BŞEÜ Fen Bilimleri Dergisi
9(1), 80-91, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN: 2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

Araştırma Makalesi -Research Article

Bir Simülasyon Çalışması ile Cezalı Regresyon Yöntemlerinin Karşılaştırılması

Comparison of Penalized Regression Methods through a Simulation Study

Murat Genç[1*]

Geliş / Received: 13/09/2021 ***Revize / Revised: 25/02/2022*** ***Kabul / Accepted: 07/03/2022***

ÖZ

Veri kümesinde çoklu iç ilişki problemi olması durumunda kararlı katsayı tahminleri elde etmek için sıklıkla cezalı regresyon yöntemleri kullanılır. Ayrıca bu yöntemler uygulanan ceza teriminin yapısına bağlı olarak otomatik değişken seçimi de yapabilmektedir. Bu çalışmada literatürde yaygın kullanım alanı bulan ridge, LASSO, elastik net ve uyarlanabilir LASSO cezalı regresyon yöntemlerinin gerçek katsayı vektörünün yapısına bağlı olarak simülasyon çalışmaları yoluyla performanslarının ayrıntılı olarak karşılaştırılması yapılmıştır. Çalışmada karşılaştırma kriteri olarak test kümesi üzerinde hata kareler ortalaması, yanlış sınıflama oranı, yanlış pozitif oranı ve aktif küme büyüklükleri kullanılmıştır. Simülasyon çalışmaları, gerçek katsayı vektörünün yapısının yöntemlerin ortaya çıkardığı model performansı üzerinde önemli etkisinin olduğunu göstermektedir.

Anahtar Kelimeler- Doğrusal Regresyon, Ridge, Lasso, Elastik Net, Çoklu İç İlişki

ABSTRACT

Penalized regression methods are often used to obtain stable coefficient estimates in case of multicollinearity problems in the dataset. In addition, these methods can make automatic variable selection depending on the nature of the penalty term applied. In this study, a detailed comparison of the performances of ridge, LASSO, elastic net and adaptive LASSO penalized regression methods, which are widely used in the literature, is made through simulation studies depending on the structure of the real coefficient vector. Mean squared error on the test set, misclassification rate, false positive rate and active set sizes are used as comparison criteria in the study. Simulation studies show that the structure of the real coefficient vector has a significant effect on the model performance revealed by the methods.

Keywords- Linear Regression, Ridge, Lasso, Elastic Net, Multicollinearity

[1]*Sorumlu yazar iletişim: muratgenc@tarsus.edu.tr (https://orcid.org/0000-0002-6335-3044)
Yönetim Bilişim Sistemleri Bölümü, Tarsus Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Takbaş Mahallesi Kartaltepe Sokak 33400 Tarsus, Mersin, Türkiye

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

I. GİRİŞ

Verinin oluşmasını sağlayan sistem hakkında bilgi edinmek ve yeni gözlemlerin ön tahminini yapmak veri analizinin temel hedeflerindendir. Veri analizinde kullanılan en yaygın yöntemlerden biri doğrusal regresyon modellemesidir. Doğrusal regresyon modellerinin doğruluğu ve yorumu genellikle regresyon katsayılarının tahminine bağlıdır y, $n \times 1$ boyutlu yanıt değişkenlerin gözlem vektörü; X, $n \times p$ boyutlu açıklayıcı değişkenlerin gözlem matrisi; β, $p \times 1$ boyutlu bilinmeyen regresyon katsayıları vektörü ve ε, $n \times 1$ boyutlu 0 ortalamalı ve σ^2I varyans-kovaryans matrisli hata terimleri vektörü olmak üzere

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon \tag{1}$$

doğrusal regresyon modeli ele alınsın. Regresyon katsayılarının klasik en küçük kareler tahmin edicisi (EKK)

$$\hat{\beta}_{ekk} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T\mathrm{y} \tag{2}$$

olarak tanımlanır. Klasik varsayımlar sağlandığında EKK, yansızdır ve tüm doğrusal yansız tahmin ediciler içinde en küçük varyansa (Best Linear Unbiased Estimator, BLUE) sahiptir. Fakat veri kümesinde açıklayıcı değişkenler arasında bulunan doğrusal bağıntı olarak tanımlanan çoklu iç ilişkinin bulunması durumunda EKK tahminleri gerçek katsayı değerlerinden uzak ve tahminlerin varyansı çok büyük olur. Çoklu iç ilişki problemi azaltılabilir; ancak tamamıyla ortadan kaldırılamaz [1]. Veri kümesinde çoklu iç ilişki olması durumunda ridge regresyon [2], kısıtlı en küçük kareler tahmin edicisi [3] ve kısıtlı ridge tahmin edici [4] EKK'ya alternatif olarak önerilen yanlı tahmin ediciler arasında gösterilebilir. Ayrıca ridge regresyona dayalı olarak daha iyi performansa sahip tahmin ediciler de önerilmiştir [5, 6].

Regresyon analizinde modele alınan açıklayıcı değişkenler model performansını önemli ölçüde etkilemektedir. Modele alınacak açıklayıcı değişkenlerin belirlenmesi için farklı yöntemler önerilmiştir. En iyi alt küme seçim ve adımsal alt küme seçim yöntemlerinin [7] yanı sıra bridge regresyon [8], LASSO (least absolute shrinkage and selection operator) [9], elastik net [10] ve uyarlanabilir (adaptive) LASSO [11] gibi sürekli bir ceza teriminin kullanıldığı yöntemler bu yöntemler arasında gösterilebilir. Bu yöntemlerden en iyi alt küme seçim yönteminde açıklayıcı değişkenlerin tüm kombinasyonlarına dayalı olarak oluşturulan regresyon modelleri dikkate alınarak yanıt değişkeni üzerinde güçlü etkisi olan açıklayıcı değişkenlerin modele dahil edilmesi hedeflenir. Dolayısıyla hesaplama maliyeti yüksek olan bir yöntemdir. Bridge regresyon, $\|\beta\|_1^\gamma$, $\gamma > 0$ ceza fonksiyonuna sahip bir cezalı regresyon yöntemleri sınıfıdır. LASSO, bridge regresyonun ceza fonksiyonunda $\gamma = 1$ durumuna karşılık gelir ve konveks bir ceza fonksiyonuna sahiptir. Elastik net, LASSO ceza fonksiyonuna yeni bir cezalandırma teriminin eklenmesi ile elde edilir. Bu bakımdan elastik net, LASSO'nun bir genelleştirilmesi olarak görülebilir. Uyarlanabilir LASSO ise LASSO'nun ceza fonksiyonunun uyarlanabilir ağırlıklar kullanılarak güncellenmesine dayalı iki aşamalı bir yöntemdir.

Cezalı regresyon yöntemleri hem gözlem sayısının açıklayıcı değişken sayısından büyük olduğu klasik veri kümelerinde hem de açıklayıcı değişken sayısının gözlem sayısını aştığı yüksek boyutlu veri kümelerinde katsayı tahmini için yaygın olarak kullanılan yöntemlerdir (bkz. [12, 13, 14]). Literatürde simülasyona dayalı olarak farklı doğrusal regresyon modellerinin karşılaştırılmasına dair çeşitli çalışmalar bulunmaktadır. [15], çalışmasında alt modellerin entegre edilmesinin etkisini incelemek amacıyla bir simülasyon çalışması yapmıştır. [16], yüksek boyutlu veri analizi için en küçük kareler ve temel bileşenler regresyonda bootstrap yöntemine dayalı olarak değişken seçimi üzerine bir simülasyon çalışması gerçekleştirmiştir.

Bu çalışmada klasik veri kümelerinde gerçek katsayı vektörünün yapısı, sıfır değerli katsayıların miktarı ve konumuna bağlı olarak EKK tahmin edicisi ile konveks ceza fonksiyonuna sahip ridge, LASSO, elastik net ve uyarlanabilir LASSO cezalı regresyon yöntemlerinin doğrusal regresyon modelinin performansı üzerindeki etkisinin incelenmesi amacıyla bir simülasyon çalışması yapılmıştır.

Çalışmanın 2. bölümünde ridge, LASSO, elastik net ve uyarlanabilir LASSO cezalı regresyon yöntemleri tanıtılmıştır. Daha sonra katsayı tahmininde kullanılan çapraz geçerlilik yöntemi açıklanmış ve yöntemlerin karşılaştırılmasında kullanılan kriterlere değinilmiştir. Son olarak simülasyon çalışmalarının kurgusu verilmiştir. Çalışmanın 3. bölümünde yöntemlerin karşılaştırılması için yapılan simülasyon çalışmalarının bulguları verilmiş ve sonuçlar ayrıntılı bir biçimde irdelenmiştir. Çalışmanın 4. bölümünde ise çalışma sonlandırılmıştır.

II. MATERYAL VE METOT

A. Doğrusal Regresyonda Katsayı Tahmin Yöntemleri

Doğrusal regresyon analizinde modelin ön tahmin performansının artırılması ve doğru değişkenlerin seçilmesi için çeşitli yöntemler önerilmiştir. Bu yöntemler arasında bulunan en iyi alt küme seçim yöntemi açıklayıcı değişkenlerin veri kümesini iyi temsil eden bir alt kümesinin belirlenmesi ilkesine dayanır. Bu yöntem model yorumlanabilirliği açısından iyi sonuçlar verir. Fakat çok sayıda açıklayıcı değişken içeren modellerde hesaplama zorlukları nedeniyle makul değildir. Ayrıca bu yöntem verideki küçük değişimler karşısında kararlı olmayan sonuçlar verir. Dolayısıyla tahminleri tutarlı değildir.

En iyi alt küme seçim yönteminde karşılaşılan zorlukların aşılabilmesi için cezalı regresyon yöntemleri önerilmiştir. Bu yöntemler son zamanlarda oldukça yaygın bir şekilde kullanılmaktadır [17]. Cezalı regresyon yöntemlerinde otomatik model seçimi katsayıların tahmini ile eşanlı olarak yapılır. Bu yöntemlerle kararlı katsayı tahmin değerleri elde edilmektedir. x_i, i. gözlem için $p \times 1$ boyutlu açıklayıcı değişken vektörü, y_i, i. gözleme ait yanıt değeri olmak üzere $(\mathrm{x}_1, y_1), (\mathrm{x}_2, y_2), \dots, (\mathrm{x}_n, y_n)$ bağımsız ve aynı dağılımlı veri kümesi verilsin. Denklem (1)'de verilen doğrusal regresyon modeli için cezalı (negatif) log-olabilirlik fonksiyonu

$$Q(\beta) = \frac{1}{2n}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda \cdot p(\beta) \tag{3}$$

şeklinde tanımlanır. Burada, $j = 1,2,\dots,p$ için β_j, β katsayı vektörünün j. terimi, $p(\cdot)$ ceza fonksiyonu ve λ ayar parametresidir. Cezalı regresyonda katsayı tahminleri Denklem (3) ile verilen cezalı log-olabilirlik fonksiyonunun minimize edilmesi ile elde edilir. Literatürde birçok ceza fonksiyonu bulunmaktadır. Ridge regresyon için ceza fonksiyonu $p(\beta) = \|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ (L_2 türü ceza fonksiyonu), LASSO regresyon için ceza fonksiyonu $p(\beta) = \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ (L_1 türü ceza fonksiyonu), $\alpha \in [0,1]$ yeni bir ayar parametresi olmak üzere elastik net regresyon için ceza fonksiyonu $p(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2$ ve w, uyarlanabilir ağırlıklar vektörü olmak üzere uyarlanabilir LASSO ceza fonksiyonu $p(\beta) = \sum_{j=1}^{p} w_j|\beta_j|$ bu ceza fonksiyonları arasında gösterilebilir.

Gauss Markov teoremine göre Denklem (2) ile verilen EKK tahmin edicisi tüm yansız tahmin ediciler sınıfı içinde en küçük varyansa sahiptir. Ancak veri kümesinde çoklu iç ilişkinin varlığında bu varyans çok büyük olur. Bu problemin aşılabilmesi için [2] tarafından yanlı fakat EKK'ya göre daha küçük varyansa sahip ridge regresyon tahmin edicisi önerilmiştir. Ridge regresyonda fazla büyük olan katsayılar cezalandırılarak bu katsayıların büzülmesi sağlanır. Böylece "katsayıların varyansının aşırı şişmesi" probleminin aşılması hedeflenir. Ridge regresyon katsayı tahminleri

$$\hat{\beta}_R = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \right\} \tag{4}$$

probleminin çözülmesi ile elde edilir. Burada $\lambda > 0$, ayar parametresi büzülmenin miktarını belirler; λ büyüdükçe büzülme miktarı artar. Denklem (4)'teki problemin çözülmesi ile ridge tahmin edicisi

$$\hat{\beta}_R = \left(\mathrm{X}^T\mathrm{X} + \lambda n \mathrm{I}_p\right)^{-1}\mathrm{X}^T\mathrm{y} \tag{5}$$

olarak bulunur. Burada I_p, $p \times p$ boyutlu birim matristir. Bir regresyon modelinde çok sayıda korelasyonlu değişken varsa EKK, regresyon model katsayılarını kötü bir şekilde tahmin eder ve katsayı tahminleri yüksek varyanslı olur. Ridge regresyon yöntemi kullanılarak bu problem aşılabilir. Dolayısıyla ridge regresyon EKK'ya göre daha doğru ön tahmin değerleri verir. Fakat ridge regresyonda katsayıların sıfır olarak tahmin edilmesi söz konusu değildir. Dolayısıyla ridge regresyon otomatik değişken seçimi yapamaz.

Ridge regresyonun model yorumlanabilirliği konusundaki eksikliği göz önüne alınarak [9]'da LASSO yöntemi önerilmiştir. LASSO'da katsayı tahminleri

$$\hat{\beta}_L = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\} \tag{6}$$

probleminin çözülmesi ile elde edilir. LASSO regresyonda yeterince büyük λ değerleri için kısıtın doğası gereği bazı regresyon katsayı tahminleri sıfır olur. Dolayısıyla LASSO sürekli parametre uzayında bir alt küme seçim yöntemidir. Buna göre LASSO regresyonun hem ön tahmin doğruluğu hem de model yorumlanabilirliği açısından iyi sonuçlar verme potansiyeli vardır. Literatürde LASSO regresyonun performansına ve çeşitli alanlara uygulanmasına dair birçok çalışma bulunmaktadır. [18], Cox modelde değişken seçimi için LASSO'yu kullanmıştır. [19], zamansal veya uzamsal yapının dikkate alınması gereken zaman serileri veya görüntü tabanlı veriler gibi veri kümeleri için kullanışlı olan kaynaşmış (fused) LASSO'yu önermiştir. [20], LASSO'nun model seçiminde tutarlı olması için gerek ve yeter bir koşul vermiştir. [21], Tukey ağırlık fonksiyonuna dayalı bir dayanıklı (robust) LASSO tahmin edicisi önermiştir.

Modelde yüksek korelasyonlu değişkenler bulunması durumunda LASSO tek başına yeterli bir yöntem olmayabilir [9]. Ayrıca yüksek korelasyonlu değişkenlerin kendi aralarında grup oluşturduğu göz önüne alındığında LASSO bu gruptaki değişkenlerin birini modele almakta diğerlerini modelden çıkarmaktadır [10]. Bu nedenle [10]'da LASSO'ya alternatif olarak elastik net yöntemi önerilmiştir. Elastik net yönteminde ceza terimi L_1 ve L_2 türü ceza terimlerinin konveks bir bileşimidir. $0 \leq \alpha \leq 1$ olmak üzere elastik net regresyonda katsayı tahminlerine

$$\hat{\beta}_E = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2) \right\} \tag{7}$$

probleminin çözülmesi ile ulaşılır. $\alpha = 0$ durumunda elastik net ridge regresyona indirgenirken $\alpha = 1$ durumunda elastik net LASSO'ya indirgenir. [10]'da bir simülasyon çalışması ile elastik netin bazı durumlarda LASSO'dan daha iyi sonuç verdiği gösterilmiştir. Grup etkisinin bulunduğu veri kümelerinde yüksek korelasyona sahip değişkenler ya birlikte modele dahil olur veya birlikte model dışında kalır. Elastik net modelde grup etkisi olması durumunda da doğru çözümler üretmektedir.

LASSO'nun değişken seçiminde tutarlı olmadığı bazı senaryolar söz konusudur. Bu gibi durumlarda [11] değişken seçiminde tutarlı bir yöntem olarak uyarlanabilir LASSO'yu önermiştir. Uyarlanabilir LASSO'nun katsayı tahminlerine

$$\hat{\beta}_{U-L} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|\mathrm{y} - \mathrm{X}\beta\|_2^2 + \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{8}$$

problemi çözülerek ulaşılır. Uyarlanabilir LASSO'da uyarlanabilir ağırlıkların uygun bir şekilde seçilmesi kritik öneme sahiptir. β^*, β için tutarlı bir tahmin edici olmak üzere uyarlanabilir ağırlıklar vektörü $\widehat{w} = 1/|\beta^*|$ olarak seçilebilir. Burada mutlak değer fonksiyonu β^* vektöründeki her bir terime uygulanmaktadır. [11], uyarlanabilir ağırlıkların hesaplanmasında $\hat{\beta}_{ekk}$'nın kullanılmasını önermiştir. Dolayısıyla uyarlanabilir LASSO iki aşamalı bir yöntem olarak tanımlanır. Birinci aşamada EKK katsayı tahminleri ve buna bağlı olarak uyarlanabilir ağırlıklar elde edilir. İkinci aşamada ise LASSO ceza fonksiyonu birinci aşamada elde edilen uyarlanabilir ağırlıklar ile yeniden ağırlıklandırılarak Denklem (8)'deki problem çözülür ve uyarlanabilir LASSO katsayı tahminlerine ulaşılır.

B. Ayar Parametresi Tahmininde Kullanılan Değerlendirme Yöntemleri

Değerlendirme yöntemleri regresyon modelinin ön tahmin performansını değerlendirerek ayar parametresi tahmininde kullanılan yöntemlerdir.

K katlı çapraz geçerlilik, yeni örneklem hatasının tahmin edilmesinde kullanılan ve yeniden örneklemeye dayalı bir yöntemdir. Yöntemin uygulanabilmesi için veri kümesi yaklaşık olarak eşit K kata (alt veriye) ayrılır. Her bir aşamada veri kümesinin $K-1$ kata karşılık gelen kısmı üzerinde model oluşturulur ve dışarıda kalan kat üzerinde ön tahmin hatası hesaplanır. Daha açık bir ifadeyle veri kümesi D ile gösterilsin ve D kümesi $D_1, D_2, \ldots, D_K$ şeklinde K tane ayrık kümeye bölünsün. Her bir D_i bir kat olarak adlandırılır. $k = 1,2, \ldots, K$ olmak üzere $D_{(-k)} = \bigcup_{j \neq k} D_j$ olsun. $D_{(-k)}$ kümesi üzerinde tahmin edilen model katsayıları $\beta_{(-k)}$ olsun. Tahmin edilen model için çapraz geçerlilik (*CV*, cross-validation)

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

$$CV = \frac{1}{n}\sum_{k=1}^{K}\left\{\frac{1}{n_k}\sum_{i\in D_k}\left(y_i - x_i^T\beta_{(-k)}\right)^2\right\} \quad (9)$$

şeklinde bulunur. Farklı ayar parametresi değerleri için CV, ilgili ayar parametresinde test hatası değerinin bir tahminini verir. Bu değerin minimuma ulaştığı noktaya karşılık gelen ayar parametresi değeri, modelin ayar parametresinin tahmini olarak alınır.

Bir diğer teknik veri kümesinin eğitim değerlendirme ve test kümelerine ayrıştırılması ilkesine dayanır [10, 22]. Bu teknikte cezalı regresyon yöntemleri tüm ayar parametresi değerleri için eğitime tabi tutulur. Eğitime tabi tutulan tüm modeller için değerlendirme kümesi üzerinde hata kareler ortalaması değerleri hesaplanır. Tahmin edilen model en küçük hata kareler ortalamasını veren model olarak seçilir. Bu modelin performansı ise test kümesi üzerinde hesaplanan hata kareler ortalaması ile belirlenir.

C. Simülasyon Çalışmaları ve Karşılaştırma Kriterleri

Bu çalışmada EKK, ridge, LASSO, elastik net ve uyarlanabilir LASSO yöntemleri simülasyon çalışmaları ile karşılaştırılmıştır. Simülasyon çalışmalarında kullanılan veri kümeleri Denklem (1)'de verilen doğrusal regresyon modeline göre [9]'da betimlenen veri üretme yöntemiyle elde edilmiştir. Simülasyon Çalışması 1-4'te katsayı vektörünün sıfıra eşit terimleri, sıfırdan farklı terimlerinden sonra gelmektedir. Bu simülasyon çalışmalarında sıfırdan farklı katsayı değerleri, ilk terimi 0.5 ve ortak farkı 0.5 olan bir aritmetik dizinin terimlerinden oluşmaktadır. Simülasyon Çalışması 5-6'da ise sıfırdan farklı değerler eşit seçilmiş olup sıfır değerli terimlerin adedi ve konumu değiştirilerek tahmin edicilerin farklı seyreklik (sparsite) ve konum durumunda gösterdikleri performans incelenmiştir. Simülasyon çalışmalarının ayrıntıları aşağıdaki gibidir:

Simülasyon Çalışması 1: Bu çalışmada, 20 tahmin ediciden oluşan 100 veri kümesi üretilmiştir. Gerçek katsayı vektörü $\beta = \left[0.5,1,1.5,2.2.5,\underbrace{0,0,\dots,0}_{15}\right]'$ şeklindedir. Bu simülasyon çalışmasında, x_i ve x_j açıklayıcı değişkenleri arasındaki korelasyon $r_{ij} = \rho^{|i-j|}$ olacak şekilde açıklayıcı değişkenler matrisi oluşturulmuştur. Burada çoklu iç ilişki durumunda tahmin edicilerin performansını incelemek için ρ'nun iki değeri, sırasıyla orta ve yüksek düzeyde korelasyonu temsil etmek üzere, 0.7 ve 0.9 olarak alınmıştır. β'nın sıfırlardan oluşan alt bloğu, sıfır olmayan terimlerden oluşan alt bloğundan sonra olacak şekilde seçilmiştir. Bu simülasyon çalışmasında gerçek katsayı değerlerinin %75'i sıfır olarak belirlenmiş olup seyrek (sparse) model durumunda tahmin edicilerin performansının karşılaştırılması hedeflenmektedir.

Simülasyon Çalışması 2: Bu çalışmada, Simülasyon Çalışması 1'deki kurgu gerçek katsayı vektörü $\beta = \left[0.5,1,1.5,\dots,5,\underbrace{0,0,\dots,0}_{10}\right]'$ olmak üzere yeniden oluşturulmuştur. Gerçek katsayı değerlerinin %50'si sıfır olarak belirlenmiştir.

Simülasyon Çalışması 3: Bu çalışmada, Simülasyon Çalışması 1'deki kurgu gerçek katsayı vektörü $\beta = \left[0.5,1,1.5,\dots,7.5,\underbrace{0,0,\dots,0}_{5}\right]'$ olmak üzere yeniden oluşturulmuştur. Gerçek katsayı değerlerinin %25'si sıfır olarak belirlenmiştir.

Simülasyon Çalışması 4: Bu çalışmada, Simülasyon Çalışması 1'deki kurgu gerçek katsayı vektörü $\beta = [0.5,1,1.5,\dots,10]'$ olmak üzere yeniden oluşturulmuştur. Gerçek katsayı değerlerinin tamamı sıfırdan farklıdır.

Simülasyon Çalışması 5: Bu çalışmada, Simülasyon Çalışması 1'deki kurgu gerçek katsayı vektörü $\beta = \left[\underbrace{0,\dots 0,3}_{5},\underbrace{0,\dots,0,3}_{5},\underbrace{0,\dots,0,3}_{5},\underbrace{0,\dots,0,3}_{5}\right]'$ olmak üzere yeniden oluşturulmuştur. β vektörü beşer birimlik $[0,\dots 0,3]$ alt bloklarından oluşmaktadır. Her bir blokta sıfırdan farklı katsayı son terime gelecek şekilde yazılmış olup katsayıların %80'i sıfırdır. Çalışmada seyrek model durumunda sıfırdan farklı katsayıya sahip açıklayıcı değişkenlerin konumunun analiz sonuçlarına etkisinin incelenmesi hedeflenmiştir.

Simülasyon Çalışması 6: Bu çalışmada, Simülasyon Çalışması 1'deki kurgu gerçek katsayı vektörü $\beta = [3,0,3,0,\dots,3,0]'$ olmak üzere yeniden oluşturulmuştur. β vektörü ikişer birimlik $[3,0]$ alt bloklarından

oluşmaktadır. Katsayıların %50'sinin sıfır olduğu bu çalışmada Simülasyon 5'tekine göre daha yoğun (dense) bir model durumu incelenmiştir.

Her bir simülasyon çalışması için bağımsız eğitim, değerlendirme ve test kümesinden oluşan veri kümeleri üretilmiştir. Eğitim kümesi ve değerlendirme kümesi 50 gözlemden, test kümesi ise 200 gözlemden oluşmaktadır. Bu çalışmada Denklem (1)'de verilen doğrusal regresyon modelinde hataların standart sapması [9]'da verildiği gibi $\sigma = 3$ alınmıştır. Eğitim kümesi, model katsayılarını tahmin etmek için kullanılırken değerlendirme kümesi, ayar parametrelerini tahmin etmek için kullanılmıştır. Elastik net $\alpha = 0$ için ridge regresyona, $\alpha = 1$ için LASSO regresyona karşılık geldiğinden burada elastik netin LASSO ve ridge regresyondan farklılık düzeyini gözlemlemek için $\alpha = 0.5$ alınmıştır. Uyarlanabilir LASSO için V, test kümesindeki açıklayıcı değişkenler matrisine karşılık gelen kovaryans matrisi ve $\tilde{\beta}$, karşılaştırılan yöntemden elde edilen katsayı vektörü olmak üzere model performansını ölçmek için test kümesi üzerinde "test hata kareler ortalaması" (test mean squared error)

$$TMSE = \left(\beta - \tilde{\beta}\right)^T \mathrm{V} \left(\beta - \tilde{\beta}\right)$$

değeri hesaplanmıştır. Her bir yöntem için sıfırdan farklı katsayı değerine sahip olan açıklayıcı değişkenlerin kümesi olan aktif kümeler elde edilmiştir. Bu bağlamda ilgili yöntem tarafından hatalı bir şekilde aktif kümeye alınan veya aktif kümenin dışında bırakılan açıklayıcı değişkenlerin oranı olarak tanımlanan yanlış sınıflama oranı ve gerçekte aktif kümenin dışında olan değişkenlerin aktif kümede tahmin edilme oranı olarak tanımlanan yanlış pozitif oranı yöntemlerin modele katkı sağlayan değişkenleri tespit etme performansının bir ölçüsü olarak verilmiştir. Ayrıca her bir yöntem için aktif küme büyüklükleri rapor edilmiştir.

III. BULGULAR VE TARTIŞMA

Simülasyon Çalışması 1-4'ün sonuçları Tablo 1'de özetlenmiştir. Simülasyon Çalışması 4 dışındaki tüm simülasyon çalışmalarında EKK en büyük medyan TMSE değerine sahiptir. Tüm katsayı değerlerinin sıfırdan farklı olduğu Simülasyon Çalışması 4'te ise uyarlanabilir LASSO'nun TMSE bakımından kötü bir performans gösterdiği söylenebilir. Simülasyon Çalışması 1-2'de model büyük ölçüde seyrektir. Bu simülasyon çalışmalarında korelasyon miktarı orta düzeyde iken TMSE değerleri bakımından LASSO yöntemi diğer yöntemlere göre daha iyi bir performans göstermiştir. Ancak seyreklik düzeyi azalıp korelasyon düzeyi arttıkça LASSO'nun performansı elastik nete göre zayıflamaktadır. Ridge regresyon ve uyarlanabilir LASSO bu iki yönteme göre daha kötü bir sonuç vermiştir. Seyrekliğin azaldığı durumda (Simülasyon Çalışması 3) elastik net, TMSE açısından daha iyi bir sonuç vermiştir. Bu çalışmada TMSE performansı bakımından ridge regresyon, LASSO ve uyarlanabilir LASSO yöntemine göre daha iyidir. Tümüyle yoğun olan model durumunda (Simülasyon Çalışması 4) ise ridge regresyon en iyi sonuçlara sahiptir. Bu durumda ridge regresyonu elastik net takip etmektedir. LASSO ve uyarlanabilir LASSO ise bu iki yönteme göre daha büyük bir medyan TMSE değeri vermiştir. Yanlış sınıflama oranı ve yanlış pozitif oranı bakımından Simülasyon Çalışması 1-3'te LASSO ve uyarlanabilir LASSO yöntemleri birbirine yakın olup diğer yöntemlere göre daha iyi bir sonuç vermiştir. Buna göre simülasyon çalışmalarında doğru açıklayıcı değişkenlerin belirlenmesi konusunda LASSO ve uyarlanabilir LASSO'nun diğer yöntemlere göre daha iyi olduğu söylenebilir. Yanlış sınıflama performansı bakımından elastik net, LASSO ve uyarlanabilir LASSO'yu takip etmektedir. EKK ve ridge regresyon değişken seçimi yapamadığı için yanlış sınıflama oranı bakımından iyi değildir. Simülasyon Çalışması 4'te hiçbir gerçek katsayı değeri sıfır olmadığı için bu çalışmada yanlış sınıflama oranı ve yanlış pozitif oranı verilmemiştir. Aktif küme büyüklükleri bakımından uyarlanabilir LASSO'nun diğer yöntemlere göre daha seyrek bir model verdiği söylenebilir.

Tablo 1. Simülasyon Çalışması 1-4 için tahmin edicilerin kalite ölçüleri

	ρ	Yöntem	Medyan TMSE	Standart Sapma	Yanlış Sınıflama Oranı	Yanlış Pozitif Oranı	Aktif Küme Büyüklüğü
Simülasyon 1	0.7	EKK	5.7484	0.28	0.75	1.00	20
		Ridge	2.2790	0.13	0.75	1.00	20
		LASSO	1.5067	0.11	0.20	0.20	7
		Elastik Net	1.8278	0.12	0.40	0.53	13
		Uyarlanabilir LASSO	2.1482	0.16	0.20	0.13	6

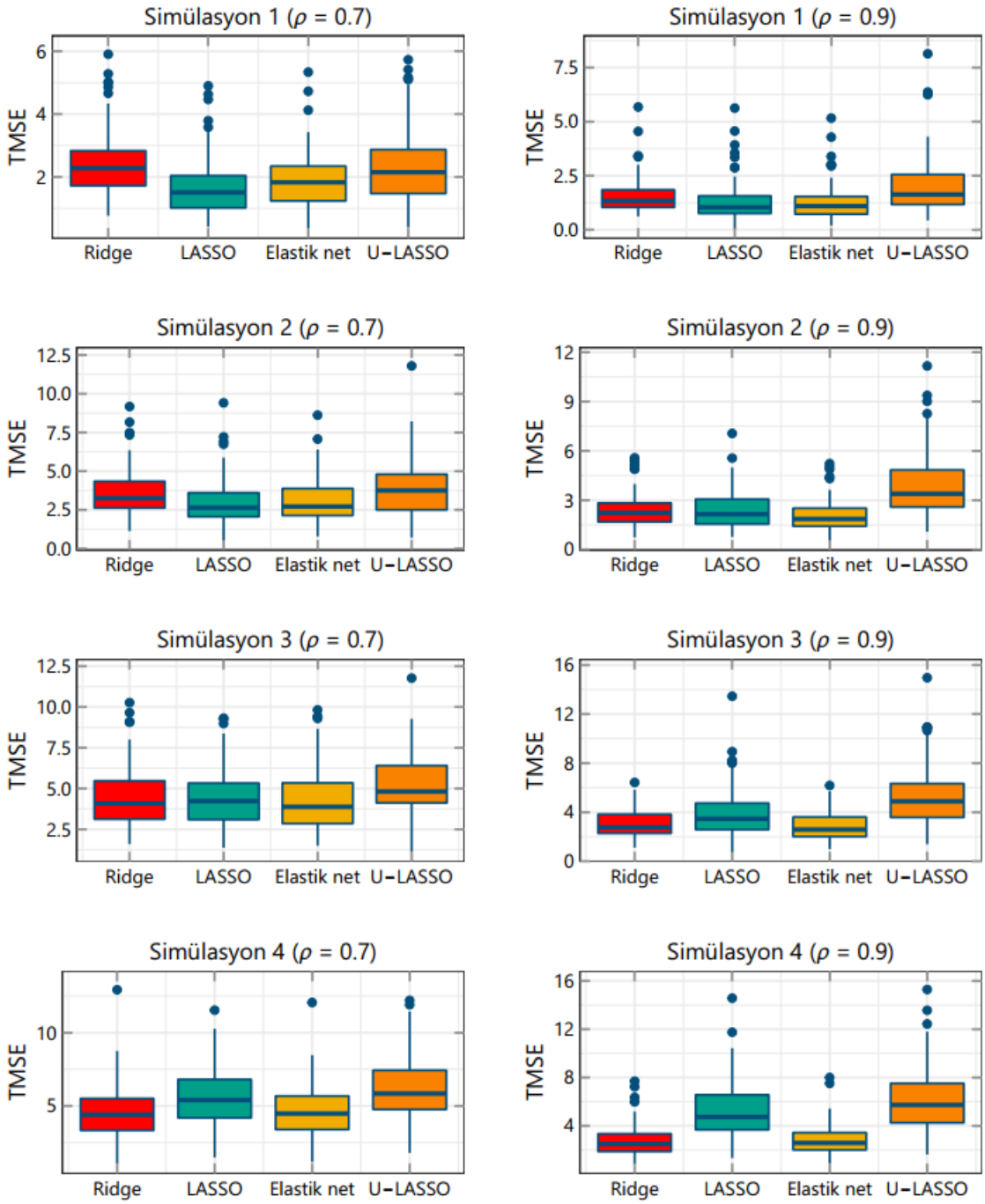
Tablo 1. Devam

	0.9	EKK	5.7462	0.29	0.75	1.00	20
		Ridge	1.3316	0.07	0.75	1.00	20
		LASSO	1.0416	0.08	0.20	0.20	7
		Elastik Net	1.0950	0.06	0.35	0.47	12
		Uyarlanabilir LASSO	1.6340	0.09	0.25	0.13	5
Simülasyon 2	0.7	EKK	5.7495	0.28	0.50	1.00	20
		Ridge	3.2506	0.25	0.50	1.00	20
		LASSO	2.6412	0.12	0.20	0.30	13
		Elastik Net	2.7168	0.16	0.40	0.80	17
		Uyarlanabilir LASSO	3.7515	0.23	0.20	0.30	11
	0.9	EKK	5.7173	0.30	0.50	1.00	20
		Ridge	2.2259	0.10	0.50	1.00	20
		LASSO	2.1475	0.13	0.20	0.30	12
		Elastik Net	1.8593	0.09	0.35	0.70	17
		Uyarlanabilir LASSO	3.3914	0.20	0.30	0.40	12
Simülasyon 3	0.7	EKK	5.7388	0.27	0.25	1.00	20
		Ridge	4.0820	0.18	0.25	1.00	20
		LASSO	4.2270	0.28	0.15	0.40	17
		Elastik Net	3.8767	0.19	0.25	1.00	19
		Uyarlanabilir LASSO	4.8149	0.20	0.15	0.40	16
	0.9	EKK	5.6790	0.31	0.25	1.00	20
		Ridge	2.7626	0.15	0.25	1.00	20
		LASSO	3.4672	0.19	0.15	0.40	16
		Elastik Net	2.5856	0.17	0.20	0.80	19
		Uyarlanabilir LASSO	4.8877	0.36	0.20	0.40	16
Simülasyon 4	0.7	EKK	5.7343	0.27	-	-	20
		Ridge	4.3928	0.17	-	-	20
		LASSO	5.4001	0.21	-	-	20
		Elastik Net	4.4781	0.22	-	-	20
		Uyarlanabilir LASSO	5.8433	0.33	-	-	19
	0.9	EKK	5.5923	0.33	-	-	20
		Ridge	2.4960	0.14	-	-	20
		LASSO	4.7243	0.31	-	-	19
		Elastik Net	2.5747	0.10	-	-	20
		Uyarlanabilir LASSO	5.7281	0.41	-	-	19

Şekil 1, Simülasyon Çalışması 1-4'te ridge, LASSO, elastik net ve uyarlanabilir LASSO (U-LASSO) yöntemleri ile elde edilen TMSE değerlerinin kutu grafiğini göstermektedir. Kutu grafikleri, Tablo 1'de verilen Simülasyon Çalışması 1-4'ten elde edilen çıkarımları desteklemektedir.

Şekil 2, Simülasyon Çalışması 1-4'te her bir tahmin edicinin katsayı izleri grafiğini göstermektedir. Karşılaştırmada kolaylık sağlaması bakımından katsayılar $\|\beta\|_1/\max\|\beta\|_1$ değerlerinin fonksiyonu olarak

çizilmiştir. Şekil 2'de değişken seçimi yapabilen yöntemlerin değişkenleri modele alma düzenleri görülmektedir. Genel olarak LASSO ve uyarlanabilir LASSO'nun daha seyrek bir model ürettiği, buna karşın α parametresinin etkisiyle elastik netin daha yoğun bir model tahmini verdiği söylenebilir.

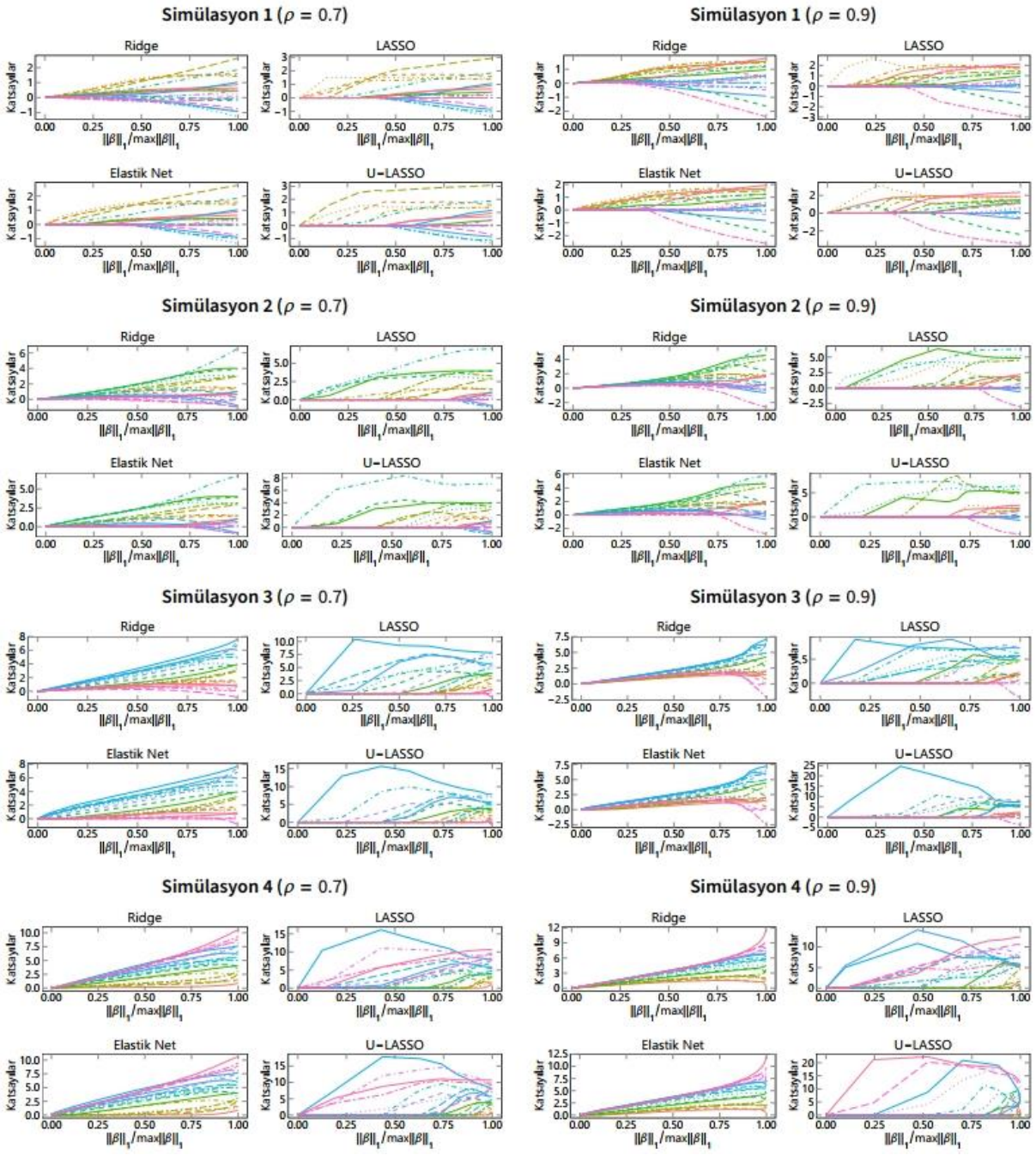


Şekil 1. Simülasyon 1-4 için tahmin edicilerin test kümesindeki TMSE değerlerinin kutu grafikleri.

BŞEÜ Fen Bilimleri Dergisi
9(1), 80-91, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)



Şekil 2. Simülasyon 1-4 için tahmin edicilerin katsayı izleri grafikleri.

Simülasyon Çalışması 5-6'nın sonuçları Tablo 2'de verilmiştir. Her iki simülasyon çalışmasında da EKK diğer yöntemlere göre daha büyük TMSE değeri vermiştir. Bu simülasyon çalışmalarında korelasyon orta düzeyde iken uyarlanabilir LASSO, yüksek düzeyde iken LASSO en iyi TMSE değerine sahiptir. Yanlış sınıflama oranı ve yanlış pozitif oranı bakımından uyarlanabilir LASSO diğer yöntemlere göre daha iyi bir performans göstermiştir. Bu kriterler bakımından uyarlanabilir LASSO'yu LASSO takip etmektedir. Aktif küme büyüklüğü bakımından en gerçek seyreklik düzeyine en yakın modelleri uyarlanabilir LASSO vermiştir.

Tablo 2. Simülasyon Çalışması 5-6 için tahmin edicilerin kalite ölçüleri

	ρ	Yöntem	Medyan TMSE	Standart Sapma	Yanlış Sınıflama Oranı	Yanlış Pozitif Oranı	Aktif Küme Büyüklüğü
Simülasyon 5	0.7	EKK	5.7472	0.28	0.80	1.00	20
		Ridge	3.5887	0.21	0.80	1.00	20
		LASSO	2.1289	0.18	0.35	0.44	11
		Elastik Net	2.9636	0.19	0.60	0.75	16
		Uyarlanabilir LASSO	1.4091	0.16	0.13	0.16	6.5
	0.9	EKK	5.7427	0.29	0.80	1.00	20
		Ridge	2.6341	0.14	0.80	1.00	20
		LASSO	1.9799	0.15	0.33	0.38	10
		Elastik Net	2.3163	0.12	0.65	0.81	17
		Uyarlanabilir LASSO	2.0179	0.21	0.23	0.25	7
Simülasyon 6	0.7	EKK	5.7449	0.28	0.50	1.00	20
		Ridge	4.5364	0.27	0.50	1.00	20
		LASSO	3.6415	0.22	0.30	0.60	16
		Elastik Net	4.1278	0.28	0.45	0.90	19
		Uyarlanabilir LASSO	3.4314	0.25	0.15	0.30	13
	0.9	EKK	5.7202	0.28	0.50	1.00	20
		Ridge	2.9387	0.08	0.50	1.00	20
		LASSO	3.4245	0.17	0.30	0.60	15
		Elastik Net	2.9464	0.09	0.50	1.00	20
		Uyarlanabilir LASSO	4.5968	0.30	0.25	0.40	13

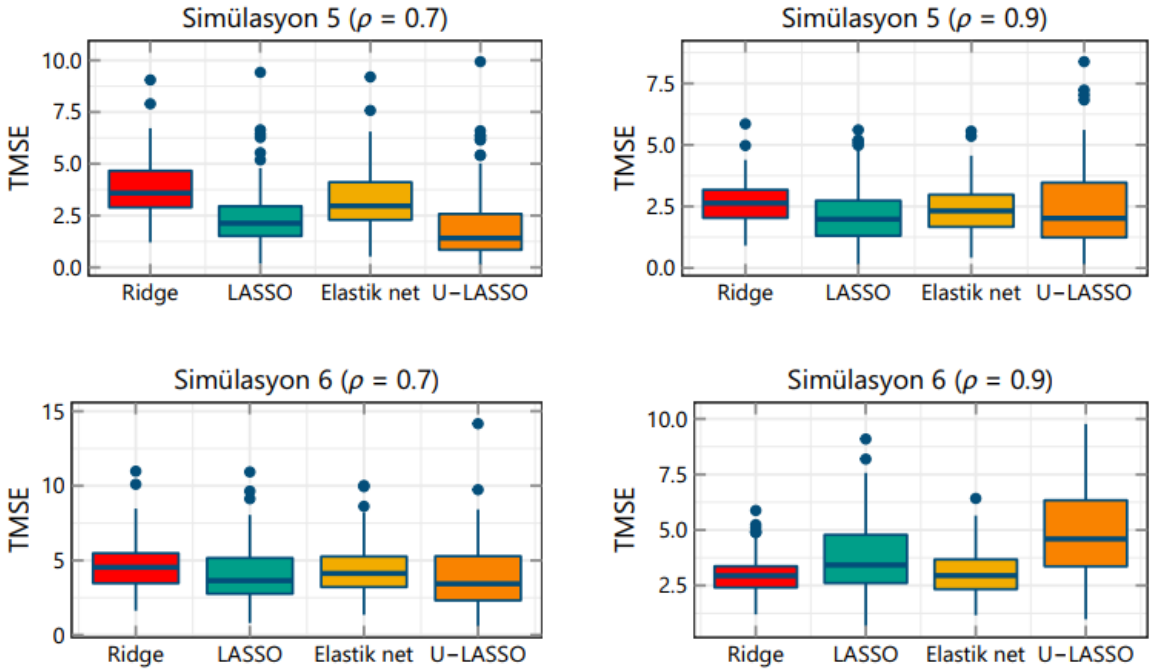
Simülasyon Çalışması 5-6'da ridge, LASSO ve elastik net yöntemleri ile elde edilen TMSE değerlerinin kutu grafikleri Şekil 3'te verilmiştir. LASSO, iki simülasyon çalışmasında da en iyi performansı göstermiştir. Kutu grafikleri, Tablo 2'den elde edilen sonuçları desteklemektedir.

Şekil 4, Simülasyon Çalışması 5-6'daki her bir tahmin edicinin katsayı izleri grafiğini göstermektedir. Bu şekilde değişken seçimi yapabilen yöntemlerin değişkenleri modele alma düzenleri görülmektedir. Elastik netin değişken seçme özelliği dışında ridge regresyon ve elastik net, benzer bir yapı göstermektedir. LASSO ve uyarlanabilir LASSO'nun sıfır değerli katsayıları belirleme düzenlerindeki farklılık görülmektedir.

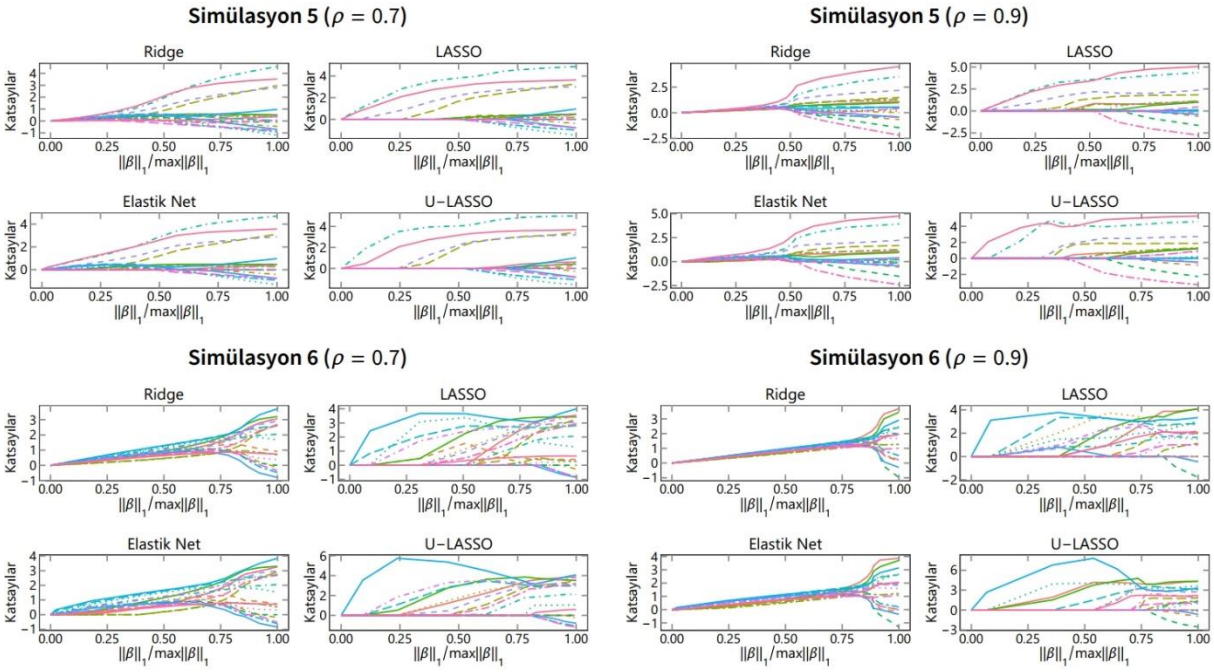
BŞEÜ Fen Bilimleri Dergisi
9(1), 80-91, 2022

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)



Şekil 3. Simülasyon 5-6 için tahmin edicilerin test kümesindeki MSE değerlerinin kutu grafikleri.



Şekil 4. Simülasyon 5-6 için tahmin edicilerin katsayı izleri grafikleri.

IV. SONUÇLAR VE ÖNERİLER

Bu çalışmada doğrusal regresyon modellerinde gerçek katsayı vektörünün yapısına bağlı olarak konveks cezalı regresyon yöntemlerinin model performansı üzerindeki etkisi incelenmiştir. Modelin yapısına bağlı olarak otomatik değişken seçimi yapabilen LASSO, elastik net ve uyarlanabilir LASSO yöntemlerinin birbirine ve ridge regresyona göre üstün veya zayıf olduğu durumlar incelenmiştir. Karşılaştırma kriteri olarak kullanılan TMSE, yanlış sınıflama oranı, yanlış pozitif oranı ve aktif küme büyüklüğü ölçülerine göre cezalı regresyon yöntemleri klasik EKK tahmin edicisine göre daha iyi sonuçlar vermiştir. Cezalı regresyon yöntemlerinin kendi içindeki üstünlüğü modelin seyreklik düzeyi ve sıfırdan farklı katsayı değerlerinin konumuna bağlı olarak değişmektedir.

BSEU Journal of Science
https://doi.org/10.35193/bseufbd.994181

e-ISSN:2458-7575 (https://dergipark.org.tr/tr/pub/bseufbd)

Çalışmanın simülasyon sonuçları LASSO, elastik net ve uyarlanabilir LASSO yöntemlerinin farklı yapılarda seyreklik gösteren modellerde tahmin konusunda başarılı olduğunu göstermektedir.

KAYNAKLAR

[1] Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021). *Introduction to linear regression analysis*, John Wiley & Sons.

[2] Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*12 (1), 55-67.

[3] Rao, C. R. & Toutenburg, H. (1995). *Linear models*, Springer.

[4] Sarkar, N. (1992). A new estimator combining the ridge regression and the restricted least squares methods of estimation. *Communications in statistics-theory and methods*21 (7), 1987-2000.

[5] Kaçıranlar, S., Sakallıoğlu, S., Akdeniz, F., Styan, G. P. & Werner, H. J. (1999). A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland cement. *Sankhyā: The Indian Journal of Statistics, Series B*, 443-459.

[6] Özkale, M. R. & Kaçıranlar, S. (2007). The restricted and unrestricted two-parameter estimators. *Communications in Statistics-Theory and Methods*36 (15), 2707-2725.

[7] Miller, A. (2002). *Subset selection in regression*, CRC Press.

[8] Frank, L. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*35 (2), 109-135.

[9] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*58 (1), 267-288.

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

[11] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418-1429.

[12] Sirimongkolkasem, T., & Drikvandi, R. (2019). On regularisation methods for analysis of high dimensional data. *Annals of Data Science* 6(4), 737-763.

[13] Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The annals of statistics* 37(1), 246-270.

[14] Yüzbaşı, B., Arashi, M., & Ejaz Ahmed, S. (2020). Shrinkage Estimation Strategies in Generalised Ridge Regression Models: Low/High- Dimension Regime. *International Statistical Review* 88(1), 229-251.

[15] Ahmed, S. E., Kim, H., Yıldırım, G., & Yüzbaşı, B. (2016). High-Dimensional Regression Under Correlated Design: An Extensive Simulation Study. In *International Workshop on Matrices and Statistics* (pp. 145-175). Springer, Cham.

[16] Shahriari, S., Faria, S., & Gonçalves, A. M. (2015). Variable selection methods in high-dimensional regression—A simulation study. *Communications in Statistics-Simulation and Computation* 44(10), 2548-2561.

[17] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[18] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine*16 (4), 385-395.

[19] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*67 (1), 91-108.

[20] Zhao, P. & Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research*7, 2541-2563.

[21] Chang, L., Roberts, S. & Welsh, A. (2018). Robust lasso regression using Tukey's biweight criterion. *Technometrics*60 (1), 36-47.

[22] Hussami, N., & Tibshirani, R. J. (2015). A component lasso. *Canadian Journal of Statistics* 43(4), 624-646.