Necmettin Erbakan Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, c.1,s.2,2019
Necmettin Erbakan University Journal of Science and Engineering, v.1,n.2,2019
e-ISSN: 2667-7989

Ridge Tahminine Dayalı Kantil Regresyon Analizinde Yanlılık Parametresi Tahminlerinin Performanslarının Karşılaştırılması

Murat ERİŞOĞLU[1*], **Nurullah YAMAN**[2]

[1] Necmettin Erbakan Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 42090, Meram, Konya, Türkiye
[2] Necmettin Erbakan Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, 42090, Konya, Türkiye

ARAŞTIRMA MAKALESİ/RESEARCH ARTICLE

(Geliş/*Received*: xx.xx.2019; Kabul/*Accepted*: xx.xx.2019; Online baskı/*Published online:* 09.12.2019)

ÖZET

Bu çalışmada aykırı gözlemlerin varlığında en küçük kareler regresyonuna alternatif olarak kullanılan kantil regresyonunda çoklu bağlantı probleminin çözümü ele alınmıştır. Kantil regresyonunda çoklu bağlantı probleminin çözümünde ridge regresyon yaklaşımı kullanılmıştır. Ridge tahminine dayalı kantil regresyonunda bazı yanlılık parametre tahminlerinin performansı hata kareler ortalamasına göre karşılaştırılmıştır. Simülasyon çalışması sonuçlarına göre Hocking, Speed ve Lynn (1976) ile Kibria (2003) tarafından önerilen yanlılık parametre tahmin edicileri daha başarılı bir performans göstermişlerdir.

Anahtar Kelimeler: Kantil regresyon, Ridge regresyonu, Çoklu bağlantı, Aykırı gözlem, Çapraz doğruluk

A Comparison of Performances of the Estimations of the Bias Parameter in the Quantile Regression Analysis Based on Ridge Estimation

ABSTRACT

In this study, the solution of the multicollinearity problem was investigated in the quantile regression which is used as an alternative to the least squares regression in case the outliers. The ridge regression approach was used to solve the multicollinearity problem in quantile regression. In the quantile regression based on ridge estimation, the performance of some bias parameter estimates was compared according to the mean error squares. According to the results of the simulation study, the bias parameter estimators proposed by Hocking, Speed and Lynn (1976) and Kibria (2003) showed a more successful performance.

Key Words: Quantile regression, Ridge regression, Multicollinearity, Outliers, Cross validation

1. GİRİŞ (INTRODUCTION)

Çoklu doğrusal regresyon modeli,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

eşitliği ile ifade edilir. Eşitlikte yer alan $\mathbf{y}$: $n \times 1$ boyutlu gözlem vektörünü, $\mathbf{X}$: $n \times p$ boyutlu standartlaştırılmış açıklayıcı değişkenlerin $n \times p$ boyutlu bilinen tasarım matrisini, $\boldsymbol{\beta}$: $p \times 1$ boyutlu bilinmeyen regresyon katsayılar vektörünü ve $\boldsymbol{\varepsilon}$: $n \times 1$ boyutlu 0 ortalama vektörü ve $\sigma^2\mathbf{I_n}$ varyans kovaryans matrisi ile çok değişkenli normal dağılıma sahip rastgele hata vektörünü göstermektedir. Regresyon katsayılarının tahmininde en yaygın kullanılan yöntem hataların kareleri toplamını,

$$\sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^2 \qquad (2)$$

en küçüklemeyi amaçlayan en küçük kareler (EKK) yöntemidir. Regresyon katsayılar vektörü $\boldsymbol{\beta}$'nın EKK tahmini

[*] Sorumlu Yazar/Corresponding Author: merisoglu@erbakan.edu.tr / Tel: +90 332 323 8220-(5815)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3)$$

ile elde edilir. Model varsayımlarının geçerli olması durumunda EKK tahminleri yansız ve en küçük varyansa sahip tahmin edicidir. Çoklu doğrusal regresyon modelinde açıklayıcı değişkenlerin karşılıklı ilişkisiz olduğu varsayılır. Ancak uygulamada açıklayıcı değişkenler arasında yüksek ya da güçlü doğrusal ilişkiler olabilir ve bu durum çoklu bağlantı problemine neden olur. Çoklu bağlantı probleminde EKK tahminlerinin yansızlık özelliği korunsa bile en küçük varyansa sahip olma yani etkinlik özelliği bozulur. Çoklu bağlantı problemi, regresyon katsayılarının tahminlerinin varyansının büyük olmasına ve buna bağlı olarak istatistiksel çıkarımların yanlış sonuçlar üretmesine neden olmaktadır[1]. Çoklu bağlantı probleminin çözümünde en yaygın kullanılan yöntemlerden biri Hoerl ve Kennard [2] tarafından önerilen ridge regresyonudur. Çoklu bağlantı probleminde, ridge regresyonu yan ekleyerek varyansı daha küçük tahminler elde etme için kullanılır. Regresyon katsayılar vektörü $\boldsymbol{\beta}$'nın tahmininde,

$$\sum_{i=1}^{n}(y_i - \mathbf{x_i}\boldsymbol{\beta})^2 + k\boldsymbol{\beta}'\boldsymbol{\beta} \quad (4)$$

şeklinde ifade edilen fonksiyonu en küçüklenmeyi amaçlayan ridge tahmini,

$$\widehat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I_p}\right)^{-1}\mathbf{X}'\mathbf{y} \quad (5)$$

eşitliği ile elde edilir. Eşitlikte yer alan k yanlılık parametresi olarak ifade edilmektedir. Yanlılık parametresi tahminlerin varyansının değişiminde önemli bir rol oynamaktadır. Bu nedenle yanlılık parametresinin seçimi oldukça önemlidir [3]. Yanlılık parametresinin seçimi için birçok yöntem önerilmiştir.

Bu çalışmada aykırı gözlem, hata terimlerinin normal olmayan bir dağılım göstermesi veya farklı varyanslılık gibi durumlarda kullanılan alternatif regresyon yöntemlerinden kantil regresyon analizinde çoklu bağıntı probleminin çözümü incelenmiştir. Çoklu bağıntı probleminin çözümünde ridge tahminine dayalı kantil regresyonu kullanılmıştır. Yanlılık parametresinin tahmini için önerilen bazı tahmin yöntemlerinin ridge tahminine dayalı kantil regresyondaki performansları hata kareler ortalaması ile karşılaştırılmıştır. Çalışmanın ikinci bölümünde kantil regresyonu, üçüncü bölümünde ridge tahminine dayalı kantil regresyonu verilmiştir. Çalışmanın dördüncü bölümünde, seçili parametre değerlerine göre yanlılık parametresinin tahmininde kullanılan tahmin edicilerin performansları simülasyon çalışması ile incelenmiştir. Çalışmanın beşinci bölümünde aykırı gözlem ve çoklu bağıntı problemi içeren tobacco veri setinde tekrarlı k katmanlı çapraz doğrulama ile yanlılık parametre tahminlerinin etkinliği karşılaştırılmıştır. Son olarak çalışmadan elde edilen bulgular sonuç bölümde verilmiştir.

2. KANTİL REGRESYONU (QUANTILE REGRESSION)

Hata terimlerinin normal olmayan bir dağılıma sahip olması veya aykırı gözlemlerin olması durumunda EKK tahminlerinin alternatifi olarak kullanılan sağlam (robust) yöntemlerden biri de kantil regresyonudur[4]. Kantil regresyonu aykırı gözlemlere veya hata terimlerinin normal olmayan bir dağılıma sahip olması durumlarına karşı hassas olmadığından dolayı sağlam yöntem olarak ifade edilmektedir. EKK yönteminde açıklayıcı değişkenler ile cevap değişkeni arasındaki ilişki, açıklayıcı değişkenler bilindiğinde cevap değişkeninin koşullu ortalaması olarak modellenir[5]. Koenker ve Bassett [6] tarafından önerilen kantil regresyon yönteminde ise koşullu kantil ile modelleme gerçekleştirilir[7]. Kantil regresyon yönteminde hata terimlerinin varyans yapısına ilişkin herhangi bir varsayım bulunmamaktadır[8]. Bundan dolayı değişen varyans durumunda da kantil regresyonu, EKK yöntemi için alternatif bir yöntem olarak kullanılmaktadır. Kantil regresyonunda,

$$\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x_i}\boldsymbol{\beta}) \quad (6)$$

eşitliği ile ifade edilen amaç fonksiyonu en küçüklenmeye çalışılır. Eşitlikte yer alan $\rho_\tau(.)$ fonksiyonu,

$$\rho_\tau(.) = \begin{cases} \tau(y_i - \mathbf{x_i}\boldsymbol{\beta}) & \text{eğer } (y_i - \mathbf{x_i}\boldsymbol{\beta}) \geq 0 \\ (\tau - 1)(y_i - \mathbf{x_i}\boldsymbol{\beta}) & \text{eğer } (y_i - \mathbf{x_i}\boldsymbol{\beta}) < 0 \end{cases} \quad (7)$$

seçilen τ $(0 < \tau < 1)$ kantil değerine göre bir mutlak değer fonksiyonu olarak değerlendirilebilir. Kantil regresyonunda Eşitlik (6) ile ifade edilen amaç fonksiyonunun analitik çözümü yoktur. Kantil regresyonunda katsayılarının tahmini için iteratif algoritmalar veya doğrusal programlama yaklaşımı kullanılır [9]. Regresyon katsayılarının tahmini için kullanılan algoritmalarda yaygın olarak katsayılar başlangıç tahmin vektörü olarak EKK yöntemi ile elde edilen regresyon katsayılar tahmin vektörü kullanılmaktadır.

3. RİDGE TAHMİNİNE DAYALI KANTİL REGRESYONU (QUANTILE REGRESSION BASED ON RIGE ESTIMATION)

Kantil regresyon yönteminde çoklu bağlantı probleminin çözümünde yanlı tahmin yöntemleri uygulanabilir[10-12]. Ridge tahminine dayalı kantil regresyon yönteminde Eşitlik (8) ile belirtilen amaç fonksiyonu en küçüklenmeye çalışılır.

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x_i}\boldsymbol{\beta}) + k\boldsymbol{\beta}'\boldsymbol{\beta} \qquad (8)$$

Ridge tahminine dayalı kantil regresyon yönteminde, katsayılar tahmin vektörünün elde edilmesinde kullanılan algoritmalarda başlangıç katsayılar tahmin vektörü olarak çoklu regresyonda ridge yöntemi ile elde edilen katsayılar vektörü kullanılabilir. Ridge yaklaşımında katsayıların tahmininde k yanlılık parametresinin seçimi oldukça önemlidir. Literatürde yanlılık parametresinin seçimi için birçok yöntem önerilmiştir. Önerilen yöntemler Eşitlik (1) ile ifade edilen regresyon modelinin,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \qquad (9)$$

eşitliği ile ifade edilen kanonik formuna dayalıdır. Eşitlikte yer alan $\mathbf{Z} = \mathbf{XD}$, $\mathbf{D}'\mathbf{D} = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ ve $\boldsymbol{\alpha} = \mathbf{D}'\boldsymbol{\beta}$ olur. Burada $\lambda_1, \dots, \lambda_p$ değerleri $\mathbf{X}'\mathbf{X}$ matrislerinin özdeğerlerini göstermektedir ve $\mathbf{D}$ ortogonal bir matristir. Literatürde önerilen bazı yanlılık parametresi tahmin değerleri aşağıda verilmiştir.
Hoerl ve Kennard [2] yanlılık parametresi k tahmini için,

$$\hat{k}_{HK} = \frac{\hat{\sigma}^2}{\hat{\alpha}^2_{enb}} \qquad (10)$$

eşitliğini önermişlerdir. Burada $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p)$ ve $\hat{\alpha}_{enb}$ en büyük $\hat{\alpha}$ değerini göstermektedir.

Hoerl, Kennard ve Baldwin [13] yanlılık parametresinin tahmininde, her $\hat{\alpha}_i$ için elde edilecek $\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ tahminlerinin harmonik ortalamasının kullanımını önerdiler.

$$\hat{k}_{HKB} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p}\hat{\alpha}_i^2} = \frac{p\hat{\sigma}^2}{\hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}} \qquad (11)$$

Lawless ve Wang [14] yanlılık parametresinin tahmininde bayesci bir yaklaşım kullanarak

$$\hat{k}_{LW} = \frac{p\sigma^2}{\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2} = \frac{p\sigma^2}{\hat{\boldsymbol{\alpha}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\alpha}}} \qquad (12)$$

eşitliğini önermişlerdir.

Hocking, Speed ve Lynn [15] yanlılık parametresinin tahmini için

$$\hat{k}_{HSL} = \hat{\sigma}^2 \frac{\sum_{i=1}^{p}(\lambda_i\hat{\alpha}_i)^2}{(\sum_{i=1}^{p}\lambda_i\hat{\alpha}_i^2)^2} \qquad (13)$$

eşitliğini önermişlerdir.

Kibria [16] yanlılık parametresinin tahmininde, $\hat{k}_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}$ tahminlerinin aritmetik ortalaması ($\hat{k}_{AM}$), geometrik ortalaması ($\hat{k}_{GM}$) ve medyanını ($\hat{k}_{MED}$) kullanmayı önermiştir.

$$\hat{k}_{AM} = \frac{1}{p}\sum_{i=1}^{p}\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \qquad (14)$$

$$\hat{k}_{GM} = \frac{\hat{\sigma}^2}{(\prod_{i=1}^{p}\hat{\alpha}_i^2)^{\frac{1}{p}}} \qquad (15)$$

$$\hat{k}_{MED} = \text{Median}\left\{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\right\}, \quad i = 1,2,\dots,p \tag{16}$$

Khalaf ve Shukur [17], yanlılık parametresinin tahmini için Hoerl ve Kennard (1970) tarafından önerilen $\hat{k}_{HK}$ tahmininin modifikasyonuna dayanarak

$$\hat{k}_{KS} = \frac{\lambda_{max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \lambda_{max}\hat{\alpha}_{max}^2} \tag{17}$$

tahminini önermişlerdir.

4. SİMÜLASYON (SIMULATION)

Bu bölümde ridge tahminine dayalı kantil regresyonunda yanlılık parametresi k tahminlerinin performansını karşılaştırmak için bir simülasyon çalışması gerçekleştirilmiştir. Simülasyon çalışması için çoklu bağıntı problemi ve aykırı gözlemler içeren yapay veri setleri oluşturulmuştur. Regresyon modelinde yer alan açıklayıcı değişkenler çoklu bağıntı içerecek şekilde

$$x_{ij} = (1-\rho^2)^{1/2} w_{ij} + \rho w_{ip} \qquad i = 1,\dots,n \text{ ve } j = 1,\dots,p \tag{18}$$

eşitliği ile üretilmiştir [18, 19]. Eşitlikte yer alan ρ açıklayıcı değişkenler arasındaki korelasyon katsayısını, w_{ij} gösterimi ise standart normal dağılımından üretilen rasgele değeri göstermektedir. Simülasyon çalışmasında açıklayıcı değişkenler, $X'X$ matrisi korelasyon formunda olacak şekilde standartlaştırılmıştır. $\boldsymbol{\beta}$ regresyon katsayılar vektörü MSE değerini en küçükleyecek şekilde $X'X$ matrisinin en büyük özdeğerine karşılık gelen özvektör olarak seçilmiştir [16]. Çoklu regresyon modelinde yanıt değişkeninin değerleri

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{19}$$

eşitliği ile oluşturulmuştur. Hata terimleri ε_i, $N(0,\sigma^2)$ dağılımından üretilmiştir. Veri setlerinde aykırı gözlem oluşturmak için iki yanıt değişkeninin değeri $y_i^* = y_i + 10\sigma$ eşitliği ile dönüştürülmüştür. Böylece simülasyon çalışmasında %20, %6.7 ve %2 aykırı gözlem oranları ile çalışılmıştır. Bu çalışmada açıklayıcı değişken sayısı $p = 4$ alınmıştır ve diğer model parametreleri aşağıdaki gibi belirlenmiştir.

$\tau = 0.25, 0.50$ ve 0.75
$n = 10, 30$ ve 100
$\sigma = 0.2$ ve 0.5
$\rho^2 = 0.95$ ve 0.99

Ridge tahminine dayalı kantil regresyonunda üretilen her yapay veri seti için yanlılık parametresinin tahmininde Eşitlik (10) ile (17) arasında tanımlanan sekiz tahmin edici kullanılmıştır. 10000 tekrar ile gerçekleştirilen simülasyon çalışmasında, yanlılık parametresinin tahminlerinin performansları, yanlılık parametresi tahminine dayalı elde edilen tahmin edicilerin toplam hata kareler ortalaması

$$\text{MSE}(\hat{\beta}_k) = \frac{1}{10000}\sum_{i=1}^{10000}\sum_{j=1}^{4}\left(\beta_j - \hat{\beta}_{k,ij}\right)^2 \tag{20}$$

kriterine göre değerlendirilmiştir. Seçili parametre değerlerine göre gerçekleştirilen simülasyon çalışmasından elde edilen regresyon katsayısı tahminlerinin toplam MSE değerleri Tablo 1-3'de verilmiştir.

Tablo 1-3'deki değerler incelendiğinde, açıklayıcı değişkenler arasındaki ρ korelasyon katsayısı değeri büyüdüğünde beklenildiği gibi tahmin edicilerin toplam MSE değerlerinin büyüdüğü, aynı şekilde σ^2 varyans değeri arttırıldığında tahmin edicilerin toplam MSE değerlerinin arttığı görülmüştür. Simülasyon çalışmasında örneklem hacmi büyüdükçe genel olarak tahmin edicilerin toplam MSE değerleri azalmıştır. Bu sonuçlar simülasyon çalışmasının başarılı olduğu göstermektedir.

Aykırı gözlem ve çoklu bağıntı problemi içeren veri setlerinde klasik kantil regresyon yöntemi ile ridge tahminine dayalı kantil regresyon yöntemi karşılaştırıldığında ridge tahminine dayalı kantil regresyonun daha başarılı olduğu görülmüştür.

Ridge tahminine dayalı kantil regresyonunda yanlılık parametresi k'nın tahminlerine göre elde edilen regresyon katsayı tahminlerinin toplam MSE değerlerine göre grafiksel karşılaştırılması Şekil 1-3'de verilmiştir.

Tablo 1. Seçili parametre değerlerine göre tahmin edicilerin toplam MSE değerleri ($\tau = 0.25$)

*(**Table 1.** Total MSE values of the estimators according to selected parameter values ($\tau = 0.25$))*

σ	ρ^2	n	Kantil regresyonu	Ridge tahminine dayalı kantil regresyonu							
				$\hat{k}_{HK}$	$\hat{k}_{HKB}$	$\hat{k}_{LW}$	$\hat{k}_{HSL}$	$\hat{k}_{AM}$	$\hat{k}_{GM}$	$\hat{k}_{MED}$	$\hat{k}_{KS}$
0.2	0.95	10	69.203	32.065	15.825	11.35	1.059	1.252	3.374	2.169	32.288
0.2	0.95	30	7.659	1.997	1.616	1.008	0.734	0.634	0.903	0.874	2.284
0.2	0.95	100	6.574	1.650	1.514	1.402	0.706	0.472	0.674	0.665	2.427
0.2	0.99	10	360.09	170.14	84.709	54.17	0.531	1.697	11.389	8.943	168.91
0.2	0.99	30	38.840	9.600	7.371	4.647	0.382	0.968	2.811	2.860	9.983
0.2	0.99	100	33.557	7.578	6.871	5.503	0.377	0.746	2.134	1.785	8.442
0.5	0.95	10	548.23	98.841	36.356	29.79	0.633	0.803	4.041	3.536	99.134
0.5	0.95	30	43.866	5.555	3.366	3.060	0.398	0.430	0.924	0.952	6.164
0.5	0.95	100	40.663	5.035	3.373	3.509	0.386	0.395	0.775	0.776	7.444
0.5	0.99	10	2803.55	536.08	194.38	151.63	0.391	0.577	11.639	18.51	535.57
0.5	0.99	30	249.34	29.997	17.354	15.063	0.228	0.300	2.285	4.017	30.601
0.5	0.99	100	211.48	26.013	16.595	14.825	0.244	0.268	1.958	3.214	28.510

Tablo 2. Seçili parametre değerlerine göre tahmin edicilerin toplam MSE değerleri ($\tau = 0.50$)

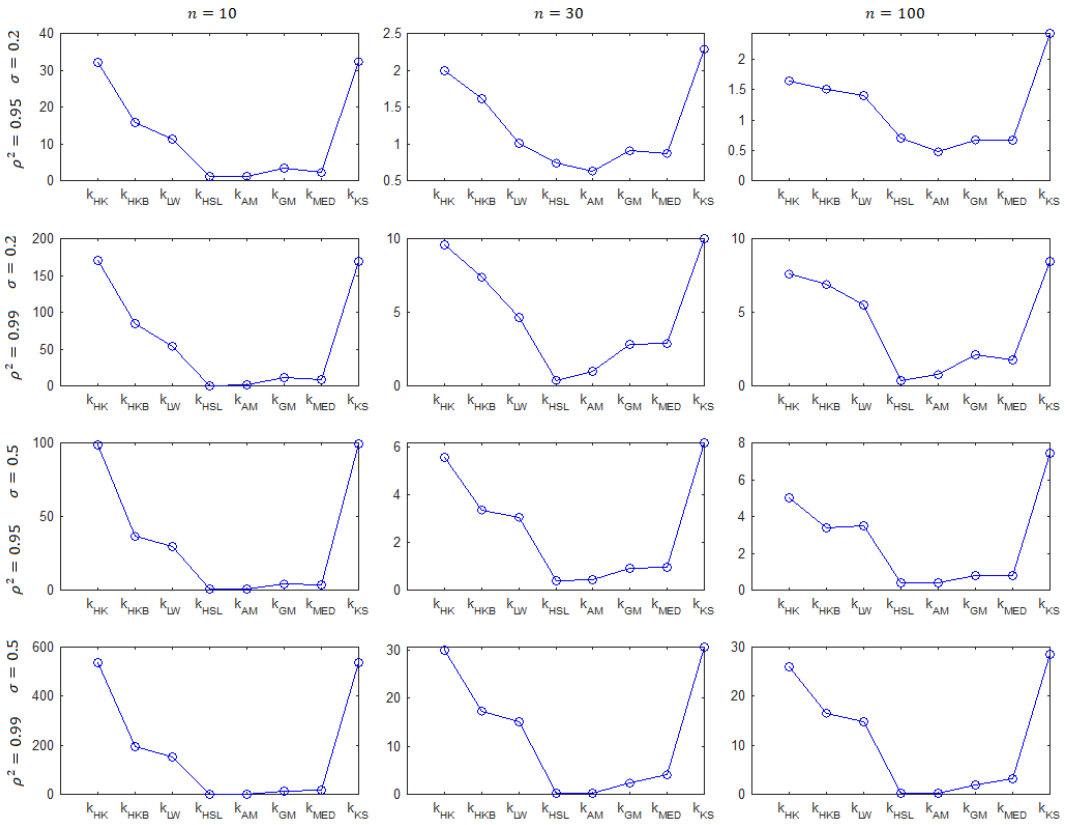
*(**Table 1.** Total MSE values of the estimators according to selected parameter values ($\tau = 0.50$))*

σ	ρ^2	n	Kantil regresyonu	Ridge tahminine dayalı kantil regresyonu							
				$\hat{k}_{HK}$	$\hat{k}_{HKB}$	$\hat{k}_{LW}$	$\hat{k}_{HSL}$	$\hat{k}_{AM}$	$\hat{k}_{GM}$	$\hat{k}_{MED}$	$\hat{k}_{KS}$
0.2	0.95	10	18.787	8.361	6.976	3.647	3.426	1.461	3.318	3.527	8.437
0.2	0.95	30	12.619	5.004	3.348	3.214	0.658	0.591	1.048	0.858	5.235
0.2	0.95	100	7.926	1.725	1.549	1.289	0.641	0.444	0.649	0.628	2.581
0.2	0.99	10	106.74	48.623	38.075	16.519	2.009	2.627	10.717	5.582	48.794
0.2	0.99	30	69.483	25.069	15.622	14.486	0.270	0.712	3.043	2.819	25.090
0.2	0.99	100	38.817	7.421	6.842	6.219	0.304	0.632	1.906	1.652	8.271
0.5	0.95	10	170.66	39.485	22.263	13.811	2.811	1.307	4.543	2.788	39.793
0.5	0.95	30	66.503	12.411	6.489	6.315	0.431	0.412	1.218	1.266	13.055
0.5	0.95	100	49.380	6.056	3.944	3.979	0.348	0.368	0.789	0.786	8.443
0.5	0.99	10	971.28	246.86	121.45	71.103	1.171	1.014	11.069	5.280	246.49
0.5	0.99	30	411.58	74.240	34.472	35.220	0.233	0.258	2.963	5.596	73.629
0.5	0.99	100	249.94	30.741	19.086	19.166	0.240	0.250	1.868	3.209	33.124

Tablo 3. Seçili parametre değerlerine göre tahmin edicilerin toplam MSE değerleri ($\tau = 0.75$)

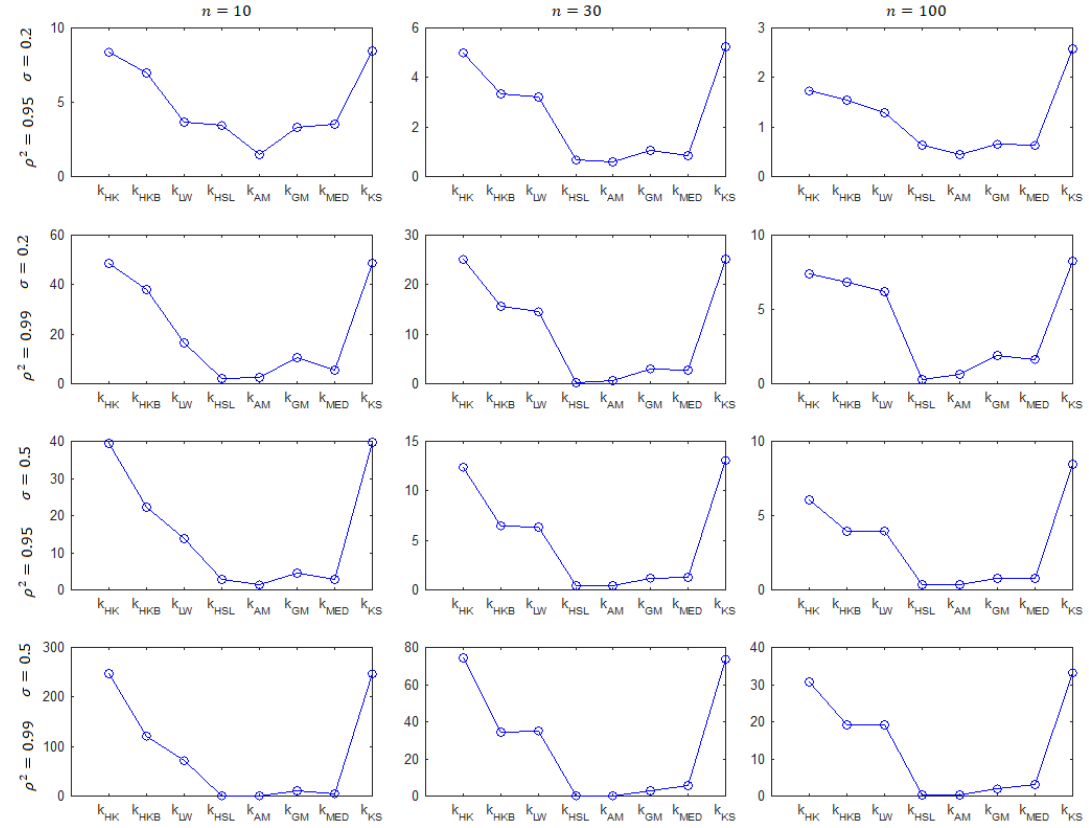
*(**Table 3.** Total MSE values of the estimators according to selected parameter values ($\tau = 0.75$))*

σ	ρ^2	n	Kantil regresyonu	Ridge tahminine dayalı kantil regresyonu							
				$\hat{k}_{HK}$	$\hat{k}_{HKB}$	$\hat{k}_{LW}$	$\hat{k}_{HSL}$	$\hat{k}_{AM}$	$\hat{k}_{GM}$	$\hat{k}_{MED}$	$\hat{k}_{KS}$
0.2	0.95	10	33.079	9.761	8.836	2.517	3.187	2.488	4.456	4.115	9.866
0.2	0.95	30	13.065	4.051	2.847	2.797	0.604	0.474	0.826	0.672	4.266
0.2	0.95	100	6.680	1.954	1.546	1.485	0.678	0.484	0.717	0.702	2.737
0.2	0.99	10	178.08	53.774	49.181	13.837	2.588	5.244	16.816	14.392	53.423
0.2	0.99	30	77.039	22.415	15.387	15.036	0.258	0.641	2.662	1.982	22.617
0.2	0.99	100	34.002	9.165	7.299	7.105	0.301	0.686	2.238	2.008	9.970
0.5	0.95	10	280.76	44.962	29.306	11.893	3.696	3.596	8.945	7.701	45.612
0.5	0.95	30	71.010	10.409	6.696	6.713	0.391	0.397	1.051	0.903	11.353
0.5	0.95	100	41.055	5.393	3.404	3.551	0.357	0.377	0.809	0.872	7.933
0.5	0.99	10	1490.7	265.19	169.52	68.173	1.621	3.350	28.096	33.232	262.37
0.5	0.99	30	455.34	68.480	35.191	36.406	0.226	0.264	2.749	3.924	68.383
0.5	0.99	100	212.82	27.476	16.888	17.643	0.265	0.251	1.969	3.709	29.330



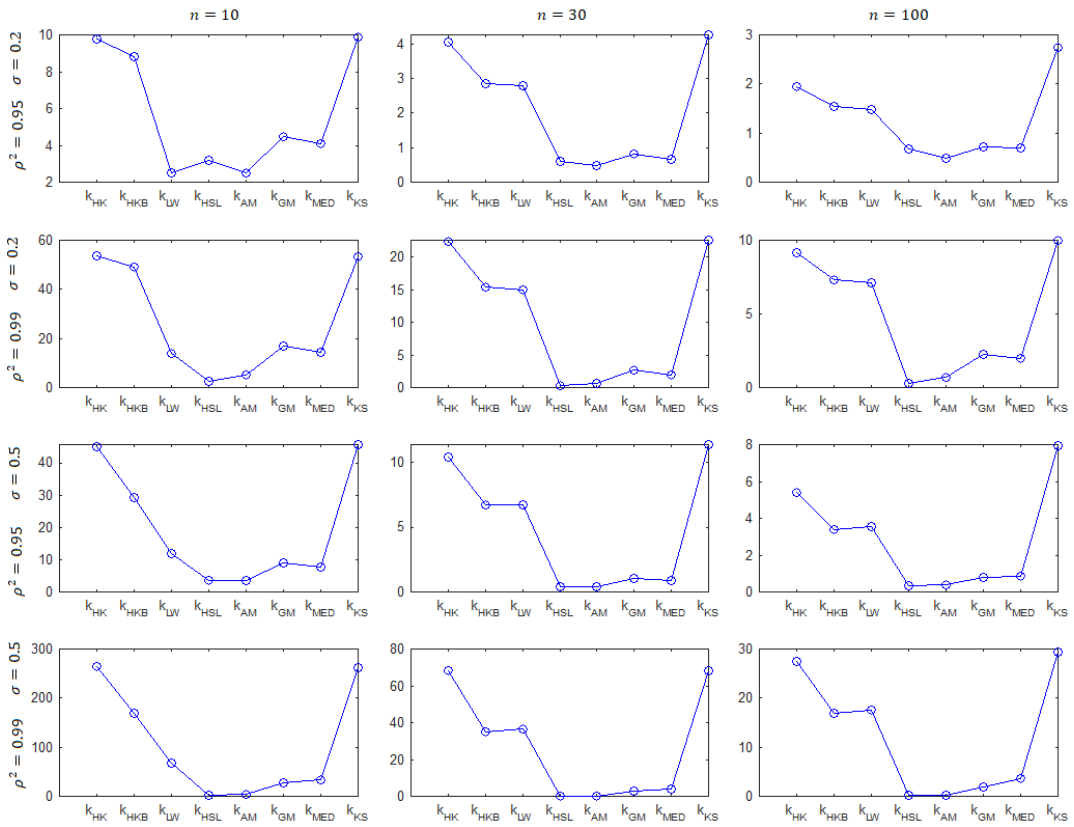
Şekil 1. $\tau = 0.25$ için yanlılık parametresi tahminlerine göre elde edilen regresyon katsayı tahminlerinin toplam MSE değerlerinin çizgi grafikleri

(**Figure 1**. Line plots of the total MSE values of the regression coefficient estimates obtained according to the bias parameter estimates for $\tau = 0.25$)



Şekil 2. $\tau = 0.50$ için yanlılık parametresi tahminlerine göre elde edilen regresyon katsayı tahminlerinin toplam MSE değerlerinin çizgi grafikleri

(**Figure 2**. Line plots of the total MSE values of the regression coefficient estimates obtained according to the bias parameter estimates for $\tau = 0.50$)



Şekil 3. $\tau = 0.75$ için yanlılık parametresi tahminlerine göre elde edilen regresyon katsayı tahminlerinin toplam MSE değerlerinin çizgi grafikleri

(**Figure 3**. Line plots of the total MSE values of the regression coefficient estimates obtained according to the bias parameter estimates for $\tau = 0.75$)

Ridge tahminine dayalı kantil regresyonunda yanlılık parametresi k'nın tahmin edicileri karşılaştırıldığında Hocking, Speed ve Lynn [15] tarafından önerilen $\hat{k}_{HSL}$ ve Kibria [16] tarafından önerilen $\hat{k}_{AM}$ tahminlerinin daha başarılı olduğu gözlemlenmiştir. Hoerl ve Kennard [2] tarafından önerilen $\hat{k}_{HK}$ ile Khalaf ve Shukur [17] tarafından önerilen $\hat{k}_{KS}$ tahminleri diğer yanlılık tahmin edicilerine göre daha yüksek MSE değerine sahip olmuş ve incelenen k yanlılık tahmin ediciler arasında en kötü performansı göstermiştir.

5. UYGULAMA (APPLICATION)

Tobacco verisi çoklu iç ilişki ve aykırı değer problemi taşıyan bir veri setidir [20]. Tobacco verisi bir yanıt değişkeni ve dört açıklayıcı değişken içeren 30 birimden oluşan bir veri setidir. Bu veri setinde kantil regresyonu ile ridge tahminine dayalı kantil regresyonunun performansı tekrarlı k katmanlı çapraz doğrulama tekniği ile karşılaştırılacaktır. Uygulamada tobacco verisi standartlaştırılmıştır.

Tekrarlı k katmanlı çapraz doğrulama tekniğinde, veri seti rastgele k parçaya bölünür ve bölünen her parça sırayla test verisi, geri kalan k-1 parçadan oluşan veri seti de eğitim verisi (traning data) olur. Böylece oluşturulan her k parça test verisi olarak kullanılmış olur. Eğitim verisi ile oluşturulan modelin etkinliği test verisinde ölçülür. Verinin rastgele k parçaya ayrılmasındaki rastgeleliğin model belirlemedeki etkisini azaltmak için bu işlemler tekrarlanarak tekrarlı k katmanlı çapraz doğrulama gerçekleştirilmiş olur. Veri seti k parçaya bölündükten sonra her parçadaki gözlem sayısı r olmak üzere test verisindeki toplam hata değeri

$$E(k) = \sum_{i=1}^{r} \left(y_{i,test} - \hat{y}_{i,test}\right)^2 \qquad (21)$$

eşitliği ile elde edilir. Eşitlikte yer alan $\hat{y}_{i,test}$ değeri k-1 parçadan oluşan eğitim verisi ile elde edilen parametre tahmin değerlerine göre test verisindeki i. yanıt değişkeninin tahmini değerini göstermektedir. Veri setinin bölünmesi ile oluşan k parçanın her biri test verisi olarak kullanıldığından k tane toplam hata değeri hesaplanır. Tekrar sayısı t olmak üzere çapraz doğruluk hatası

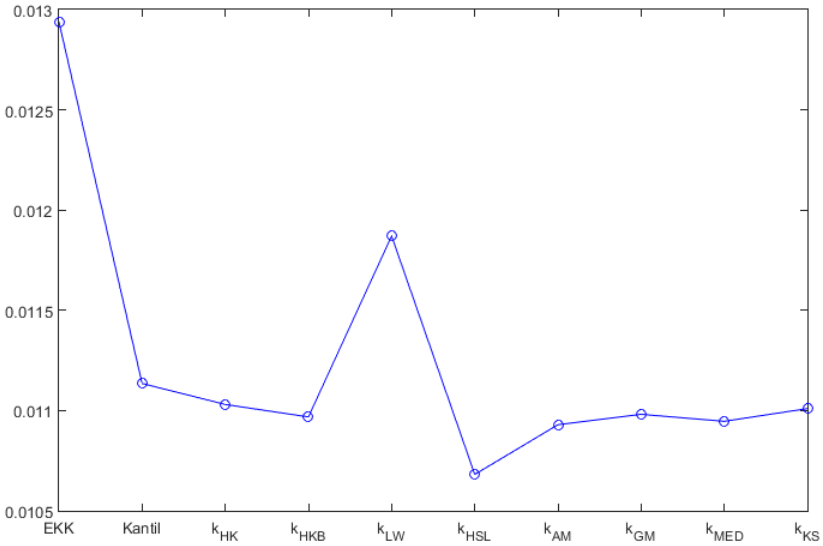
$$CV = \frac{1}{t}\sum_{i=1}^{t}\frac{1}{k}\sum_{m=1}^{k} E_i(m) \quad (22)$$

eşitliği ile elde edilir. Elde edilen değerin küçük olması parametre tahmin yönteminin performansının iyi olduğu anlamına gelir.

Uygulamada, tekrarlı k katmanlı çapraz doğrulama tekniğinde k=5 ve tekrar sayısı 1000 alınmıştır. Tobacco verisi için EKK, Kantil ve ele alınan yanlılık parametresi tahminlerine göre ridge tahminine dayalı kantil regresyon analizlerinden elde edilen katsayı tahminleri, çoklu belirlilik katsayısı R^2 ve tekrarlı k katmanlı çapraz doğrulma tekniği ile elde edilen CV değerleri Tablo 4'de verilmiştir. Tablo 4'de ayrıca elde edilen yanlılık parametresi tahminleri de parantez içerisinde verilmiştir. Tekrarlı k katmanlı çapraz doğrulma tekniği ile elde edilen CV değerlerine ait çizgi grafiği Şekil 4'de verilmiştir.

Tablo 4. Farklı tahmin yöntemlerine göre regresyon katsayı tahminleri, R^2 ve CV değerleri
(**Table 4.** Regression coefficient estimations according to different estimation methods, R^2 and CV values)

	EKK	Kantil regresyon	Ridge tahminine dayalı kantil regresyonu							
			$\hat{k}_{HK}$ (0.000790)	$\hat{k}_{HKB}$ (0.001677)	$\hat{k}_{LW}$ (0.105750)	$\hat{k}_{HSL}$ (0.006693)	$\hat{k}_{AM}$ (0.003208)	$\hat{k}_{GM}$ (0.002312)	$\hat{k}_{MED}$ (0.002532)	$\hat{k}_{KS}$ (0.000786)
$\hat{\beta}_1$	1.50739	1.50151	1.27887	1.15996	0.34391	0.40921	0.36151	1.10929	0.34930	1.28448
$\hat{\beta}_2$	-0.52107	-0.52475	-0.54479	-0.55466	-0.16386	-0.40675	-0.34839	-0.48318	-0.34573	-0.54454
$\hat{\beta}_3$	-0.84160	-0.88288	-0.73362	-0.67789	0.20556	-0.03318	-0.31753	-0.71234	-0.30900	-0.73434
$\hat{\beta}_4$	0.82171	0.87158	0.96596	1.03944	0.60515	1.01369	1.28394	1.06697	1.28541	0.96081
R^2	0.95720	0.95717	0.95697	0.95668	0.94799	0.95262	0.95290	0.95641	0.95280	0.95698
CV	0.01294	0.01114	0.01103	0.01097	0.01187	0.01068	0.01093	0.01098	0.01095	0.01101



Şekil 4. Tekrarlı k katmanlı çapraz doğrulma ile elde edilen CV değerlerine ait çizgi grafiği
(**Figure 4.** Line plots of CV values obtained by repeated k-fold cross validation)

Tablo 4 ve Şekil 4 incelendiğinde Hocking, Speed ve Lynn [15] tarafından önerilen $\hat{k}_{HSL}$ yanlılık parametre tahminine dayalı kantil regresyon yöntemi ile en küçük CV değeri elde edilmiştir. Simülasyon sonuçları ile uyumlu bir şekilde CV kriterine göre en başarılı iki yanlılık tahmin yöntemi Hocking, Speed ve Lynn [15] tarafından önerilen $\hat{k}_{HSL}$ ve Kibria [16] tarafından önerilen $\hat{k}_{AM}$ tahmin edicileridir. Çoklu bağlantı ve aykırı gözlem problemleri içeren veri setinde EKK tahminleri en yüksek CV değeri ile en başarısız tahmin edici olmuştur.

6. SONUÇ (CONCLUSION)

Bu çalışmada çoklu bağlantı problemi ve aykırı gözlem varlığında kantil regresyonunda ridge tahminine dayalı çözüm incelenmiştir. Literatürde yaygın kullanıma sahip sekiz yanlılık tahmin edicisinin performansı tahmin edicilerin toplam MSE kriterine göre değerlendirilmiştir. Simülasyon ve gerçek veri seti ile gerçekleştirilen uygulama sonuçları çoklu bağlantı problemi ve aykırı gözlem varlığında kantil regresyonunda ridge tahmin

yaklaşımının kullanılabileceğini göstermiştir. Simülasyon çalışması sonucunda Hocking, Speed ve Lynn [15] tarafından önerilen $\hat{k}_{HSL}$ ve Kibria [16] tarafından önerilen $\hat{k}_{AM}$ yanlılık tahmin edicileri en başarılı tahmin ediciler olarak belirlenmiştir. Tobacco veri setinde tekrarlı k katmanlı çapraz doğrulama tekniği ile gerçekleştirilen karşılaştırma sonucunda en başarılı yanlılık tahmin edicisi $\hat{k}_{HSL}$ olmuştur.

TEŞEKKÜR (ACKNOWLEDGEMENT)

Bu çalışma Doç. Dr. Murat ERİŞOĞLU'nun danışmanlığında Nurullah YAMAN'ın Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalında devam eden "Kantil Regresyonda Yanlı Tahmin Edicilerin Performanslarının İncelenmesi" başlıklı Yüksek Lisans tezinden üretilmiştir.

KAYNAKLAR (REFERENCES)

[1] R. C. Pfaffenberger, T. E. Dielman, A comparison of regression estimators when both multicollinearity and outliers, içinde: K. D. Lawrence, J. L. Arthur (Ed.) Robust Regression: Analysis and applications, Marcer Dekker Inc. New york and Basel, (1990) 243-270.

[2] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics,* 12(1) (1970) 55-67. doi: 10.1080/00401706.1970.10488634

[3] M. S. Suhail, Chand, B. M. G. Kibria, Quantile based estimation of biasing parameters in ridge regression model, *Communications in Statistics-Simulation and Computation*, (2019) doi: 10.1080/03610918.2018.1530782

[4] A. A. Yavuz, E. G. Aşık, Kantil Regresyon, *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi*, 9(2) (2017) 137-146. doi: 10.29137/umagd.352530

[5] C. Chen, An introduction to quantile regression and the QUANTREG procedure, *In Proceedings of the Thirtieth Annual SAS Users Group International Conference,* SAS Institute Inc. Cary, NC. 2005, 213-30.

[6] R. Koenker, G. Basset, Regression Quantiles, *Econometrica* 46 (1)(1978) 33-50. doi: 10.2307/1913643.

[7] R. Koenker, K. F. Hallock, Quantile Regression, *Journal of Economic Perspectives* 15 (4)(2001) 143-156. doi: 10.1257/jep.15.4.143.

[8] D. Baur, M. Saisana, N. Schulze, Modelling the effects of meteorological variables on ozone concentration: a quantile regression approach, *Atmospheric Environment*, 38(28) (2004) 4689-4699. doi: 10.1016/j.atmosenv.2004.05.028

[9] İ. Altındağ, Quantile regresyon ve bir uygulama, Yüksek Lisans Tezi, *Selçuk Üniversitesi Fen Bilimleri Enstitüsü*, İstatistik Ana Bilim Dalı, Konya, 2010.

[10] A. S. Bager, Ridge Parameter in Quantile Regression Models: An Application in Biostatistics, *International Journal of Statistics and Applications*, 8(2) (2018) 72-78. doi: 10.5923/j.statistics.20180802.06

[11] H. Zaikarina, A. Djuraidah, A. H. Wigena, Lasso and Ridge Quantile Regression using Cross Validation to Estimate Extreme Rainfall, *Global Journal of Pure and Applied Mathematics*, 12 (3) (2016) 3305–3314.

[12] Z. Zeebari, Developing ridge estimation method for median regression, *Journal of Applied Statistics,* 39(12) (2012) 2627-2638. doi: 10.1080/02664763.2012.724663

[13] A. E. Hoerl, R. W. Kennard, K. F. Baldwin, Ridge regression: Some simulations. *Communications in Statistics,* 4(2) (1975) 105-123. doi:10.1080/03610927508827232

[14] J. F. Lawless, P. Wang, A simulation study of ridge and other regression estimators, *Communications in Statistics – Theory and Methods,* 5(4)(1976) 307-323. doi: 10.1080/03610927608827353

[15] R. R. Hocking, F. M. Speed, M. J. Lynn, (1976). A class of biased estimators in linear regression, *Technometrics,* 18(4) (1976) 55-67. doi:10.1080/00401706.1976.10489474

[16] B.M. G. Kibria, Performance of some new ridge regression estimators, *Communications in Statistics – Simulation and Computation*, 32(2) (2003) 419-435. doi: 10.1081/SAC-120017499

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

[18] D. G. Gibbons, A Simulation Study of Some Ridge Estimators, Journal of the American Statistical Association, 76 (1981) 131-139. doi: 10.1080/01621459.1981.10477619

[19] G. C. McDonald, D. I. Galarneau, A Monte Carlo Evaluation of Some Ridge-Type Estimators, Journal of the American Statistical Association, 70 (350) (1975) 407-416. doi: 10.1080/01621459.1975.10479882

[20] R. H. Myers, Classical and modern regression with applications, Second Edition, *Belmont, CA: Duxbury press,*1990.