İstatistik Araştırma Dergisi Cilt 04 No 03 Sayfa 45-51 Aralık 2005

AYKIRI GÖZLEM SAYISININ BELİRLENMESİ

Ufuk EKİZ * Müslim EKNİ *

ÖZET

Clarke and Lewis (1998)'e göre aykırı gözlem, diğer gözlemlerle aynı merkezi parametreli fakat farklı varyanslı benzer bir dağılım tarafından üretilir. Bu tanımdan hareketle Wen-Liang Hung, Jong-Wuu Wu (2005), örnekteki aykırı gözlemlerin sayısını hata kareler toplamını en küçükleyerek belirleyen bir yöntem önermişlerdir. Bu yöntem Clarke ve Lewis tarafından tanımlanan R istatistiğine göre daha basit ve kolay hesaplanabilir. Ayrıca normal dağılımdan geldiği düşünülen örnekteki alt ve üst aykırı gözlemlerin sayısını belirlemekte gizleme ve yanılgıya-düşürme problemlerinden etkilenmediği söylenmektedir. Bu çalışmada, yöntemin ne kadar sağlıklı sonuçlar verdiğini görmek için bir simülasyon çalışması yapılmıştır. Sonuçlar örnek çapı büyüdükçe aykırı gözlem sayısını doğru belirleme oranının düştüğünü gösterdiğinden, yöntem gizleme ve yanılgıya-düşürme problemlerinden etkilenmektedir.

Anahtar Kelimeler : *Aykırı Gözlem, En Küçük Kareler, Gizleme, Monte Carlo, Sıra İstatistiği, Yanılgıya-Düşürme.*

1. GİRİŞ

Son yıllarda örnekte yer alabilecek aykırı gözlemlerin hangisi olduğu ya da bunların sayısının belirlenmesi problemleri ile pek çok yazar ilgilenmiştir (Guttman, I. (1973b), Pearson ve Sekar (1936), Cook ve Weisberg. (1982)). Aykırı gözlemlerle ilgili yapılmış çalışmaların kapsamlı bir özeti Barnett ve Lewis (1994)'te yer almaktadır.

Bu çalışmada, Clarke ve Lewis tarafından tanımlanmış olan problemle ilgilenilecektir. Bu problem aşağıdaki gibi tanımlanabilir.

Örneğin k tane üst aykırı gözlem içerdiği varsayılıyor olsun, tesadüfi örneğine ilişkin yokluk hipotezi,

* Gazi Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Teknikokullar, Ankara, TÜRKİYE, e-mail: ufukekiz@gazi.edu.tr, mekni@gazi.edu.tr

$$H_0 = \chi_i \approx N(\mu, \sigma^2 / \lambda_i), \quad i=1,2,\ldots, n \tag{1.1}$$

şeklinde tanımlanmaktadır. Burada μ ve σ^2 sırasıyla konum ve yayılım parametrelerini, λ_i ise yayılımdaki değişim (shift in dispersion) parametresini ifade etmektedir. Karşıt hipotezde $X_1, X_2, ..., X_n$ tesadüfi örneğinde yer alan k tane tesadüfi değişkenin merkezi parametreleri birbirlerinden ve μ'den farklı normal dağılıma sahip olduğu şeklinde tanımlanmaktadır. Yani n çaplı örnekte yer alan gözlemlerden k tanesinden her biri, ortalamaları n-k tane gözlemin geldiği dağılımın ortalaması olan μ'den ve diğerlerinin sahip olduğu dağılımın ortalamasından farklı bir ortalamaya sahip normal dağılımdan geldiği şeklinde tanımlanmaktadır. Diğer bir ifadeyle karşıt hipotez,

$$H_k = \chi_r \approx N(\mu, \sigma^2 / \lambda r) \qquad \chi_t \approx N(\upsilon t, \sigma^2 / \lambda_t) \tag{1.2}$$

şeklinde tanımlanmaktadır. Burada r ve t indisleri r=1,2,...,n-k , t=1,2,...k değerlerini almaktadır (bu çalışmada örnekte yer alan gözlemlerden en büyük k tanesinin aykırı olabileceği önem taşımaktadır.). Ayrıca μ, σ, υ_t ve k $(0 \le k \le n - n/2 - 1)$ 'nın bilinmediği λ_r ve λ_t 'nin ise bilindiği varsayılmaktadır.

$X_1, X_2, ..., X_n$ tesadüfi örneğinin içerebileceği aykırı gözlem sayısını belirlemekte sıra istatistiklerinden faydalanarak yukarıdaki probleme çözüm getirebilecek en küçük karelere dayalı bir yöntem Wen-Liang Hung ve Jong-Wuu Wu (2005) tarafından ileri sürülmektedir. Bu yöntem aşağıdaki gibi tanımlanabilir. $X_1 \le X_2 \le ... \le X_n$, $X_1, X_2, ..., X_n$ tesadüfi örneğine ilişkin sıra istatistikleri olmak üzere, $X_{(n-k+1)}, X_{(n-k+2)}, ..., X_{(n)}$ sıra istatistiklerinin dağılımları tarafından üretilmiş $x_{(n-k+1)}, x_{(n-k+2)}, ..., x_{(n)}$ gözlem değerlerinin k tane üst aykırı gözlem olduğu varsayılsın. Bu durumda, hata kareler toplamının en küçüklenmesine dayalı en küçük kareler yöntemi ile k'nın değerinin ne olabileceğine karar verilebilir. Bu yöntemin uygulanışı diğer yöntemlere göre çok daha basit ve kolaydır. Ayrıca gizleme ve yanılgıya düşürme problemlerinden de etkilenmediği düşünülmektedir. Üst aykırı gözlem sayısının belirlenmesinin yanı sıra alt aykırı gözlem sayısının belirlenmesi veriye negatif dönüşümün uygulanması ile mümkün olur. Bölüm 2'de en küçük karelere dayalı olarak k'nın değerinin belirlenmesine ayrıntılı olarak yer verilecektir. Son bölümde simülasyon çalışması ile yöntemin doğru k sayısını belirleme oranı üzerinde durulacak ve bu oranın örnek çapının büyüklüğünden ve parametre tahmin değerlerinden nasıl etkilendiği tartışılacaktır.

2. EN KÜÇÜK KARELER

$X_1, X_2, ..., X_n$, Bölüm 1 'deki gibi tanımlanmış k tane aykırı gözlemi içeren tesadüfi bir örnek ve

$$\hat{Z}(X_{(i)}) = \frac{\sqrt{\lambda_i}\left(X_{(i)} - \mu\right)}{\sigma} \quad , \quad i = 1,2,...,n-k \tag{2.1}$$
$$\hat{Z}(X_{(i)}) = \frac{\sqrt{\lambda_i}\left(X_{(i)} - \upsilon_i\right)}{\sigma} \quad , \quad i = n-k+1, n-k+2,...,n$$

şeklinde bir tesadüfi değişken tanımlanmış olsun. Burada λ_i ve υ_i sırasıyla, $X_{(i)}$ sıra istatistiğinin dağılımına ilişkin yayılımdaki değişim ve konum parametrelerdir. F standart normal dağılımın birikimli dağılım fonksiyonu olmak üzere, $F^{-1}\left(i/(n+1)\right)$, $i = 1,2,...,n$, $\hat{Z}(X_{(i)})$ istatistiğine ilişkin dizinin yakınsama noktası olarak değerlendirilmektedir. En küçük karelerden hareketle k'nın değerine karar vermekte aşağıdaki kriterden yararlanılmaktadır..

$$\Phi_k(\mu,\sigma,\upsilon_{n-k+1},...,\upsilon_n) = \sum_{i=1}^{n-k}\left\{F^{-1}\left(i/(n+1)\right) - \frac{\sqrt{\lambda_i}\left(X_{(i)} - \mu\right)}{\sigma}\right\}^2 + \sum_{i=n-k+1}^{n}\left\{F^{-1}\left(i/(n+1)\right) - \frac{\sqrt{\lambda_i}\left(X_{(i)} - \upsilon_i\right)}{\sigma}\right\}^2 \tag{2.2}$$

ve

$$\Phi_0(\mu,\sigma,\upsilon_{n-k+1},...,\upsilon_n) = \sum_{i=1}^{n}\left\{F^{-1}\left(i/(n+1)\right) - \frac{\sqrt{\lambda_i}\left(X_{(i)} - \mu\right)}{\sigma}\right\}^2 \tag{2.3}$$

olmak üzere, $\hat{\Phi}_k(\hat{\mu},\hat{\sigma},\hat{\upsilon}_{n-k+1},...,\hat{\upsilon}_n)$ 'nın en küçüklenmesi ile üst aykırı gözlem sayısı olan k' nın optimal çözümünü elde edebiliriz ($\Phi_k(\mu,\sigma,\upsilon_{n-k+1},...,\upsilon_n)$'nın en küçüklenmesi ile $\mu,\sigma,\upsilon_{n-k+1},...,\upsilon_n$ 'nin en küçük kareler tahminleri sırasıyla $\hat{\mu},\hat{\sigma},\hat{\upsilon}_{n-k+1},...,\hat{\upsilon}_n$ olur). Yani $\hat{\Phi}_k(\hat{\mu},\hat{\sigma},\hat{\upsilon}_{n-k+1},...,\hat{\upsilon}_n)$ k 'nın her değeri için elde edilecek ve en küçük $\hat{\Phi}_k(\hat{\mu},\hat{\sigma},\hat{\upsilon}_{n-k+1},...,\hat{\upsilon}_n)$ değerinin elde edildiği k örneğin içerdiği üst aykırı gözlem sayısı olarak nitelenecektir.

Eğer X tesadüfi değişkeni normal dağılıma sahipse, (-X) tesadüfi değişkeni de normal dağılıma sahiptir. Bundan dolayı örneğin alt aykırı gözleme sahip olması durumunda, örneğe negatif bir dönüşüm uygulayarak yine aynı yöntemin uygulanması ile üst aykırı gözlem sayısı belirlenebilir. (-X) tesadüfi değişkeni üzerinden belirlenmiş üst aykırı gözlem sayısı, X tesadüfi değişkeni için alt aykırı gözlem sayısı anlamına gelmektedir. Dolayısıyla örnekte yer alabilecek üst ve alt aykırı gözlemlerin sayısına ilişkin bir tahmine, bu yöntem ile ulaşılabilir.

3. BENZETİM ÇALIŞMASI

Yöntemin örnekte yer alan aykırı gözlemlerin sayısını hangi oranda doğru olarak belirlediğini ortaya koymak amacına uygun bir Monte Carlo deney düzeni hazırlanmıştır. Bu deney düzenine göre, n çaplı bir örnek k tane üst aykırı gözlem içerecek şekilde normal dağılımdan üretilmektedir. Bu deney düzeninde n 'nin 10, 30, 50 ve 100 değerlerinden her biri için, k=0 (hiç üst aykırı gözlem olmaması), k=1 (örneğin bir tane üst aykırı gözlem içermesi), k=2 (örneğin iki tane üst aykırı gözlem içermesi), durumlarından her biri ayrı ayrı incelenmektedir. Örneğin, n=30 ve k=1 olması durumu için 500 adet örnek üretilecek ve birer tane üst aykırı gözlem içerdiği bilinen bu 500 örnekten kaç tanesinde, yöntemin gerçekten de bir tane üst aykırı gözlem tespit ettiği belirlenecektir. Böylece yöntemin üst aykırı gözlemlerin sayısını doğru olarak belirleme oranı elde edilecektir. Üretilecek örnekte yer alacak gözlemler ve k tane üst aykırı gözlem aşağıdaki gibi tanımlanmaktadır.

k=0 durumu için,

$\chi_i \approx N(1,1/\lambda_i)$, i = 1,2...,n

k=1 durumu için,

$\chi_i \approx N(1,1/\lambda_i)$ i = 1,2,..., n-1

$\chi_n \approx N(\upsilon_n,1/\lambda_i)$

k=2 durumu için,

$\chi_i \approx N(1,1/\lambda_i)$ i = 1,2,..., n-2

$\chi_{n-1} \approx N(\upsilon_{n-1},1/\lambda_{n-1})$ $\chi_n \approx N(\upsilon_n,1/\lambda_i)$, j=n-1, n

Farklı υ_j ve λ_j değerleri için, Monte Carlo simülasyon tekniğinin (tekrar sayısı 500) uygulanması ile yöntemin üst aykırı gözlem sayısı k'yı doğru belirleme oranları Tablo 1, Tablo 2, Tablo 3 ve Tablo 4' te yer almaktadır.

Tablo1. $\upsilon_n = 2$, $\lambda_1 = \lambda_2 = ... = \lambda_{n-1} = 5$, $\lambda_{n-1}, \lambda_n = 10$ için k' nın doğru belirlenme oranları

N	k=0	k=1	k=2
10	1	0.9458	1
30	1	0.7640	0.9380
50	1	0.7040	0.8220
100	1	0.5920	0.6640

Tablo 2. $\upsilon_n = 2$, $\lambda_1 = \lambda_2 = ... = \lambda_{n-1} = 5$, $\lambda_{n-1}, \lambda_n = 5$ için
k' nın doğru belirlenme oranları

n	k=0	k=1	k=2
10	1	0.8720	0.9760
30	1	0.7560	0.8715
50	1	0.6400	0.7640
100	1	0.5840	0.6980

Tablo 3. $\upsilon_n = 3$, $\lambda_1 = \lambda_2 = ... = \lambda_{n-1} = 5$, $\lambda_{n-1}, \lambda_n = 5$ için
k' nın doğru belirlenme oranları

n	k=0	k=1	k=2
10	1	1	1
30	1	1	1
50	1	0.9940	1
100	1	0.9960	1

Tablo 4. $\upsilon_n = 1.5$, $\lambda_1 = \lambda_2 = ... = \lambda_{n-1} = 5$, $\lambda_{n-1}, \lambda_n = 5$ için
k' nın doğru belirlenme oranları

n	k=0	k=1	k=2
10	0.9960	0.4920	0.5900
30	1	0.2960	0.2687
50	0.9980	0.1746	0.1701
100	1	0.1620	0.0924

4. SONUÇ

Tablo 1, Tablo 2, Tablo 3 ve Tablo 4'te yer alan sonuçlara bakıldığında, λ_i ve υ_n 'nin farklı değerleri için örnek çapı arttıkça örnekte yer alan üst aykırı gözlem sayısı k'nın ileri sürülen yöntem tarafından doğru belirlenmesi oranı düşmektedir. Bunun sebebi örnekte bir tane aykırı gözlem varken, bu gözlemin $\hat{\Phi}_k(\hat{\mu}, \hat{\sigma}, \hat{\upsilon}_{n-k+1}, ..., \hat{\upsilon}_n)$ üzerindeki etkisinin örnek çapının büyüklüğüne bağlı olarak azalmasıdır. Bu durum ise yöntemin gizleme probleminden etkilenmediği söylemiyle ters düşmektedir. Yani gerçekte aykırı bir gözlem içeren büyük çaplı bir örnek üzerinde bu yöntemin uygulanması sonucunda, gerçekte aykırı olan gözlemin aykırı olmayan n-1 gözlem

tarafından gizlenmesi ihtimali yükselmektedir. Ayrıca Tablo 3 ve Tablo 4'te yer alan sonuçlar karşılaştırıldığında, yöntem $\hat{\upsilon}_n$'nin $\mu = 1$ değerinden uzak değerleri için yüksek, yakın değerler için düşük doğru belirleme oranı vermektedir. Bu ise yöntemi, aykırı gözlemin geldiği düşünülen dağılımın parametrelerinin, μ ve σ^2/λ_i büyük ölçüde farklı olmasına bağlı kılmaktadır. $\hat{\upsilon}_n$'nın μ'den çok büyük değerleri için örnek çapı çok büyük olsa da yöntemin doğruyu belirleme oranı yüksek olacaktır. Tablo 1, Tablo 2, Tablo 3 ve Tablo 4'te k=2 durumuna ilişkin sonuçlar, aykırı gözlem sayısının örnek çapına oranı yüksek oldukça, yöntemin doğru sonuç vermesi oranının yüksek olacağını göstermektedir.

KAYNAKLAR

BARNETT V., LEWIS T. (1994). *Outlier in Statistical Data, third ed., Wiley*, Chichester.

CLARKE B. R., LEWIS T., (1998). *An Outlier Problem in the Determination of Ore Grade, Journal of Applied Statistics* 25, 751-762.

COOK, R. D., and WEISBERG, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

GUTTMAN, I. (1973b). *Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity- a Bayesian Approach. Technometrics*, 15, 723-738.

PEARSON, E. S, and CHANDRA SEKAR, C. (1936). *The Efficieny of Statistical Tools and a Criterion Forthe Rejection of Outling Observations. Biometrika*, 28, 308-320.

WEN-LIANG HUNG, JONG-WUU WU, (2005). *A Note on Determining the Number of Outliers in a Normal Sample with Unequal Variances by Least Squares Procedure. Applied Mathematics and Computation* 162 ,1007-1012.

DETECTING THE NUMBER OF OUTLIERS

ABSTRACT

According to Clarke and Lewis (1998) an outlier observation is generated by a similar distribution as of other observations, with the same location parameter but with different variance. Starting with this definition, Wen-Liang Hung, Jong-Wuu Wu (2005) have proposed a method based on minimizing square root error to determine the number of outliers in the sample. This method is more advantages compared to R statistic defined by Clarke and Lewis for its calculation is more simple and easier. Furtermore, it is said that this method is not affected from masking and swamping problems, in determining the number of lower and upper outlier observations in the sample generated from normal distribution. In this work, a simulation study is performed to detect how reliable results obtained by the method. The method is affected by the masking and swamping problems for the results showed that as the sample size is increased, the ratio of accurately determining the number of outlier observations decreases.

Key Words : *Least Square Method, Masking, Monte Carlo, Order Statistics, Outlier Observations, Swamping.*