İstatistik Araştırma Dergisi Cilt 02 No 01 Sayfa 11-20 Nisan 2003

Çok Değişkenli Normal Dağılıma Sahip Örneklerdeki Aykırı Gözlemlerin Belirlenmesi İçin Bayesgil Bir Yaklaşım

Ufuk EKİZ*

ÖZET

Bu çalışmada çok değişkenli normal dağılıma sahip örneklerdeki aykırı gözlemlerin belirlenmesi için önerilen Bayesgil bir metod tanıtılmaktadır.Örnekleme teorisinde de kullanılan karesel formların dağılımından yararlanarak aykırı gözlemleri belirlemek fikrinden hareketle , gerçekleşmiş hataların karesel formlarının sonsal (posterior) dağılımı , bu gözlemlerin belirlenmesi için kullanılmaktadır.Uygulamada , gerçek bir veri üzerinde her bir gözleme ilişkin sonsal olasılıklar elde edilmiş ve aykırı (outlier) gözlem(ler) belirlenmiştir.

Anahtar Kelimeler: *Baskın Olmayan Önsel , Aykırı Gözlem , Bayes Faktör , Önsel İhtimal , Sonsal İhtimal .*

1. GİRİŞ

Aykırı gözlemler , verinin genel özelliklerini göstermeyen gözlemlerdir. Bazı temel modellere göre aşırılık (extremeness) gösterirler. Barnett ve Lewis ,1994,sh.7 aşırılığın , gözlem değerlerinin "sıralanmasının" bazı formlarından kaynaklanabileceğini vurgulamışlardır.Ayrıca aşırı gözlemleri ifade etmek için uygun bir alt sıralama özelliğinin benimsenmesine ihtiyaç olduğunu tartışmışlardır. Barnett ,1976 çok değişkenli problemlerde alt-sıralama kurallarının özelliklerini göz önünde bulundurmakta ve onları dört sınıfa ayırmaktadır. Bu dört alt-sıralama sınıfları ; düşürülmüş , marjinal , kısmi ve koşullu olarak isimlendirilmiştir . Düşürülmüş alt-sıralama çok değişkenli problemlerde aykırı gözlemleri belirlemekte kullanılan neredeyse tek örnektir. Bu alt-sıralama , p boyutlu $Y_1, Y_2, ..., Y_n$ rasgele değişkenlerinin gözlenmiş değerlerinden oluşan örneği , tek değişkenli $R(Y)$ istatistiğinin değerleri bakımından sıralayarak uygulanmaktadır. Şayet

$$R(Y_i) = \max\{R(Y_j); j = 1,2,...,n\}$$

ise , i. gözlemden aykırı olarak şüphelenilebilir.

Ayrıca , göz önünde bulundurulan temel model altında $R(Y_i)$, istenmeyecek ölçüde dağılımın konumundan(location) uzakta ise , Y_i, aykırı gözlem olarak açıklanabilir.

* Gazi Ünv.Fen-Ed.Fak.İstatistik Böl. Teknikokullar/ANKARA

Literatürde genellikle , η örneğin ya da yığının konumunu ve V'de örneğin veya yığının yayılımını(scale) ifade etmek üzere , çok değişkenli aykırı gözlemlerin belirlenmesine ilişkin olarak örnekleme teorisinde ;

$$R(Y_j,\eta,V)=(Y_j-\eta)^T V^{-1}(Y_j-\eta) \qquad (1)$$

formundaki istatistikler kullanılmaktadır. Eğer ilgilenilen dağılımın ortalama vektörü μ ve kovaryans matrisi Σ , daha önceden biliniyorsa , (1)

$$R(Y_j,\mu,\Sigma)=(Y_j-\mu)^T \Sigma^{-1}(Y_j-\mu) \qquad (2)$$

şeklinde yazılabilir.Bu karesel formda , μ ve Σ çoğunlukla bilinmedikleri için (2) de onların yerine örnekten elde edilen tahmin edicileri kullanılır. Bu tahminler aykırı gözlemlerden etkilenebilmektedirler. Bunun için bazı yazarlar μ ve Σ için sağlam (robust) tahminlerinin kullanılmasını önermektedirler (Campell ,1980 ; Rousseeuw ve Vant Zameren ,1990).

Çok değişkenli aykırı gözlemlerin belirlenmesine yönelik olarak Guttman ,1973 Bayesgil bir yaklaşım öne sürmektedir. Bu yaklaşım , n birimlik rasgele bir örnekteki pek çok gözlemin parametreleri μ ve Σ , geriye kalan gözlemlerin ise parametreleri $\mu+\alpha$ ve Σ olan normal dağılımdan geldiğini varsaymaktadır. Guttman , aykırı gözlemleri belirlemek için α'nın sonsal dağılımını kullanmayı öngörmektedir. Buna göre , örnekteki tüm gözlemlere ilişkin kovaryans matrisinin determinantı ile ters orantılı olarak j.gözleme bir c_j ağırlığı atanır.Eğer bir aykırı gözlem var ise buna karşılık gelen ağırlığın incelenmesi çok belirleyici olacaktır.

Çok değişkenli aykırı gözlemlerin belirlenmesine ilişkin diğer çalışmalar Gnanadesikan ve Kettenring ,1972 , Hawkins ,1980 , Rousseuw ve Leroy ,1987 , Justel ve Pena ,2001 nın makalelerinde yer almaktadır.

Bu çalışmada çok değişkenli normal dağılıma sahip bir örnekteki aykırı gözlemlerin belirlenmesine yönelik önerilen Bayesgil bir yaklaşım tanıtılmaktadır. Bu , tek değişkenli doğrusal modellerde Chaloner ve Brant ,1988 tarafından önerilen aykırı gözlemlerin belirlenmesine yönelik bir yaklaşımın genellemesidir. Chaloner ve Brant , aykırı gözlemleri belirlemek için modele ilişkin , rasgele hataların sonsal dağılımını kullanmışlardır. Burada sunulan Bayesgil yaklaşımda , (2) de tanımlanan $R(Y_j,\mu,\Sigma)$ karesel formuna ilişkin sonsal dağılım , aykırı gözlemlerin belirlenmesinde kullanılmaktadır.

Bölüm 2 'de çok değişkenli normal dağılıma sahip p boyutlu $Y_1, Y_2, \ldots, Y_n$ rasgele değişkenlerinin gözlenmiş değerlerinden oluşan bir örnekteki aykırı gözlemlerin tanımlanması ele alınmakta , Bölüm 3'te de yapılan tanım için gerekli olan sonsal olasılıkların elde edilmesi anlatılmaktadır. Uygulamada , gerçek bir verideki herhangi bir gözlemin aykırı olup olmadığını test etmekte kullanılan bayes faktör değerlerinin

hesaplanması için gerekli sonsal olasılıklar , Monte Carlo simülasyon tekniği ile elde edilmekte ve olası aykırı gözlemler belirlenmektedir.

2. ÇOK DEĞİŞKENLİ NORMAL DAĞILIMA SAHİP YIĞINDAN ÇEKİLMİŞ RASGELE ÖRNEKLERDE AYKIRI GÖZLEMLERİN BELİRLENMESİ

Bu bölümde çok değişkenli normal dağılıma sahip rasgele bir örneğin aykırı gözlemlerinin belirlenmesine ilişkin Bayesgil bir yaklaşım sunulmaktadır. p değişkenli, $Y \sim N(\mu,\Sigma)$ olan dağılımın n çaplı rasgele bir örneği $Y_1, Y_2,..., Y_n$ olsun.

$$\varepsilon_i = \Sigma^{-1/2}(Y_i - \mu) \quad , \; i=1,2,...,n$$
$$\delta_i = \varepsilon_i^T \varepsilon_i = R(Y_i,\mu,\Sigma) \tag{3}$$

şeklinde tanımlansın. Burada $\varepsilon_i \sim N(0,I)$ dır.Eğer δ_i nin sonsal dağılımı üzerinden $\delta_i > k$ olasılığı , uygun k değerini belirlemek için kullanılacak olasılıktan büyük ise , i.gözlemin aykırı olduğu söylenebilir. Herhangi bir gözlemin aykırı olarak nitelenmesine izin vermeyecek kadar büyük bir önsel(prior) olasılığa karşılık gelecek şekilde bir k değeri belirlenebilir. Eğer olasılık 0.95 alınırsa ;

$$0.95 = pr(\delta_1 \le k).pr(\delta_2 \le k)...pr(\delta_n \le k)$$

yazılabilir.Diğer bir ifade ile

$$0.95 = pr(\delta_i \le k, \text{tüm.i'ler.için}) = \{F_p(k)\}^n \tag{4}$$

dır.Burada , $F_p(.)$, p serbestlik dereceli merkezsel ki-kare dağılımına sahip rasgele değişkenin dağılım fonksiyonudur. Yukarda verilen eşitliğin çözümünden ;

$$k = F_p^{-1}(0.95^{1/n})$$

elde edilir. δ_i'nin sonsal dağılımının elde edilmesi ve $p_i = pr(\delta_i > k / Y)$ sonsal olasılıklarının hesaplanabilmesi için μ ve Σ 'nın bilgi vermeyen(noninformative) önselleri kullanılacaktır. Bulunan p_i olasılık değerlerinden , muhtemel aykırı gözlemler belirlenebilir. Bu p_i olasılık değerleri sıralanabilirler ve en büyük p_i olasılık değerine sahip gözlemler aykırı gözlemler olarak ifade edilebilir. Bu çalışmada , i.gözlemin aykırı olup olmadığına ilişkin ,

$$H_{0i} : \delta_i > k \quad , \; (Y_i \text{ aykırı bir gözlem ise})$$
$$H_{1i} : \delta_i \le k \quad , \; i = 1,2,...,n$$

hipotezi test etmek için Bayes Faktörleri kullanılmaktadır. H_{1i} 'ye karşı H_{0i} hipotezinin testi için B_i Bayes faktörü ;

$$B_i = \frac{p_i F_p(k)}{(1-p_i)\{1-F_p(k)\}}$$

H_{0i}' ye ilişkin sonsal olasılıkların(posterior odds) , önsel olasılıklara(prior odds) oranıdır. Kass ve Raftery ,1995 , B_i değerinin 10 'dan büyük olmasının H_{0i} için güçlü , 100 'den büyük olmasının ise çok güçlü bir kanıt olduğunu öne sürmektedir.Bayes faktör değerlerinin elde edilmesinde gerekli p_i sonsal olasılıkları ,μ ve Σ'nın bilgi vermeyen önsellerinin kullanılması ile Σ ve Y 'nin biliniyor olması koşulu altında bulunacak δ_i 'nin sonsal dağılımından elde edilecektir.

3. μ VE Σ İÇİN BASKIN OLMAYAN ÖNSEL FONKSİYONU VE p_i SONSAL DEĞERLER

δ_i ' nin sonsal(posterior) dağılımını elde etmek ve $p_i = pr(\delta_i > k / Y)$ olasılığını hesaplamak için μ ve Σ 'nın bilgi vermeyen bileşik önsel fonksiyonu , μ ve Σ 'nın yaklaşık olarak bağımsız oldukları varsayımı altında;

$$p(\mu,\Sigma) = p(\mu).p(\Sigma)$$

dır. Box ve Tiao ,1973, sh.426 çok değişkenli normal dağılımda μ ve Σ ' nın bilgi vermeyen önsel fonksiyonlarını sırası ile ;

$$p(\mu) \propto 1$$
$$p(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$$

ve birleşik önsel fonksiyonu da ,

$$p(\mu,\Sigma) = p(\mu).p(\Sigma) \propto |\Sigma|^{-\frac{p+1}{2}}$$

şeklinde vermektedir.

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \quad , \quad S = \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_i - \overline{Y})^T$$

olarak tanımlanırsa , gözlenen değerlere ilişkin olabilirlik fonksiyonu ;

$$f(Y/\mu,\Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\{iz(\Sigma^{-1}S)+n(\overline{Y}-\mu)^T\Sigma^{-1}(\overline{Y}-\mu)\}\right]$$

formunda ifade edilir. Bileşik önsel fonksiyonu ve Y 'nin biliniyor olması koşulu altında μ ve Σ 'ya ilişkin olabilirlik fonksiyonlarını kullanarak (μ,Σ) nın bileşik sonsal fonksiyonu;

$$p(\mu,\Sigma/Y) \propto p(\mu,\Sigma).f(\mu,\Sigma/Y)$$

$$\propto |\Sigma|^{-\frac{n+p+1}{2}} \exp\left\{-\frac{1}{2}iz(\Sigma^{-1}S)\right\}\exp\left\{-\frac{n}{2}(\mu-\overline{Y})^T\Sigma^{-1}(\mu-\overline{Y})\right\}$$

olarak elde edilir. Bileşik sonsal fonksiyon

$$p(\mu,\Sigma/Y) = p(\mu/\Sigma,Y).p(\Sigma/Y)$$

şeklinde ifade edildiğinden ,

$$\mu/\Sigma,Y \sim N(\overline{Y},n^{-1}\Sigma)$$

$$\Sigma/Y \sim W^{-1}(S,p,n-p)$$

bulunur. Burada $W^{-1}(S,p,n-p)$ ters (inverted) Wishart dağılımıdır (Box ve Tiao ,1973 , sh.460-464).

$\mu/\Sigma,Y$ ve Σ/Y 'nin sonsal fonksiyonlarını kullanarak $\varepsilon_i/\Sigma,Y$ ve $\delta_i/\Sigma,Y$ ' nin koşullu sonsal fonksiyonları bulunabilir. ε_i'ler , Σ verildiğinde μ 'nün doğrusal fonksiyonlarıdır. Y ve Σ 'nın biliniyor olması koşulu altında μ'ye ilişkin sonsal dağılım ;

$$p(\mu/\Sigma,Y) = c.\left[\exp\left\{-\frac{n}{2}(\mu-\overline{Y})^T\Sigma^{-1}(\mu-\overline{Y})\right\}\right]$$

olarak ifade edilir.

$$\varepsilon_i = \Sigma^{-\frac{1}{2}}(Y_i-\mu)$$

$$\mu = Y_i - \Sigma^{\frac{1}{2}}\varepsilon_i$$

dönüşümü uygulandığında , dönüşüme ilişkin Jacobian terimi bir sabit olacağından c ifadesi içinde yer alır. Y ve Σ 'nın biliniyor olması koşulu altında ε_i'nin sonsal fonksiyonu ;

$$p(\varepsilon_i / Y,\Sigma) = c.\left[\exp\left\{-\frac{n}{2}(Y_i - \Sigma^{\frac{1}{2}}\varepsilon_i - \overline{Y})^T\Sigma^{-1}(Y_i - \Sigma^{\frac{1}{2}}\varepsilon_i - \overline{Y})\right\}\right]$$

$$= c.\left[\exp\left\{-\frac{n}{2}[\varepsilon_i - \Sigma^{-\frac{1}{2}}(Y_i - \overline{Y})]^T[\varepsilon_i - \Sigma^{-\frac{1}{2}}(Y_i - \overline{Y})]\right\}\right]$$

elde edilir. $\gamma_i = \Sigma^{-\frac{1}{2}}(Y_i - \overline{Y})$ biçiminde tanımlanırsa, $\varepsilon_i / \Sigma, Y \sim N(\gamma_i, n^{-1}I)$ olur. Burada c , fonksiyona ilişkin sabit terimdir. Böylece Σ ve Y verildiğinde $W_i = n\delta_i$, serbestlik derecesi p , merkezsel olmama parametresi $\lambda_i = n(Y_i - \overline{Y})^T\Sigma^{-1}(Y_i - \overline{Y})$ olan merkezsel olmayan bir ki-kare dağılımıdır.

Yukarıda verilen dağılımdan ,

$$p_i = E_{\Sigma/Y}[pr(\delta_i > k / Y,\Sigma)]$$

$$= E_{\Sigma/Y}[pr(W_i > nk / Y,\Sigma)] \quad ; i = 1,2,...,n$$

yazılabilir. Hesaplama kolaylığı bakımından p_i'leri $\Sigma^{-1} = V$ metriğinde yazmak daha uygundur.

$$p_i = E_{V/Y}[pr(W_i > nk / Y, V)] \tag{5}$$

yazılabilir. Burada $\lambda_i = n(Y_i - \overline{Y})^T V(Y_i - \overline{Y})$ ve

$$V / Y \sim W(S^{-1}, p, n - p)$$

dir. Her bir i için p_i' nin tahmin değerlerinin elde edilmesinde Monte Carlo teknikleri uygulanabilir.

4. UYGULAMA

Aşağıda , Tablo 1 'de verilen çok değişkenli normal dağılıma sahip rasgele değişkenlere ilişkin gözlenen değerler Khattree R. ve Naik D.N.,1999 Appendix B den alınmıştır. Burada $n = 28$, $p = 4$ tür.

Tablo 1. (Cork Data)

	Y_1	Y_2	Y_3	Y_4		Y_1	Y_2	Y_3	Y_4
1	72	66	76	77	**15**	91	79	100	75
2	60	53	66	63	**16**	56	68	47	50
3	56	57	64	58	**17**	79	65	70	61
4	41	29	36	38	**18**	81	80	68	58
5	32	32	35	36	**19**	78	55	67	60
6	30	35	34	26	**20**	46	38	37	38
7	39	39	31	27	**21**	39	35	34	37
8	42	43	31	25	**22**	32	30	30	32
9	37	40	31	25	**23**	60	50	67	54
10	33	29	27	36	**24**	35	37	48	39
11	32	30	34	28	**25**	39	36	39	31
12	63	45	74	63	**26**	50	34	37	40
13	54	46	60	52	**27**	43	37	39	50
14	47	51	52	43	**28**	48	54	57	43

Hiçbir gözlemin aykırı olmaması olasılığı 0.95 olarak seçildiğinde $pr(\delta_i < k) = F_k(.) = 0.999$ ve k=17.1211 bulunur. MAT LAB (6.0) paket programında , Monte Carlo simülasyon tekniği kullanılarak yazılan programdan, (5) 'teki p_i olasılıklarına ilişkin tahminler elde edilmiştir. Her bir gözlem için , $W(S^{-1},4,24)$ dağılımından rasgele bir matris üretilmiş ve varyansın biliniyor olması koşullu , merkezsel olmayan ki-kare olasılığı hesaplanmıştır ($pr(W_i > nk / Y, V)$). Bu döngü 20.000 defa tekrarlanmış ve p_i bu koşullu olasılıkların ortalaması olarak alınmıştır. Her bir gözlem için elde edilen Bayes Faktör ve p_i olasılık değerleri aşağıda verilen Tablo 2 'de yer almaktadır.

Tablo 2. Bayes Faktör (B_i) ve p_i Olasılık Değerleri

İ	B_i	p_i	İ	B_i	p_i
1	0.5457	0.0010	15	3.5766	0.0066
2	2.1698E-015	3.9801E-018	16	14.0862	0.0265
3	1.2283E-014	2.2532E-017	17	1.9158E-008	3.5142E-011
4	5.9138E-014	1.0848E-016	18	1.8242	0.0034
5	5.2341E-014	9.6012E-017	19	0.1000	1.8338E-004
6	7.0795E-014	1.2986E-016	20	1.0472E-013	1.9209E-016
7	3.6632E-014	6.7196E-017	21	7.9652E-014	1.4611E-016
8	2.6495E-009	4.8601E-012	22	1.1157E-013	2.0465E-016
9	1.7219E-014	3.1586E-017	23	5.9686E-014	1.0948E-016
10	2.9206E-013	5.3575E-016	24	4.1208E-009	7.5591E-012
11	4.2923E-014	7.8737E-017	25	9.8705E-014	1.8106E-016
12	0.0050	9.2551E-006	26	4.6401E-006	8.5116E-009
13	5.8215E-014	1.0679E-016	27	0.0015	2.8117E-006
14	3.5794E-014	6.5659E-017	28	1.5924E-009	2.9211E-012

Sadece on altıncı gözleme ilişkin Bayes Faktör değeri $B_{16} = 14.0862$, 10 'dan büyük çıkmıştır. Buna göre on altıncı gözlemin aykırı olarak düşünülmesi için çok güçlü bir kanıt mevcuttur. Khattree R. ve Naik D.N. , bu verideki aykırı gözlemlerin belirlenmesi için örnekleme teorisindeki farklı yöntemleri kullanmışlar ve aynı sonucu elde etmişlerdir.

5. SONUÇ

Bu çalışmada , önerilen Bayesgil yaklaşımda , çok değişkenli normal dağılıma sahip örneklerdeki aykırı gözlemleri belirlemek için hata terimlerinin karesel formlarının koşullu sonsal dağılımı kullanılmıştır. Eğer parametre değerleri bilinmiyorsa (2)'deki istatistiğin iyi bir alternatifi yoktur. Parametrelerdeki belirsizlik , (2) nin sonsal dağılımının kullanılması ile açıklanmış olur. Bu yaklaşımın en az iki avantajı bulunmaktadır. Birincisi, ifade etmesi kolaydır ve ikinciside , tahmin edicilerin aykırı gözlemlerden etkilenmesi bu analizde daha az etkili olur . Parametreler için bir önsel dağılım mevcut değilse Bilgi vermeyen önseli seçmek mantıklıdır.Bilgi veren (Informative) önselleri kullanmakta mümkün olabilir. Çok değişkenli normal dağılımda , ortalama parametresi için normal önseli ve kovaryans parametresi için ters Wishart önseli kullanıldığında elde edilen bileşik önsel fonksiyon uygun parametre değerleri için , bilgi vermeyen bileşik önsel fonksiyon ile aynı sonucu vermektedir (Peter M.Lee ,1989.sh.73). Eğer başka bilgi veren önseller kullanılırsa , hatalara ilişkin sonsal dağılım bilinen bir formda ortaya çıkmayabilir . Bu durumda Marcov Chain Monte Carlo (MCMC) teknikleri p_i 'lerin tahmini için gerekli olur.

KAYNAKLAR

BARNETT,V.(1976),*The ordering multivariate data (withdiscussion),* J.Roy.Statist.Soc.Ser.A,139,318-54.

BARNETT,V.,LEWIS,T(1994),*Outlier in Statistical Data,*3rd.ed.Chichester:John Wiley & Sons.

BOX,G.E.P.,TIAO,G.C.(1973),*Bayesian Inference in StatisticalAnalysis,* Reading, MA:Addison-Wesley .

CAMPBELL,N (1980) , *Robust Procedures in Multivariate analysis I :*Robust Covariance Estimation., Applied Statistics, 29, 231-37.

CHALONER,K.,BRANT,R.(1988),*A Bayesian Approach to Outlier Detection and Residual Analysis.* , Biometrika,75,651-9.

GNANADESIKAN,R.,KETTENRING,J.(1972),*RobustEstimates,Residuals and Outlier Detection With Multiresponse Data., Biometrics,28,81-124.*

GUTTMAN,I.(1973),*Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity-a Bayesian Approach.,* Technometrics.15.723-38.

HAWKINS,D.(1980) , *Identification of Outliers , London ,* Chapman and Hall.

JUSTEL,A.,PENA,D.(2001),*Bayesian unmasking in Linear Models,* ComputationalStatistics & Data Analysis.36,69-84.

PETER M.LEE.(1989) , *Bayesian Statistics ,* London,Chapman and Hall.

KASS,R.,RAFTERY,A.(1995) , *Bayes Factors ,* JASA , 90 , 773-95.

KHATTREE R.,NAIK N.D.(1999),*Applied Multivariate Statistics with SAS* Software , SAS Institute.

ROUSSEEUW,P.,LEROY,A.(1987),*Robust Regression and Outlier Detection,* NewYork:Wiley.

ROUSSEEUW,P.,VAN ZAMEREN,B.(1990),*Unmasking Multivariate Outliers and Leverage Points ,* JASA , 85 , 633-9.

A Bayesian Method to Indentification of Outlier Observations in Multivariate Normal Distribution

ABSTRACT

In this work , a method of Bayesian which is suggested for determine of outlier observation in the samples which has multivariate normal distribution is introduced. Outlier observation is introduced using distribution of quadratic forms in also sampling theory. Posterior odds of quadratic forms of errors are used to determine of observations. In aplication , posterior odds in the real data corresponding to every observation are found and outlier observations are determined.

Key Words: *Noninformative Prior , Outlier Observation , Bayes Factor ,Prior Odds , Posterior Odds.*