SDÜ FEN EDEBİYAT FAKÜLTESİ
FEN DERGİSİ (E-DERGİ). 2008, 3(2) 219-229

REGRESYON ANALİZİNDE KULLANILAN EN KÜÇÜK KARELER VE EN KÜÇÜK MEDYAN KARELER YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Özlem GÜRÜNLÜ ALMA, Özgül VUPA
Dokuz Eylül Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, İzmir
email: ozlem.gurunlu@deu.edu.tr
Alınış: 08 Ağustos 2008, Kabul: 2 Ekim 2008

Özet: İstatistiksel yöntemler içerisinde yer alan regresyon çözümlemesi en çok kullanılan yöntemlerden biridir. Olası birçok regresyon yöntemlerinin dışında, genellikle matematiksel hesaplamalardaki kolaylığından dolayı, En Küçük Kareler yöntemi (EKK) en uygun tahmin yöntemi olarak kullanılmaktadır. Veri analizi ve ekonometri uygulamalarında EKK kestiricileri yaygın olarak tercih edilmektedir. Bununla birlikte EKK kestiricileri sapan değerlere karşı oldukça hassas olduğundan, veri kümesinin sapan değerler içermesi durumunda veriler hakkında EKK kestiricileriyle yapılacak yorumlamalar geçersiz ve yanıltıcı olabilmektedir. Bu gibi durumlarda sapan değerler için önerilen güçlü regresyon yöntemlerini tercih etmek, sonuçların güvenirliliği açısından daha uygundur. İstatistiksel çözümlemelerde kullanılan bu güçlü yöntemlerden biri de En Küçük Medyan Kareler yöntemidir (EKMK). Bu çalışmada, benzetim yoluyla oluşturulan veri kümelerinden yararlanılarak basit doğrusal regresyon modeli için EKK ve EKMK yöntemlerinden elde edilen model kestirim değerleri ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2) karşılaştırılmıştır.

Anahtar kelimeler: En Küçük Kareler Yöntemi, En Küçük Medyan Kareler Yöntemi, güçlü regresyon, sapan değer, benzetim çalışması

THE COMPARISON OF LEAST SQUARES AND LEAST MEDIAN SQUARES ESTIMATION METHODS WHICH ARE USED IN LINEAR REGRESSION ANALYSIS

Abstract: Regression analysis is one of the most commonly used statistical techniques. Out of many possible regression techniques, the Least Squares Method (LSM) has been generally adopted because of tradition and ease of computation. In data analysis and trend modelling applications the least squares (LS) estimator is widely used and LS regression is, in most cases, the method of choice. However, the crucial fact that the LS estimator is very sensitive to outlying observations may lead to unreliable results in the regression estimates and, hence, to a misleading interpretation of the data. To remedy this problem, some statistical techniques have been developed that are not so easily affected by outliers. These are the robust methods, the results of which remain trustworthy even if a certain amount of data is outlier. One of them is the least median squares method which is using in statistical analysis. In this study, estimation of Least Square and Least Median Square has been given. LS and LMS methods are applied and compared on differrent sample that can be produced by simulation study. To find whether there is important difference between methods are compared their estimations ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2).

Key words: Least Squares, Least Median of Squares, robust regression, outlier, simulation study

GİRİŞ

EKK regresyon yöntemi hata kareler toplamını en küçük yapmayı amaçlayan istatistiksel bir yöntemdir. Bu yöntem, gözlemlenen verilerin normallik, sabit varyanslılık, sapan değer içermeme gibi bazı varsayımların sağlandığı durumlarda güvenilir tahminler elde edilmesini sağlamaktadır (NETER vd. 1996, FOX 1997). İstatistiksel çözümlemelerde EKK yöntemi, matematiksel işlemlere en uygun tahmin yöntemi olarak kullanılsa da varsayımların ihlaline karşı olan dayanıksızlığı nedeniyle eleştirilmekte ve alternatif olarak daha güçlü yöntemler önerilmektedir (NETER vd. 1996, WILCOX 1997, ORTIZ vd. 2006, MOHEBBI vd. 2007). Regreyon çözümlemesinde varsayımların sağlanmadığı durumlardan biri de veri kümesinin sapan değer içermesidir. Sapan değer, bir veri kümesinde gözlemlerin çoğunun sahip olduğu dağılıma veya modele uymayan gözlemler olarak ifade edilebilir (BARNETT & LEWIS 1994). Sapan değer içeren veri kümesinde varsayımların sağlanamamasından dolayı kurulan regresyon modelinden alınan sonuçlarda yanıltıcı olmaktadır (GOODAL 1983, RYAN 1997). Bu nedenle regresyon çözümlemesinde veri analizi oldukça önemli bir yer tutmaktadır. Sapan değerlerin veri kümesinden çıkartılması regresyon denklemini tamamen veya kısmen değiştirebilmektedir. Bu nedenle büyük artık değerlere sahip olan gözlemler, regresyon çözümlemesinde oldukça etkilidirler. Böyle durumlarda sapan değerlerin tespiti ve sonuçların güvenirliliği için güçlü regresyon yöntemlerini tercih etmek daha uygundur (ROUSSEUW & LEROY 1987). Bu güçlü yöntemlerden biri de EKMK yöntemidir.

Bu çalışmada, EKK ve EKMK yöntemlerinin parametre kestirimleri üzerindeki etkinliği incelenmiştir. Bu doğrultuda, basit doğrusal regresyon modelinde bağımlı değişkenin farklı oranlarda sapan değerler içerdiği küçük örneklemler oluşturulmuştur. Bu örneklemlere ait regresyon modelinden elde edilen parametre kestirim değerleri karşılaştırılarak, EKK ve EKMK yöntemlerinin etkinliği araştırılmıştır.

MATERYAL VE METOT

Regresyon çözümlemesi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanarak o konu ile ilgili tahminler ya da kestirimler yapabilmek amacıyla kullanılan istatistiksel bir yöntemdir. Bu çözümleme yönteminde iki veya daha fazla değişken arasındaki ilişki açıklamak için matematiksel bir model kurulur ve bu model regresyon modeli olarak adlandırılır (BIRKES & DODGE 1993). İstatistiksel açıdan model kurulduktan sonra o modelin geçerliliğini araştırmak regresyon çözümlemesinin önemli bir parçasıdır. Kestirilen modelin gerçek modele ne kadar yaklaştığını belirleyebilmek için, kullanılan EKK yönteminin regresyon çözümlemesinin varsayımlarını sağlayıp sağlamadığının kontrolünün yapılması gerekmektedir. Eğer kurulan regresyon modeli veriye uygun değilse alınan sonuçlar da yanıltıcı olacaktır (WILCOX 1997).

Y bağımlı değişkeni, X_1 bağımsız değişkeni, β_1 bu değişkenin bilinmeyen parametresini ve ε_i gözlenemeyen hata terimlerini göstermek üzere kitle için basit doğrusal regresyon (BDR) denklemi

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \qquad i = 1,2,...,n \tag{1}$$

şeklinde yazılır. BDR çözümlemesinde bulunacak olan regresyon denklemlerinin kestirim amaçlı kullanılabilmesi için; hata terimlerinin $(\varepsilon_i = Y_i - \hat{Y}_i)$ rassal olup normal dağılım göstermesi, hataların beklenen değerinin 0 ve varyanslarının da sabit olup σ^2'e eşit olması, hataların birbirinden bağımsız olması $(\text{cov}(\varepsilon_i, \varepsilon_j) = 0)$, hata terimleri ile bağımlı değişken arasında korelasyonun olmaması gibi bazı varsayımların sağlanması gerekmektedir (FOX 1997). Bu varsayımlardan birisinin sağlanamaması durumunda EKK kestiricileri, gözlemler ve ön kestiriciler üzerindeki kararlı ve küçük varyanslı olma özelliğini kaybederek yanlı, tutarsız veya etkisiz olacaktır.

EN KÜÇÜK KARELER YÖNTEMİ

Günümüzde β_0 ve β_1 parametrelerinin tahmini için kullanılan en yaygın yöntemlerden birisi EKK yöntemidir. Kitle regresyon denkleminde yer alan β_0 ve β_1 parametrelerinin örneklemden elde edilen kestirimleri $\hat{\beta}_0$ ve $\hat{\beta}_1$ olarak ele alındığında, tek değişkenli regresyon doğrusunun denklemi

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}, \qquad i = 1,2,...,n \tag{2}$$

biçimindedir. Denklemde yer alan $\hat{\beta}_0$ ve $\hat{\beta}_1$ terimlerinin değerlerini bulmak için kullanılan EKK yönteminin temelini, toplam sapmaların karelerinin toplamını en küçük yapacak değerlerin bulunması oluşturmaktadır. Hata terimlerini, gözlemlenen Y_i değerleri ile beklenen $\hat{Y}_i$ değerleri arasındaki farklar oluşturmaktadır (RYAN 1997).

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \tag{3}$$

3. eşitlikte verilen ifade ile hesaplanan hata terimleri pozitif, negatif veya sıfır değerine sahip olurken bu farkların toplamı

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0 \tag{4}$$

olur. EKK yöntemi, β_0 ve β_1 parametrelerinin kestirimleri olan $\hat{\beta}_0$ ve $\hat{\beta}_1$'nın farkını en küçük yapacak biçimde aşağıdaki gibi belirler

$$\text{en küçük} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \text{en küçük} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \ . \tag{5}$$

Burada regresyon katsayılarının EKK tahminlerini elde edebilmek için 6. eşitlikte $\hat{\beta}_0$ ve $\hat{\beta}_1$'ya göre kısmi türevler alınıp sıfıra eşitlendiğinde 7. ve 8. eşitliklerdeki gibi I. ve II. normal eşitlikleri elde edilir. Bu eşitlikler üzerinden gerekli çözümlemeler yapıldığında

β_0 ve β_1 parametrelerinin kestirimleri olan $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerlerinin bulunabileceği eşitlikler 9 ve 10'da ki gibi elde edilir.

$$\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i}\right)\right)^2 = L \tag{6}$$

$$\sum_{i=1}^{n} Y_i = \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^{n} X_i \tag{7}$$

$$\sum_{i=1}^{n} X_{1i} Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^{\ 2} \tag{8}$$

$\hat{\beta}_1$, $\hat{\beta}_0$ ve regresyon belirtme katsayısının hesaplanması ise aşağıdaki gibidir.

$$\hat{\beta}_1 = \frac{n\left[\sum_{i=1}^{n} X_{1i} Y_i\right] - \left(\sum_{i=1}^{n} X_{1i}\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\left(\sum_{i=1}^{n} X_{1i}^2\right) - \left(\sum_{i=1}^{n} X_{1i}\right)^2} = \frac{\sum_{i=1}^{n}(X_{1i} - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_{1i} - \overline{X})^2} \tag{9}$$

$$\hat{\beta}_o = \frac{\sum_{i=1}^{n} Y_i + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}}{n} = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{10}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{11}$$

EN KÜÇÜK MEDYAN KARELER YÖNTEMİ

Varsayımların sağlanmadığı durumlarda güçlü regresyon tahmin edicileri EKK yöntemine alternatif olarak kullanılabilmektedir. Güçlü regresyon yöntemleri içerisinde EKMK en çok kullanılan tahmin yöntemlerinden biridir (ERICKSON vd. 2006). Rousseeuw 1984 yılında yapmış olduğu çalışmasında birçok örnekle gösterdiği gibi, veri kümesinde bir tane sapan değerin bulunması durumunda bile bu sapan değer, diğer bütün verilerden elde edilen bilgiye engel olmakta ve istatistikleri güvenilmez yapmaya başlamaktadır. DAVIES & GATHER (1993) tarafından geliştirilen, ortalama standart sapma ve aşırı student sapmaya bağlı olan Extreme Studentized Deviate testi, veri kümesinde sadece bir tane sapan değer olduğu durumlarda kullanılır. Ancak veri kümesinin birden fazla sapan değer içerdiği durumlarda bu değerler bazen birbirlerini maskeleyebilmekte ve hatta bu değerler klasik tahmin yöntemlerinde güvenilir verilerin bile sapan değer olarak görünmesine sebep olabilmektedir. EKMK yöntemi veri kümesinin %50'ye kadar sapan değer içerdiği durumlarda da iyi tahmin değerleri veren güçlü bir regresyon yöntemi olarak kullanılmaktadır (ROUSSEUW & LEROY 1987). Ancak EKMK yöntemi artıkların medyan değerini en küçük yapmayı amaçlarken geriye

kalan (n–1) adet gözlemi dikkate almaz. Bundan dolayı örneklem büyüklüğü arttıkça regresyon katsayılarının kestiriminde EKMK yöntemi EKK yöntemi kadar etkili olmamaya başlar (RYAN 1997).

WALD (1940), iki değişkenli bir örneklem kümesinde x gözlem değerlerinin medyanını temel alarak, bu gözlem değerlerinin serpme diyagramında veri setini sol ve sağ bölge olmak üzere ikiye ayıran basit bir yöntem önermiştir. Ayrılan her bölgenin x ve y gözlem değerlerinin ortalaması $((\overline{x}_{sag}, \overline{y}_{sag}), (\overline{x}_{sol}, \overline{y}_{sol}))$ şeklinde gösterilirken bu ortalamaların hesaplanması yalnızca o bölgeye ait x ve y gözlem değerleri kullanılarak elde edilir.

NAIR & SHRIVASTAVA (1942) tarafından önerilen yöntemde ise ilk olarak iki değişkenli veri setindeki x ve y değişkenleri kendi içlerinde sıralanır. Daha sonra sıralanan bu değişkenlerin ($x_1 \leq x_2 \ldots \leq x_n$), birbirine yakın olan değerleri aynı parçada olacak şekilde üç bölgeye ayrılır. Son olarak WALD (1940) yönteminde olduğu gibi ayrılan her bölgenin x ve y gözlem değerlerinin ortalaması $((\overline{x}_{sag}, \overline{y}_{sag}), (\overline{x}_{sol}, \overline{y}_{sol}))$, yalnızca o bölgeye ait x ve y gözlem değerleri

$$\overline{x}_{sol} = \frac{x_1 + x_2 + \ldots + x_{nsol}}{n_{sol}} \qquad \overline{y}_{sol} = \frac{y_1 + y_2 + \ldots + y_{nsol}}{n_{sol}} \tag{12}$$

$$\overline{x}_{sag} = \frac{x_{n-n_{sag+1}} + x_2 + \ldots + x_n}{n_{sag}} \qquad \overline{y}_{sag} = \frac{y_{n-n_{sag+1}} + y_2 + \ldots + y_n}{n_{sag}} \tag{13}$$

ifadeleriyle hesaplanır. Eşitliklerde, n_{sol}: ilk gruba ait gözlem sayılarını, n_{sag}: ikinci gruba ait gözlem sayılarını göstermektedir. Geriye kalan ($n-n_{sol}-n_{sag}$) adet gözlem veri kümesinden atılır. Burada birinci ve ikinci gruba ait gözlem sayıları (n/3) değerine yaklaşacak şekilde bir tamsayı değer olup her iki gruptaki gözlem sayıları da birbirine eşittir. Bu eşitlik $n_{sol} = n_{sag}$ olacak şekilde gösterilir. EKMK yönteminin uygulandığı basit doğrusal regresyona ait parametre kestirimleri

$$\hat{\beta}_0 = \overline{y}_{sol} - \hat{\beta}_1 \overline{x}_{sol} = \overline{y}_{sag} - \hat{\beta}_1 \overline{x}_{sag} \tag{14}$$

$$\hat{\beta}_1 = \frac{\overline{y}_{sag} - \overline{y}_{sol}}{\overline{x}_{sag} - \overline{x}_{sol}} \tag{15}$$

biçiminde hesaplanır. ROUSSEEUW (1984) EKMK tahminini 5. eşitlikte verilen amaç fonksiyonunda "Σ" yerine "medyan" koymak olarak tanımlar.

$$\underset{\hat{\beta}}{\text{en küçük medyan}} [Y_i - \hat{Y}_i]^2 \tag{16}$$

Ancak 16. eşitlikte verilen amaç fonksiyonundan analitik bir çözüm elde etmek oldukça güç olduğundan, β_i parametre tahminlerinin değerleri bilgisayar iterasyonları ile bulunabilir.

ROUSSEEUW (1984), ROUSSEEUW & LEROY (1987), EDELSBRUNNER & SOUVAINE (1990), OLSON (1997), MOUNT vd. (2007) parametre tahminlerinin elde edilmesini sağlayan iterasyonlar için çeşitli algoritmalar önermişlerdir. Yaygın olarak kullanılan ROUSSEEUW (1984) algoritmasında, n elemanlı bir veri kümesinin tüm mümkün p elemanlı alt kümelerine $\binom{n}{p}$ EKK yöntemi uygulanır ve her biri için artıkların medyan değeri hesaplanır. Bu medyan değerleri içerisinde en küçük medyan değerine sahip olan alt kümenin EKK tahminleri EKMK tahmini olarak kabul edilir. Küçük veri kümeleri için EKMK tahminlerinin kesin değerlerini hesaplamak bu şekilde mümkün olsa da, büyük veri kümelerinde mümkün olan tüm altkümelerin taranması ve EKK uygulanması işlem yükü açısından oldukça zor olacaktır. Bu durumda veri içerisinden bazı altkümelerin rastlantısal olarak çekilmesi ve amaç fonksiyonunun bu altkümelerde uygulanması düşünülebilir. ROUSSEEUW & LEROY (1987), belirli kısıtlar altında veriden rastlantısal olarak çekilecek en az bir altkümenin istenilen sonucu verme olasılığının 1'e yakın olduğunu ispatlamıştır. Buna göre bir veri kümesinden p elemanlı k tane altküme seçtiğimizde p tane aşırı olmayan değer içeren en az bir altkümeye rastlama olasılığını (n/p)'nin çok büyük değerleri için aşağıdaki ifadeye eşit olacağını belirtmiştir.

$$P_{\text{sapandeger\_icermeyen}} = 1 - [1 - (1 - \varepsilon)^{p}]^{k} \tag{17}$$

17. eşitlikte veri kümesinin kirlilik oranı ε ile gösterilmektedir. Bu ifade yardımıyla kirlenme oranının ε olduğu bir veriden p birimlik k tane alt kümeler çektiğimizde bunlardan en az birinin sapan değer içermeyen gözlemlerden oluşma olasılığı hesaplanır. Kirlenme oranının ε = %50 olduğu bir veriden 15 birimlik alt kümeler çektiğimizde bunlardan en az birinin sapan değer içermeyen gözlemlerden oluşma olasılığını $1 - [1 - (1 - \varepsilon)^{p}]^{k} = 0.95$ olması için çekmemiz gereken 15 birimlik alt kümelerin sayısı 0.98'dir.

EKMK yöntemi için standart sapma kestirimi ve regresyon modelinin açıklayıcılık katsayısı aşağıdaki gibi ifade edilebilir. s^0, gözlem sayısı ve açıklayıcı değişken sayısına (n: gözlem sayısı, m: açıklayıcı değişken sayısı) bağlı bir düzeltme çarpanıyla çarpılmasından elde edilir.

$$s^0 = 1{,}4826\left(1 + \frac{5}{n - m}\right)\sqrt{\underset{i=1,\ldots,n}{\text{med}}\,\sigma_i^2} \tag{18}$$

s^0 kestirimiyle standartlaştırılmış r_i / s^0 artıkları hesaplanır ve aşağıdaki gibi i. gözlemin w_i ağırlığını tanımlamada kullanılır.

$$w_i = \begin{cases} 1, & \left| r_i / s^0 \right| \le 2.5 \\ 0, & d.d \end{cases} \tag{19}$$

EKMK regresyonu için standart sapma kestirimi 20 no'lu eşitlikte verilen ifade ile hesaplanır.

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{n} w_i \sigma_i^2}{\sum_{i=1}^{n} w_i - m}} \qquad (20)$$

Burada σ^*, %50 kirlilik oranına sahip bir veri kümesi için standart sapma kestirimini gösterir. EKMK regresyonu için bağımlı değişkendeki değişimin ne kadarının model tarafından açıklandığını tanımlayan belirleyicilik katsayısı ise regresyon modelinin sabit terim içerdiği ve içermediği duruma göre aşağıdaki ifadelerden hesaplanmaktadır.

Sabit Terimli Regresyon Modeli

$$R^2 = 1 - \left(\frac{\text{med}|r_i|}{\text{mad}(y_i)} \right)^2$$

Sabit Terimi Olmayan Regresyon Modeli

$$R^2 = 1 - \left(\frac{\text{med}|r_i|}{\text{med}(y_i)} \right)^2 \qquad (21)$$

Burada mad=medyanın mutlak sapması (median absolute deviation) kısaltması olup $\text{mad}(y_i) = \left| y_i - \underset{i}{\text{med}}(y_i) \right|$ ile hesaplanır.

BULGULAR

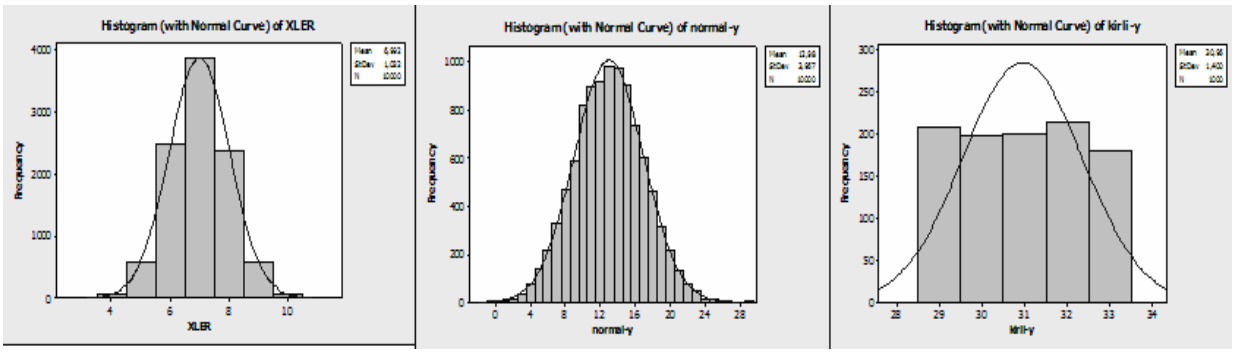
Basit doğrusal regresyon çözümlemesinde, küçük örneklemler üzerinde EKK ve EKMK yöntemlerinin etkinliğini karşılaştırmak için yapılan benzetim çalışmasında kullanılan kestiricilerin elde edilmesi bazı koşullar altında gerçekleştirilmiştir. Çalışmada, basit doğrusal regresyon modeli $Y_i = 1 + 2X_{1i} + \varepsilon_i$ olarak seçilmiş olup, bağımlı ve bağımsız değişken ile hata terimleri benzetim çalışması yapılarak türetilmiştir. Başlangıçta sapan değer içermeyen bağımlı değişkenin Y_i değerleri, X_{1i} bağımsız değikeni $N\sim(7,\ 1^2)$ parametreli normal dağılıma, hata terimleri ise standart normal dağılıma ($\varepsilon_i \sim N(0,1)$), sahip olacak şekilde 10,000 adet Minitab programı kullanılarak türetilmiştir. Böylece sapan değer içermeyen bağımlı değişken Y_i normal dağılıma sahip olacak şekilde elde edilmiştir ($Y_i \sim N(13,\ 4^2)$).

ROUSSEEUW & LEROY'in (1987) bir veri kümesinin içermiş olduğu sapan değerin yüzdesine bağlı olarak çekilecek olan örneklem büyüklüğüne göre seçilecek olan örneklem sayısını belirlemiş ve bu ifade 17. eşitlikte verilmişti. Bu eşitlik dikkate alınarak veri kümesinin %15 ve %25 oranında sapan değer içerdiği örneklem sayıları; 5, 10 ve 15 birimlik örneklem büyüklüklerine bağlı olarak belirlenmiş olup bu değerler Tablo 1'de verilmiştir.

Tablo 1. %15 ve %25'lik sapan değer içeren örneklemlerden birim sayısına göre çekilecek örneklem sayısı ve buna bağlı olarak elde edilen sapan değer sayıları

		Sapan Değer Yüzdesi			
		%15		%25	
		Çekilecek Örneklem Sayısı	Sapan Değer Sayısı	Çekilecek Örneklem Sayısı	Sapan Değer Sayısı
Örneklem Birim Sayısı	5	5	1	11	2
	10	14	2	52	3
	15	33	3	222	4

Tablo 1'deki değerler dikkate alınarak her bir örneklemin içermesi gereken sapan değer sayısının belirlenmesi ile bağımlı değişken, belirlenen gözlem sayılarına bağlı olarak kirletilmiştir. Kirletme işlemi için türetilecek olan sapan değerler bağımlı değişkenin ortalama değerinden en az 3σ uzaklıkta olacak şekilde oluşturulmuştur. Bu amaçla Y_i^* sapan değerleri, U~(29,33) parametreli Tekdüze dağılımdan gelecek şekilde türetilmiştir. Bağımsız değişken X_{1i}'nin, bağımlı değişken Y_i'nin ve sapan değer içeren Y_i^*'in histogramları Şekil 1'de verilmiştir.



Şekil 1. X_i, Y_i ve Y_i^* Değerleri Histogramları

EKK ve EKMK yöntemlerinin karşılaştırılması için gerekli veriler türetildikten sonra, EKK yöntemi Minitab programı, EKMK yöntemi ise SYSTAT programı kullanılarak uygulanmıştır. Her iki yöntem sonucunda da elde edilen parametre kestirimleri ($\hat{\beta}_0$, $\hat{\beta}_1$), modelin varyansı ($\hat{\sigma}^2$) ve belirtme katsayısı değerleri (R^2) arasında anlamlı bir farkın olup olmadığını karşılaştırmak için bağımlı t testi yapılmıştır. Bu testlere ait hipotezler Tablo 2'de belirtildiği gibidir.

Tablo 2. EKK ve EKMK yöntemlerinin parametre kestiricilerini karşılaştırmak için kurulan hipotezler

Hipotezler				
	$H_0 : \beta_{0.EKK} = \beta_{0\cdot EKMK}$	$H_0 : \beta_{1.EKK} = \beta_{1\cdot EKMK}$	$H_0 : \sigma^2_{EKK} = \sigma^2_{EKMK}$	$H_0 : R^2_{EKK} = R^2_{EKMK}$
	$H_1 : \beta_{0\cdot EKK} \neq \beta_{0\cdot EKMK}$	$H_1 : \beta_{1\cdot EKK} \neq \beta_{1\cdot EKMK}$	$H_1 : \sigma^2_{EKK} \geq \sigma^2_{EKMK}$	$H_1 : R^2_{EKK} \leq R^2_{EKMK}$

Karşılaştırma sonucunda elde edilen bağımlı t testi sonuçları Tablo 3'de verilmiştir.

Tablo 3. EKK ve EKMK yöntemlerinin parametre kestirimlerini karşılaştıran bağımlı t testi sonuçları

	Sapan Değer Yüzdesi			
	% 15		**% 25**	
Örneklem Birim Sayısı	**t**	**p-değeri**	**t**	**p-değeri**
	$\hat{\beta}_0$			
5	-0,84	0,450	0,36	0,725
10	2,02	0,065	2,27	0,027
15	0,01	0,989	1,16	0,931
	$\hat{\beta}_1$			
5	0,70	0,521	-0,07	0,945
10	-1,67	0,118	-1,61	0,114
15	0,82	0,418	-2,02	0,090
	$\hat{\sigma}^2$			
5	1,58	0,095	3,41	0,003
10	8,79	0,020	16,83	0,000
15	18,66	0,000	18,27	0,000
	$\mathbf{R^2}$			
5	-8,71	0,000	-8,63	0,000
10	-10,61	0,000	-14,84	0,000
15	-17,53	0,000	-21,32	0,000

Tablo 3'de verilen bağımlı t testi sonuçlarına göre, bağımlı değişkenin %15 ve %25 oranında sapan değer içermesi durumunda EKK ve EKMK yöntemlerinin parametre kestirimleri arasında anlamlı bir fark bulunamamıştır. Ancak modelin standart hatası ile modelin açıklayıcılık yüzdeleri arasında anlamlı bir farkın olduğu test sonuçlarından gözlemlenmiştir.

TARTIŞMA VE SONUÇ

Parametre kestirimi için kullanılan EKK yöntemi; hataların normal dağılması, sabit varyanslılık, bağımsızlık varsayımları ile değişkenlerin hatasız bir şekilde ölçüldüğü varsayımlarına dayanmaktadır. Ancak gerçek hayatta incelenecek olan veri kümesi için bu varsayımların her zaman sağlanması mümkün olmayabilir. Özellikle de veriler elde edilirken veya kaydedilirken meydana gelen hatalar, veri giriş hatası, skorlama hatası gibi nedenlerle veri kümelerinde diğer verilerden farklı gözlem değerlerine dönüşür. Dönüşen bu gözlemler literatürde sapan değerler olarak adlandırılır. Veri kümesinin sapan değer içermesi durumunda uygulanacak olan yöntemin daha güvenilir sonuçlar

vermesi için bu verilerin etkilerinin giderilmesi gerekmektedir. EKMK yöntemi veri kümesinin sapan değer içermesi durumunda EKK yöntemine göre daha güvenilir sonuçlar veren güçlü bir yöntemdir. Bu çalışmada her iki yöntemin etkinliğinin araştırılması için küçük örneklemler üzerinde bir benzetim çalışması yapılmıştır. Yapılan benzetim çalışmasında bağımlı değişkenin %15 ve %25 oranında sapan değer içerdiği durumlarda, kurulan $Y_i = 1 + 2X_{1i} + \varepsilon_i$ regresyon modelinde EKMK ve EKK yöntemlerinin parametre kestirimleri karşılaştırılmıştır. α = 0.05 anlam düzeyinde yapılan parametre kestirimleri karşılaştırmalarının bağımlı t testi sonuçları Tablo 3'de verilmiştir. Elde edilen bu sonuçlara göre:

- $\hat{\beta}_0$: için bağımlı değişkenin %15 oranında sapan değer içerdiği durumda örneklem büyüklüğünün n = 5, 10 ve 15 birim olduğu durumlarda her iki yöntemin parametre kestirim değerleri arasında anlamlı bir fark görülmemiştir. Benzer şekilde; bağımlı değişkenin %25 oranında sapan değer içermesi durumunda örneklem büyüklüğü n = 5 ve 15 için her iki yöntemin parametre kestirim değerleri arasında anlamlı bir fark görülmemiştir.
- $\hat{\beta}_1$: için bağımlı değişkenin sapan değer yüzdesinin ve örneklem büyüklüğünün parametre kestirimleri üzerinde her iki yöntem için anlamlı bir fark oluşturmadığı görülmüştür.
- $\hat{\sigma}^2$**:** regresyon modelinin varyans karşılaştırılmaları için kurulan $H_0 : \sigma^2_{EKK} = \sigma^2_{EKMK}$ ve $H_1 : \sigma^2_{EKK} \geq \sigma^2_{EKMK}$ hipotezlerine göre bağımlı değişkenin %15 ve %25 oranında sapan değer içermesi durumunda örneklem büyüklüğü n = 5, 10 ve 15 birim için her iki yöntem arasında varyans kestirim değerleri arasında anlamlı bir fark olup, genel olarak EKK yönteminden elde edilen model varyans değerinin EKMK yönteminden elde edilen değere göre daha büyük olduğu görülmektedir.
- R^2: belirtme katsayısı için kurulan $H_0 : R^2_{EKK} = R^2_{EKMK}$ ve $H_1 : R^2_{EKK} \leq R^2_{EKMK}$ hipotezlerine göre bağımlı değişkenin %15 ve %25 oranında sapan değer içerdiği durumda örneklem büyüklüğünün n = 5, 10 ve 15 birim için her iki yöntem arasında belirtme katsayılarının değerleri arasında anlamlı bir fark bulunamamış olup, genel olarak EKMK yönteminden elde edilen model belirtme katsayısı değerlerinin EKK yöntemine göre daha büyük olduğu görülmektedir.

Küçük örneklemler için EKK ve EKMK yöntemlerini karşılaştırmak amacıyla yapılan bu çalışmada genel olarak $\hat{\beta}_0$ ve $\hat{\beta}_1$ arasında anlamlı bir fark bulunamazken, EKMK regresyon modellerinin daha küçük varyansa sahip olduğu ve model açıklayıcılığını gösteren belirtme katsayısı değerlerininde daha büyük olduğu görülmüştür. Belirtme katsayılarından elde edilen bu anlamlı farklar nedeniyle veri kümesinin sapan değer içermesi durumlarında küçük örneklemler için EKMK parametre kestirim değerlerinin modeli daha iyi açıkladığı söylenebilir.

Sonuç olarak hata terimlerinin normal dağılmadığı veya bağımlı değişkenin sapan değer içermesi durumlarında küçük örneklemler için regresyon modelinde, EKMK yönteminin EKK yöntemine göre daha az etkilendiğini belirtebilir ve EKMK parametre kestirim değerlerinin regresyon modelini daha iyi açıkladığını söyleyebiliriz.

KAYNAKLAR

BARNETT V, LEWIS T, 1994. *Outliers in Statistical Data*. John Wiley Sons, Canada, pp.7–25.

BARRETO H, 2001. An Introduction to Least Median of Squares, www.wabash.edu/econexcel

BIRKES D, DODGE Y, 1993. *Alternative Methods of Regression*. John Wiley Sons, New York, pp.80–140.

DAVIES PL, GATHER U, 1993. The Identification of Multiple Outliers. *Journal of Statistical Planning and Inference*, 122, 65–78.

EDELSBRUNNER H, SOUVANIE L, 1990. Computing Least Median of Squares Regression Lines and Guided Topological Sweep. *Journal of the American Statistical Association*, 85(409), 115–119.

ERICKSON J, HAR-PELED S, MOUNT DM, 2006. On the Least Median Square Problem. *Discrete Comptutational Geometry.* 36, 593–607.

FOX J, 1997. *Applied Regression Analysis: Linear Models and Related Methods*. Sage Publication, USA, pp.123–240.

GOODAL C, 1983. Examining Residuals. In: HOAGLIN D & TUKEY J (Eds.) *Understanding Robust and Exploratory Data Analysis*. John Wiley Sons, Canada, pp.211–242.

KLEINBAUM, KUPPER, MULLER, and NIZAM, 1998. *Applied Regression Analysis and Other Multivariate Methods*. Duxbury, USA.

MOHEBBI M, NOURIJELYANI K, ZERAATI H, 2007. A Simulation Study on Robust Alternatives of Least Squares Regression. *Journal of Applied Sciences,* 7(22), 3469–3476.

MONTGOMERY D, HINES W, 1990. *Probability and Statistics in Engineering and Management Science*, John Wiley Sons, Canada.

MOUNT DM, NETANYAHU N, ROMANIK K, SILVERMAN R, WU AY, 2007. A Pratical Approximation Algorithm for The LMS Line Estimator. *Computational Statistics and Data Analysis,* 51, 2461–2486.

NAIR KR, SHRIVASTAVA MP, 1942. On a Simple Method of Curve Fitting. *Sankhaya,* 6, 121–132.

NETER J, KUTNER M, NACHTSHEIM C, and WASSERMAN W, 1996. *Applied Lineear Regression Models*, Irwin, USA.

OLSON CF, 1997. An Approximation Algorithm for Least Median of Squares Regression. *Information Processing Letters,* 63, 237–241.

ORTIZ M, SARABIA L, and HERRERO A, 2006. Robust Regression Techniques: A Useful Alternative for the Detection Data in Chemical Analysis. *Talanta,* 70, 499–512.

ROUSSEEUW JP, 1984. Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388), 871–880.

ROUSSEEUW P, LEROY A, 1987. *Robust Regression and Outlier Detection*. John Wiley Sons, Canada, pp. 84–143.

RYAN TP, 1997. *Modern Regression Methods.* John Wiley Sons, New York.

WALD A, 1940. The Fitting of Straight Lines if Both Variables are Subject to Error. A*nnals of Mathematical Statistic,* 11, 282–300.

WILCOX RR, 1997. *Introduction to Robust estimation and Hypothesis Testing.* Academic Press. San Diego.