + \beta_1 KBGB_i + \beta_2 EGD_i + \beta_3 MH\dot{I}H_i + \beta_4 MH\dot{I}T_i + \beta_5 GHNTH_i + \beta_6 AKH_i + \beta_7 CHD_i + \varepsilon_i$, $i = 1,2,\ldots,46$.

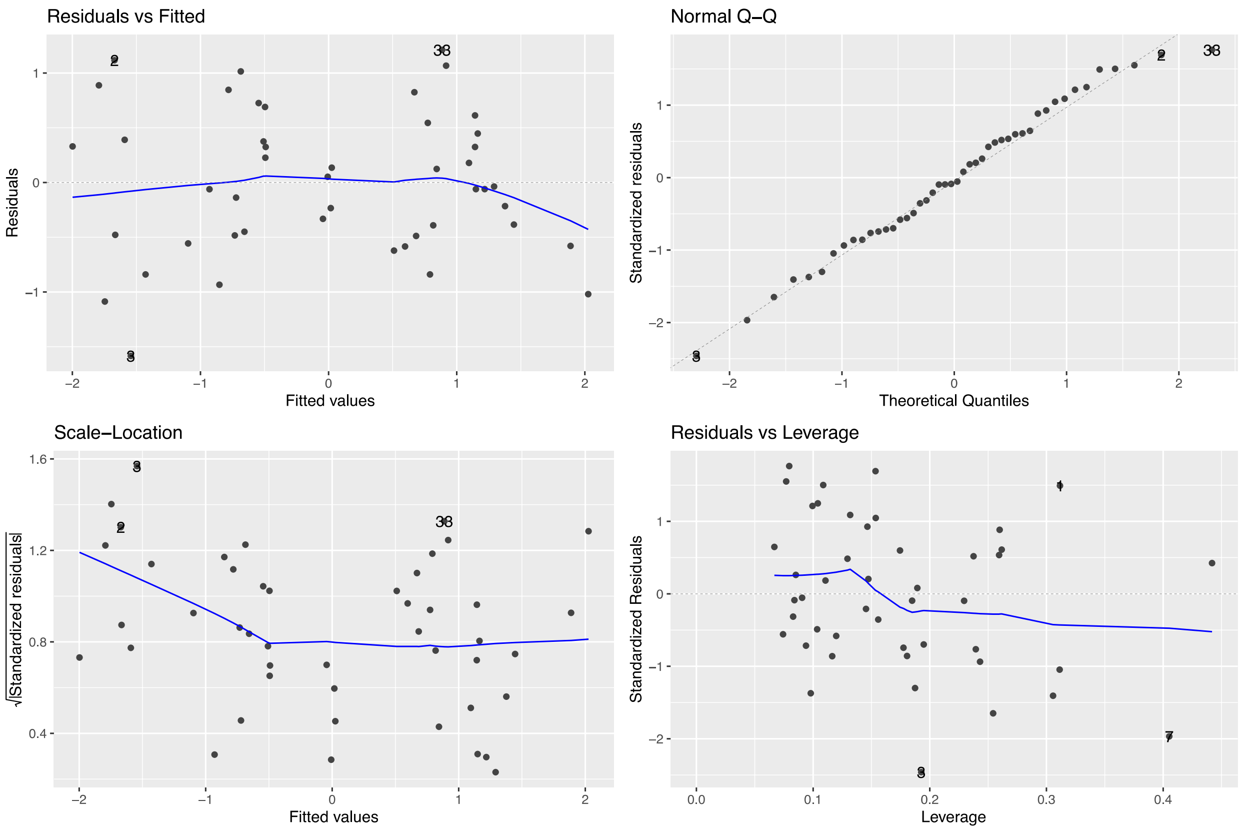
Kurulan regresyon modelindeki bağımlı ve bağımsız değişkenlerin açıklamaları Çizelge 6'da detaylı bir şekilde gösterilmiştir;

Çizelge 6: Değişkenler ve Açıklamaları

DEGİŞKENLER	**AÇIKLAMALAR**
Bağımlı Değişkenler	
DYY	Doğrudan yabancı yatırımlar, net girişler (GSYİH'nin yüzdesi)
Bağımsız Değişkenler	
KBGB	Kişi başına GSYİH büyümesi (Yıllık %)
EGD	Enflasyon, GSYİH deflatörü (Yıllık %)
MHİH	Mal ve hizmet ihracatı (GSYİH'nin yüzdesi)
MHİT	Mal ve hizmet ithalatı (GSYİH'nin yüzdesi)
GHNTH	Genel hükümet nihai tüketim harcaması (GSYİH'nin yüzdesi)
AKH	Alınan kişisel havaleler (GSYİH' nin yüzdesi)
CHD	Cari hesap dengesi (GSYİH' nin yüzdesi)

4.1.Model varsayımların kontrolü

Doğrusal regresyon modelinde veriler hakkında bazı çeşitli varsayımlar bulunur. Bunları Şekil 1'de özetleyebiliriz. Şekil 1'in üst ve sol tarafında verilen grafikte tahmin edilen bağımlı değişken (x-ekseninde) ve kalıntılar (y-ekseninde) olmak üzere doğrusallık varsayımının sınanması yapılmaktadır. Burada, mavi renkteki doğru yatay eksene yakın olması doğrusallık varsayımın sağlandığını göstermektedir. Şekil 1'in üst ve sağ tarafında verilen grafikte kalıntıların normal dağılıp dağılmadığını incelemek için kullanılır. Burada, kalıntı noktaların düz kesikli çizgiyi takip etmesi normallik varsayımının dağlandığını gösterir ki, varsayım sağlanmış olur. Şekil 1'in alt ve sol tarafında verilen grafikte kalıntıların varyansının homojenliğini kontrol etmek için kullanılır. Mavi renkteki düz çizginin yatay ekseni izlemesi değişen varyans problemimizin olmadığını gösterir ki, burada varsayım sağlanmış olur. Son olarak, Şekil 1'in alt ve sağ tarafında verilen grafikte kalıntılara karşı kaldıraç grafiği olup, regresyon modelinde etkili gözlemleri tanımlamamıza yardımcı olur. Burada herhangi bir nokta Cook'un mesafesinin (normalde mavi kesikli çizgi olarak görünmeli) dışında kalmadığı için, varsayım sağlanmış olup aykırı bir değer yoktur. Ayrıca, Kolmogorov-Smirnov ve Breusch-Pagan testleri ile sırasıyla kalıntıların normallik ve sabit varyans varsayımlarının sağlandığı kontrol edilmiştir.



Şekil 1. Model varsayım grafikleri

ÇDB'yı belirleyebilmek için birçok yöntem vardır. Bunlardan bazıları korelasyon matrisinin incelenmesi, varyans artış faktörleri (VIF), öz değerler ve öz vektör analizidir. Bu yöntemler sayesinde sorunun boyutunu ve hangi değişkenden kaynaklandığı görülmektedir. İki bağımsız değişken arasındaki korelasyon katsayısının 1'e yakın olması doğrusal bağlantının olduğuna güçlü bir işarettir. MHİT ve MHİH bağımsız değişkenlerinin VIF değerlerini 10'dan büyük oluğu Çizelge 7'de belirtilen analiz sonucunda tespit edilmiş ve çoklu doğrusal bağlantının varlığı ispat edilmiştir. Ayrıca $CI = 19.86$ değerine bakıldığında 10'dan büyük olduğu görülmektedir [22]. İhracat ve ithalat verileri arasındaki kuvvetli korelasyon çoklu doğrusal bağlantı problemine neden olduğu görülmektedir.

Çizelge 7: VIF değerleri

KBGB	EDG	MHİT	GHNTH	KBGB	MHİH	CHD
7.88	2.2	45.18	3.57	2.32	54.78	7.05

Çizelge 8'de bağımsız değişkenler arasındaki korelasyon değerleri verilmiştir. Korelasyon değerleri -1 ile $+1$ arasında değerler alır. Korelasyon değerleri 0 ile 1 arasında ise pozitif yönde, -1 ile 0 arasında ise negatif yönde bir korelasyon söz konusudur. Genel olarak korelasyon katsayısının 0.6 ile 0.8 arasında olması yüksek korelasyon, 0.8'in üzerinde olması ise çok yüksek korelasyon olduğunu göstermektedir. Grafik incelendiğinde, bağımsız değişkenlerden MHİT ve MHİH arasında 0.938 oranında bir korelasyon olduğu görülmektedir. Buna göre ithalat ve ihracat arasındaki korelasyonun pozitif yönde oldukça yüksek olduğu görülmektedir. Diğer değişkenler arasında korelasyon değerleri ve saçılım grafiklerine bakıldığında, düşük ve orta düzeyde zayıf bir korelasyon olduğu gözlemlenmiştir. Değişkenlere bakıldığında, MHİT ile AKH arasında -0.718, GHNTH ile AKH arasında -0.773 ve MHİH ile AKH arasında -0.717 oranında çıkan sonuçlara göre yüksek oranda ters yönde negatif bir korelasyon olduğu görülmektedir. EDG ile AKH arasında 0.566, CHD ile EDG arasında 0.606, GHNTH ile MHİT arasında 0.552 ve MHİH ile GHNTH arasında 0.515 oranında çıkan sonuçlara göre pozitif yönde orta düzeyde bir korelasyon olduğu görülmektedir.

Çizelge 8. Açıklayıcı değişkenler arasındaki korelasyon değerleri

	AKH	EDG	MHİT	GHNTH	KBGB	MHİH
EDG	0.566					
MHİT	-0.717	-0.339				
GHNTH	-0.773	-0.443	0.552			
KBGB	-0.167	-0.342	0.278	-0.063		
MHİH	-0.718	-0.251	0.938	0.515	0.069	
CHD	0.401	0.606	-0.246	-0.365	-0.513	0.024

Analize başlamadan önce veri seti rastgele olacak şekilde, 23 gözlemli eğitim ve 23 gözlemli test seti olarak iki parçaya bölünmüştür. Bağımlı ve bağımsız değişkenleri, eğitim veri setine dayalı olarak ortalarız, yani $\bar{X}_{eğitim} = \left(\bar{X}_{1,eğitim}, \dots, \bar{X}_{p,eğitim}\right)$ ve $\bar{Y}_{eğitim} = (Y_{eğitim} - \bar{Y}_{eğitim})$. Model eğitim verisi kullanılarak kurulmuştur. Test veri setini kullanarak, herhangi bir $\hat{\beta}^*$ tahmin edicisi için TH değeri şu şekilde hesaplanmıştır: $TH(\hat{\beta}^*) = \frac{1}{n_{test}} r'_{test} r_{test}$ ise $r_{i,test} = Y_{test} - \left(\bar{Y}_{eğitim} + \left(X_i + \bar{X}_{eğitim}\right)' \hat{\beta}^*\right)$. Veri setini ikiye bölme işlemi rastgele bir süreç olduğu için, rastgeleliği en az düzeye indirmek için, bu süreç 1000 kez tekrar edilmiştir ve 1000 adet TH'nin ortalaması alınarak sonuç raporlanmıştır. Çizelge 9'da yapılan analiz sonuçlarının TH ve Rölatif TH (RTH) değerleri verilmiştir. Burada $RTH(\hat{\beta}^*) = \frac{TH(\hat{\beta})}{TH(\hat{\beta}^*)}$ şeklinde tanımlanmıştır. Eğer *RTH* değeri 1'den büyük ise, bu $\hat{\beta}^*$ tahmin edicisinin EKK tahmin edicisinden daha iyi bir performans gösterdiğini gösterir.

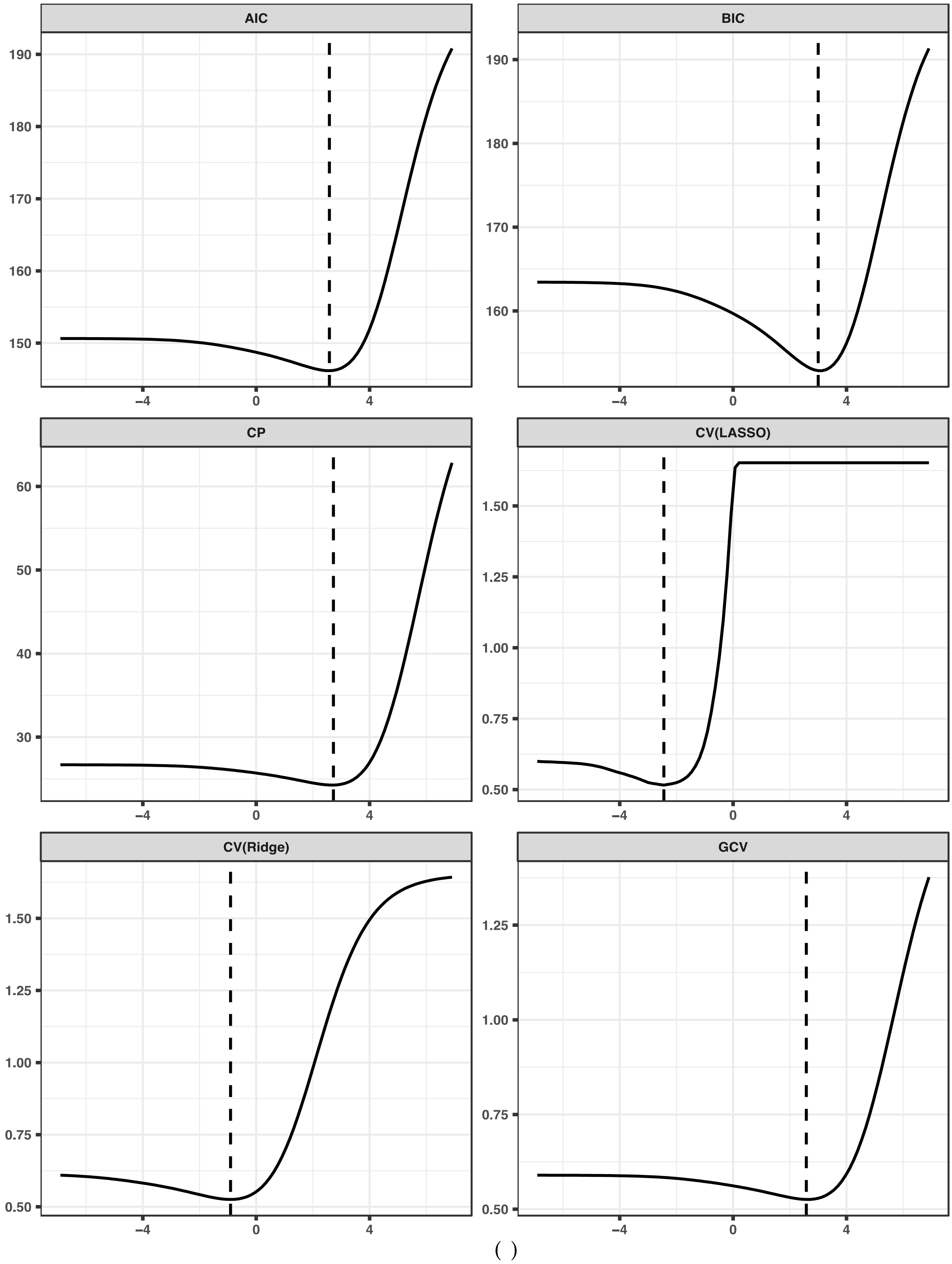
Çizelge 9: Parametre tahminleri ve tahmin edicilerin performansları

	EKK	RIDGE(GCV)	RIDGE(AIC)	RIDGE(BIC)	RIDGE(CP)	RIDGE(CV)	LASSO(CV)
AKH	-0.239	-0.209	-0.209	-0.209	-0.209	-0.209	0.197
EDG	-0.141	-0.096	-0.096	-0.091	-0.94	-0.095	0.023
MHİT	0.705	0.370	0.370	0.342	0.361	0.364	0.390
GHNTH	0.084	0.140	0.140	0.144	0.142	0.141	0.087
KBGB	-0.088	-0.028	-0.028	-0.012	-0.017	-0.018	0.000
MHİH	0.161	0.369	0.369	0.347	0.363	0.365	0.423
CHD	0.104	0.025	0.025	0.019	0.023	0.024	0.000
k	-	13.219	13.219	20.092	15.199	0.404	0.087
TH	0.973	0.676	0.785	0.666	0.676	0.656	0.685
RTH	1.000	1.440	1.240	1.462	1.440	1.483	1.421

Çizelge 9'a göre, tüm değerler için *RTH* değerleri üzerinden yorum yapılacak olursa en iyi tahmini RIDGE(CV) 1.483 değeriyle verdiği görülmektedir. Kurulan regresyon modelindeki bağımsız değişkenlerin bağımlı değişkene etkilerini analiz sonuçlarına göre yorumlamak da mümkündür. En iyi tahmini veren RIDGE(CV) kriterine bakıldığında, değişkenlerdeki %1 oranındaki artışta bağımlı değişken olan DYY etkileri açık olarak görülmektedir. EKK'ya göre en iyi sonucu veren bilgi kriteri RIDGE(CV)'ye göre analizi yorumlanırsa, bağımsız değişkenlerin %1'lik artışta bağımlı değişken olan DYY'yi etkilerini görmek mümkündür. %1'lik artışta bağımsız değişkenler AKH %0,20 oranında, EDG %0.09 oranında, KBGB %0.01 oranında DYY'yi azaltmaktadır. MHİT %0.36 oranında, GHNTH %0.14 oaranında, MHİH %0.36 oranında ve CHD %0.02 oranında DYY'yi arttırdığı görülmektedir. Analiz yorumlarına bakıldığında, pozitif yönde DYY'yi etkileyen en iyi bağımsız değişkenler MHİT ve MHİH olduğu görülmüştür. Ekonomik olarak bakıldığında, doğrudan yabancı yatırımı etkileyen en iyi değişkenler ithalat ve ihracat olduğu söylenilebilir. Çizelge 10'da tahmin edilen modeller verilmiştir;

Çizelge 10: Tahmin edilen modeller

EKK	DYY = -0.239x*AKH* - 0.141x*EDG* + 0.705x*MHİT* + 0.084x*GHNTH* - 0.088x*KBGB* + 0.161x*MHİH* + 0.104x*CHD*
RIDGE(GCV)	DYY = -0.209x*AKH* - 0.096x*EDG* + 0.370x*MHİT* + 0.140x*GHNTH* - 0.020x*KBGB* + 0.369x*MHİH* + 0.025x*CHD*
RIDGE(AIC)	DYY = -0.209x*AKH* - 0.096x*EDG* + 0.370x*MHİT* + 0.140x*GHNTH* - 0.020x*KBGB* + 0.369x*MHİH* + 0.025x*CHD*
RIDGE(BIC)	DYY = -0.209x*AKH* - 0.091x*EDG* + 0.342x*MHİT* + 0.144x*GHNTH* - 0.012x*KBGB* + 0.347x*MHİH* + 0.019x*CHD*
RIDGE(CP)	DYY = -0.209x*AKH* - 0.094x*EDG* + 0.361x*MHİT* + 0.142x*GHNTH* - 0.017x*KBGB* + 0.363x*MHİH* + 0.023x*CHD*
RIDGE(CV)	DYY = -0.209x*AKH* - 0.095x*EDG* + 0.364x*MHİT* + 0.141x*GHNTH* - 0.018x*KBGB* + 0.365x*MHİH* + 0.024x*CHD*
LASSO(CV)	DYY = 0.197x*AKH* + 0.023x*EDG* + 0.390x*MHİT* + 0.087x*GHNTH* + 0.000x*KBGB* + 0.423x*MHİH* + 0.000x*CHD*



( )

Şekil 2. Seçim kriterleri için log*k* ayar parametresi grafikleri

Şekil 2'de *k* ayar parametresi seçimi için kullanılan kriterlerin (AIC, BIC, CP, CV(LASSO), CV(RIDGE), GCV) optimum noktaları verilmiştir.

5. Sonuç ve öneriler

Çoklu doğrusal regresyon analizinde parametrelerin tahmini için genellikle kullanılan EKK yöntemi ile ancak belirli varsayımların sağlanması durumunda güvenilir sonuçlar elde edilir. Bu temel varsayımlardan biri olan bağımsız değişkenler arasında ÇDB'nın olmamasıdır. Değişkenler arasında ÇDB'nın varlığı durumunda EKK yöntemi ile elde edilen sonuçlar güvenilir olamamaktadır. ÇDB'nın EKK tahmin edicileri üzerindeki birçok olumsuz etkisi bulunduğundan, bu problemin saptanmasına yarayan yöntemler ve bu problemin düzeltilmesi veya en aza indirgenmesi amacıyla önerilen Ridge regresyon ayrıntılarıyla incelenmiştir.

Bu çalışmada hem Monte Carlo simülasyon çalışması hem de gerçek veri uygulaması yapılmış ve sonuçlar R programı yardımı ile elde edilmiştir. Yapılan simülasyon çalışmasının sonucunda, en iyi değeri veren RIDGE(CV) kriteri olduğu gözlemlenmiştir. Gerçek veri setinde açıklayıcı değişkenler arasındaki yüksek derecede korelasyon değerlerinin olmasından dolayı EKK ile uygun bir çözüm elde edilememiştir. Ayrıca açıklayıcı değişkenler için koşul indeksi, VIF değerleri ÇDB sorununu göstermiştir. Dolayısıyla çoklu doğrusal regresyon modeline alternatif Ridge regresyon tahmin edicisi önerilmiştir. Ridge regresyon EKK yöntemine göre daha iyi sonuçlar vermiştir. Gerçek veri seti ile yapılan çalışmada tüm kriterler için RTH değerleri üzerinden yapılan yorumlar sonucunda RIDGE(CV)'nin en iyi sonucu verdiği gözlemlenmiştir.

Bundan sonraki çalışmalar için, araştırmacılara genelleştirilmiş lineer modeller için Ridge regresyon tahmin edicisi için ayar parametresinin seçimini açık problem olarak bırakıyoruz.

Teşekkür

Yazarlar, dergi Editörü ve iki hakeme, makaleyi ayrıntılı olarak okudukları ve makalenin önemli ölçüde iyileştirilmesine yol açan değerli yorum ve önerileri için teşekkür eder. Bu çalışma, İnönü Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı'nda Doç. Dr. Bahadır Yüzbaşı danışmanlığında Mustafa Pala tarafından "Ridge Regresyon Parametre Seçimi: Türkiye'nin Doğrudan Yabancı Yatırım Örneği" başlığı ile tamamlanarak 17.01.2022 tarihinde savunulan Yüksek Lisans tezinden türetilmiştir.

Kaynaklar

[1] P. Kennedy, 2003, *A guide to econometrics*. 5th. Edition. Cambridge.
[2] R. Frisch, 1934, Statistical confluence analysis by means of complete regression systems, *Norway, Institute of Economics Oslo*.
[3] A. E. Hoerl, R. W. Kennard, 1970a, Ridge regression: Applications to non-orthogonal problems, *Technometrics*, 12, 1, 69-82.
[4] A. E. Hoerl, R. W. Kennard, 1970b, Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics*, 12, 1, 55-67.
[5] G. C. McDonald, D. I. Galarneau, 1975, A Monte Carlo evaluation of some Ridge-type estimators. *Journal of the American Statistical Association*, *70*, 350, 407-416.
[6] J. F. Lawless, P. Wang, 1976, A simulation study of Ridge and other regression estimators, *Communications in Statistics*, 5, 307-323.
[7] G. H. Golub, M. Heath, G. Wahba, 1979, Generalized cross-validation as a method for choosing a good Ridge parameter. *Technometrics*, 21,2, 215-223.
[8] G. Khalaf, G. Shukur, 2005, Choosing Ridge parameter for regression problems. Communications in Statistics-Theory and Methods, 34(5), 1177-1182.
[9] M. Alkhamisi, G. Khalaf, G. Shukur, 2006, Some modifications for choosing Ridge parameters. *Communications in Statistics-Theory and Methods*, 35, 11, 2005-2020.
[10] M. A. Alkhamisi, G. Shukur, 2007, A Monte Carlo study of recent Ridge parameters. *Communications in Statistics-Simulation and Computation*, 36, 3, 535-547.
[11] 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.

[12] A. T. Owolabi, K. Ayinde, O. O. Alabi, 2022. A new Ridge-type estimator for the linear regression model with correlated regressors. *Concurrency and Computation: Practice and Experience*, 34,15, 6933.

[13] S. Karakaş, 2008, *Çoklu doğrusal bağlantı problemi ve yanlı regresyon tahmincileri*, İstanbul Üniversitesi, İstanbul.

[14] H. Akaike, 1998, Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, Springer, New York, NY.

[15] G. Schwarz, 1978, Estimating the Dimension of a Model. Annals of Statistics, 6, 461-464. http://dx.doi.org/10.1214/aos/1176344136.

[16] A. Küçük, 2019, *Doğrusal regresyonda Ridge, Liu ve LASSO tahmin edicileri üzerine bir çalışma,* Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara.

[17] R. Tibshirani, 1996, Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.

[18] H. Zou, T. Hastie, 2005, Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67,2, 301-320.

[19] G. Tutz, J. Ulbricht, 2009, Penalized regression with correlation-based penalty. *Statistics and Computing*, 19,3, 239-253.

[20] M. Arashi, Y. Asar, B. Yüzbaşı, 2021, SLASSO: SLASSO: A scaled LASSO for multicollinear situations. *J. Stat. Comput. Simul*, 91,15, 3170-3183, https://doi.org/10.1080/00949655.2021.1924174.

[21] B. Yüzbaşı, S. E. Ahmed, 2020, Ridge type shrinkage estimation of seemingly unrelated regressions and analytics of economic and financial data from “fragile five” countries. *Journal of Risk and Financial Management*, 13,6, 131, https://doi.org/10.3390/jrfm13060131.

[22] E. Karaoğlan, 2019, Regresyon analizinde çoklu doğrusal bağlantı probleminin incelenmesi: Temel bileşenler*, Ondokuz Mayıs Üniversitesi, Samsun*.