İstatistik Araştırma Dergisi Cilt 04 No 01 Sayfa 35-41 Nisan 2005

ÇOKLU DOĞRUSAL BAĞLANTI DURUMUNDA RİDGE REGRESYON VE TEMEL BİLEŞENLER REGRESYON YÖNTEMLERİNİN BENZETİM ÇALIŞMASI İLE KARŞILAŞTIRILMASI

Neslihan ORTABAŞ* Serdar KURT**

ÖZET

Bu çalışmada, çoklu doğrusal regresyon modelinde, çoklu doğrusal bağlantı sorununu ortadan kaldırmak için kullanılan yöntemlerden, temel bileşenler regresyon ve ridge regresyon yöntemleri incelenmiştir.

Çoklu doğrusal regresyon modelinin varsayımlarından biri de bağımsız değişkenler arasında tam ilişki olmamasıdır. Bağımsız değişkenler arasında önemli derecede ilişki olması, çoklu doğrusal bağlantı olarak adlandırılır. Çoklu doğrusal bağlantı olması durumunda uygulanan en küçük kareler yöntemi ile parametre tahminleri büyük standart hatalara sahip olmakta ve hipotez testleri çelişkili sonuçlar vermektedir. Bu sorunu ortadan kaldırmak için kullanılan çeşitli yöntemler vardır. Bu yöntemlerden yanlı regresyon yöntemleri, hem çoklu doğrusal bağlantı yapısının açıklanabildiği hem de standart hatası daha küçük hata kareler ortalamalı tahminlerin bulunabildiği yöntemlerdir.

Çalışmada, yanlı regresyon yöntemlerinden temel bileşenler ile ridge regresyon yöntemi kuramsal açıdan incelenmiş, benzetim çalışması ile hangi yöntemin daha iyi sonuç verdiği araştırılmıştır.

Benzetim çalışmasında, genişlikleri 40, 80 ve 120 olan örneklemlerin her birisi için 50 tekrar yapılmış ve bu örneklemlere en küçük kareler, ridge ve temel bileşenler yöntemi uygulanarak regresyon katsayılarının tahminleri hesaplanmıştır. Tahmin ediciler arasında yapılan karşılaştırmalarda kriter olarak tahminlerin ortalaması ve standart hatası dikkate alınmıştır. Yapılan karşılaştırmalara göre, temel bileşenler regresyon yönteminin diğerlerinden daha iyi sonuçlar verdiği gözlemlenmiştir.

Anahtar Kelimeler : *Regresyon, Çoklu Doğrusal Bağlantı, Yanlı Regresyon Yöntemleri, Temel Bileşenler Regresyon, Ridge Regresyon*

* Dokuz Eylül Üniversitesi Fen-Edebiyat Fak. İstatistik Böl. Buca İZMİR
e-mail: neslihan.ortabas@deu.edu.tr. (Haberleşme adresi)

** Prof. Dr., Dokuz Eylül Üniversitesi Fen-Edebiyat Fak. İstatistik Böl. Buca İZMİR

1. GİRİŞ

Çoklu doğrusal regresyon modelinin varsayımlarından biri de bağımsız değişkenler arasında tam ilişki olmamasıdır. Bağımsız değişkenler arasında önemli derecede ilişki olması, çoklu doğrusal bağlantı olarak adlandırılır.

Çoklu doğrusal bağlantı sorunu; veri toplama yönteminden, kitledeki bir kısıtın örnekleme yansımasından, model seçiminden veya modeldeki açıklayıcı değişken sayısının, örneklem genişliğinden daha fazla olması gibi durumlardan kaynaklanabilir.

Çoklu doğrusal bağlantı sonucunda, en küçük karelerle elde ettiğimiz tahmincilerin standart hataları büyük çıkacağından, güven aralıkları genişler ve katsayılar önemsiz çıkar. Bunun yanında model geçerli ve R^2 belirtme katsayısı oldukça yüksek çıkar. Verilerdeki küçük değişimlerden en küçük kareler tahmincileri oldukça güçlü şekilde etkilenir.

Bir veri setinde, çoklu doğrusal bağlantı sorunu bulunduğunu; R^2'nin yüksek, katsayıların önemsiz, modelin geçerli çıkmasından anlaşılabileceği gibi, çiftli korelasyon, kısmi korelasyon katsayıları, yardımcı regresyon kriteri, özdeğerler, şartlı indeks, tolerans ve varyans şişirme faktörü gibi kriterlere bakarak da anlayabiliriz.

Çoklu doğrusal bağlantı sorununu ortadan kaldırmak için kullanılan çeşitli yöntemler aşağıda verilmektedir.

1. Ek veri toplama yöntemi
2. Modelin yeniden seçilmesi
3. Ridge regresyon
4. Genelleştirilmiş ridge regresyon
5. Temel bileşenler regresyonu
6. Gizli kök regresyon analizi

Bu yöntemlerden ridge regresyon, genelleştirilmiş ridge regresyon, temel bileşenler regresyonu ve gizli kök regresyon analizi yanlı regresyon yöntemleridir. Yanlı regresyon yöntemleri, hem çoklu doğrusal bağlantı yapısının açıklanabildiği hem de standart hatası daha küçük hata kareler ortalamalı tahminlerin bulunabildiği yöntemlerdir. Bu çalışmada, yanlı regresyon yöntemlerinden ridge regresyon ve temel bileşenler regresyon yöntemleri incelenmiştir.

Ridge regresyon yöntemi, $X'X$ matrisinin köşegen elemanlarına k $(0 \le k \le 1)$ gibi sabit bir değerinin eklenmesidir. Böylece ridge regresyon yöntemi ile katsayılar

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

formulü ile hesaplanır. Köşegen elemanlarına ilave edilen k değeri, sapmayı ifade edeceğinden seçimi çok önemlidir. k=0 olduğunda en küçük kareler yöntemi ile aynı sonucu verir. k'nın seçimi için çeşitli yöntemler önerilmiştir. Ridge izi, varyans şişirme

faktörü ve Hoerl, Kennard, Baldwin yöntemleri bunlardan birkaçıdır. (Montgomery ve Peck, 1992)

Temel bileşenler regresyon yönteminde model denklemine

$$Y = Z\alpha + \varepsilon$$

dönüşümü yapılır. Burada

$$Z = XT,\ \alpha = T'\beta,\ T'X'XT = Z'Z = \Lambda\ [\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)]$$

pxp boyutlu T matrisinin kolonları özvektörlerden oluşmaktadır. Z matrisinin kolonları

$$Z = [Z_1, Z_2, \ldots, Z_p]$$

şeklinde gösterilir ve temel bileşenler regresyon analizi sonucunda çıkan yeni regresyon katsayılarını verir. (Montgomery ve Peck, 1992)

Temel bileşenler regresyon analizinde özdeğerler küçükten büyüğe sıralanır $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Bu özdeğerlerden yaklaşık *l* tanesi sıfıra yakındır. Geriye kalan *p-l* tanesi ise sıfırdan büyük değerler alır. Sıfıra yakın olan özdeğerler için temel bileşenler modelden kaldırılır ve analize kalan bileşenlere en küçük kareler uygulanarak devam edilir. Temel bileşenler analizi sonucu regresyon katsayılarını

$$\hat{\beta}_{TB} = T\hat{\alpha}_{TB}$$

formülünü kullanarak elde edebiliriz. Burada

$$\hat{\alpha}_{TB} = B\hat{\alpha}$$

$b_1=b_2=\ldots=b_{p-l}=1$, $b_{p-l+1}=b_{p-l+2}=\ldots=b_p=0$'dır.

Bu çalışmada, düzenlenen benzetim çalışması ile yanlı regresyon yöntemlerinden ridge regresyon ve temel bileşenler regresyon yöntemleri kullanılarak regresyon katsayıları tahminlenmiş ve hangi yöntemin daha iyi sonuç verdiği araştırılmıştır.

2. KİTLENİN YARATILMASI VE BENZETİM ÇALIŞMASI

Minitab istatistiksel paket programı kullanılarak amacımıza uygun kitle türetildi. Kitlemizde X_1 ve X_2 olmak üzere iki bağımsız değişkenimiz bulunmaktadır. X_1 10,20,30,...,100 olmak üzere 10 farklı değer almakta, X_2'de X_1'in her değerine karşılık 4 farklı değer almaktadır. Her X_1 ve X_2 değerine, 25 tane Y değeri karşılık gelmekte ve Y ortalaması E(Y) ve varyansı $\sigma^2=25$ olan normal dağılıma uymaktadır. Türetilen bu kitle için korelasyon matrisine bakıldığında (Tablo 1) X_1 ve X_2 arasında oldukça güçlü bir ilişki olduğu görülür.

Tablo 1. Değişkenler Arasında Korelasyon Katsayıları

	Y	X_1
X_1	0.97632	0.97822
X_2	0.99530	

Kitle yaratıldıktan sonra sırasıyla aşağıdaki işlemler yapıldı.

1. n=40 birimlik örneklem kitleden çekildi.
2. Örnekleme sırasıyla en küçük kareler, ridge ve temel bileşenler regresyonu uygulandı.
3. 2. adıma geri dönüldü ve bu işlem 50 kez yapıldı.
4. 2.,3. ve 4. adımlar, 80 ve 120 birimlik örneklem genişlikleri için tekrar edildi.
5. Bu yöntemlerden elde edilen değerler karşılaştırılarak, hangi yöntemin en iyi sonucu verdiği araştırıldı.

3. BENZETİM ÇALIŞMASI İLE ELDE EDİLEN SONUÇLAR

Benzetim çalışmasında, genişlikleri 40, 80 ve 120 olan örneklemlerin her birisi için 50 tekrar yapılmış ve bu örneklemlere en küçük kareler, ridge ve temel bileşenler yöntemi uygulanarak regresyon katsayılarının tahminleri hesaplanmıştır. Tahmin ediciler arasında yapılan karşılaştırmalarda kriter olarak tahminlerin ortalaması ve standart hatası dikkate alınmıştır. Karşılaştırma tabloları üç farklı örneklem genişliği için ayrı ayrı aşağıda verilmiştir.

Tablo 2. n=40 için En Küçük Kareler, Ridge ve Temel Bileşenler Regresyon Katsayılarının Karşılaştırılması

	En Küçük Kareler		Ridge Regresyon		Temel Bileşenler Regresyonu	
i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	55.58887	1.80166	58.44952	1.72302	54.16945	1.40697
1	0.04858	0.28385	0.35609	0.01752	0.40423	0.01243
2	0.37816	0.14089	0.18514	0.01076	0.20117	0.00618

Tablo 3. n=80 için En Küçük Kareler, Ridge ve Temel Bileşenler Regresyon Katsayılarının Karşılaştırılması

	En Küçük Kareler		Ridge Regresyon		Temel Bileşenler Regresyonu	
i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	55.82650	1.27708	58.29188	1.08844	54.26610	1.04973
1	0.01337	0.16200	0.35776	0.00999	0.40436	0.00774
2	0.39580	0.08043	0.18695	0.00597	0.20123	0.00385

Tablo 4. n=120 için En Küçük Kareler, Ridge ve Temel Bileşenler Regresyon Katsayılarının Karşılaştırılması

	En Küçük Kareler		Ridge Regresyon		Temel Bileşenler Regresyonu	
i	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$	$\bar{\hat{\beta}}_i$	$S_{\hat{\beta}_i}$
0	56.31434	1.12874	58.87980	0.94752	54.69707	0.92746
1	-0.00511	0.15327	0.35223	0.00909	0.40012	0.00750
2	0.40079	0.07535	0.18401	0.00512	0.19912	0.00373

Tablo 2,3 ve 4'te de görüldüğü gibi yanlı regresyon yöntemlerinin standart hatası en küçük karelere göre daha düşüktür. Regresyon katsayıları açısından baktığımızda, ridge ve temel bileşenler regresyonu birbirine yakın sonuçlar vermektedir. Bu üç yöntem arasında temel bileşenler regresyonu en küçük standart hata değerlerini veren yöntemdir.

Artan örneklem genişlikleri için standart sapma değerleri daha da küçülmüştür. En küçük standart hata değerleri 120 birimlik örneklem genişliğinde temel bileşenler regresyonu yöntemiyle elde edilmiştir.

5. SONUÇ

Literatürde de belirtildiği gibi regresyon katsayılarının standart hataları, yanlı regresyon yöntemlerinde en küçük kareler yöntemine göre daha küçük bulunmuştur. Farklı örneklem genişlikleri için üç yöntemin karşılaştırıldığı tablolarda da bu durum görülebilmektedir.

Üç farklı örneklem genişliği için karşılaştırmalı tablolara bakıldığında, regresyon katsayılarının standart hatasını en küçük veren yöntem temel bileşenler regresyonudur. Temel bileşenler regresyonunu yakın bir farkla ridge regresyon takip etmektedir.

Sonuç olarak, farklı örneklem genişlikleri için en küçük standart hata değerlerini veren temel bileşenler regresyonu tercih edilebilir. Aynı zamanda, ridge regresyon yöntemi de temel bileşenler regresyonuna yakın sonuçlar vermektedir.

KAYNAKLAR

Aldrin, M. (1997), *Length Modified Ridge Regression,* Computational Statistics & Data Analysis, 25, 377-398.

Boneh, S. and Mendieta, G.R. (1994), *Variable Selection in Regression Models Using Principal Components,* Commun. Statist. - Theory Meth., 23(1), 197-213.

Gujarati, D.N. (1995). *Basic Econometrics,* (3rd ed.). New York, MCGraw-Hill, Inc.

Hoerl, A.E., Kennard R.W. and Baldwin K.F. (1975), *Ridge Regression:Some Simulations,* Communications in Statistics, 4(2), 105-123.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, Canada, John-Wiley&Sons, Inc.

Johnson, D.E. (1998), *Applied Multivariate Methods for Data Analysis*, California, Duxburry Press.

Johnson R.A. and Bhattacharyya G.K. (1992), *StatisticsPrinciples and Methods*, Canada, John-Wiley&Sons, Inc.

Lawless, J.F. and Wang, P. (1976), *A Simulation Study of Ridge and Other Regression Estimators*, Commun. Statis.-Theor. Meth., A5(4), 307-323

Mansfield, E.R., Webster, J.T. and Gunst, R.F. (1977), *An Analytic Variable Selection Technique for Principal Component Regression*, Appl. Statist., 36, 34-40

Marquardt, D.W. and Snee, R.D. (1975), *Ridge Regression in Practice*, Am. Statist. 29(1), 3-20

Mason, R.L. (1975), *Regression Analysis and Problems of Multicollinearity*, Communication Statististics, 4(3), 277-292

Montgomery D.C. and Peck E. (1992), *Introduction to Linear Regression Analysis,* Canada, John Wiley & Sons, Inc.

Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, California, Wadsworth&Brooks

THE COMPARISON OF RIDGE REGRESSION AND PRINCIPAL COMPONENTS REGRESSION METHODS IN THE PROBLEM OF MULTICOLLINEARITY BY SIMULATION

ABSTRACT

In this study, principal components regression and ridge regression are examined among the methods used to remedy multicollinearity problem in multiple linear regression model.

One of the assumptions in multiple linear regression is that there must be no perfect linear relations among the regressors. The relationship among the regressors is called multicollinearity. In case of multicollinearity, parameter estimations by least square method have large variances and hypothesis tests result in contradictory. There are various methods for dealing with multicollinearity problem. Biased regression methods (BRM) are the ones that can explain the

structure of multicollinearity and provide small standard errors among the methods used.

In this study two of biased regression methods; principal components regression and ridge regression are examined as theoretically and researched which methods give the best consequence by simulation.

In the application, 50 repetitions have been generated for each of the sample sizes of 40, 80 and 120. Least squares, ridge and principal components regression are used for each sample. Regression coefficients for each estimator were computed and the mean and the standard deviation of the estimates were used as statistical comparison criteria. According to comparisons among the estimators the principal components regression has been found to provide better estimates.

Key Words: *Regression, Multicollinearity, Biased Regression Methods, Principal Components Regression, Ridge Regression*