İSTANBUL ÜNİVERSİTESİ
İKTİSAT FAKÜLTESİ
EKONOMETRİ VE İSTATİSTİK
DERGİSİ

Ekonometri ve İstatistik Sayı:15 2011 38-52

KISMİ EN KÜÇÜK KARELER REGRESYONU YARDIMIYLA OPTİMUM BİLEŞEN SAYISINI SEÇMEDE MODEL SEÇME KRİTERLERİNİN PERFORMANS KARŞILAŞTIRMASI

Elif BULUT* **Özlem GÜRÜNLÜ ALMA****

Özet

Regresyon modellerinin çok sayıda açıklayıcı değişkene sahip olması, gözlem sayısının açıklayıcı değişken sayısından daha az olması ve açıklayıcı değişkenler arasında çoklu doğrusal bağlantı probleminin varlığı gibi durumlar, regresyon analizindeki problemlerden bazılarıdır. Bu problemler en küçük kareler yöntemi varsayımlarını bozmaktadır. Kısmi en küçük kareler regresyonu (KEKKR), bu varsayımların bozulduğu durumlarda regresyon analizi yapmaya olanak sağlayan: kısmi en küçük kareler (KEKK) ve çoklu doğrusal regresyon yöntemlerinden oluşan çok değişkenli istatistiksel bir metottur. Bu çalışmada, çoklu doğrusal bağlantı probleminin olduğu veri setlerinde KEKKR tarafından elde edilen gizli değişkenler ile model kurulup, gizli değişkenlerin optimum sayısını saptamak için ise MAIC (Bedrick & Tsai, 1994), MAIC (Bozdogan,2000), MA_opt(PRESS) ve Wold's R model seçme kriterleri kullanılmıştır. Model seçme kriterlerinin optimum sayıda gizli değişkeni bulma performanslarını karşılaştırmak amacıyla k-çapraz geçerlilikte benzetim çalışması yapılmıştır. Benzetim çalışması sonucunda; kriterlerin küçük boyutlu veri setlerinde doğru bir şekilde gizli değişken sayısını bulduğu fakat veri setlerinin boyutu arttıkça kriterlerin optimum sayıdan daha fazla sayıda gizli değişken seçme eğiliminde oldukları görülmüştür. Ayrıca, M_{AKAKIE} ve $M_{BEDRICK}$ kriterlerinin hemen hemen aynı sonuçları bulmakta olduğu fakat regresyon modellerinin boyutu büyütüldüğünde optimum sayıda gizli değişkenleri bulamadığı saptanmıştır. MA_opt(PRESS) kriteri ve Wold's R kriteri yaklaşık olarak aynı sonuçları vermekte olup diğer kriterlere göre daha doğru iyi bir performansla optimum sayıda gizli bileşenleri bulmaktadırlar.
***Anahtar Kelimeler**:* Parasal Kısmi En Küçük Kareler, MAIC (Çok Değişkenli Akaike Bilgi Kriteri), PRESS (Tahmin Edilen Artık Kareler Toplamı), Wold's R.
***Jel Sınıflaması**:* C15, C30, C49, C63

Abstract

Partial Least Squares Regression (PLSR) is a multivariate statistical method for constructing predictive models when the variables are many and highly collinear. Its goal is to predict a set of response variables from a set of predictor variables. This prediction is achieved by extracting a set of orthogonal factors called latent variables from the predictor variables. This study investigated the performances of model selection criteria in selecting the optimum number of latent variables from PLSR models for data sets that have various observations and variable numbers. Their performances have been compared in a simulation study with k-fold cross validation. This simulation has been performed to compare the performance of MAIC (Bedrick & Tsai, 1994), MAIC (Bozdogan, 2000), MA_opt(PRESS) and Wold's R criterion in finding the optimum number of latent variables. The simulation results show that all the criteria achieved the optimum number of latent variables for a small-sized design matrix. But when the data dimensions get bigger, M_{AKAKIE} and $M_{BEDRICK}$ could not find the optimum number of latent variables. MA_opt(PRESS) and Wold's R criteria gave almost the same results and found the optimum number of latent variables with a better performance than the MAIC's.

Keywords: Partial Least Squares, MAIC (Multivariate Akaike Information Criterion), PRESS (Predicted Residual Sum of Squares), Wold's R.
Jel Classification: C15, C30, C49, C63

* Yrd. Doç. Dr., Ondokuz Mayıs Üniversitesi, İktisadi ve İdari Bilimler Fak. İşletme Bölümü, Kurupelit Kampüsü/SAMSUN, E-Mail: bulut.elif@hotmail.com (Sorumlu Yazar)
** Yrd. Doç. Dr., Muğla Üniversitesi, Fen Fakültesi İstatistik Bölümü / MUĞLA, E-Mail: ozlem.gurunlu@hotmail.com

1. GİRİŞ

KEKKR, çok sayıda bağımlı ve açıklayıcı değişkenle çalışma olanağı sağlayan çok değişkenli istatistiksel bir metottur. KEKKR' nin gelişimi ile ilgili çok sayıda çalışma yapılmıştır. Bu konuda yapılmış en kapsamlı çalışmalardan biri Geladi ve Kowalski tarafından 1986 yılında yayımlanmıştır. Höskuldsson (1988) ve Geladi (1988), KEKKR nin tarihsel gelişimini gözler önüne sererek KEKKR yi istatistiksel bir çerçevede yorumlamıştır Yakın zamanlı çalışmalar ise Helland (1990), Garthwaite (1994), Wold vd. (2001), Tobias (2003) ve Abdi (2007) tarafından verilmiştir.

KEKKR; KEKK ve çoklu doğrusal regresyon yöntemlerinin her ikisini de içeren bir metottur. KEKKR algoritmalarında gizli değişken hesaplama ve regresyon adımı bütünleşmiş şekildedir. Gizli değişkenler tekil değer veya özdeğer ayrıştırması kullanılarak boyut indirgeme ile elde edilmektedir. Boyut indirgemeden sonra gizli değişkenler, yeni açıklayıcı değişkenler olarak, regresyon analizinde kullanılmaktadır. Gizli değişkenler, açıklayıcı değişkenlerin doğrusal bir birleşimi olup aralarında doğrusal bağlantıya sahip değildir.

Kısmi en küçük kareler yönteminde boyut indirgemeyi takiben, bağımlı değişkendeki değişimi açıklamada en etkili olan gizli değişkenleri saptamada bazı model seçme kriterleri kullanılmaktadır. Bu kriterlerden bazıları PRESS, Wold's R ve Akaike bilgi kriteri ve BIC olarak sayılabilir.

Bu çalışmada Li vd. (2002) çalışması temel alınmış; farklı sayıda gözlem sayısının ve açıklayıcı değişken sayısının olduğu durumlar için bir benzetim çalışması gerçekleştirilerek, çalışmaları genişletilmiştir. Çalışmalarında yer alan model seçme kriterleriyle birlikte; Bozdoğan tarafından geliştirilen çok değişkenli Akaike bilgi kriteri ve MA_opt (PRESS) model seçme kriterleri de incelenmiştir. Burada amaç, KEKKR varsayımlarına göre türetilen

verilere model seçme kriterlerinin uygulanması ve doğru sayıda gizli değişken sayısının bulunmasında kriterlerin performanslarının karşılaştırmasıdır. Çalışmada MAIC, MA_opt (PRESS) ve Wold's R kriterlerinin performansları k çapraz geçerlilikte incelenmiştir. Tüm performans karşılaştırmaları Matlab ortamında yazılan program aracılığıyla yapılmıştır.

Çalışma şu bölümlerden oluşmaktadır. Bölüm 2'de, KEKKR modeli ve benzetim çalışmasında kullanılan KEKKR algoritması hakkında kısaca bilgi verilmektedir. Bölüm 3 ve 4' de, model seçme kriterleri hakkında özet bir bilgi verilerek, gözlem sayısının ve açıklayıcı değişken sayısının farklı değerleri için verilerin nasıl türetildiği anlatılmaktadır. Benzetim çalışmasının sonuçları ve yorumları Bölüm 5 ve 6 da yer almaktadır.

2. KISMİ EN KÜÇÜK KARELER REGRESYONU

KEKKR de açıklayıcı ve bağımlı değişken için ayrı ayrı modelleme yapılmaktadır.

$$\mathbf{X} = \sum_{j=1}^{A} \mathbf{t}_j \mathbf{p}_j' + \mathbf{E} \text{ ve } \mathbf{Y} = \sum_{j=1}^{A} \mathbf{u}_j \mathbf{q}_j' + \mathbf{F} \tag{1}$$

Burada $\mathbf{t}_j$,, $\mathbf{u}_j$ ler gizli değişkenler olup, $\mathbf{t}_j$ ler birbirine ve aynı zamanda sonraki $\mathbf{u}_j$ ye diktir. KEKKR model kurmak için optimum sayıda gizli değişken sayısını, açıklayıcı ve bağımlı değişkenler arasındaki kovaryansı maksimum yapacak şekilde elde eder (Carrascal,vd. 2009:682; Höskuldsson, 1988:213). KEKKR metodunda açıklayıcı ve bağımlı değişkenlerin her biri bir değişken bloğu olarak göz önüne alınır ve sonrasında birbirine dik, karşılıklı bağımsız ve orijinal açıklayıcı değişkenlerin doğrusal kombinasyonları olacak şekilde gizli değişkenler elde edilir.Elde edilen bu gizli değişkenler yeni açıklayıcı değişkenler olarak bağımlı değişkeni tahminleme de kullanılır. KEKKR; bağımlı değişken kümesi, $\mathbf{Y}_{N\times K}$ ve açıklayıcı değişkenler kümesi, $\mathbf{X}_{N\times M}$ arasındaki ilişkiyi doğrusal modelleyen bir metottur. Burada N: gözlem sayısını, K: bağımlı değişken sayısını ve M: açıklayıcı değişken sayısını göstermektedir (Lindgren, vd. 1993). KEKKR algoritmaları yardımıyla M tane açıklayıcı değişken daha az sayıda gizli değişkene indirgenmektedir. En çok kullanılan KEKKR algoritmaları: NIPALS, UNIPALS, KERNEL, SAMPLS ve SIMPLS'dır. İlk çalışmalar NIPALS algoritması ile yapılmış olup diğer algoritmalar bu

algoritmadan geliştirilmiştir. SIMPLS algoritması De Jong (1990), KERNEL algoritmaları ise farklı versiyonları olacak şekilde; Lingren vd. (1990), De jong vd. (1994) ve Rannar vd. (1994) tarafından çalışılmıştır. Algoritmanın seçimi çoğunlukla çalışılacak veri matrisinin şekline dayanmaktadır. Bazı çalışmalarda gözlem sayısı değişken sayısından daha büyüktür. Böyle bir durumda varyans-kovaryans matrisi, gözlem sayısından bağımsız olacak şekilde bir algoritma ile çalışmada kolaylık sağlar. Tersi bir durum için ise, açıklayıcı değişken sayısının gözlem sayısından daha fazla olduğu durumda, açıklayıcı değişken sayısından bağımsız bir matrisle çalışacak bir algoritma seçmek en iyi seçenek olacaktır (Lindgren, vd. 1998:107). Bu çalışmada, Dayal ve MacGregor (1997) tarafından geliştirilen Modified Kernel#2 algoritması üzerinde çalışılmıştır. Bu algoritma, gözlem sayısından bağımsız olacak şekilde $M \times M$ boyutlu $X'YY'X$ kernel matrisi ile çalışmaktadır. Çalışmamızda, farklı sayılarda gözlem ve açıklayıcı değişkenlerle çalışılmış olmasına rağmen, veri matrislerinin her durumunda gözlem sayısı açıklayıcı değişken sayısından büyük olduğu için (N>M), bu algoritma tercih edilmiştir. Modified Kernel#2 algoritması $\mathbf{X}$ ve $\mathbf{Y}$ orijinal veri matrislerine müdahale etmeden $\mathbf{X'Y}$ varyans-kovaryans matrisini $M \times M$ boyutlu $(\mathbf{I} - \mathbf{wp}')$ matrisi ile çarparak güncellemektedir. Bu güncelleme matrisinde $\mathbf{I}$ birim matrisi, $\mathbf{w}$ ve $\mathbf{p}$ ise sırasıyla açıklayıcı değişkenin değişkene katkısı olan ağırlık vektörünü ve değişkenin açıklayıcı değişken üzerindeki etkisini ifade eden yük vektörünü göstermektedir.

3. MODEL SEÇME KRİTERLERİ

Model seçme ve geçerlilik, regresyon modelinin tahminleme gücünü etkileyen önemli bileşenlerdir. İyi bir modelin kurulabilmesi tamamen doğru değişkenlerin seçimine dayanmaktadır. Böylece, model tahmin hatası minimuma inecek ve model istenmeyen değişkenlerden korunmuş olacaktır. Regresyon modelinde yer alacak değişken sayısına karar vermede ve doğrusal regresyon modelinin tahmin geçerliliğini incelemede genellikle çapraz-geçerlilik kullanılır (Stone and Brooks 1990:238; Wold, vd.1984:737). Kavramsal olarak anlaşılması kolay fakat modeli en iyileme de yoğun hesaplamalara sahip bir metottur. Bunun yanı sıra, çapraz geçerlilik bir modelin doğruluğunu tahminleme de sıkça kullanılan ve mevcut örneklemin deneme ve geçerlilik seti olarak bölmeye dayanan bir yöntemdir (Last, 2006:275). Tahminlenen ve gözlemenen değerler arasındaki farkın kareler toplamı o model için PRESS değerini vermektedir. PRESS kriteri bir model için tahminlenen değerin

açıklayıcı değişkenin gözlemlenen bağımlı değişkeni nasıl tahminlediğini gösteren bir kriterdir. i. nci gözlem için PRESS değeri, $\text{PRESS} = \sum_{i=1}^{N} \left(y_i - \hat{y}_{i(i)} \right)^2$ eşitliği ile elde edilir (Neter, vd., 1996). $\hat{y}_{i(i)}$, tahminlenen değeri göstermektedir. İlk alt indis i.nci durum için tahminlenen değeri ve ikinci alt indis (i) ise regresyon fonksiyonu uygulanırken i. nci durumun ihmal edildiğini göstermektedir. PRESS değerinin küçük olması, o modelin tahminleme için iyi bir model olduğunu göstermektedir. Çalışmada kullanılan MA_opt(PRESS) olarak isimlendirilen kriter birden fazla bağımlı değişken için PRESS değerlerinin normu alınarak hesaplanmaktadır. PRESS değerleri kullanılarak bazı model seçme kriterlerini hesaplamak mümkündür. Bunlardan bir tanesi de bu çalışmada kullanılan Wold'S R kriteridir ve aşağıdaki şekilde hesaplanmaktadır:

$$\text{Wold's R} = \frac{\text{PRESS}(a+1)}{\text{PRESS}(a)}, \tag{2}$$

burada a gizli değişken sayısını göstermektedir. Bu çalışmada kullanılan diğer bir kriter ise Akaike bilgi kriteridir. Hirotsugu Akaike tarafından AIC bilgi kriteri ismi altında 1971 yılında geliştirilmiştir. Tahmin edilen istatistiksel modelin bir uyum iyiliği ölçüsü olup, olası mevcut modeller kümesinden uygun modelin seçimini sağlar. Modeller içerisinden en küçük AIC değerine sahip olan model en iyi model olarak nitelendirilir. Birden fazla bağımlı değişken için (K>1), çok değişkenli AIC kriteri Bedrick ve Tsai (1994) tarafından önerilmiş olup (3) eşitliğinde verilmektedir:

$$\text{MAIC}(a) = N\left(\log\left|\hat{\Sigma}\right| + K\right) + 2d\left[Ka + K(K+1)/2\right], \tag{3}$$

burada $d = N/[N-(a+K+1)]$ ve $\hat{\Sigma}$ ise Σ'nın en çok olabilirlik kestiricisidir.

Çalışmada kullanılan Akaike bilgi kriterinin bir diğer çok değişkenli versiyonu ise Bozdogan (2000) tarafından önerilmiş olup (4) eşitliğinde verilmiştir:

$$\text{MAIC} = N\log(2\pi a + N\log\left|\hat{\Sigma}\right| + Na + 2\left[Ka + \frac{K(K+1)}{2}\right]. \tag{4}$$

Bu çalışmada Bozdoğan (2000) ve Bedrick vd. (1994) tarafından tanımlanan MAIC çok değişkenli model seçme kriterleri sırasıyla M_{AKAIKE} ve $M_{BEDRICK}$ olarak gösterilmiştir.

4. KEKKR MODELİ VERİLERİNİN TÜRETİLMESİ

Gözlem sayısı ve açıklayıcı değişken sayısının farklı değerleri için model seçme kriterlerinin performanslarını karşılaştırmak amacıyla benzetim çalışması yapılmıştır. Bu benzetim çalışmasında KEKKR modelleri Li vd.'nin (2002) çalışması temel alınarak türetilmiştir. Benzetim çalışmasında: Açıklayıcı değişken matrisi **X:** boyutları $N \times M$ olarak gösterilmiş olup, M=6, 8, 10, 12 olarak seçilmiştir. Bağımlı değişken matrisi **Y:** boyutları ise $N \times K$ olarak belirtilmiş olup, K=4 olarak seçilmiştir. N gözlem sayısını ifade ederken, örneklem büyüklükleri N=100, 250 ve 500 olacak şekilde seçilmiştir. Benzetim çalışmasında gizli değişkenlerin optimum (doğru) sayısı A^* ile gösterilmiştir. A değeri ise gizli değişkenlerin alabileceği maksimum sayıyı göstermektedir ki bu sayı açıklayıcı değişkenlerin sayısına eşittir. Burada $A^* \leq A$ dır. Türetilen tüm KEKKR modelleri, kısmi en küçük kareler model varsayımlarına dayanmaktadır. Çalışmada açıklayıcı değişken matrisleri gizli değişken sayısı (doğru, optimum) $A^* = 4$ olacak şekilde eşitlik (5) kullanılarak türetilmiştir:

$$\mathbf{X} = \sum_{a=1}^{A^*} \mathbf{r_a} \xi'_a + \mathbf{E} \qquad (5)$$

Burada $\mathbf{r_a}$, $\mathbf{E}$; karşılıklı bağımsız normal değişkenler, ξ_a; normalleştirilmiş ortogonal vektörler ve $\text{var}(\mathbf{r_1}) + \text{var}(\mathbf{e_m})$ değerinin; $\text{cov}(\mathbf{X})$, in en yüksek özdeğeri olacak şekilde türetilmiştir. **X** matrisinin değişkenleri Tablo 1 ve Tablo 2 de verilmektedir. Tablo 2'de ξ'_a değerleri ortogonal, birim vektörlerdir.

Tablo 1: X matrisi için R ve E matrisleri

Veri matrisinin boyutu	**R**	**E**
	$\mathbf{R}=[\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \mathbf{r}_4]$	$\mathbf{E}=[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_j]$,
	$\mathbf{r}_1 \sim N(0,10)$	$j=1,\ldots,M$
$N \times M$	$\mathbf{r}_2 \sim N(0,5)$	$\mathbf{e}_j \sim N(0,0.01)$
	$\mathbf{r}_3 \sim N(0,2)$	
	$\mathbf{r}_4 \sim N(0,0.5)$	

Tablo 2: ξ_i için türetilmiş ortogonal vektörler

$N \times 6$

$$\begin{bmatrix} 0.4082 & 0.7071 & 0.4082 & 0.2887 \\ 0.4082 & -0.7071 & 0.4082 & 0.2887 \\ 0.4082 & 0 & -0.8165 & 0.2887 \\ 0.4082 & 0 & 0 & -0.8660 \\ 0.4082 & 0 & 0 & 0 \\ 0.4082 & 0 & 0 & 0 \end{bmatrix}$$

$N \times 8$

$$\begin{bmatrix} 0.1612 & 0.3030 & 0.4082 & 0.4642 \\ 0.3030 & 0.4642 & 0.4082 & 0.1612 \\ 0.4082 & 0.4082 & 0 & -0.4082 \\ 0.4642 & 0.1612 & -0.4082 & -0.3030 \\ 0.4642 & -0.1612 & -0.4082 & 0.3030 \\ 0.4082 & -0.4082 & 0 & 0.4082 \\ 0.3030 & -0.4642 & 0.4082 & -0.1612 \\ 0.1612 & -0.3030 & 0.4082 & -0.4642 \end{bmatrix}$$

$N \times 10$

$$\begin{bmatrix} 0.1201 & 0.2305 & 0.3223 & 0.3879 \\ 0.2305 & 0.3879 & 0.4221 & 0.3223 \\ 0.3223 & 0.4221 & 0.2305 & -0.1201 \\ 0.3879 & 0.3223 & -0.1201 & -0.4221 \\ 0.4221 & 0.1201 & -0.3879 & -0.2305 \\ 0.4221 & -0.1201 & -0.3879 & 0.2305 \\ 0.3879 & -0.3223 & -0.1201 & 0.4221 \\ 0.3223 & -0.4221 & 0.2305 & 0.1201 \\ 0.2305 & -0.3879 & 0.4221 & -0.3223 \\ 0.1201 & -0.2305 & 0.3223 & -0.3879 \end{bmatrix}$$

$N \times 12$

$$\begin{bmatrix} 0.0939 & 0.1823 & 0.2601 & 0.3228 \\ 0.1823 & 0.3228 & 0.3894 & 0.3667 \\ 0.2601 & 0.3894 & 0.3228 & 0.0939 \\ 0.3228 & 0.3667 & 0.0939 & -0.2601 \\ 0.3667 & 0.2601 & -0.1823 & -0.3894 \\ 0.3894 & 0.0939 & -0.3667 & -0.1823 \\ 0.3894 & -0.0939 & -0.3667 & 0.1823 \\ 0.3667 & -0.2601 & -0.1823 & 0.3894 \\ 0.3228 & -0.3667 & 0.0939 & 0.2601 \\ 0.2601 & -0.3894 & 0.3228 & -0.0939 \\ 0.1823 & -0.3228 & 0.3894 & -0.3667 \\ 0.0939 & -0.1823 & 0.2601 & -0.3228 \end{bmatrix}$$

Bağımlı değişken matrisi **Y**, aşağıda verilen (6) ve (7) eşitlikleri kullanılarak türetilmiştir.

$$\mathbf{Y} = \sum_{a=1}^{A^*} \mathbf{z}_a \boldsymbol{\eta}'_{\mathbf{A}_a^*} + \varphi = \sum_{a=1}^{A^*} \mathbf{r}_a \boldsymbol{\eta}'_{\mathbf{A}_a^*} + \mathbf{F}_{A^*} \tag{6}$$

$$\mathbf{F} = \sum_{a=1}^{A^*} \mathbf{f}_a \boldsymbol{\eta}'_{\mathbf{A}_a^*} + \varphi \,. \tag{7}$$

Burada **E** ve **F** ilk A adet gizli değişken çıkartıldıktan sonraki **X** ve **Y** matrisleri için artık değerleri göstermektedir. φ çok değişkenli normal dağılımdan türetilirken $\boldsymbol{\eta}_{\mathbf{A}_a^*}$ ise normalleştirilmiş ortogonal vektörlerden türetilmiştir. $\mathbf{z}_a = \mathbf{r}_a + \mathbf{f}_a$ ve $\mathbf{f}_a$ bağımsız normal değişkenler olarak türetilmiştir.

Tablo 3: φ için türetilmiş değerler

Veri matrisinin boyutu	φ
	$\varphi = [\varphi_1, \varphi_2, \varphi_3, \varphi_4]$
$N \times K$	$\varphi \sim MN\left(0, \begin{bmatrix} 0.00010 & 0.00006 & 0.00006 & 0.00006 \\ 0.00006 & 0.00010 & 0.00006 & 0.00006 \\ 0.00006 & 0.00006 & 0.00010 & 0.00006 \\ 0.00006 & 0.00006 & 0.00006 & 0.00010 \end{bmatrix}\right)$

Tablo 4: $N \times M$ veri matrisinde kullanılan $\boldsymbol{\eta}_i$ için türetilmiş ortogonal vektörler

$N \times 6$	$N \times 8$
$\begin{bmatrix} 0.500 & 0.500 & 0.500 & 0.500 \\ 0.7071 & -0.7071 & 0 & 0 \\ 0.4082 & 0.4082 & -0.8165 & 0 \\ 0.2887 & 0.2887 & 0.2887 & -0.8660 \end{bmatrix}$	$\begin{bmatrix} 0.2887 & 0.500 & 0.5774 & 0.500 \\ 0.500 & 0.500 & 0 & -0.500 \\ 0.5774 & 0 & -0.5774 & 0 \\ 0.500 & -0.500 & 0 & 0.500 \end{bmatrix}$
$N \times 10$	$N \times 12$
$\begin{bmatrix} 0.3717 & 0.6015 & 0.6015 & 0.3717 \\ 0.6015 & 0.3717 & -0.3717 & -0.6015 \\ 0.6015 & -0.3717 & -0.3717 & 0.6015 \\ 0.3717 & -0.6015 & 0.6015 & -0.3717 \end{bmatrix}$	$\begin{bmatrix} 0.6935 & 0.5879 & 0.3928 & 0.1379 \\ 0.5879 & -0.1379 & -0.6935 & -0.3928 \\ 0.3928 & -0.6935 & 0.1379 & 0.5879 \\ 0.1379 & -0.3928 & 0.5879 & -0.6935 \end{bmatrix}$

Tablo 5: Y matrisi için türetilmiş F veri matrisi

Veri matrisinin boyutu	**F**
$N \times K$	$\mathbf{F}=[\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4]$ $\mathbf{f}_1 \sim N(0,0.25)$ $\mathbf{f}_2 \sim N(0,0.125)$ $\mathbf{f}_3 \sim N(0,0.05)$ $\mathbf{f}_4 \sim N(0,0.0125)$

Tablo 1 – Tablo 5; KEKKR modellerinin nasıl türetildiklerini göstermektedir. Elde edilen bu modellerin açıklayıcı değişkenleri arasında çoklu doğrusal bağlantının olup olmadığının tespiti amacıyla $N \times 6$ boyutlu veri matrisi için varyans şişirme değerlerine (VIF) bakılmıştır. VIF değerleri Tablo 6'da görüldüğü gibidir.

Tablo 6: $N \times 6$ veri matrisi için VIF değerleri

Açıklayıcı değişkenler	X_1	X_2	X_3	X_4	X_5	X_6
VIF	635.6	439.9	990.8	765.8	626.9	839.0

Algoritma yardımıyla elde edilen kümülatif varyans değerleri Tablo 7'de verilmektedir. Bu değerler $N \times 6$ boyutlu veri matrisi için gizli değişkenlerin hem açıklayıcı hem de bağımlı değişkendeki değişimi açıklama yüzdelerini ifade etmektedir. Tablodan da görüldüğü gibi optimum değişken sayısı $A^* = 4$'tür. Böylece türetilen KEKKR modellerinin doğru bir şekilde türetildiği doğrulanmaktadır.

Tablo 7: $N \times 6$ veri matrisi için X ve Y matrislerinin göreli kümülatif varyans değerleri

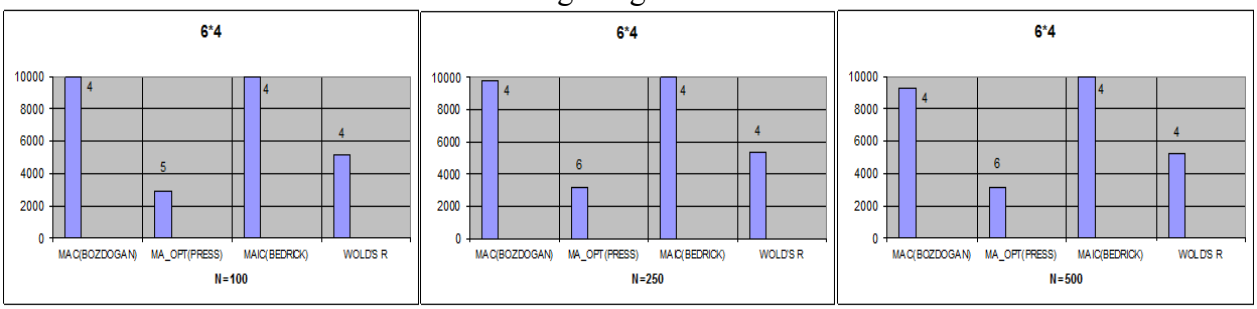
		Gizli değişken sayısı					
Doğru model	Bloklar	1	2	3	4	5	6
A*=4	**X**-blok	0.54196	0.88199	0.95207	0.99998	1.00000	1.00000
	Y-blok	0.93226	0.96577	0.97492	0.97530	0.97530	0.97519

Diğer veri matrisleri için de algoritma çalıştırıldığında görülmüştür ki optimum gizli değişken sayısı A*=4 olarak bulunmaktadır. Böylelikle doğru bir şekilde elde edilen KEKKR modellerine, model seçme kriterleri uygulanarak performans karşılaştırmaları yapılabilir. Çapraz geçerlilik; model seçme ve geçerlilik de kullanılan bir teknik olup, doğrusal regresyon modelinin tahmin geçerliliğini incelemede de kullanılmaktadır. Çalışmamızda Modified Kernel#2 algoritması 5-kat çapraz geçerlilikte kullanılmıştır. Her bir veri matrisi için 10.000 adet veri seti türetilmiş olup model seçme kriterlerinin gizli değişken sayısını doğru bir şekilde bulma performansları karşılaştırılmıştır.

5. BULGULAR

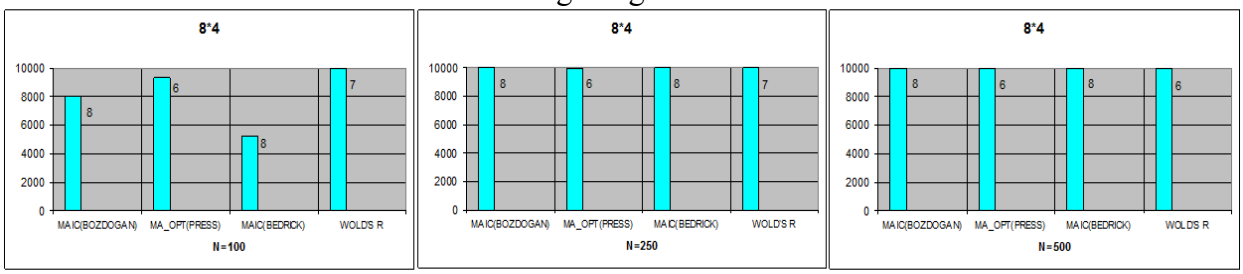
Bu çalışmada, KEKKR varsayımlarına göre türetilen verilerde açıklayıcı değişken ve gözlem sayısının farklı değerleri için model seçme kriterlerinin istenilen gizli değişken sayısını doğru bir şekilde seçebilme performanslarının karşılaştırması yapılmıştır. Gözlem sayısı N=100, 250 ve 500 için, M_{AKAKIE}, $M_{BEDRICK}$, MA_opt(PRESS) ve Wold's R model seçme kriterlerinin performans sonuçları Şekil 1 - Şekil 4' de verilmiştir. Bu şekillerde, tüm model seçme kriterlerinin en fazla tekrarla kaç gizli değişken buldukları gösterilmektedir.

Şekil 1: N× 6 için tüm model seçme kriterlerinin en fazla tekrarla kaç sayıda gizli değişken bulduğunu gösterir

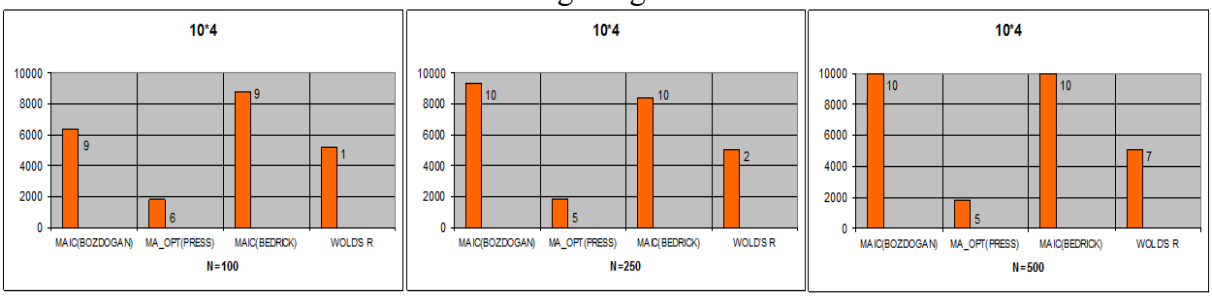


Şekil 1' de görüldüğü gibi, 6 açıklayıcı değişkene sahip veri seti için, MA_opt(PRESS) dışındaki kriterler optimum sayıda (4) gizli değişkeni bulmaktadır. Bu veri seti için model seçme kriterleri kabul edilebilir bir performans sergilemekte ve gizli değişken sayısını doğru bir şekilde bulmaktadırlar. Tüm veri matrisleri için model seçme kriterlerinin performansları Şekil 2- Şekil 5'de verilmiştir.

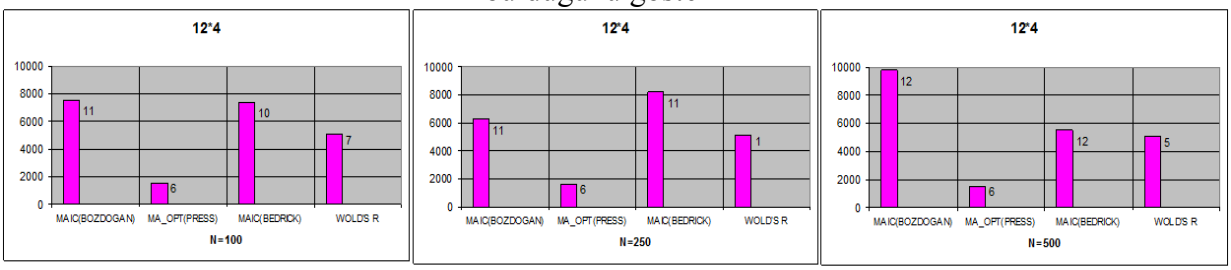
Şekil 2: $N \times 8$ için tüm model seçme kriterlerinin en fazla tekrarla kaç sayıda gizli değişken bulduğunu gösterir



Şekil 3: $N \times 10$ için tüm model seçme kriterlerinin en fazla tekrarla kaç sayıda gizli değişken bulduğunu gösterir

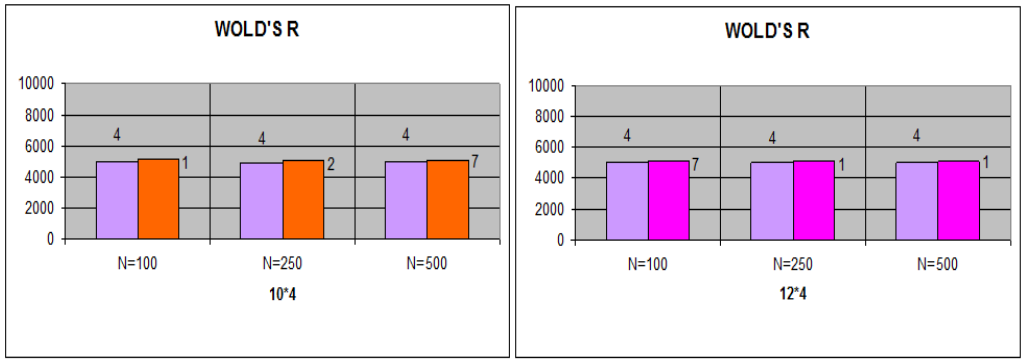


Şekil 4: $N \times 12$ için tüm model seçme kriterlerinin en fazla tekrarla kaç sayıda gizli değişken bulduğunu gösterir



Şekil 2 - Şekil 4 'de görülmektedir ki model seçme kriterleri optimum sayıdan daha fazla sayıda gizli değişken bulmaktadır. Yani, her bir model seçme kriteri için overfite doğru bir eğilim vardır ve veri matrislerinin boyutları büyüdükçe daha fazla sayıda gizli değişken seçtikleri gözlemlenmiştir. Fakat genel olarak söylenebilir ki açıklayıcı değişken sayısı arttıkça her bir model seçme kriteri gizli değişken sayısını açıklayıcı değişken sayısına yakın olarak bulmaktadır. Elde edilen bir diğer sonuçsa istenilen gizli değişken sayısına yakın değer veren kriterler MA_opt(PRESS) ve Wold's R kriterleridir.

Şekil 5: Wold's R kriterinin performansı



Şekil 5'de Wold's R kriterinin optimum gizli değişken sayısını bulma performansı görülmektedir. Wolds'R kriteri en fazla tekrar sayısı ile bir (1) gizli değişkeni bulurken, ikinci en fazla tekrar sayısı ile de optimum değişken sayısını (4) bulabilmektedir.

6. TARTIŞMA VE SONUÇ

Bu çalışmada ki temel amaç; optimum gizli değişken sayısını saptamada M_{AKAKIE}, $M_{BEDRICK}$, PRESS ve Wold's R kriterlerinin performanslarının karşılaştırılmasıdır. Veri türetilirken kriterlerin bulması gereken gizli değişken sayısı yani optimum değişken sayısı 4 olarak belirlenmiş ve veri türetimi bu doğrultuda yapılmıştır. Fakat benzetim çalışması sonuçları göstermiştir ki tüm kriterler optimum gizli değişken sayısını sadece küçük boyutlu veri matrisinde bulabilmektedir. Bu durum Li vd. (2002) çalışmalarıyla benzerlik göstermektedir. Bu sonuca dayanarak artan gözlem sayısı ve açıklayıcı değişken sayısı için veri türetildiğinde küçük boyutlu veri matrisi dışındaki diğer veri matrislerinde sonuçlar değişmekte ve farklı sayıda gizli değişken bulmaktadırlar. Genel olarak söylenebilir ki, veri matrisinin boyutları büyüdükçe KEKKR için model seçme kriterleri daha fazla sayıda gizli değişkenle çalışan bir model oluşturmaktadır. Benzetim sonuçları ayrıca göstermiştir ki Wold's R kriteri tüm boyutlu veri matrislerinde en tutarlı çalışanıdır. Özellikle küçük boyutlu veri matrisinde Li vd., (2002) çalışmasına göre veri türetildiğinde aynı sonuca ulaşılmıştır. Fakat amacımız, veri matrislerinin boyutu büyütüldüğünde kriterlerinin doğru dizli bileşen sayısını bulmadaki performanslarının hangi eğilimde olduğunu saptamaktır. M_{AKAKIE} ve $M_{BEDRICK}$ hemen hemen aynı sonuçları bulmakta fakat veri matrisinin boyutları büyüdüğünde

istenilen gizli değişken sayısını bulamamaktadırlar. MA_opt(PRESS) ise Wold's R kriteri ile aynı veya benzer sonuçları vermektedir.

Sonuç olarak, benzetim çalışmaları sonucunda: 6*4 veri matrisinde tüm kriterler çok iyi performans sergilemektedirler. Buna rağmen büyük boyutlu veri matrisleri için tüm kriterler optimum gizli değişken sayısını bulamamakta ve de gizli değişken sayısını açıklayıcı değişken sayısına yakın bulmaktadırlar. Bu sonuçlar göstermektedir ki MAIC, PRESS ve Wold's R kriterleri doğru modeli seçmede, regresyon modelinin boyutundaki ve gözlem sayısındaki değişmelerden etkilenmektedir.

KAYNAKÇA

Abdi, H. 2007. Partial Least Square Regression (PLS regression).–In: Salkind, N. J. (ed.), Encylopedia of Measurement and Statistics. Sage.

Akaike, H. 1971. Autoregressive Model Fitting for Control. Ann. Inst. Statistics Math 23: 163-180.

Akaike, H. 1974. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control 19:716–723.

Bedrick, E. J., and Tsai, C. L. 1994. Model Selection for Multivariate Regression in Small Samples. Biometrics 50:226-231.

Bozdogan, H. 2000. Akaike's Information Criterion and Recent Developments in Information Complexity. Journal of Mathematical Psychology 44:62-91.

Carrascal, L. M, and Galva´n I Gordo O., 2009. Partial Least Squares Regression as an Alternative to Current Regression Methods Used in Ecology. The Authors. Journal Compilation, Oikos 118:681-690.

Dayal, B., and MacGregor., J. 1997. Improved PLS algorithms. Journal of Chemometrics 11:73-85.

De Jong, S. 1993. SIMPLS: An Alternative Approach to Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 18:251-263.

De Jong, S., and Ter Braak, C. J. F. 1994. Comments on the PLS kernel algorithm. Journal of Chemometrics 8:169–174.

Garthwaite, P. H. 1994. An Interpretation of Partial Least Squares. Journal of the American Statistical Association 89:122-127.

Geladi, P., and Kowalski, R. 1986. Partial Least Squares Regression: A Tutorial. Analytica Chimical Acta 185:1-17.

Geladi, P. 1988. Notes on the History and Nature of Partial Least Squares (PLS) Modelling. Journal of Chemometrics 2:231-246.

Helland, I. S. 1990. Partial Least Squares Regression and Statistical Models. Scandinavian Journal of Statistics 17:97-114.

Höskuldson, A. 1988. PLS Regression Methods. Journal of Chemometrics 2:211-228.

Last, M. 2006. The Uncertainty principle of cross-validation. Member, IEEE:275-280.

Li, B., Morris, J., and Martin, B. 2002. Model Selection for Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 64:79-89.

Lindgren, F., Geladi, P., and Wold, S. 1993. The Kernel Algorithm for PLS. Journal of Chemometrics 7:45-59.

Lindgren, F., and Rannar, S. 1998. Alternative Partial Least-Squares (PLS) Algorithm. Perspective in Drug Discovery and Design 12/13/14:105-113.

Martens, H. and Naes, T. 1989. Multivariate Calibration. John Wiley & Sons.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1996. Applied Linear Regression Models. Times Mirror Higher Education Group, Inc.

Rännar, S., Lindgren, F., Geladi, P., and Wold, S. 1994. A PLS Kernel Algorithm For Data Sets With Many Variables and Fewer Objects. Part 1: Theory and Algorithm. Journal of Chemometrics 8:111-125.

Shao, J. 1993. Linear Model Selection by Cross-validation. J.Amer.Stat.Assoc. 88:486-494

Stone, M., and Brooks, R. J. 1990. Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression (with discussion). Journal of the Roya1 Statistical Society, Ser. B 52:237-269.

Tobias, R. D. 2003. An Introduction to Partial Least Squares Regression from:www.ats.ucla.edu/stat/sas/library/pls.pdf. (08.10.2011)

Wold, S., Ruhe, A., Wold, H., and Dunn, W. 1984. The Collinearity Problem in Linear Regression. The PLS Approach to Generalised Inverses. Journal of Scientific Statistical Computing, SIAM 5:735-743.

Wold, S., Sjöström, M., and Eriksson, L. 2001. PLS-regression:a Basic Tool of Chemometrics. Chemometrics and Intelligent Laboratory Syste