Uludağ Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Cilt XXII, Sayı 1, 2003, s. 281-292

BOOTSTRAP YÖNTEMİNİN RIDGE REGRESYONDA UYGULANMASI

Dilek ALTAŞ [*]
Murat ÇİNKO [**]

Özet

İstatistiksel analizlerde, en yaygın olarak kullanılan yöntemlerden biri regresyon analizidir. Gauss-Markov varsayımları altında en küçük kareler yöntemi ile, en iyi takdir ediciler elde edilmesine rağmen, bağımsız değişkenler arasında ilişki olması durumunda, en küçük varyansa sahip takdir edici özelliğini kaybetmektedir. Bu duruma önerilen çözüm yöntemlerinden biri ridge regresyondur. Ridge regresyon ile elde edilen takdir edicilerin varyansı en küçük kareler yöntemi ile bulunan varyanslara göre daha küçüktür.

Bu çalışmada ridge regresyonda parametrelerin tahmin edilmesinde Efron(1979) tarafından önerilen ve bir örneklemden tekrarlı örneklemlerin seçilmesi esasına dayanan Bootstrap yöntemi kullanılarak takdir edicilerin dağılımları ortaya konulmaya çalışılmıştır. Çoklu bağlantı olan veri setinde ridge regresyon yöntemi kullanılarak elde edilecek olan hata terimlerinden tekrarlı örneklemler seçilerek en iyi parametre takdir edicileri elde edilmiş ve uygulama sonucunda elde edilen sonuçlar karşılaştırılmıştır.

Anahtar Kelimeler: *Ridge Regresyon, Bootstrap.*

Abstract

Regression analysis is the most frequently used tool in statistics. Under the Gauss Markov assumption maximum likelihood estimators are the best estimator.

---

[*] *Yard. Doç. Dr.; Marmara Üniversitesi Ekonometri Bölümü İstatistik Anabilim Dalı.*
[**] *Araş. Gör.; Marmara Üniversitesi İngilizce İşletme Bölümü Sayısal Yöntemler Anabilim Dalı.*

However, these estimators do not have the minimum variance if the independent variables have multicollinearity. The remedy for this problem is the Ridge regression. The estimators obtained by Ridge regression have less variance compare to the maximum likelihood estimation. In this study ridge regression estimator are going to be estimated by bootstrap techniques. At the end of the study the distribution of the estimators are going to be shown.

Keywords: *Ridge Regression, Bootstrap.*

1. GİRİŞ

Değişkenler arasındaki ilişkiyi araştıran en önemli istatistiksel yöntemlerden biri regresyon analizidir. Regresyon analizinin hayatımızın bir çok alanında kullanılması mümkündür: mühendislik, fizik, iktisat, biyoloji, sosyal bilimler bunlardan sadece bir kaçıdır. Regresyon analizinin amacı bilinmeyen model parametreleri hakkında takdirde bulunmaktır. Model kurulduktan sonra modelin doğruluğunun test edilmesi gerekmektedir. Unutulmaması gereken en önemli konu ise regresyon analizi sebep-sonuç ilişkisini incelemez, sadece değişkenler arasında var olan ilişkinin modellenmesini ortaya koyar. Modelin oluşturula bilmesi için Gauss-Markov varsayımlarının gerçekleşmesi gerekmektedir. Varsayımlardan herhangi birinin gerçekleşmemesi durumunda geçerli sonuçlara ulaşmak mümkün olmaz. Böyle bir durumla karşılaşıldığında her bir durum için farklı çözüm yoları mevcuttur.

Örnekden elde edilen takdir değerlerinin anakütle gerçek değerlerine, α ve β, yakın olabilmesi için bazı varsayımların, Gauss-Markov varsayımları, yapılması gerekmektedir. Berry (1993:12) bu varsayımları şöyle sıralamaktadır:

1. Bağımsız değişkenlerin hepsi nicel veya nitel olarak ölçülmüş olması, bağımlı değişken Y'nin ise nicel ve sürekli olması gerekmektedir. X ve Y değişkenleri doğru olarak ölçülmelidir.
2. Bütün bağımsız değişkenlerin varyansının sıfırdan farklı olması gerekmektedir.
3. Bağımsız değişkenler arasında doğrusal bir ilişkinin olmaması gerekmektedir.
4. Hata terimleri ortalaması sıfırdır: $E(\varepsilon_i) = 0$.
5. Bağımsız değişkenler ve hata terimi arasında korelasyon olmamalıdır.
6. $E(\varepsilon_i X_i) = 0$.
7. Hata terimlerinin varyansı sabit olmalıdır, Homoskedasticity: $E(\varepsilon^2_i) = \sigma^2$.

8. Hata terimleri arasında korelasyon olmamalıdır $E(\varepsilon_j \varepsilon_i)=0$ ($i \neq j$).
9. Hata terimleri ε_i, normal dağılmalıdır.

Gauss-Markov varsayımlarının gerçekleşmesi durumunda maksimum benzerlik yöntemine göre takdir edilen değerler: yansızlık, yeterlilik gibi özelliklere sahip olmakta ve istatistiksel testlerin yapılması güven aralıklarının oluşturulmasını mümkün olamamaktadır. Bu çalışmanın amacı varsayımlardan üçüncüsü olan, bağımsız değişkenler arasında doğrusal bir bağlantı olmamalıdır (çoklu bağlantı), sorunuyla karşılaşılırsa nasıl çözümleneceğini göstermektir. Çoklu bağlantı olduğunda önerilen çözüm yollarından ilki arasında çoklu bağlantı olan bağımsız değişkenlerden sadece bir tanesinin alınması iken, ikinci yöntem Ridge regresyonudur. Eğer çoklu bağlantı var ise takdir edicilerin varyansları büyümektedir buda takdir edicilerin yanlı olmasına ve tutarlı olmamasına sebep olmaktadır. Ridge regresyonun kullanılması sonucunda elde edilen takdir ediciler en küçük kareler yönteminde elde edilen takdir edicilere göre daha küçük varyansa sahip olmaktadır.

Bu çalışmanın ikinci kısımda çoklu bağlantı kavramı açıklanırken, üçüncü bölümde ridge regresyon yöntemi anlatılacak, dördüncü bölümde bootstrap tekniği anlatılacak, beşinci bölümde ise uygulama ve sonuçları verilecek ve sonuç bölümü ile çalışma sonuçlandırılacaktır.

2. ÇOKLU BAĞLANTI KAVRAMI

Çoklu regresyon modelini

$$y = X\beta + \varepsilon \qquad (1)$$

şeklinde yazdığımızda

y'nin bağımlı değişkeni *nx1*,

X bağımsız değişkenleri *nxp*,

β bilinmeyen parametrelerin *px1*,

ε ise *nx1*'lik artıkların

matrisini oluşturmaktadır.

X bağımsız değişkenlerinin herhangi bir sütunu diğer sütunların doğrusal bir bileşeni olarak ifade edilebilirse $(X'X)$ matrisinin rankı bağımsız değişken sayısından az olacak ve $(X'X)^{-1}$ değeri hesaplanamayacaktır. Eğer bağımsız değişkenler arasında doğrusala yakın bir ilişki var ise $(X'X)$ matrisinde problemler oluşacaktır. Bağımsız değişkenler arasında tam yada yüksek derecede ilişkilerin bulunması durumu çoklu bağlantı kavramı ile açıklanmaktadır.

(Montgomery ve Peck,1992: 305-324) çoklu doğrusal bağlantıyı tespit etmenin yöntemini şu şekilde açıklamaktadırlar.

- Bağımsız değişkenler arasındaki korelasyonun incelenmesi,
- Genel modelde yüksek F değerine rağmen katsayı testlerindeki t değerinin küçük çıkması.
- VIF (Varyans Büyütme Faktörü) bakılarak anlaşılabilir. VIF değeri bağımsız değişkenler arasında yapılan regresyon sonucunda bulunan bir katsayıdır. Her bir bağımsız değişken bağımlı değişkenmiş gibi düşünülür diğer değişkenler kullanılarak regresyon modeli kurulup belirlilik katsayısı (determination of coefficient) bulunur. $VIF_J = 1/(1- R^2_j)$ değeri hesaplanır. Bu değerin 10'dan büyük olması durumunda çoklu bağlantıdan şüphe edilebilir.
- Özdeğerlerin terslerinin toplamı bağımsız değişken sayısından büyük olduğunda.
- Özdeğerlerin en büyüğünün en küçüğüne oranının 100'den büyük olması durumunda.

Çoklu bağlantının olduğu durumlarda modelin takdir değerlerinin doğru sonuç vermesi için kullanılabilecek bazı yöntemler şu şekilde sıralanabilir.

- Daha fazla veri toplamak
- Oluşturulan modelden kaynaklanan problemlerde modelin değiştirilmesi söz konusu olabilir. Mesela x_1,x_2,x_3 değişkenleri arasında güçlü bir korelasyon var ise yeni bir değişken tanımlanabilir $x=(x_1+x_2)/x_3$ veya $x=x_1*x_2*x_3$ olarak tanımlanıp modele x değişkeni konur.
- Ridge regresyonu kullanılarak model parametrelerinin yanlı tahin edilmesi sağlanabilir.

Bu çalışmada üçüncü çözüm yöntemi olan ridge regresyonu kulanılacaktır.

2.1. Çoklu Bağlantının Etkileri

En küçük kareler takdir edicilerinin varyans kovaryans matrisi $\sigma^2(X'X)^{-1}$ olarak tanımlanmaktadır. Diagonal elemanlarının toplanması ise ağırlıklandırılmamış ortalama hata karelerinin toplamını (OHK) vermektedir.

$$OHK(b)=\sigma^2 Tr(X'X)^{-1} \tag{2}$$

$X'X$ matrisinin özdeğerleri türünden yazarsak

$$\text{OHK(b)}= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1} > \sigma^2 \lambda_p^{-1} \qquad (3)$$

olacaktır.

Çoklu bağlantının etkileri şu şekilde sıralanabilir:

- b_i'lerin varyanslarının yüksek olması.
- katsayıların işaretlerinin ters olması.
- regresyon katsayıları durağan, kararlı olmaması.
- belirlilik katsayısının yorumlanmasında hatalı sonuçlar vermesi.

3. RIDGE REGRESYON

Çoklu bağlantı olan veri setine en küçük kareler yöntemi uygulandığında takdir edilen değerler şişmektedir. Bunun anlamı ise takdir edilen değerlerin mutlak değerlerinin büyüklük ve işaret olarak farklı örnekler için farklı sonuçlar bulunması olacaktır. Buda en küçük kareler yöntemi ile takdir edilen değerlerin en küçük varyansa sahip olduğu yolundaki sonucun ihlal edilmesi anlamına gelmektedir. Hoerl ve Kennard, (1970a,b: 55-66; 69-82) yaptıkları çalışmada takdir edicilerin aşağıdaki şekilde bulunmasını önermektedirler.

$$\hat{\beta}_R = (X'X + kI)^{-1} X'y \qquad (4)$$

Ridge takdir edicileri aslında en küçük kareler yöntemine göre elde edilen takdir edicilerin doğrusal bir kombinasyonudur.

$$\begin{aligned} \hat{\beta}_R &= (X'X + kI)^{-1} X'y \\ &= (X'X + kI)^{-1} (X'X)\hat{\beta} \\ &= Z_k \hat{\beta} \end{aligned} \qquad (5)$$

$\hat{\beta}_R$ beklenen değeri $E(\hat{\beta}_R) = E(Z_k \hat{\beta}) = Z_k \beta$, β'nın yanlı takdir edicisi $\hat{\beta}_R$ 'dır.

k değerinin seçilmesinde kullanılan yöntemler:

1. Hoerl, Kennard ve Baldwin (1975: 105-123) tarafından önerilen

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{6}$$

2. McDonald ve Galarneau (1975: s.16) tarafından önerilen

$$\hat{\beta}_R^{'}\hat{\beta}_R = \hat{\beta}_R^{'}\hat{\beta}_R^{'} - \hat{\sigma}^2 \sum_{i=1}^{p} \left(\frac{1}{\lambda_j} \right) \tag{7}$$

3.1 Ridge İzi

Çoklubağlantı olduğunda katsayıların seçilen örneğe göre farklı takdir değerlerinin bulunacağı daha önceden belirtilmişti. Bu takdir değerlerinin karasız olduğu bu yüzden de k gibi sayının eklenerek takdir edicilerin bulunacağı belirtilmişti. Ridge izi yatay eksende k değerlerinin, düşey eksende ise her bir k değeri için takdir edilen katsayıların olduğu grafiğe denir. Bu grafik sayesinde hangi takdir değerlerinin duyarlı olduğunun bulunması için araştırmacıya yardımcı olur. Bu grafik sayesinde de, zor olmasına rağmen, k parametresinin ne olması gerektiğine karar verilebilir. Takdir değerlerinin kararlı olduğu k değeri ridge parametresi olacaktır. Şekil 1 bu çalışma için risge izi grafiğini göstermektedir.

4. BOOTSTRAP YONTEMİ

Efron (1979:1-26) yeniden örneklemeyi, anakütleden elde edilen örneklemden alt örneklemlerin seçilmesi olarak tanımlamaktadır. Regresyon analizinde hata terimleri ve bağımsız değişkenlerle ilgili varsayımların gerçekleşmemesi durumunda bir düzeltme işlemi amacıyla da kullanılan bootstrap yöntemi, daha küçük tahmin hatalarının elde edilmesi, standart sapmaların küçülmesi ve buna bağlı olarak da daha güvenilir parametre tahmincilerinin elde edilmesi ve güven aralıklarının oluşturulması amacıyla geliştirilmiştir. (Efron ve Tibshirani, 1993: 45) Bootstrap yöntemi ile elde edilen tahmincilerin etkinlikleri varsayımdan sapmalardan etkilenmemektedir. (Shou ve Tu, 1995:292)

Çalışmada regresyon analizinde hata terimlerinin yeniden örneklemesine dayanan bootstrap yaklaşımı kullanılmıştır.

$Y=X\beta+\varepsilon$ doğrusal regresyon modelinde, hata terimlerinin dağılımı ($F\varepsilon$), $e_i - \bar{e}$ değerlerine 1/n olasılığı verilerek, (Shou ve Tu, 1995:289)

$$\hat{\mathbf{F}}_{\varepsilon}(\mathbf{x}) = \#\left\{\hat{\mathbf{e}}_{\mathbf{i}}^{\,2} \leq \mathbf{x}\right\} / \mathbf{n} \tag{8}$$

şeklinde ifade edilen deneysel dağılım($\hat{\mathbf{F}}_{\varepsilon}(\mathbf{x})$) ile tahmin edilir (Stine, 1985: 1027).

Burada $e_i=Y_i-\hat{\mathbf{Y}}_{\mathbf{i}}$ i. hata terimidir ve $\bar{\mathbf{e}} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{e}_{\mathbf{i}}}{\mathbf{n}}$ ’dir. $\hat{\mathbf{F}}_{\varepsilon}$ ’den hata terimlerinin bootstrap örnekleri ($\varepsilon^*_{1}, \ldots, \varepsilon^*_{n}$) elde edilerek, $Y_i^*=X\hat{\beta}+\varepsilon^*_i$ olmak üzere bootstrap Y değerleri hesaplanır. Y^* ve X’den hareketle β’nın bootstrap tahmini, en küçük kareler yöntemi ile,

$$\hat{\beta}^{*} = (\mathbf{X}^{/}\mathbf{X})^{-1}\mathbf{X}^{/}\mathbf{Y}^{*} \tag{9}$$

şeklinde elde edilir. (Liu, 1988:1706)

Bootstrap yönteminin bu uygulamasında hata terimlerinden tekrarlı örneklemler seçilerek, tahmin değerlerine($\hat{\mathbf{Y}}_{\mathbf{i}}$) eklendiği için hata terimlerinin normal dağılıma sahip olduğu varsayılmaktadır (Fox, 1997: 506).

$$E_*(\varepsilon^*_i)=0 \text{ olduğu için,} \tag{10}$$

$$E_*(\hat{\beta}^{*}) = (X^{/}X)^{-1}X^{/}\, E_*(Y^*)= \hat{\beta}$$

olmaktadır.

5. UYGULAMA

Uygulama için Matlab programında aralarında yüksek korelasyona sahip olan bir veri seti yaratılmıştır. Dört adet bağımsız değişken yaratılarak 100 adet veri elde edilmiştir. Veri seti standardize edildikten sonra ridge regresyon uygulanmıştır.

5.1 Yeniden Örnekleme Yönteminin Kullanılması

Çalışmada yeniden örnekleme yöntemi iki şekilde kullanılmıştır. Öncelikle standardize edilmiş veri setine ridge regresyon uygulanarak ve ***k*** parametresi bulunmuştur. Bu işlemden sonra elde edilecek olan hata terimleri (11) modeli kullanılarak elde edilecek olan yeni y değerleri sayesinde β takdir edicileri hesaplanmıştır.

$$Y^* = x\beta + \varepsilon^* \qquad (11)$$

Burada ε^* yeniden örnekleme ile elde edilecek olan hata terimleri, Y^* yeniden örnekleme sonucu elde edilen hata terimlerinin eklenmesi sonucu elde edilen bağımlı değişkeni göstermektedir.

Çoklu bağlantının tespiti için bir çok yöntem olduğu daha önceden açıklanmıştı. Bağımsız değişkenler arasındaki korelasyon matrisi Tablo-1'de verilmiştir.

Tablo 1. Bağımsız Değişkenler Arasındaki Korelasyon

	x1	**x2**	**x3**	**x4**
X1	1.000			
X2	0.997	1.000		
X3	0.992	0.996	1.000	
X4	0.991	0.995	0.998	1.000

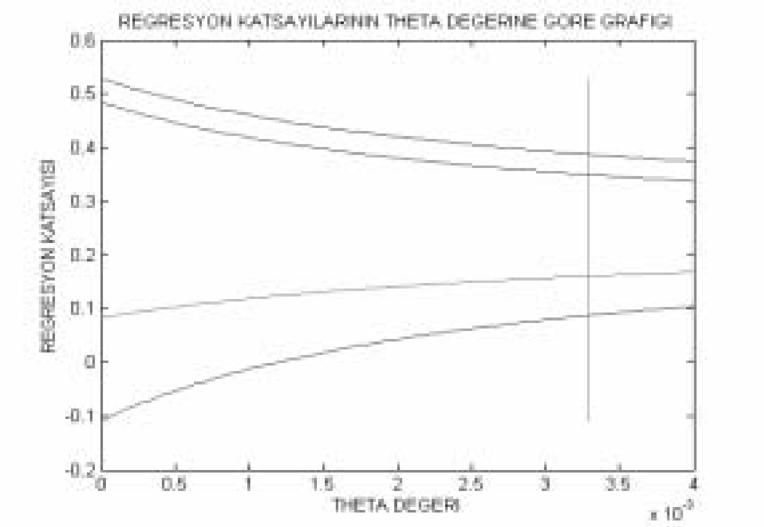
Tablodan da görüleceği üzere bağımsız değişkenler arasında yüksek bir korelasyon vardır.

Tablo 2. Varyans Büyütme Faktörü

	x1	**x2**	**x3**	**x4**
VBF	146,25	317,88	349,25	231,41

Tablodan da görüleceği üzere VBF 10 sayısından oldukça büyüktür. Bu da bağımsız değişkenler arasında çoklu bağlantı olduğunu göstermektedir. Çoklu bağlantının varlığını göstermemenin bir başka yolu da özdeğerlerin en büyünün en küçüğüne bölünmesi sonucunda 100'den daha büyük bir rakam bulmak olarak belirtilmişti. En büyük özdeğerin, 654,15, en küçük özdeğere, 0,2316, oranı ise 2824,48'dir. Bu oranda çoklu bağlantı kavramını doğrulamaktadır.

Ridge izinin grafiğini çizecek olursak 0,0033 değerinin optimum değer olduğu hesaplanmış grafikten de bu değerden sonra katsayıların tutarlı kaldığı gözlemlenmektedir.

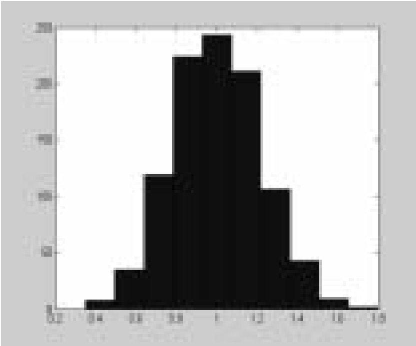


Şekil 1. *Ridge İzi*

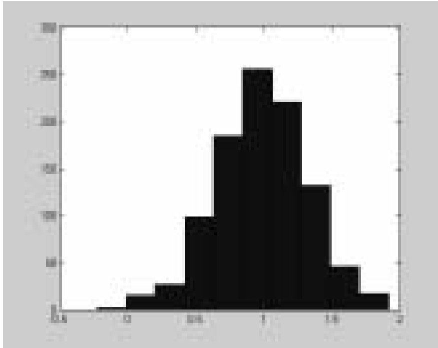
k parametresinin bulunması hata terimlerinin bu değeri kullanarak hesaplanmasını sağlamış ve hesaplanan hata terimlerini kullanarak bağımlı değişkeni yeniden bularak β takdir değerleri yeniden bulunmuştur. Aşağıdaki tabloda en küçük kareler yöntemi, ridge regresyonu ve 250, 1000, 2000 kere yapılan yeniden örnekleme ile elde edilen β değerleri, ve standart hataları verilmiştir. 1000 kere yapılan yeniden örnekleme sonucu elde edilen β değerlerinin histogramları da aşağıda verilmiştir.

Tablo 3. Takdir Değerleri ve Standart Hataları

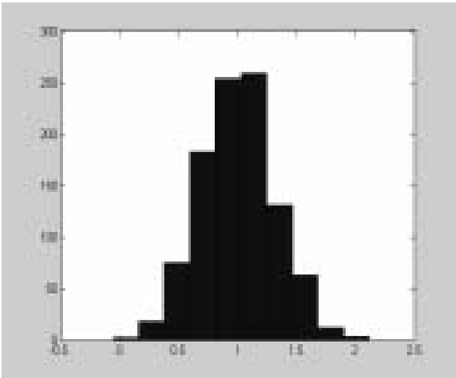
	En küçük Kareler Yöntemi		Ridge Regresyon Yöntemi		Bootstrap Yöntemi Sonuçları N=250		Bootstrap Yöntemi Sonuçları N=1000		Bootstrap Yöntemi Sonuçları N=2000	
	β	Std. Hata	β	Std Hata	β	Std Hata	β	Std Hata	Ortalama β	Std. Hata
(Sabit)	0.065	.019	0.0243	0.0016	0.9997	0.0011	1.0001	0.0005	0.9996	0.0004
X1	0.532	.226	0.0473	0.0058	1.0036	0.0144	1.0008	0.0069	0.9940	0.0049
X2	-0.091	.276	0.1506	0.1038	0.999	0.0209	1.0030	0.01	1.0122	0.0071
X3	0.073	.305	0.2286	0.2472	1.0141	0.0193	0.9868	0.0107	0.998	0.0075
X4	0.286	.167	0.4394	0.2481	0.9831	0.0164	1.01	0.0084	0.9973	0.0062



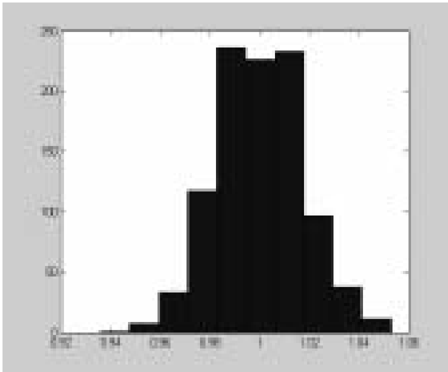
β_0 takdir değerlerinin dağılımı



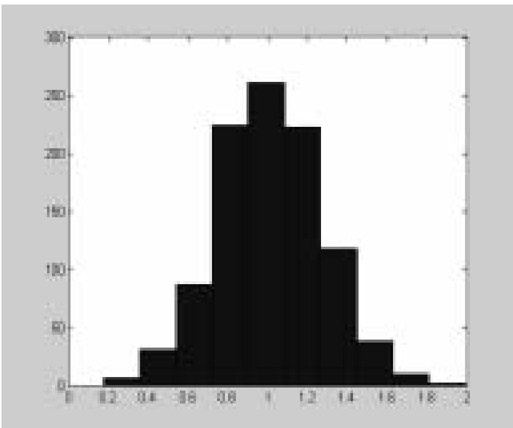
β_1 takdir değerlerinin dağılımı



β_2 takdir değerlerinin dağılımı



β_3 takdir değerlerinin dağılımı



β_4 takdir değerlerinin dağılımı

Şekil 2. *Yeniden Örnekleme Sonucu Elde Edilen Takdir Değerlerinin Dağılımı*

Sonuçlara bakıldığında en küçük kareler yöntemi ile bulunan katsayıların standart hatalarının ridge yöntemine göre takdir edilen değerlerin standart hatalarından daha büyük olduğu görülmektedir. Yeniden örnekleme yöntemi ile elde edilen takdir edicilerin standart hatalarının ise her iki yöntem ile elde edilen standart hatalardan daha küçük olduğunu göstermektedir. Yeniden örnekleme yönteminin sağladığı fayda bir örnek yerine binlerce örnek varmış gibi davranarak takdir değerlerinin dağılımlarını görmemizi sağlamaktır.

Aşağıdaki grafikler her bir takdir değerinin örneklem dağılımını göstermektedir.

6. SONUÇ

Uygulamada kullanılan veri setinde bağımsız değişkenler arasındaki korelasyon katsayılarının yüksek olduğu görülmektedir. Bağımlı değişken ise β katsayıları bir olacak şekilde türetilmiştir. En küçük kareler yöntemi, ridge regresyonu ve 250, 1000, 2000 kere yapılan yeniden örnekleme ile elde edilen β değerleri ve standart hataları Tablo 3'te verilmiştir. 1000 kere yapılan yeniden örnekleme sonucu elde edilen β değerlerinin histogramları da uygulama bölümünde sunulmuştur. En küçük kareler yöntemine göre yapılan tahmin değerlerinin parametre değerlerinden çok farklı olduğu ve standart hataların ise ridge regresyon analizi ile elde edilen sonuçlara göre yüksek olduğu görülmektedir. Ridge regresyon analizi ile elde edilen değerler parametre değerlerinden farklı olmasına rağmen standart hatalarının küçük olduğu görülmektedir. Bootstrap tekniği ile elde edilen değerlerde ise parametre değerlerine daha yakın sonuçlar elde edildiği belirlenmiştir. Bu da yeniden örnekleme tekniğinin diğer yöntemlere göre üstünlüğünü göstermektedir.

KAYNAKLAR

Berry W. D. (1993), Understanding Regression Assumptions, Sage publications.

Efron B. (1979), "Bootstrap Methods: Another Look at the Jackknife; *Annals of Statistics;* 7.

Efron B., Tibshirani, R. (1993), "An Introduction to the Bootstrap; Chapman&Hall; New York.

Fox, J. (1997), Applied Regression Analysis Linear Models and Related Methods, Sage Publication, USA.

Hoerl, A.E., Kennard R.W., (1970a), "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics,* Vol: 12-No:1.

Hoerl, A.E., Kennard R.W., (1970b), “Ridge Regression: Applications to Nonorthogonal Problems”, *Technometrics*, Vol: 12-1.

Hoerl, A.E., Kennard R.W., Baldwin K.F., (1975), “Ridge Regression: Some Simulation”, *Communications in Statistics*, 4.

Liu, Y. R, (1988), “Bootstrap Procedures Under Some Non-I.I.D. Models” Annals of Statistics, Vol. 16, No.4.

McDonald G.C., Galarneau D. I. (1975), “A Monte Carlo Evaluation of some Ridge Type Estimators”, *Journal of American Statistical Association*, 70-407.

Montgomery D.C., Peck E., (1992), Introduction to Linear Regression Analysis, John Wiley & Sons, s. 305-324.

Shao, J., Tu, D. (1995), The Jackknife and Bootstrap, Springer-Veriag; New York.

Stine, Robert, (1985), Bootstrap Prediction Intervals For Regression, *Journal of American Statistical Association*, Vol. 80, No: 392.

Vinod H.D., Ullah A. (1981), Recent advances in regression methods, Marcel Dekker.