ISSN:1306-3111
e-Journal of New World Sciences Academy
2010, Volume: 5, Number: 4, Article Number: 3A0027

PHYSICAL SCIENCES
Received: August 2010
Accepted: October 2010
Series : 3A
ISSN : 1308-7304

Ayşe İşi
Gazi University
ayseisi@gazi.edu.tr
Ankara-Turkey

RİDGE VE LATENT ROOT REGRESYON TAHMİN EDİCİLERİNİN KOMBİNASYONU

ÖZET

Bu çalışmada En Küçük Kareler (EKK), Ridge Regresyon (RR) ve Latent Root Regresyon (LRR) tahmin edicilerinin kombinasyonu olan yeni bir tahmin edici sınıfı tanımlandı. Yeni tahmin edici sınıfının tanımlanabilmesi için öncelikle Latent Root Regresyonun tahmin denklemi matris notasyonuyla gösterildi. Tanımlanan yeni tahmin edici sınıfının geçerliliği, bir örnek üzerinde uygulandı.
Anahtar Kelimeler: Çoklubağlantı, Ridge Regresyon, Latent Root Regresyon, M-K Sınıflı Tahmin Ediciler, Hald’s Data

COMBINING RIDGE AND LATENT ROOT REGRESSION ESTIMATORS

ABSTRACT

In this study, a new class of estimator, which is the combining of the Ordinary Least Squares (OLS), the Ridge Regression (RR) and the Latent Root Regression (LRR) estimators, is introduced. In order to be able to identify the new estimator, first of all the estimation equation of Latent Root Regression was shown by matris notation. A new estimator is applied on the example.
Keywords: Multicollinearity, Ridge Regression, Latent Root Regression, M-K Class Estimators, Hald’s Data

1. GİRİŞ (INTRODUCTION)

Çoklu regresyon denkleminin yorumu, bağımsız değişkenlerin kuvvetli bir şekilde ilişkili olmaması varsayımına bağlıdır. Bu varsayımın bozulması, yani bağımsız değişkenler arasında bir ya da daha fazla sayıda doğrusal bağıntının olması, çoklubağlantı sorununu gündeme getirir (Alpar, 1997:254).

Çoklubağlantının giderilmesi için önerilen yöntemlerden biri, EKK yönteminden başka tahmin yöntemlerinin yani yanlı tahmin edicilerin kullanılmasıdır. Yanlı tahmin yöntemlerinde genel amaç, EKK yönteminde parametre tahminleri için büyük olan varyans alanını, küçük bir yan karşılığında daraltmaktır (Hoerl and Kennard, 1970,a:61).

Yanlı tahmin ediciler arasında en çok bilinen yöntem olan Ridge Regresyon, Hoerl ve Kennard tarafından (4,5)'te sunulmuş ve geliştirilmiştir. Webster, Gunst ve Mason (8)'de, Latent Root Regresyon üzerinde çalışmış, bu yöntemi, "Değiştirilmiş EKK Yöntemi" olarak tanımlamış ve EKK ile LRR yöntemlerini karşılaştırmışlardır.

Çoklubağlantı problemini gidermede alternatif yöntem olan yanlı tahmin edicilerin birleştirilmesi ile yeni tahmin edici sınıfları tanımlanmıştır. (2)'de Baye ve Parker, özel durumlarda EKK, RR ve Temel Bileşenler Regresyonu (TBR) tahmin edicilerini veren r-k sınıflı tahmin edicileri tanımlamıştır. (7)'de Kaçıranlar ve Sakallıoğlu, özel durumlarda EKK, TBR ve Liu tahmin edicisini veren r-d sınıflı tahmin edicileri tanımlamışlar ve yeni tahmin edici sınıfı ile EKK, TBR ve Liu tahmin edicilerini Hata Kareler Ortalaması değerlerine göre karşılaştırmışlardır.

2. ÇALIŞMANIN ÖNEMİ (RESEARCH SIGNIFICANCE)

Bu çalışmada, EKK, LRR ve RR tahmin edicilerinin kombinasyonu olan ve çoklubağlantı problemini gidermede alternatif bir yöntem olabilecek yeni bir tahmin edici sınıfı tanımlandı.

3. YENİ TAHMİN EDİCİ SINIFI (THE NEW CLASS OF ESTİMATOR)

3.1. Ridge Regresyon Tahmin Modeli (The Model of the Ridge Regression Estimator)

Bir doğrusal regresyon modeli,

$$Y = X\beta + \varepsilon \tag{1}$$

ile verilir. Bu doğrusal regresyon modelinde X, n×p (n>p) boyutlu standartlaştırılmış bağımsız değişkenler matrisi; Y, n×1 boyutlu, örnek ortalamasından sapması alınmış bağımlı değişken vektörü; β, p×1 boyutlu parametre vektörü; ε, $E(\varepsilon) = 0$ ve $Var(\varepsilon) = \sigma^2 I$ olmak üzere, n×1 boyutlu hata vektörüdür.

Doğrusal regresyon modelinde parametre vektörünün EKK tahmin edicisi,

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

olarak tanımlanır.

Ridge Regresyon yönteminde $X'X$, korelasyon matrisi olmak üzere, $X'X$ matrisinin köşegen ögelerine küçük k değerlerinin (genelde $0 \le k \le 1$) eklenmesiyle Ridge Regresyon tahmin denklemi elde edilir. $\tilde{\beta}_R = b(k)$, Ridge Regresyonun tahmin edicisi olmak üzere aşağıdaki denklemle tanımlanır (Hoerl and Kennard, 1970,a:57).

$$\tilde{\beta}_R = b(k) = (X'X + kI)^{-1} X'Y, \quad k \ge 0 \tag{3}$$

Burada k=0 değeri için denklem, EKK tahmin edicisini verir. $b(k=0) = \hat{\beta}$
Ridge Regresyonun kestirim değerleri ise aşağıdaki denklem ile bulunur.

$$\tilde{Y} = X\tilde{\beta}_R \tag{4}$$

3.2. Latent Root Regresyon Tahmin Modeli (The Model of the Latent Root Regression Estimator)

Latent Root Regresyon, genel yapı itibariye TBR'ye benzemektedir. Latent Root Regresyon yöntemi, TBR'deki $X'X$'in özdeğerleri ve özvektörleri gibi, bağımlı ve bağımsız değişkenlerin korelasyon matrisinin latent root ve latent vektörlerinden yaralanılarak uygulanır.

Y_i^* ve X_i^*, değişkenlerin gözlenen değerleri olmak üzere bağımlı ve bağımsız değişkenler, "birim uzunluk dönüşümü" yöntemiyle standartlaştırılır.

$$X_i = \frac{X_i^* - \overline{X}_i^*}{\left(\sum_{i=1}^{n}(X_i^* - \overline{X}_i^*)^2\right)^{1/2}} \text{ ve } Y_i = \frac{Y_i^* - \overline{Y}_i^*}{\left(\sum_{i=1}^{n}(Y_i^* - \overline{Y}_i^*)^2\right)^{1/2}}, \quad \eta = \left(\sum_{i=1}^{n}(Y_i^* - \overline{Y}_i^*)^2\right)^{1/2} \tag{5}$$

$A = [Y : X]$, n×(p+1) boyutlu standartlaştırılmış bağımlı ve bağımsız değişkenler matrisi olmak üzere; $A'A$, bağımlı ve bağımsız değişkenlerin korelasyon matrisidir ve $|A'A - \lambda_j I| = 0$ denklemiyle tanımlanan latent vektörlere sahiptir. j. latent vektörün elemanları $\gamma_j' = (\gamma_{0j}\ \gamma_{1j} \cdots \gamma_{pj})$ ile tanımlanır. γ_j^0, γ_j'nin ilk sütunu dışındaki bütün elemanlarını içermek üzere, $\gamma_j^{0'} = (\gamma_{1j}\ \gamma_{2j} \cdots \gamma_{pj})$ olarak tanımlanır. Ayrıca $\lambda_j^* = \dfrac{\lambda_j}{\gamma_{0j}^2}$ olmak üzere; değişken elemesi yapıldıktan sonra elde edilen LRR parametre tahminleri aşağıdaki denklemle ifade edilir (Webster, Mason and Gunst, 1974:515).

$$\tilde{\beta}_{LR} = b_m = -\eta\left(\sum_{l=m}^{k}\lambda_l^{*-1}\right)^{-1}\sum_{j=m}^{p}\gamma_{0j}\lambda_j^{-1}\gamma_j^0 \tag{6}$$

(6) denklemi ile verilen ifade, her bir parametre tahmin değerini tek tek bulmaya yöneliktir. Ancak, m-k sınıflı tahmin edicileri tanımlamak için LRR tahmin edicisinin matris notasyonuyla tanımlanması gerekmektedir. Bu nedenle LRR'nin TBR'ye benzerliği dikkate alınarak ve Baye and Parker tarafından (2)'de tanımlanan r-k sınıflı tahmin edicilerin prosedürü izlenerek matris notasyonuna dönüştürme işlemi uygulandı.

LRR yöntemi için, (1) denklemiyle belirtilen $Y = X\beta + \varepsilon$ genel doğrusal regresyon modeli ve (2) denklemiyle belirtilen $\hat{\beta} = (X'X)^{-1}X'Y$ EKK tahmin edicisi kullanılır.

$A = [Y : X]$, n×(p+1) boyutlu standartlaştırılmış bağımlı ve bağımsız değişkenler matrisi; Γ, (p+1)×(p+1) boyutlu latent vektörler matrisi; W, $W = A\Gamma$ olarak tanımlanan n×(p+1) boyutlu doğrusal bileşenler matrisi; β, (p+1) ×1 boyutlu parametre vektörü; ε, $E(\varepsilon) = 0$ ve $Var(\varepsilon) = \sigma^2 I$ olmak üzere, n×1 boyutlu hata vektörü olmak üzere,

LRR'nin tahmin edicisi $\tilde{\beta}_{LR}$, aşağıdaki denklemle tanımlanır.

$$\tilde{\beta}_{LR} = (W'W)^{-1}W'Y \tag{7}$$

$W = A\Gamma$ eşitliği gözönüne alınarak,

$$\tilde{\beta}_{LR} = (\Gamma'A'A\Gamma)^{-1}\Gamma'A'Y \tag{8}$$

denklemi elde edilir. Burada, $\Gamma' A' A \Gamma$ ifadesi Λ olarak tanımlanır ve bu matris, köşegen elemanları λ_j'ler olan köşegen bir matristir. $\Lambda = \Gamma' A' A \Gamma = diag(\lambda_0, \lambda_1, \cdots, \lambda_p)$.

$W = A\Gamma$, $\alpha = \Gamma' \beta$ ve α'nın EKK tahmini, $\hat{\alpha} = (W'W)^{-1} W'Y$ veya $\hat{\alpha} = \Lambda^{-1} W'Y$ olmak üzere standartlaştırılmış parametre tahminleri aşağıdaki denklemle elde edilir.

$$\begin{aligned}\tilde{\beta}_{LR(s)} &= \Gamma \alpha \\ &= \Gamma \Lambda^{-1} W'Y \\ &= \Gamma (\Gamma' A' A \Gamma)^{-1} \Gamma' A' Y\end{aligned} \qquad (9)$$

Γ_m, Γ'nın (p+1-m) sütunu çıkarıldıktan sonra $A'A$'nın değişmeyen latent vektörleri olmak üzere, $\Lambda_m = \Gamma'_m A' A \Gamma_m$ olur. Bu durumda standartlaştırılmış regresyon katsayıları da, değişken elemesi yapıldıktan sonra (9) denklemi yeniden yazılarak,

$$\tilde{\beta}_{LR(s)} = \Gamma_m (\Gamma'_m A' A \Gamma_m)^{-1} \Gamma'_m A' Y \qquad (10)$$

denklemi elde edilir. Bu eşitlik daha sonra m-k sınıflı tahmin edicileri tanımlamak için kullanılacaktır.
Bu tanımlamalardan sonra, (8) denklemi ile verilen genel parametre tahminleri de değişken elemesi yapıldıktan sonra yeniden düzenlenirse,

$$\tilde{\beta}_{LR} = b_m = (\Gamma'_m A' A \Gamma_m)^{-1} \Gamma'_m A' Y, \qquad m \le p+1 \qquad (11)$$

olur.
Latent Root Regresyonun kestirim değerleri ise aşağıdaki denklem ile bulunur.

$$\tilde{Y} = W \tilde{\beta}_{LR} \qquad (12)$$

$$\tilde{Y} = \bar{Y} + \eta (b_1 W_1 + \cdots + b_m W_m) \qquad (13)$$

$W = A\Gamma$ eşitliği göz önüne alınarak W_j'ler aşağıdaki şekilde hesaplanır.

$$W_j = \Gamma_{1j} Y + \Gamma_{2j} X_1 + \cdots + \Gamma_{mj} X_{m-1} \qquad (14)$$

Yukarıdaki denklem yardımıyla standartlaştırılmış değişkenlere dönüşüm yapılmış olur.
Böylece LRR tahmin ve kestirim denklemlerinin matris notasyonuyla gösterimi tamamlanmış olur.

3.3. m-k Sınıflı Tahmin Edicilerin Tanımlanması (The Introduction of m-k Class Estimator)

Bölüm 3.1 ve Bölüm 3.2'de verilen açıklamalardan sonra m-k sınıflı tahmin ediciler tanımlanabilir.

Burada m, LRR parametresi; k, RR parametresi olmak üzere; m-k sınıflı tahmin ediciler, sırasıyla (3) ve (10) denklemleriyle verilen Ridge Regresyon ve Latent Root Regresyon tahmin edicilerinin uygun bir kombinasyonu olmakta ve $b_m(k)$, m-k sınıflı tahmin edicileri simgelemek üzere aşağıdaki denklemle tanımlanmaktadır.

$$b_m(k) = \Gamma_m (\Gamma'_m A' A \Gamma_m + k I_m)^{-1} \Gamma'_m A' Y \,, \quad m \le p+1 \; ve \; 0 \le k \le 1 \qquad (15)$$

Burada k=0 değeri için denklem, LRR tahmin edicisini verir.

$$b_m(k=0) = \tilde{\beta}_{LR} \qquad (16)$$

4. UYGULAMA (APPLICATION)

m-k sınıflı tahmin edicilerin geçerliliği Hald's Data (Draper and Smith,1984:365-366) üzerinde uygulandı. Örnek, SPSS paket programının Syntax editöründe m-k sınıflı tahmin edicilerin programı yazılarak çözümlendi.

Bağımlı ve bağımsız değişkenler birim uzunluk dönüşümü yöntemiyle standartlaştırıldıktan sonra birleştirilmiş matris olan A matrisi oluşturularak işlemlere başlanır. Korelasyon matrisi olan $A'A$ matrisine ilişkin latent rootlar sırasıyla; $\lambda_4 = 3{,}2116$, $\lambda_3 = 1{,}5761$, $\lambda_2 = 0{,}1990$, $\lambda_1 = 0{,}0117$, $\lambda_0 = 0{,}0016$ olarak bulunur. Bu latent rootlara ilişkin latent vektörler de aşağıda verilmiştir.

	λ_4	λ_3	λ_2	λ_1	λ_0
γ_{0j}	0,5534	-0,0034	0,2112	-0,8047	0,0408
γ_{1j}	0,4012	0,5125	0,5809	0,4129	-0,2617
γ_{2j}	0,4682	-0,4096	-0,3899	0,1884	-0,6523
γ_{3j}	-0,3189	-0,608	0,6747	-0,0531	-0,2657
γ_{4j}	-0,4603	0,4471	-0,1039	-0,3791	-0,6586

Latent root ve latent vektörler belirlendikten sonra değişken elemesi yapılır. Değişken elemesi için kullanılacak yöntemlerden biri, geriye doğru seçim yöntemidir. Diğer bir yaklaşım da λ_j ve γ_{0j}'nin değerlerin büyüklüklerine göre eleme yapılmasıdır. Sözkonusu değişkenin modele alınabilmesi için, $\lambda_j > 0{,}05$ ve $\gamma_{0j} > 0{,}10$ olmalıdır (Webster, Gunst and Mason, 1974:518). Bu şartların birlikte sağlandığı latent vektörler işleme alınır. Bu yaklaşım, geriye doğru eleme yönteminden uygulama açısından daha kolay olduğundan örneğin çözümünde bu yaklaşım tercih edilmiştir.

$\lambda_j \le 0{,}05$ ve $\gamma_{0j} \le 0{,}10$ koşulunu sağlayan j. latent vektör işlemden çıkarılır. Yukarıdaki latent vektörler incelendiğinde bu koşulu, $\lambda_0 = 0{,}0016$ ve $\gamma_{0j} = 0{,}0408$ sağladığı için λ_0'a ilişkin sütun işlemden çıkarılır. Böylece LRR parametresi olan m'nin değeri belirlenmiş olur.

Geriye kalan m=4 tane latent vektörle yapılan işlemler sonucu, RR parametresi olan k'nın farklı değerlerine göre hesaplanan parametre tahmin değerleri ve hata kareler toplamları aşağıdaki tabloda verilmiştir.

Tablo 1. m=4 için artan k değerlerine göre parametre tahmin değerleri ve hata kareler toplamları
(Table 1. The parameters' estimations and sum of squares of error according to increasing k values for m=4)

k	$b_4(k)_0$	$b_4(k)_1$	$b_4(k)_2$	$b_4(k)_3$	$b_4(k)_4$	$\sum e_i^2$
0	0,99834	0,01067	0,02658	0,01083	0,02684	0,00721
0,02	0,58377	0,20779	0,12814	-0,02809	-0,16206	8,49759
0,04	0,48609	0,24455	0,15444	-0,04397	-0,20236	13,39614
0,06	0,44052	0,25631	0,16772	-0,05479	-0,21864	16,59811
0,08	0,41320	0,26007	0,17608	-0,06313	-0,22674	19,21579
0,1	0,39449	0,26051	0,18193	-0,06989	-0,23114	21,61775
0,2	0,34632	0,25026	0,19572	-0,09086	-0,23524	33,59236
0,3	0,32218	0,23802	0,19962	-0,10123	-0,23155	47,25816
0,5	0,29250	0,21810	0,19836	-0,10958	-0,22085	80,44261
1	0,28443	0,18474	0,18302	-0,10916	-0,19529	187,4492

Tablo 1. incelendiğinde, parametre tahmin değerlerinin artan k değerlerine göre büyük değişiklikler göstermediği; buna karşın, k değerleri arttıkça hata kareler toplamlarının da arttığı görülmektedir. Burada önemli olan en uygun k değerinin belirlenmesidir. k parametresinin belirlenmesinde nokta kestirim yaklaşımı kullanıldığında elde edilen değer $k \leq 0{,}031$ olarak hesaplanır. Ancak, çok daha iyi kestirimler elde etmek için k'nın sıfıra yakın bir değer seçilmesi daha uygun olacağı için k=0,02 olarak belirlenir.

Elde edilen verilerin karşılaştırılması için aynı örnek, r-k sınıflı tahmin ediciler ile çözümlendiğinde r=3 değeri için artan k değerlerine göre hesaplanan hata kareler toplamları aşağıdaki çizelgede verilmiştir.

Tablo 2. r=3 için artan k değerlerine göre hata kareler toplamları
(Table 2. The sum of squares of error according to increasing k values for r=3)

k	$\sum e_i^2$
0	48,5456
0,02	49,1923
0,04	50,8204
0,06	53,1191
0,08	55,9069
0,1	59,0732
0,2	78,7983

Yukarıdaki hata kareler toplamlarına bakıldığında, k=0,02 için hesaplanan hata kareler toplamı m-k sınıflı tahmin edicilerde 8,49759 iken r-k sınıflı tahmin edicilerde 49,1923 olmaktadır. Hatta m-k sınıflı tahmin edicilerde k=1 için hesaplanan hata kareler toplamı, r-k sınıflı tahmin edicilerde k'nın 0,3'ten küçük olduğu durumlarda gözlemlenmektedir. Bu durumda, bu veriler için, m-k sınıflı tahmin edicilerin, r-k sınıflı tahmin edicilerden çok daha küçük hata kareler toplamlarına sahip olduğu ve dolayısıyla daha iyi kestirim değerleri verdiği söylenebilir.

5. SONUÇLAR (CONCLUSIONS)

Bu çalışmada çoklubağlantıyı gidermek için alternatif bir yöntem olan m-k sınıflı tahmin ediciler; EKK, RR ve LRR tahmin edicilerinin uygun bir kombinasyonu olarak tanımlandı. Tanımlanan yeni tahmin edici sınıfının geçerliliği bir örnek üzerinde denendi. Belirlenen m parametresinde k değerleri arttıkça hata kareler toplamlarının da arttığı görüldü. Ancak bu artışların, r-k sınıflı tahmin edicilerle kıyaslandığında çok küçük artışlar olduğu ve çoklubağlantıyı gidermede alternatif bir yöntem olarak tanımlanan m-k sınıflı tahmin edicilerin daha iyi kestirim değerleri verdiği değerlendirildi.

KAYNAKLAR (REFERENCES)

1. Alpar, R., (1997). Uygulamalı Çok Değişkenli Yöntemlere Giriş-1. Ankara:Bağıran Yayınevi.
2. Baye, M.R. and Parker, D.F., (1984). Combining Ridge and Principal Component Regression: A Money Demand Illustration, Communications in Statistics-Theory and Methods, 13(2), pp:197-205.
3. Draper, N.R. and Smith, H., (1981). Applied Regression Analysis-Second Edition, Wiley Series in Probability and Math. Stat., NewYork.
4. Hoerl, A.E. and Kennard, R.W., (1970,a). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, pp:55-67.
5. Hoerl, A.E. and Kennard, R.W., (1970,b). Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12, pp:69-82.
6. İşi, A., (2002). Yanlı Tahmin Ediciler ve Kombinasyonları. Yüksek Lisans Tezi. Ankara: Gazi Üniversitesi Fen Bilimleri Enstitüsü.

7. Kaçıranlar, S. and Sakallıoğlu, S., (2001). Combining The Liu Estimator and the Principal Component Regression Estimator. Communications in Statistics-Theory and Methods, 30(12), pp:2699-2705.

8. Webster, J.T., Gunst, R.F., and Mason, R.L., (1974). Latent Root Regression Analysis. Technometrics, 16(4), pp:513-522.