Öneri.C.8.S.29.
Ocak 2008.199-207.

HATA TERİMLERİNİN PARETO VE WEIBULL DAĞILDIĞI DURUMDA LTS VE EKK REGRESYON KESTİRİCİLERİNİN KARŞILAŞTIRILMASI

Latif ÖZTÜRK

Kırıkkale Üniversitesi, İ.İ.B.F., İktisat Bölümü, Yardımcı Doçent Dr.

THE COMPARISON OF LTS AND OLS REGRESSION ESTIMATORS WHERE ERROR TERMS ARE DISTRUBETED AS PARETO AND WEIBULL

Abstract: *In this study, the historical developments, aims and necessity of robust estimators are investigated briefly. Then, the breakdown point of robust methods is tried to be explained. A simple linear regression model, with error terms distributed Weibull and Pareto, is constructed. By adding(with the addition of) leverage points on the direction of explained and explanatory variables, the solutions of S-Plus and unbiasedness results from simulation application with 10,50 and 100 run are taken. These result are displayed separately for both OLS estimator and robust estimator LTS. On the base of these results, when there are leverage points in data set, if the robust estimator LTS is used instead of OLS estimator, it is seen that the influence of leverage points on estimated parameters is minimised.*

Keywords: *Linear Regression, Robust Estimators, Leverage Point, Breakdown Point, Simulation*

HATA TERİMLERİNİN PARETO VE WEİBULL DAĞILDIĞI DURUMDA LTS VE EKK REGRESYON KESTİRİCİLERİNİN KARŞILAŞTIRILMASI

Özet: *Bu çalışmada sağlam kestirim yöntemlerinin tarihsel gelişimi, amaçları ve gerekliliğine kısaca değinilmiştir. Devamında, sağlam yöntemlerdeki dönüm noktası kavramına açıklık getirilmeye çalışılmıştır. Hata terimlerinin Weibull ve Pareto dağıldığı iki değişkenli doğrusal regresyon modeli oluşturulmuştur. Açıklanan ve açıklayıcı değişken yönünde aykırı değerlerin ilave edilmesi ile, S-Plus çözümleri ve benzetim uygulamasından 10,50 ve 100 döngü yapılarak yanlılık sonuçları elde edilmiştir. Bu sonuçlar hem sağlam bir kestirici olan LTS kestiricisi için hem de EKK kestiricisi için ayrı ayrı gösterilmiştir. Bu sonuçlar doğrultusunda, veri kümesinde aykırı değerlerin varlığında, EKK kestiricisi yerine sağlam bir kestirici olan LTS yönteminin kullanılmasıyla, parametrelerin kestirilmesinde aşırı değerlerin etkisinin minimize edildiği görülmüştür.*

Anahtar Kelimeler: *Doğrusal Regresyon, Sağlam Kestiriciler, Aykırı Değer, Dönüm Noktası, Benzetim*

I. GİRİŞ

İlk kez 1963 yılında Box "sağlam" (robust) sözcüğünü istatistiksel bir anlamda kullanmıştır. Bilim adamları varsayımlara bağlı olmayan, özellikle normallik varsayımına duyarsız yaklaşımları "sağlamlık" (Robustness) olarak tanımlamışlardır [1]. Sağlamlık model varsayımlarındaki sapmalarda sonucun kararlı ve değişmezliğine karşılık gelmektedir. Pratikte, verilerdeki küçük değişimin kestirimin varyansında büyük değişimlere neden olmayacağı anlamına gelmektedir. Sağlamlık kestirimdeki aykırı değerlere karşı dayanıklılığıyla tanımlanmaktadır [2].

Çoğu klasik istatistik yöntemleri için büyük hataların varlığı ve aykırı gözlem olarak da adlandırılan çeşitli hata değerlerinin var olması sorun teşkil etmektedir. Bu sorunun istatistiğin ilk zamanlarında bilinmesine rağmen, sağlam istatistiğin formüle edilmesi ve bu doğrultuda ölçekle elde edilmesi son yüzyıla rastlamaktadır. Bu geç kalışın nedeni yeterli derecede açık değilse de, büyük ve kompleks veri kümeleriyle uğraşmanın kolay olmaması nedeniyle günümüze sarktığı düşünülmektedir. Son yüzyılda ele alınmasının diğer bir nedeni de, E.S.Pearson, G.E.P.Box ve J.W.Tukey gibi bilim adamlarının çalışmalarından dolayı sağlam yöntemlere gereksinim duyulmasıdır. Günümüzde sağlamlık problemleri doğrultusunda çok sayıda çalışma vardır. Bazıları genel bazıları sağlamlık ile ilgili matematiksel görüşleri veya değişik topolojileri çoğu da parametrik olmayan istatistik alanlarındaki çalışmaları genişletmeyi yeğlemişlerdir [3]. Uzun-kuyruklu verilerin ele alınmasında ortalamadan ortancaya geçiş ve kesikli veya gruplanmış veriler için tepe değeri birer sağlam yöntemdirler.

II. AYKIRI DEĞERLERİN ATILMASI

Aykırı gözlemlerin atılmasının uygunluğu hakkındaki tartışmalar ilk kez Daniel Bernolli (1777), ile Bessel ve Baeyer (1838) tarafından yapılmıştır. Aynı zamanda Boscovich (1755)'in aykırı gözlemleri reddettiği bilinmektedir. İlk olarak aykırı gözlemlerin atılmasına karar vermek için Pierce (1852) ve Chauvenet (1863) tanımlamalar yapmışlardır. Bunları takiben, Stone (1868), Wright (1884), Irwin (1925), Student (1927), Thompson (1935), Pearson ve Chandra Sekar (1936) vd. bu doğrultuda çalışmalara devam etmişlerdir [3].

Aykırı değerlerin atılmasındaki amaç çeşitli olmasına karşın iki temel amaç göze çarpmaktadır: Birincisi, büyük hataların ortaya çıktığı gözlemlere dayanmaktadır. Açıkça aykırı gözlem olduğu belli olan basit bir büyük hata kullanılan istatistiksel yönteme çok zararlı olabileceğinden dolayı, bu büyük hata teşhis edilmeli ve haklı nedenlere dayanarak örneklemden çıkarılmalıdır. Bu gözlem değeri uygun bir gözlem değeri olsa da veri kümesinde tutulmasının maliyeti, atılmasındaki kaybolacak yarardan daha büyüktür. Bununla beraber, varsayılan model dağılımından bu derece uzak olan bir değerin uygun bir değer olma olasılığı çok düşüktür. Buradaki temel amaç istatistiksel analizin emniyetidir. Aykırı gözlemlerin atılması için kullanılan çeşitli yöntemleri ve diğer sağlam yöntemlere karşı olan bütün uzaktaki değerleri atma yaklaşımlarını, yarar ve emniyet arasındaki denge açısından dikkatlice incelemek gerekmektedir.

Kullanılan parametrik modelin yeterince doğru olmaması durumunda, uygun bir gözlem değeri kolayca modelin kapsamadığı bir aralıkta yer alabilmekte ve bu nedenle aykırı gözlem değeri olarak görülebilmektedir. Genellikle, verilerin çoğunluğuna uygun bir gözlemin atılmaması gerektiği fikri yaygın olarak benimsenmektedir. Fakat, uzaktaki uygun gözlem bize modelin doğru olmadığını göstermektedir. Eğer model uygun bir şekilde değiştirilirse, gözlemin parametre kestirimleri üzerindeki etkisinin büyük ölçüde azalacağı görülecektir. Eğer modeli değiştiremiyor veya değiştirmek istemiyorsak bu gözlemin kötü etkisini gidermek için en azından atılması sağlanarak zararsız hale getirilmelidir.

Kestirilen veya test edilen modellerin üzerinde aşırı düzeyde etkili bir gözlem bulunabilir, hatta söz konusu gözlemin modele potansiyel zararı bilinse de, atmak için haklı bir gerekçe yoktur. Çünkü, kestirilen parametre sadece bazı modellerde ortaya çıkan zararlar ile birlikte veri grubunun çoğu tarafından değil, verinin tamamı tarafından belirlenmektedir. Burada yapılacak olan işlem, aşırı düzeyde etkili olan gözlemler üzerinde kontrol sağlamak ve bunların potansiyel zararlarını ortadan kaldırmak için, onlar olmadan tecrübeye dayalı bir analiz yapmaktır. Yinede veri kümesi dışında herhangi bir ekstra bilgi olmaksızın bunları reddetmek ve etkilerini azaltmak için haklı bir neden yoktur.

Aykırı gözlemlerin atılmasındaki ikinci temel amaç ise özel işlemler için ilginç gözlemlerin toplanmasıdır. Bu gözlemler ilerleme kaydetmek için beklenmeyen veya incelenmeyen farklı gözlemler olabilirler, bunlar ilginç yeni bir modele veya yeni etkilere karar verebilirler. Bunlar daha detaylı olarak üzerinde çalışılarak düzeltilme olasılığı olan büyük hatalar olabilirler. Buradaki ikilem güvenilirlik ile etkinlik arasında olmaktan çok, azınlıktaki özel gözlemlere bakmak suretiyle verilerin özelliklerinin gözden kaçması ile çok fazla gözleme fazla zamanın ve enerjinin harcanması arasındadır [3]. Aykırı gözlemlerin veri kümesinden çıkarılmasında uygulanacak kurallar şöyle sıralanabilir:

1- Herhangi bir şekilde tam olarak aykırı gözlem olduğuna karar verilememiş gözlemlerin incelenmesi ve meydana gelecek en kötü durumun önlenmesi.

2- Daha yüksek etkinlik standardı sağlamak ve çoğu yöntemde olduğu gibi bazı gerçek durumlarda gereksiz bir şekilde en az %5-%20 etkinlik kaybını önlemek.

3- Genelde aykırı gözlemler sadece belirlenip birbirine uygun hale getirilmekten çok, yorumlanıp belki de düzeltilmeleri gereklidir.

4- Aykırı gözlemlerin belirlenmesi ve aykırı gözlem olmasından şüphe duyulan gözlemler, sağlam olmayan bir yöntemden kaynaklanıyorsa, sağlam bir yöntemden elde edilen artıklara bakılarak daha güvenli hale getirilebilir.

5- Bütün veriyi "iyi" ve "kötü" olarak iki ayrı gruba bölmek kavramsal olarak sade ve çekici olsa da, bütün özellikler dikkate alındığında aykırı gözlemlerin atılması sağlam yöntemler kadar verimli değildirler. Bunların özellikleri sağlamlık teorisinin içeriği içerisinde açıklanabilir [4].

III. SAĞLAM YÖNTEMLERİN AMAÇLARI

Yapılan çalışmalarda hata terimlerinin normal dağılmadığına sıkça rastlanmaktadır. Hata terimlerinin normal dağılmaması EKK yönteminin en küçük varyanslı etkin kestirici olma özelliğini kaybettirmemektedir. Bu durum karşısında geliştirilen alternatif yöntemler hata terimi normal dağılmadığında daha uygun ve etkin sonuçlar vermektedir. Dolayısıyla varsayım bozulmalarında EKK kestiricilerinde ortaya çıkan problemlerden kurtulmak için sağlam yöntemler bir yol olarak düşünülmektedir [5].

Regresyon doğrusu herhangi bir regresyon metoduyla kestirildiğinde, verilerdeki bir veya daha fazla aykırı değerin bulunması sonuçları taraflı hale getirmektedir. Buda regresyon katsayılarında yüksek hatalara neden olmaktadır. Dolayısıyla da, regresyon katsayılarına uygulanan istatistik testlerini önemli derecede etkilemektedir ve hatalı sonuçlara götürmektedir [6].

Sağlam yöntemlerin kullanım amaçları aşağıdaki gibi özetlenebilir:

1- Veri kümesinin önemli ve büyük bir kısmını kullanarak en iyi kestirimin yapısını tanımlamak.

2- Daha ileri analizler için aykırı gözlemleri ve altyapıyı tanımlamak.

3- Yüksek etkili gözlemleri (leverage points) tanımlamak.

4- Gözlemlerin bağımsız olduğu kabul edilip araştırmaların yapıldığı bazı bilim dalları için kuşku duyulan serisel bağımlılıklarla ilgilenmek [3].

Aykırı gözlemlerden dolayı meydana gelecek sorunları önlemek için bazı sağlam yöntemlere ihtiyaç duyulmaktadır. Genel olarak son 20 yılda geliştirilmiş iyi sağlam yöntemler %3 ile %30 veya daha fazla kazanılabilir etkinlik kaybını önlemek için gereklidirler. Veri kümesi ne kadar çok boyutlu ve kompleks olursa sıradan yöntemler o kadar az verimli olmaktadır. Dolayısıyla, modern sağlam yöntemlere daha fazla gereksinim duyulmaktadır.

IV. DÖNÜM NOKTASI

Dönüm noktası notasyonu ilk olarak Hodges tarafından 1967'de kullanılmış ve daha sonra Hampel (1968,1971) tarafından genelleştirilmiştir [7].

Şu çok iyi bilinmelidir ki, aykırı değerlerin istatistiksel analizler üzerinde bazı şeylerden mahrum edici kötü etkileri olmaktadır. Sağlam yöntemler, model varsayımlarının ihlallerinden kaynaklanan sonuçların içerilmesini ve aykırı değerleri birbirine uyumlu hale getirecek şekilde dizayn edilmişlerdir. Örneğin, örneklem medyanı konumun bir sağlam kestiricisidir. Buna karşılık örneklem ortalaması verideki aykırı değerlere daha çok duyarlıdır. Sağlamlığın aykırı değerlere karşı bir ölçüsü de dönüm noktasıdır. Bir kestirimin dönüm noktası kestirimi anlamsız yapmak için kirletilecek veya karıştırılabilecek örneklemdeki noktaların en küçük kısmıdır. Örneğin, örneklem ortalamasının sıfıra yakın bir dönüm noktası varken, medyanın 0.50'ye yakın bir dönüm noktası vardır.

İstatistiksel kestirimlerde, bir uzayda sonlu sayıdaki veri noktaları topluluğunu bir doğruya yerleştirmek temel problemdir. EKK gibi yöntemler kolay anlaşılır ve hesaplanır olmasına rağmen, az sayıdaki rasgele büyüklükte olan aykırı noktalar yerleştirilen doğruyu etkilemektedir. Bundan dolayı, sağlam kestirici olarak adlandırılan ve bahsedilen problemden fazlaca etkilenmeyen kestiricilere ilgi hızla artmaktadır. Kestiricinin gelişigüzel büyüklükte değer almasına neden olabilecek, aykırı veri noktalarının parçası (%50'ye kadar) olarak bir kestiricinin dönüm noktası tanımlanır. Bir kestiricinin dönüm noktası onun sağlamlığının ölçüsüdür. Örneğin, EKK'nın dönüm noktası asimptotik olarak sıfırdır. Çünkü, tek bir sapan değerin bile bu kestiricinin üzerinde gelişigüzel büyük bir etkisi olabilmektedir [8].

Yüksek dönüm noktasına sahip kestiriciler (HBE-High Breakdown Estimation) pek çok sayıda ve kötü pozisyonda bulunan aykırı değerler açısından güvenilir parametre kestirimleri elde etme problemine çözüm getirmektedirler. Çoklu regresyonda standart HBE'ler, LTS(Least Trimmed Squares) ve LMS(Least Median of Squares) kriterleri tarafından tanımlanan kriterlere sahiptirler. Her iki kriterde de veri kümesindeki n tane durumu yarıya bölme yoluna gitmişlerdir. Bu bölünen veri kümesinin yarısı kestirimde yer verilen veriler ve diğer yarısı ise herhangi bir aykırı gözlem içerdiği planlanan ve dikkate alınmayan kısımdır. LMS'deki kriter ele alınan durumların artık değerlerinin Chebyshew normudur, LTS'deki kriter ise ele alınan durumların artık değerlerinin kareleri toplamıdır [9].

LTS kestiricisinin çok iyi bir sağlamlık özelliği olan yaklaşık %50 dönüm noktasına sahip olma özelliği vardır (eğer h n'in doğru kesimiyse). Bir regresyon kestiricisinin dönüm noktası, sonuçtaki kestirimin Euclidean ölçüsünü $\|\hat{\beta}\|$ sonsuza göndermeden, verinin en büyük kısmını gelişigüzel büyük değerlerle değiştirmektir. Euclidean ölçüsü ise aşağıdaki gibi tanımlanmaktadır;

$$\|\hat{\beta}\| = \sum_{i=1}^{p} \hat{\beta}_i^2 \qquad (1)$$

Dönüm noktası kavramını daha iyi göstermek için kestiricinin örneklem ortalaması,

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (2)$$

olduğu, konumun kestirim problemini ele alalım. Ortalamanın kestiricisinin dönüm noktası sıfırdır, çünkü, eğer herhangi bir tek değer $y_i \to \pm\infty$ sonsuza giderse, örneklemin ortalaması da $\bar{y} \to \pm\infty$ sonsuza gidecektir. Diğer taraftan örneklem medyan'ının yaklaşık olarak %50 dönüm noktası vardır, çünkü uygunluk için örneklemin büyüklüğü n tek sayı olarak alındığında, medyanı $\pm$ sonsuza götürmeden, y_i değerlerinin $(n-1)/2$ kadarı $\pm$ sonsuza taşınabilir. Herhangi bir kestirici yaklaşık olarak %50 dönüm noktasına sahipse bu kestirici yüksek dönüm noktalı kestirici olarak adlandırılır. Bundan dolayı da, LTS kestiricisi bir yüksek dönüm noktalı kestiricidir.

V. LTS YÖNTEMİ

LTS regresyon yöntemi 1984'de Rousseeuw tarafından tanımlanmıştır. Doğrusal bir regresyon modeli kestiriminde yüksek derecede bir sağlamlığa sahiptir.

$r_i(\beta)$ 'nın i'inci artık olduğu durumda LTS kestiricisi olan $\hat{\beta}_{LTS}$, h tane kareleri alınmış en küçük artık değerlerin toplamını minimize etmektedir. Yani,

$$\sum_{i=1}^{h} r_i^2(\beta_{LTS}) \tag{3}$$

fonksiyonu minimize edilmektedir. h'in değeri genelde n değerinin yarısından biraz fazla olmasına özen gösterilmektedir. Daha detaylı bir şekilde açıklanacak olursa, ilk olarak karesi alınmış olan artık değerler aşağıdaki gibi küçükten büyüğe doğru sıralanmaktadır, $(r^2)_{1:n} \le (r^2)_{2:n} \le ..... \le (r^2)_{n:n}$ daha sonra bunların ilk h tanesi toplanmaktadır ve yukarıda da bahsedildiği gibi bu toplam,

$$\min_{\hat{\beta}} \sum_{i=1}^{h} (r)_{i:n}^2 \tag{4}$$

şeklinde minimize edilmektedir. $h = \lfloor n/2 \rfloor + 1$ şeklinde alındığında LTS kestiricisinin dönüm noktası asimptotik olarak %50'ye yaklaşmaktadır. $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ için LTS maksimum dönüm noktasına ulaşmaktadır[10]. Buna karşı EKK kestiricisi $\hat{\beta}_{LS}$ kareleri alınmış bütün artık değerlerin toplamı $\sum_{i=1}^{n} r_i^2(\beta)$ 'yı minimize etmektedir. EKK kestiricisi sağlamlık yönünden oldukça zayıftır, çünkü, herhangi bir tek değer (y_i, x_i) $\hat{\beta}_{LS}$ kestiricisinin gelişigüzel bir değer almasına neden olabilmektedir.

LTS kestiricisinin yüksek dönüm noktası, verilerin önemli esas kısmı, verinin %50 lik kısmından birazcık fazlasından oluşsa bile $x_i^T \hat{\beta}_{LTS}, i = 1,...,n$ değerlerinin verinin esas kısmına iyi uyduğu anlamına gelmektedir. Buna benzer olarak da,

$$r_i(\hat{\beta}_{LTS}) = y_i - x_i^T(\beta_{LTS}) \tag{5}$$

artık değerleri aykırı gözlemleri yeterli derecede açık olarak gösterecektir. EKK artık değerleri,

$$r_i(\hat{\beta}_{LS}) = y_i - x_i^T(\beta_{LS}) \tag{6}$$

ve M-kestirimin artık değerleri,

$$r_i(\hat{\beta}_M) = y_i - x_i^T(\beta_M) \tag{7}$$

sık sık verideki problemi gösterme açısından yetersiz kalmaktadır.

V.1. Özellikleri

1. LTS kestiricisinin her zaman için bir çözümü vardır.

2. LTS kestirim yöntemi regresyon ve ölçek olarak düzgün dönüştürülebilme özelliğine sahiptir.

3. Eğer p>1, $h = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ ve gözlemler genel düzende ise, yani, gözlemlerin p tanesi β için tek bir çözüm veriyorsa, LTS kestiriminin dönüm noktası;

$$\varepsilon^* = (\lfloor (n-p)/2 \rfloor + 1)/n \tag{8}$$

şeklindedir. Verilen koşullarda LTS kestiricisi %50 dönüm noktasına ulaşmaktadır. Genelde h, α gibi kesilmiş bir orana bağlı olmaktadır, şöyle ki, $h = \lfloor n(1-\alpha) \rfloor + 1$ olduğunda dönüm noktası $\varepsilon^* \approx \alpha$ dır.

4. Eğer p>1 ve genel düzendeki gözlemlerin $(n+p-1)/2$ den fazlasının $y_i = X_i\theta$'yı kesinlikle sağlıyorsa, diğer gözlem ler ne olursa olsun LTS çözümü θ'ya eşittir.

5. LTS kestirimcisi normal dağılımda,

$$\psi(t) = \begin{cases} t, & \text{eger} |t| < \Phi^{-1}(1-\alpha/2) \\ 0, & \text{diger durumda} \end{cases} \tag{9}$$

şeklinde M-kestiricisi olarak tanımlanan Huber-tipi atlayan (skipped) ortalama olarak da adlandırılan kestiriciyle aynı asimptotik etkinliğe sahiptir.

Yukarıdaki formüle dayalı olarak, $\alpha = 0.5$ olduğunda $(\varepsilon^* \cong \%50)$ asimptotik etkinlik ε yalnızca %7.2 olmaktadır. Aynı ε değerini L_1-kestirim yönteminde elde etmek istenildiğinde, dönüm noktası %7.5'e düşmektedir. LTS kestiriminin temel dezavantajı, amaç fonksiyonunun kareleri alınmış artık değerlerin sıralı bir şekline gereksinim duymasıdır. Bu da, ortancayla karşılaştırıldığında daha fazla işlem gerektirmektedir.

V.2. Algoritması

Buradaki algoritma p tane faklı gözlemin altkümelerini elde etmekle başlar. Bu alt kümeler $J=\{i_1,.....,i_p\}$ gösterilecek olursa, bu p noktaları ile regresyon yüzeyi belirlenip θ_J ile gösterilen regresyon katsayılar vektörü elde edilir. Bütün gözlemler kullanılarak her θ_J için LTS nin amaç fonksiyonu bulunur. Yani, $\sum_{i=1}^{h}(r^2)_{i:n}$ hesaplanır. Sonuçta bu amaç fonksiyonunu enküçük yapan katsayılar çözüm olarak alınırlar.

Genel olarak bu işlem tüm p sayıdaki altküme için tekrarlanır. Buda, C_n^p sayıda altkümeden oluşmaktadır. Bu sayı n ve p arttıkça hızlı bir şekilde artmaktadır. Çoğu uygulamada bütün alt kümeleri ele almak olanaksız hale gelmektedir. Bu durumda, ele alınacak m alt örneklemden an az bir tanesinin iyi olma olasılığının bire yakın olacağı şekilde çeşitli sayıda rastsal seçim yapılır.

Bu durum p=2 için özetlenecek olursa, (f,g), (f,h) ve (g,h) şeklinde üç adet alt küme alınıp ve f ve g ile başlanacak olursa, f ve g den geçen regresyon doğrusu;

$$\begin{aligned} y_f &= \theta_1^0 x_f + \theta_2^0 \\ y_g &= \theta_1^0 x_g + \theta_2^0 \end{aligned} \qquad (10)$$

eşitliklerinin çözümünden elde edilir. Burada, (x_f,y_f) ve (x_g,y_g) f ve g noktalarının koordinatlarıdır. Buradan elde edilen θ_1^0 ve θ_2^0 değerleri kullanılarak, artık değerler $y_i-\theta_1^0 x_i-\theta_2^0$ formülüyle bütün gözlem değerleri için elde edilir. Amaç fonksiyonumuzda artık değerler minimize edilmek istenildiğinden dolayı bulunan bu artık değerler toplamı diğer nokta çiftlerinin kullanımıyla elde edilen artık değerler toplamıyla karşılaştırılır. Karşılaştırma sonucunda en küçük değeri veren nokta çiftinden geçen doğru eniyi regresyon doğrusu olarak kabul edilir. Burada kullanılan katsayılarda eniyi çözümü veren katsayılar olarak alınır.

VI. UYGULAMANIN TANIMLANMASI

EKK ve LTS kestirim yöntemleri için benzetim programı oluşturup, bu program ile kestirim yöntemlerinin, hata terimlerinin Weibull ve Pareto dağılımlarının çeşitli parametre değerleri ile dağıldığı koşullar altında katsayılar kestirilerek bu katsayıların en iyi doğrusal yansız kestiriciler olduğunu göstermek amaçlanmıştır. Bunun içinde oluşturulan benzetim yazılımında, bir doğrusal regresyon modelinde katsayıların yansızlığı gösterilmeye çalışılmaktadır.

Weibull dağılımının olasılık yoğunluk ve birikimli yoğunluk fonksiyonları sırasıyla, $x \geq 0, a>0$ ve $b>0$ koşulları altında;

$$f(x)=\frac{b}{a}(x/a)^{b-1}e^{-(x/a)^b} \qquad (11)$$

$$F(x)=1-e^{-(x/a)^b} \qquad (12)$$

şeklinde tanımlanmaktadır. Pareto dağılımının olasılık yoğunluk ve birikimli yoğunluk fonksiyonları ise, $x \geq a>0$ ve $b>0$ koşullarıyla,

$$f(x)=\frac{b}{a}(a/x)^{b+1} \qquad (13)$$

$$F(x)=1-(a/x)^b \qquad (14)$$

şeklindedir. Birikimli yoğunluk fonksiyonlarının ters dönüşümü ve düzgün dağılım kullanılarak her iki dağılıma da karşılık gelen rastsal sayılar ise aşağıdaki gibidir.

Hata terimlerini Weibul dağılımından elde etmek için, $X=a(-\ln(1-U))^{1/b}$ ve Pareto dağılımından elde etmek için $X=a(1-U)^{-1/b}$ dönüşümleri kullanılmaktadır. Benzetim uygulamasında a parametresi için 1 ve 10, b parametresi için 2 ve 20 alınarak denemeler yapılmıştır. Dolayısıyla, parametrelerdeki değişimlerden kaynaklanacak olan yan (burada yan önceden belirlenen parametre değeri ile benzetim sonucunda elde edilen parametre değeri arasındaki farkı ifade etmektedir.) miktarları da benzetim sonucuna bakılarak görülebilecektir.

Regresyon modelindeki katsayılar basit regresyon modeli ele alındığından dolayı, regresyon sabiti θ_1 ve eğim θ_2'den oluşmaktadır. Basit doğrusal regresyon modelindeki açıklanan ve açıklayıcı iki değişkenden, açıklayıcı olan değişken verileri ele alınmakta ve bu veriler üzerinden rastsal olarak elde ettiğimiz hata terimleriyle birlikte açıklanan değişkenin değerleri elde edilmektedir. Daha sonra elde edilen bu değerlere kestirim yöntemi uygulanarak parametreler kestirilmektedir. Kestirilen bu parametreler ile benzetim başlangıcında modele dahil ettiğimiz regresyon parametreleri arasındaki fark ele alınmaktadır. Bu

farkların ($\left|\hat{\theta}_1 - \theta_1\right|$ ve $\left|\hat{\theta}_2 - \theta_2\right|$) büyük olması durumunda kestirimin yanlılığına, sıfıra yaklaşması durumunda ise kestirimin yansızlığına karar verilmektedir. Devamında, S-PLUS ile verilerin doğrusal regresyon modelinde ki katsayıları ve diğer parametreleri EKK ve LTS kestirim yöntemleri kullanılarak ayrı ayrı incelenip karşılaştırılmaktadırlar. Aynı zamanda verilerin bazı değerleri bozularak aykırı gözlem değerleri oluşturulmakta ve bu bozulan veri karşısında kestirim yöntemlerinin davranışları incelenmekte ve karşılaştırılmaktadır.

VII. BENZETİM UYGULAMASI

Benzetim uygulaması için oluşturulan yazılım QBX derleyicisi kullanılarak hazırlanmıştır. Hazırlanan bu yazılımda önce oluşturduğumuz basit regresyon modelindeki açıklayıcı değişken değerleri, hata değerleri için kullanacağımız dağılım ve bu dağılımın parametreleri ile modeldeki parametreler girilmektedir. Hata terimleri rastsal olarak düzgün bir dağılımdan çekilerek daha sonra Weibull ve Pareto dağılıma dönüştürülmektedir. Dönüştürüldükten sonra elde edilen bu rastsal hatalar ve açıklayıcı değişken değerleri de kullanılarak açıklanan değişkenin değerleri elde edilmektedir. Açıklayıcı ve açıklanan değişkenin değerleri de elde edildiğine göre, bunlara yukarıda bahsettiğimiz kestirim yöntemleri uygulanmış ve sonuçta bu uygulamanın hata terimlerinin dağılımını daha iyi yansıtması için deneme sayısı 10,50 ve 100 şeklinde artırılarak deneme yapılmış ve sonuçları alınmıştır. Bu denemeler sonucunda hesaplanan regresyon parametreleri, daha önce belirlediğimiz parametrelerle karşılaştırılmaktadırlar. LTS kestirim yönteminde girilen bütün verilerin alt kümeleri oluşturulmakta ve bu alt kümelere göre amaç fonksiyonumuzu minimum yapan alt küme tespit edilerek çözüme sokulmaktadır. Burada bahsedilen en iyi alt küme regresyon doğrusunun geçtiği iki noktadan oluşmaktadır. Bu sağlam yöntemde bütün noktalar ikişerli olarak alınıp alt kümeler oluşturulmakta ve bütün bu nokta kümelerinden regresyon doğrusu geçirilerek artık değerler hesaplanmaktadır. Artık değerleri minimum yapan regresyon doğrusunun geçtiği iki nokta en iyi alt küme olarak tanımlanmaktadır. Amaç fonksiyonu ise artık değerlerin minimum yapılmasında kullanılan ve artık değerlerin karelerinin toplamını gösteren bir fonksiyondur.

Benzetim uygulamasında kullanılacak veriler, açıklanan değişken olarak 1995 yılında Marmara bölgesinde meydana gelen trafik kazalarındaki ölüm sayısı ve açıklayıcı değişken olarak da bu bölgedeki toplam araç sayısı alınmıştır. Marmara bölgesindeki illerde meydana gelen kaza sayısı ve bu kazalardaki ölüm sayısı üzerine bir regresyon modeli kuracağımızda, illerdeki ölüm sayılarını açıklanan değişken ve bu illerde meydana gelen kaza sayılarını da açıklayıcı değişken olarak ele aldığımızda, verilerde hiçbir aykırı değer yaratmadan ve verilerde açıklanan ve açılayıcı değişken yönünde ayrı ayrı birer ve ikişer aykırı değer yaratıldığında EKK ve LTS kestirimcilerinden elde edilen sonuçlar aşağıda gösterilmiştir.

Tablo.1. 1995 Yılında Marmara Bölgesindeki İllerde Meydana Gelen Trafik Kazaları ve Ölüm Olaylarının Sayıları

İller	Kaza sayısı	Ölü sayısı
Balıkesir	3973	118
Bilecik	1020	33
Bursa	11615	203
Çanakkale	898	44
Edirne	1008	31
Yalova	494	7
İstanbul	98491	458
Kırklareli	765	53
Kocaeli	7748	131
Sakarya	3715	144
Tekirdağ	2126	82

EKK ve LTS kestiricileri için hata terimlerinin Weibull(1,2) ve (10,20) ile Pareto(1,2) ve (10,20) dağıldığı durumda orijinal değerler ve değişkenler yönündeki birer aykırı değer varlığında 10, 50 ve 100 deneme sonucunda benzetimden elde edilen parametrelerdeki yan miktarları ise aşağıdaki gibidir. Dağılımların farklı parametre değerlerinin ve deneme sayılarının farklı şekilde ele alınması kestirim üzerindeki etkilerini göstermektedir.

Tablo.2. 10 Deneme Sonucunda EKK ve LTS Benzetimlerinin Aykırı Değerler Karşısındaki Yan Miktarları

Kestirici Parametre	EKK Yan Miktarı		LTS Yan Miktarı		Hata Terimlerinin Dağılımı
	Sabit	Eğim	Sabit	Eğim	
Orijinal Değerler İçin	2.0383	0.0001	1.6313	0.0000	Pareto(1,2)
	10.585	0.0001	10.4435	0.0000	Pareto(10,20)
	0.9955	0.0001	0.7309	0.0000	Weibull(1,2)
	9.6534	0.0000	9.7186	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında	3345	1.0213	1.4181	0.0000	Pareto(1,2)
	3354	1.0213	10.4211	0.0000	Pareto(10,20)
	3344	1.0213	0.7729	0.0000	Weibull(1,2)
	3353	1.0213	9.6813	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında	106.43	1.2017	1.3529	0.0000	Pareto(1,2)
	97.98	1.2017	10.2967	0.0000	Pareto(10,20)
	107.63	1.2017	0.6154	0.0000	Weibull(1,2)
	98.59	1.2017	9.9120	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde İki Aykırı Değer Varlığında	46.035	0.0155	1.4624	0.0000	Pareto(1,2)
	37.27	0.0155	10.5648	0.0000	Pareto(10,20)
	47.155	0.0156	0.5849	0.0000	Weibull(1,2)
	38.004	0.0155	9.8462	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde İki Aykırı Değer Varlığında	10693	0.1220	1.6697	0.0000	Pareto(1,2)
	10701	0.1221	10.8408	0.0000	Pareto(10,20)
	10690	0.1220	0.9269	0.0000	Weibull(1,2)
	10701	0.1221	9.8675	0.0000	Weibull(10,20)

Tablo.2. 10 Deneme Sonucunda EKK ve LTS Benzetimlerinin Aykırı Değerler Karşısındaki Yan Miktarları (devam)

Kestirici Parametre	EKK Yan Miktarı		LTS Yan Miktarı		Hata Terimlerinin Dağılımı
	Sabit	Eğim	Sabit	Eğim	
Açıklayıcı ve Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında	13685	1.0805	1.3197	0.0000	Pareto(1,2)
	13693	1.0805	10.2569	0.0001	Pareto(10,20)
	13683	1.0805	0.5845	0.0001	Weibull(1,2)
	13692	1.0805	9.8102	0.0000	Weibull(10,20)
Açıklayıcı ve açıklanan Değişken Yönünde İki Aykırı Değer Varlığında	106.78	0.5945	1.3642	0.0000	Pareto(1,2)
	97.98	0.5945	10.3777	0.0000	Pareto(10,20)
	108.05	0.5945	0.4721	0.0000	Weibull(1,2)
	98.87	0.5945	9.7917	0.0001	Weibull(10,20)

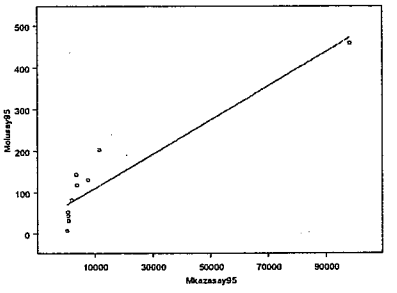
Tablo 3. 50 Deneme Sonucunda EKK ve LTS Benzetimlerinin Aykırı Değerler Karşısındaki Yan Miktarları

Kestirici Parametre	EKK Yan Miktarı		LTS Yan Miktarı		Hata Terimlerinin Dağılımı
	Sabit	Eğim	Sabit	Eğim	
Orijinal Değerler İçin	2.0424	0.0001	1.6069	0.0000	Pareto(1,2)
	10.512	0.0001	10.4279	0.0000	Pareto(10,20)
	0.8791	0.0000	0.8533	0.0000	Weibull(1,2)
	9.7538	0.0000	9.6796	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında	3345	1.0212	1.3783	0.0000	Pareto(1,2)
	3354	1.0213	10.4793	0.0000	Pareto(10,20)
	3344	1.0212	0.8741	0.0000	Weibull(1,2)
	3353	1.0213	9.7424	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında	45.61	0.0155	1.7631	0.0000	Pareto(1,2)
	37.21	0.0155	10.3651	0.0000	Pareto(10,20)
	47.18	0.0156	0.8485	0.0000	Weibull(1,2)
	38.043	0.0155	9.7046	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde İki Aykırı Değer Varlığında	106.51	1.2017	1.5542	0.0000	Pareto(1,2)
	97.97	1.2017	10.4199	0.0000	Pareto(10,20)
	107.64	1.2017	0.8241	0.0000	Weibull(1,2)
	98.72	1.0217	9.7867	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde İki Aykırı Değer Varlığında	10693	0.1220	1.6656	0.0000	Pareto(1,2)
	10701	0.1221	10.5375	0.0000	Pareto(10,20)
	10691	0.1220	0.8326	0.0000	Weibull(1,2)
	10701	0.1221	9.7750	0.0000	Weibull(10,20)
Açıklayıcı ve Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında	13685	1.0805	1.5341	0.0000	Pareto(1,2)
	13693	1.0805	10.4460	0.0000	Pareto(10,20)
	13684	1.0805	0.6588	0.0000	Weibull(1,2)
	13692	1.0805	9.8476	0.0001	Weibull(10,20)
Açıklayıcı ve Açıklanan Değişken Yönünde İki Aykırı Değer Varlığında	106.83	0.5945	1.6986	0.0000	Pareto(1,2)
	97.93	0.5945	10.1947	0.0000	Pareto(10,20)
	108.08	0.5945	0.8254	0.0000	Weibull(1,2)
	98.82	0.5945	9.9108	0.0001	Weibull(10,20)

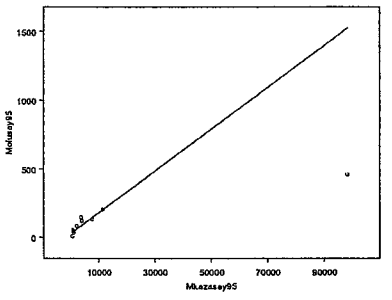
Tablo.4. 100 Deneme Sonucunda EKK ve LTS Benzetimlerinin Aykırı Değerler Karşısındaki Yan Miktarları

Kestirici Parametre	EKK Yan Miktarı		LTS Yan Miktarı		Hata Terimlerinin Dağılımı
	Sabit	Eğim	Sabit	Eğim	
Orijinal Değerler İçin	2.0982	0.0000	1.8779	0.0000	Pareto(1,2)
	10.529	0.0000	10.453	0.0000	Pareto(10,20)
	0.0403	0.0000	0.8843	0.0000	Weibull(1,2)
	9.7198	0.0000	9.8259	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında	3345	1.0213	2.0639	0.0002	Pareto(1,2)
	3354	1.0213	10.4207	0.0000	Pareto(10,20)
	3344	1.0212	0.8779	0.0000	Weibull(1,2)
	3353	1.0213	9.8641	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında	1628	0.5016	1.9001	0.0000	Pareto(1,2)
	37.22	0.0155	10.5024	0.0000	Pareto(10,20)
	1627	0.5016	0.8747	0.0000	Weibull(1,2)
	38.044	0.0155	9.7857	0.0000	Weibull(10,20)
Açıklayıcı Değişken Yönünde İki Aykırı Değer Varlığında	106	1.2017	1.8728	0.0000	Pareto(1,2)
	97.99	1.2017	10.5161	0.0000	Pareto(10,20)
	107	1.2017	0.8747	0.0000	Weibull(1,2)
	107.58	1.2017	9.7667	0.0000	Weibull(10,20)
Açıklanan Değişken Yönünde İki Aykırı Değer Varlığında	10692	0.1220	1.9304	0.0000	Pareto(1,2)
	10701	0.1221	10.5819	0.0000	Pareto(10,20)
	10690	0.1220	0.8185	0.0000	Weibull(1,2)
	10701	0.1221	9.8083	0.0000	Weibull(10,20)
Açıklayıcı ve Açıklanan Değişken Yönünde Birer Aykırı Değer Varlığında	13684	1.0805	1.8131	0.0000	Pareto(1,2)
	13693	1.0805	10.3919	0.0000	Pareto(10,20)
	13684	1.0805	0.8936	0.0000	Weibull(1,2)
	13692	1.0805	9.6689	0.0000	Weibull(10,20)
Açıklayıcı ve Açıklanan Değişken Yönünde İkişer Aykırı Değer Varlığında	106.92	0.5945	1.5577	0.0000	Pareto(1,2)
	98.05	0.5945	10.3682	0.0000	Pareto(10,20)
	107.98	0.5945	0.8578	0.0000	Weibull(1,2)
	98.82	0.5945	9.7412	0.0001	Weibull(10,20)

Yukarıdaki tablolarda da açıkça görüldüğü gibi aykırı değerlerin varlığında EKK kestiriminde yan miktarları aşırı derecede artmaktadır. LTS kestiriminde ise eğim parametresinin değeri değişmemekle birlikte sabit parametre her iki dağılımda da a parametresi kadar artmaktadır.



Şekil.1. Orijinal Veriler İçin Kaza Sayısı ve Ölüm Sayısının EKK Sonucu



Şekil.2. Orijinal Veriler İçin Kaza Sayısı ve Ölüm Sayısının LTS Sonucu

Orijinal veriler için sabit değer ve eğim parametrelerinin katsayıları ile determinasyon katsayısının değerleri ise aşağıdaki gibidir.

Tablo.5. Orijinal Değerler İçin EKK ve LTS Sonuçları

Kestirici	Sabit	Eğim	R^2
EKK	69.4986	0.0041	0.8644
LTS	26.0758	0.0152	0.8532

Aynı verilerde açıklayıcı ve açıklanan değişken yönlerinde ayrı ayrı birer ve ikişer aykırı değer yaratıldığında EKK ve LTS kestiricilerinin sonuçları aşağıdaki gibi değişmektedir.

Tablo.6. Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında

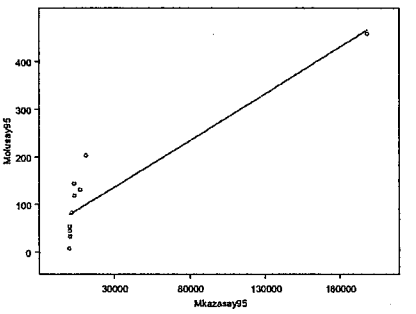
Kestirici	Sabit	Eğim	R^2
EKK	77.2190	0.0020	0.8256
LTS	26.0758	0.0152	0.8532

Tablo.7. Açıklanan değişken yönünde tek aykırı değer varlığında

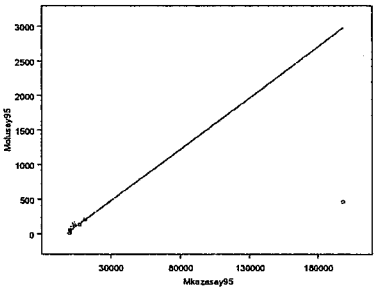
Kestirici	Sabit	Eğim	R^2
EKK	82.7904	-0.0001	0.0006036
LTS	30.1393	0.0149	0.789

Tablo.8. Açıklanan ve Açıklayıcı Değişken Yönünde Birer Aykırı Değer Varlığında

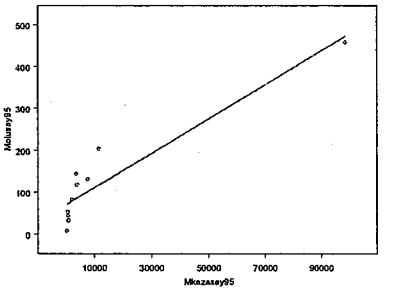
Kestirici	Sabit	Eğim	R^2
EKK	10416.815	-0.0573	0.01257
LTS	42.4240	0.0138	0.8338



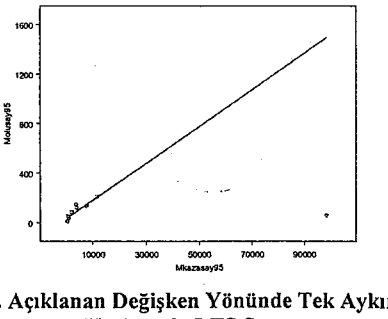
Şekil.3. Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında EKK Sonucu



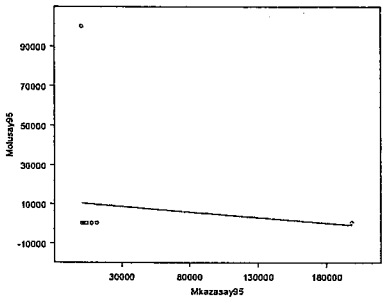
Şekil.4. Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında LTS Sonucu



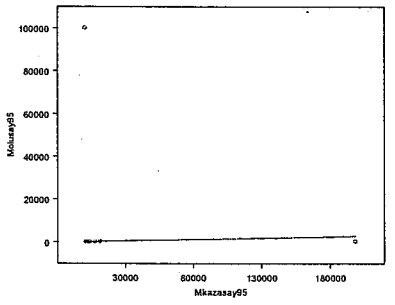
Şekil.5. Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında EKK Sonucu



Şekil.6. Açıklanan Değişken Yönünde Tek Aykırı Değer Varlığında LTS Sonucu



Şekil.7. Açıklanan ve Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında EKK Sonucu



Şekil.8. Açıklanan ve Açıklayıcı Değişken Yönünde Tek Aykırı Değer Varlığında LTS Sonucu

VIII. SONUÇ

Doğrusal regresyon çözümlemelerinde sıkça rastlanan verilerdeki aykırı değerler ve bu değerlerin kestirim üzerindeki olumsuz etkileri bilinmektedir. EKK kestirim metoduyla kestirilen parametreler bir aykırı değer varlığında bile aşırı etkilenmektedir. EKK'nın bu zaafını göstermek için orijinal veriler ve bu verilerde açıklayıcı ve açıklanan değişken yönünde aykırı değerler oluşturularak, hata terimlerinin Pareto ve Weibull dağılımlarının (1,2) ve (10,20) parametreleriyle dağıldığı durumlar için sonuçlar alınmıştır. Yukarıdaki tablolardan da açıkça görüleceği gibi her üç deneme(10,50 ve 100) sonucunda da orijinal verilerde EKK ve LTS kestirim metotlarındaki yan miktarları birbirine çok yakın olmasına rağmen, açıklanan ve açıklayıcı değişkenler yönünde aykırı değerlerin varlığında EKK metodunda eğim parametresinin yan miktarı artmasına karşılık LTS metodunda yan miktarında bir değişim söz konusu değildir. Sabit parametre de ise aynı şekilde bir artış söz konusudur. LTS'de sadece parametre değeri kadar bir yan miktarı oluşurken EKK'de bu miktar aşırı derecede artmaktadır. Aynı verilerle S-PLUS uygulaması yapılıp grafik ve R^2 değerlerinin gösterilmesiyle de daha açık bir hal almaktadır. Aykırı değerler varlığında EKK'de kestirim tamamen aykırı değerlere yönelmesine karşılık LTS bu değerlerden etkilenmemekte ve verinin aykırı değer olmayan kısmı ile kestirimi yapmaktadır. Bu uygulama farklı verilerle ve farklı aykırı değerlerin varlığında değişik sonuçlar vermesi de olasıdır.

YARARLANILAN KAYNAKLAR

[1] Stigler, S.M. (1973). Simon Newcomb, Percy Daniell and the History of Robust Estimation, 1885-1920. *Journal of the American Statistical Association*, 68(344), 872-879

[2] Western, B. (1995). Concepts and suggestions for robust regression analysis. *Amarican Journal of Political Science*, 39(3), 786-817.

[3] Hampel, F.R., Ronchetti, E. M.., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics. The Approach Based on Influence Functions. New York: John Willey & Sons.

[4] Pena, D. & Yahoi, V. (1999). A Fast Procedure for Outlier Diagnostics in Large Regression Problems. *Journal of the American Statistical Association*, 94(446), 434-445.

[5] Büyüklü, A.H. (1999). Robust Tahmin Edicilerin Kullanılma Sebep ve Sonuçları. *IV. Ulusal Ekonometri ve İstatistik Sempozyumu.* Antalya, 655.

[6] Rio, F.J.; Rio, J. & Rius, F.X. (2001). Linear regression taking into account errors in both axes in the presence of outliers. *Analytical Letters*, 34(14), 2547-2561.

[7] Yijun, Z. (2001). Some quantitative relationships between two types of finite sample breakdown point. *Statistics and Probability Letters*, 51(4), 369-375.

[8] Motoujek, J.; Mount, D.M. & Natenyahu, N.S. (1998). Efficient Randomized Algorithms For The Repeated Median Line Estimator. *Algorithmica*, 20(2), 136-150.

[9] Douglas, M.H. & David, O. (1999). Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis,* 32(2), 119-134.

[10] Rousseeuw, P.J. & Leroy, M.A. (1987). *Robust Regression and Outlier Detection.* New York: John Wiley & Sons.

Latif ÖZTÜRK (latifozturk@kku.edu.tr) has Ph.D. of Robust Estimation Techniques in Linear Regression and Their Comparisons at Mimar Sinan University in 2003. His research areas are regression, robust estimators and simulation.