İstatistik Araştırma Dergisi Cilt 02 No 03 Sayfa 1-10 Aralık 2003

Genelleştirilmiş T (Gt) Dağılımına Dayalı Regresyon Analizi

Ali İhsan GENÇ Olcay ARSLAN*

ÖZET

Bu çalışmada $y = X\beta + u$ çoklu lineer regresyon modelindeki u hata teriminin 0 ortalamalı ve σ ölçek parametreli genelleştirilmiş t (GT) dağılımından geldiği kabul edilmiştir. GT dağılımının şekil parametrelerinin bilindiği varsayımı altında regresyon modelinin parametreleri ve σ ölçek parametresi tahmin edilmiştir. Önerilen kestirim yöntemi bir takım problemli veri kümelerine uygulanmış ve alınan sonuçlar diğer dayanıklı (robust) kestirimlerle karşılaştırılmıştır.

***Anahtar Kelimeler**: Maksimum Olabilirlik; Sapan Değer; Regresyon; Dayanıklı Kestirim; Etki Fonksiyonu.*

1. GİRİŞ

Regresyon analizinin hem teorisinde hem de uygulamasında modeldeki hataların çoğunlukla normal (Gauss), bağımsız ve ortak bir varyansla dağıldığı varsayılır. Fakat bazı veri kümeleri için hataların dağılımı normalden daha kalın kuyruklu olabilir ve bu durumda normal dağılıma dayalı yapılan analizler doğru sonuç vermeyebilir. Literatürde kuyruk çeşitliliği bakımından zengin pek çok dağılım regresyonda hataların modellenmesinde normal dağılma alternatif olarak kullanılmıştır. Örneğin, Laplace dağılımı (Bloomfield ve Steiger, 1983), t dağılımı (Zellner, 1976, Lange vd., 1989, Arslan, 1992), *BT* (Box-Tiao) dağılımı (Klein ve Spady, 1984), genelleştirilmiş üstel dağılımlar (Lye ve Martin, 1993). Çalışmada hataların modellenmesi için önerilen *GT* dağılımı ise başta t dağılımı ve *BT* dağılımı olmak üzere bir çok dağılımı parametrelerinin özel durumlarında veya limit durumlarında içerdiğinden hataların modellenmesinde daha genel bir dağılım sınıfı oluşturmaktadır.

2. GENELLEŞTİRİLMİŞ T (GT) DAĞILIMI

McDonald ve Newey (1988) *GT* dağılımını normal ve t dağılımının bir alternatifi olarak regresyonda hataları modelleyip bir kısmen adapte kestirim (partially adaptive estimation) yöntemi geliştirmek için tanımlanmıştır. Kullanılan yöntem en küçük kareler (EKK), en küçük mutlak sapma (LAD-Least Absolute Deviation), L_p gibi kestiricileri özel durumları olarak içermektedir. Bu yöntem bir çok ekonomik modeli kestirmek için kullanılmıştır. Örneğin, market modelinde (McDonald ve Nelson, 1989, Butler vd., 1990), ARMA zaman serileri modellerinde (McDonald, 1989).

* Çukurova Üniversitesi Fen-Edebiyat Fakültesi Matematik Bölümü 01330 Balcalı-Adana

İstatistiksel ekonomideki kullanışlığına rağmen *GT* dağılımının parametrik özellikleri literatürde yeterince ele alınmamıştır. Arslan ve Genç (2002) tek değişkenli bir veri kümesini *GT* dağılımıyla modelleyip konum ve ölçek parametrelerinin maksimum olabilirlik kestiricilerini buldular. Ayrıca konum ve ölçek kestiricilerinin parametre uzayında tek olarak mevcut olabilmesi için $p \geq 2$, $pq \geq 1$ yeter koşulunu elde ettiler.

GT dağılımı aşağıda (1) denklemiyle verilen simetrik tek modlu bir yoğunluğa sahiptir:

$$f(x;\mu,\sigma,p,q) = \frac{pq^q}{2\sigma B(1/p,q)\sigma}\left(q + \frac{|x-\mu|^p}{\sigma^p}\right)^{-q-1/p}. \qquad (1)$$

Yoğunluk fonksiyonunda $B(.)$ beta fonksiyonu, $\sigma \in (0,\infty)$ ölçek parametresi, $-\infty < \mu < \infty$ konum parametresi ve $p>0$ ve $q>0$ da şekil parametreleridir. p ve q nun büyük değerleri yoğunluk fonksiyonunun kuyruklarını inceltirken küçük değerleri daha kalın kuyruklu yoğunluklara neden olur. Ayrıca bu parametrelerin çeşitli değerlerinde ve limit durumlarında çeşitli dağılımları elde ederiz. Örneğin, $p=2$ için t dağılımı, $q \to \infty$ için *BT* dağılımını ve $p \to \infty$ için de $(-\sigma,\sigma)$ üzerinde düzgün dağılımı elde ederiz (McDonald ve Newey, 1988).

3. PARAMETRELERİN KESTİRİMİ

y bağımlı değişken üzerindeki bir $n\times 1$ vektör, X, k regresör değişkeninin değerlerinin $n\times p$ matrisi olmak üzere

$$y = X\beta + u, \qquad (2)$$

çoklu lineer regresyonu ele alalım. Burada u gözlenemez rasgele hataların $n\times 1$ vektörü, β ise modeldeki bilinmeyen parametrelerin $p\times 1$ vektörüdür.

(2) denklemindeki u_i hatalarının bağımsız ve özdeş olarak 0 ortalamalı, bilinmeyen σ ölçek parametreli *GT* dağılımına sahip olduğunu varsayalım. Şekil parametreleri p ve q nun bilindiğini kabul edelim. O zaman x_i, X matrisinin $i.$ satırı olmak üzere y_i nin yoğunluk fonksiyonu

$$f(y_i) = \frac{pq^q}{2\sigma B(1/p,q)}\left(q + \frac{|y_i - x_i\beta|^p}{\sigma^p}\right)^{-q-1/p},\ -\infty < y_i < \infty,\ i=1,2,...,n \qquad (3)$$

ve log-olabilirlik fonksiyonu

$$l(\beta,\sigma) = -n\log\sigma - \left(q + \frac{1}{p}\right)\sum_{i=1}^{n}\log\left(q + \frac{|y_i - x_i\beta|^p}{\sigma^p}\right), \qquad (4)$$

dir. (4) denkleminin β ve σ ya göre kısmi türevlerinin alınıp 0 a eşitlenmesiyle aşağıda verilen olabilirlik kestirim denklemleri elde edilir.

$$\frac{\partial l}{\partial \beta} = (q+\frac{1}{p})\sum_{i=1}^{n} \frac{\sigma^{-p} p x_i' \mid y_i - x_i\beta \mid^{p-1} sign(y_i - x_i\beta)}{q + \frac{\mid y_i - x_i\beta \mid^p}{\sigma^p}} = 0\,,\ p>1 \qquad (5)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + (q+\frac{1}{p})\sum_{i=1}^{n} \frac{p\sigma^{-p-1} \mid y_i - x_i\beta \mid^p}{q + \frac{\mid y_i - x_i\beta \mid^p}{\sigma^p}} = 0\,. \qquad (6)$$

(5) ve (6) da verilen kestirim denklemlerinin yeniden düzenlenmesiyle aşağıdaki denklemleri elde ederiz.

$$\sum_{i=1}^{n} w_i x_i' (x_i\hat{\beta} - y_i) = 0\,, \qquad (7)$$

$$\sum_{i=1}^{n} w_i (y_i - x_i\hat{\beta})^2 - n\hat{\sigma}^2 = 0\,, \qquad (8)$$

ağırlıklar $w_i = \dfrac{(pq+1) \mid y_i - x_i\hat{\beta} \mid^{p-2} \hat{\sigma}^{2-p}}{q + \mid y_i - x_i\hat{\beta} \mid^p \hat{\sigma}^p}$ dir. Bu denklemler sıfıra azalan M-kestirim denklemleridir. (7) ve (8) denklemlerinden regresyon parametrelerinin kestirimi aşağıdaki gibi yinelemeli tekrar ağırlıklandırılan en küçük kareler kestirimi şeklinde yazılabilir.

$$\hat{\beta} = (\sum_{i=1}^{n} w_i x_i' x_i)^{-1} (\sum_{i=1}^{n} w_i y_i x_i')\,, \qquad (9)$$

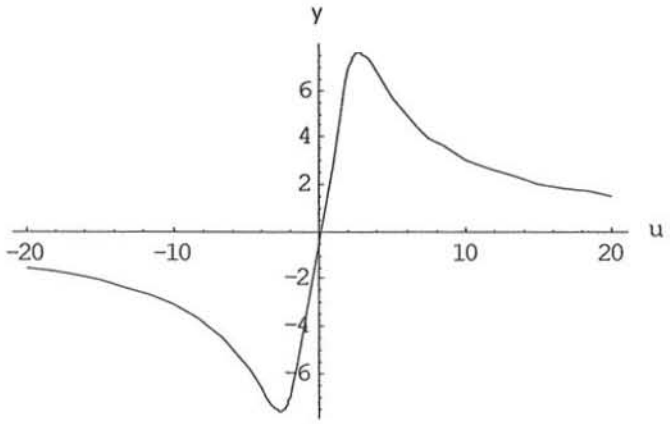
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} w_i (y_i - x_i\hat{\beta})^2\,. \qquad (10)$$

Hatalar *GT* dağılımlı olduğunda kestiricilerin hesaplanmasında kullanılacak yinelemeli yöntemin ağırlıklı en küçük kareler yönteminden farkı ağırlıkların sabit kalmayıp her iterasyonda değişecek olmasıdır. Üstelik *w* ağırlık fonksiyonu rezidülerin azalan bir fonksiyonu olduğundan büyük rezidülere küçük ağırlıklar verilecektir. Böylece veri kümemizdeki sapan değerlere küçük ağırlıklar verileceğinden elde edeceğimiz kestirimler sapan değerlerden çok az etkilenecektir.

$u_i = y_i - x_i\beta$ hatalarının ψ-fonksiyonu $p>1$ için

$$\psi(u) = \frac{(pq+1) \mid u \mid^{p-1} sign(u)}{q\sigma^p + \mid u \mid^p} \qquad (11)$$

olarak elde edilir (McDonald ve Newey, 1988) (Şekil 1). ψ -fonksiyonunun rezidülerin ağırlık fonksiyonuyla arasındaki ilişkisi $\psi(u) = w(u)u\sigma^{p-2}$ bağıntısıyla verilir. $u>0$ için ψ -fonksiyonu $(0,[(p-1)q\sigma^p]^{1/p})$ aralığında artan, $([(p-1)q\sigma^p]^{1/p},\infty)$ aralığında ise azalandır. Üstelik $\lim_{|u|\to\infty}\psi(u)=0$ olduğundan regresyon katsayılarının etki fonksiyonu sıfıra azalan (redescending) tipindedir. ψ -fonksiyonu p ve q şekil parametrelerine bağlı olduğundan küçük ağırlıklar tahsis edilecek olan sapan değer sayısı p ve q nun değerlerine göre azalacak veya artacaktır. Bu yüzden p ve q şekil parametrelerine ayar sabitleri (tuning constants) olarak bakılabilir.



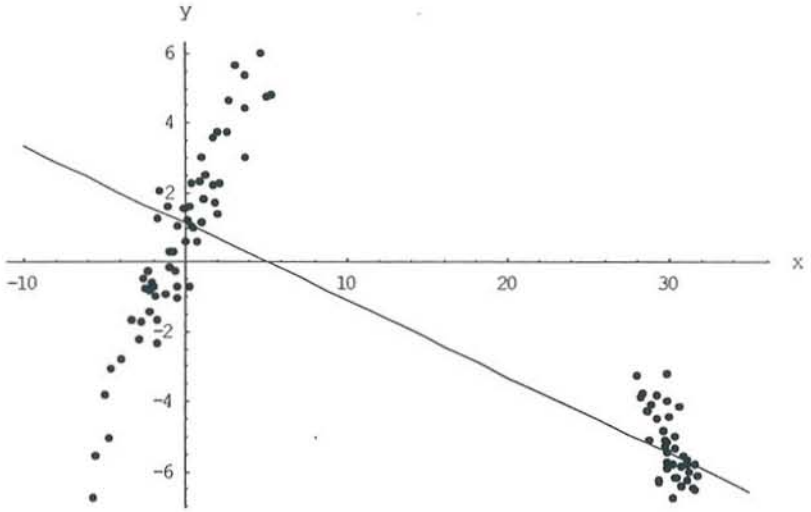
Şekil 1. ψ -fonksiyonu (p=3, q=10)

4. ÖRNEKLER

Örnek 1. Bu örnekte Arslan (2002) tarafından üretilen suni veriyi analiz edeceğiz. Veri kümesi içerisinde iki altgrubu barındıran 100 noktadan oluşmaktadır. Serpme diyagramından da görüldüğü gibi 40 noktalık sapan değerlerden oluşan bir alt grup x-uzayı yönünde bulunmaktadır (Şekil 2). Parametre kestirimleri aşağıdaki Tablo 1'de verilmiştir. İterasyona başlangıç değeri olarak en küçük kareler (EKK) kestirimleri alınmıştır. Tablo 1'den çeşitli yöntemler için benzer kestirimlerin elde edildiğini ve regresyon doğrusunun veri kümesindeki sapan değerlerden oluşan alt grubun arasından geçtiğini görürüz. (Şekil 2).

Tablo 1. Örnek 1' e ait regresyon kestirimleri (LMS: Least Median of Squares)

Yöntem	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	Log-lik
EKK	.34	-.17	2.50	-233.38
$GT(p=1.2, q=.01)$	.95	-.21	1.03	-516.52
$GT(p=1.8, q=.01)$	1.12	-.22	.46	-476.78
$GT(p=1.3, q=.1)$	1.13	-.22	.52	-300.61
$GT(p=1.01, q=1)$	.95	-.21	1.03	-69.31
t_1	.44	-.19	1.15	-234.94
t_2	.36	-.19	1.46	-228.46
LMS	1.82	-.24	1.68	
Huber	.39	-.18		
Tukey	.47	-.19		



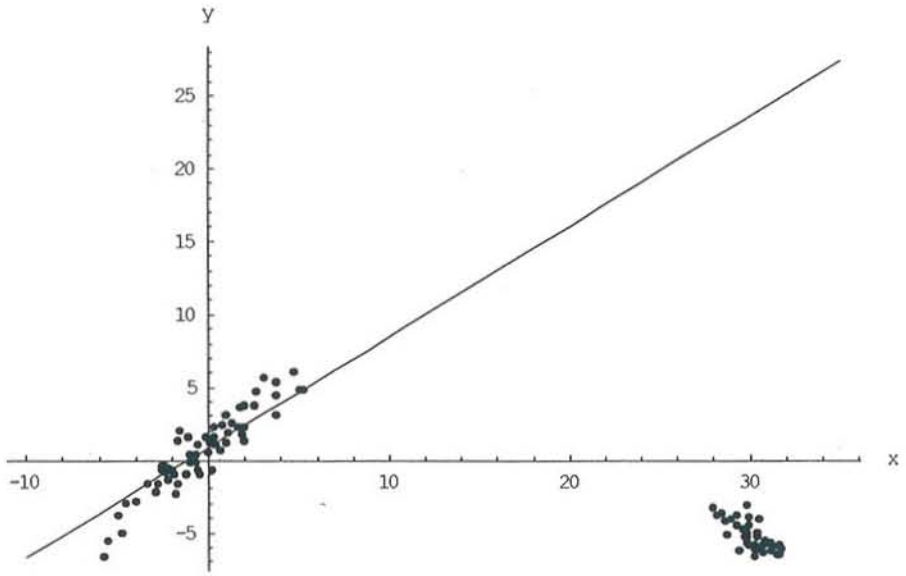
Şekil 2. Sapan değerlerin arasından geçen bir doğru.

Tablo 2. Örnek 1 için regresyon kestirimleri (*: sapan değersiz veri kümesi için)

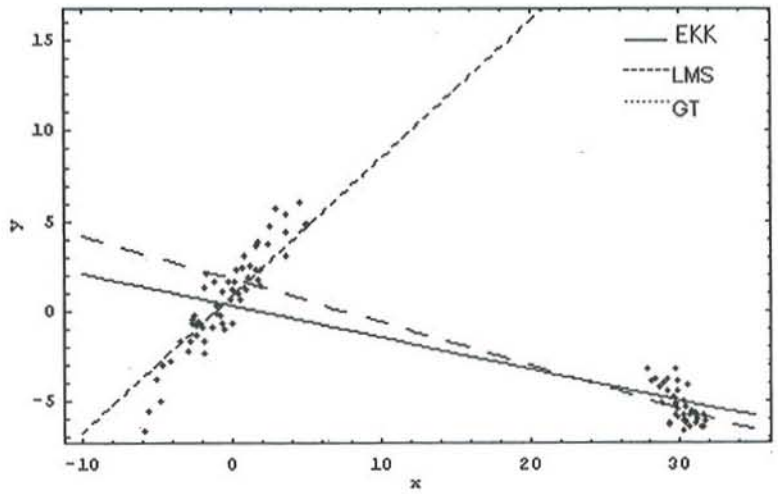
Yöntem	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	Log-lik
EKK*	.98	.97	.99	-84.43
$GT(p=1.2, q=.01)$	.77	.90	.62	-516.52
$GT(p=1.8, q=.01)$	.87	.76	.38	-476.78
$GT(p=1.3, q=.1)$	.69	.63	.74	-300.61
$GT(p=1.01, q=1)$	-.65	-.16	1.04	-69.31
t_5	1.02	-.22	.76	-250.68
t_1	.45	-.19	1.15	-234.94
t_2	.36	-.19	1.46	-228.46
Huber	.39	-.18		
Tukey	.97	.97		

Eğer iterasyona başlangıç değer olarak EKK* doğrusunu alırsak *GT* kestiricilerinin yeteri kadar küçük p and q değerleri için iyi gözlemlerin arasından geçecek şekilde doğrular ürettiğini buna karşılık bir başka sıfıra azalan M-kestiricisi Tukey'in dışındakilerinin sapan değerlerin etkisi altında kaldığını görürüz (Tablo 2 ve Şekil 3.)

Yüksek bozulma noktasına (breakdown point) sahip kestiriciler sınıfına dahil olan LMS kestiricisinin EKK gibi sapan değer alt grubundan etkilendiğini ve iyi veri grubunu modelleyemediğini görmekteyiz (Tablo 1 ve Şekil 4.)



Şekil 3. İyi veri kümesi arasından geçen doğru.



Şekil 4. EKK, LMS ve GT regresyon doğrularının karşılaştırılması.

Örnek 2. Stackloss veri kümesi.·

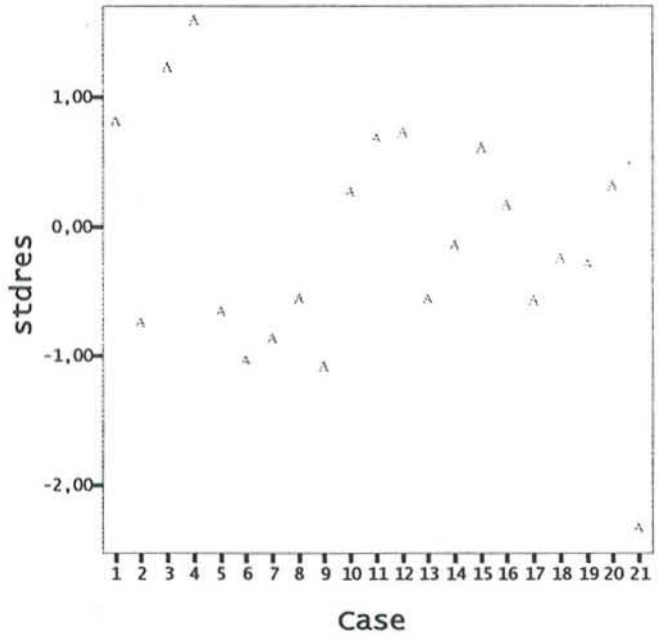
Bu veri kümesi bitkiler üzerine yapılan kimyasal çalışmalardan elde edilen gözlemlere dayanır. Her biri 21 gözlemlik 3 regresörden oluşan veri kümesine dayanıklı kestirimle alakalı olarak literatürde sık rastlıyoruz (Örneğin, Lange vd., 1989, Andrews, 1974).

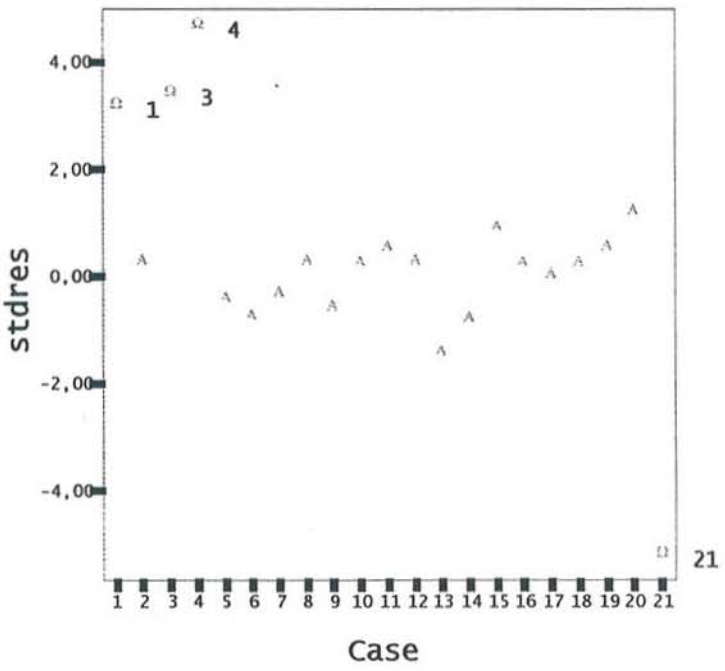
Model kurulup parametrelerin kestirimleri hesaplandığında EKK doğrusunu $\hat{y} = -39{,}92 + 0{,}72\hat{\beta}_1 + 1{,}30\hat{\beta}_2 - 0{,}15\hat{\beta}_3$ olarak buluruz. Bu doğruya ait endeks grafiği Şekil 5'te verilmiştir. Bu grafikte dikey eksendeki stdres standartlaştırılmış rezidülere karşılık gelir. Grafikten EKK standartlaştırılmış rezidülerinin tümünün (-2.5; 2.5) aralığında bulunduğunu görürüz. Yani EKK yöntemine göre veri kümesinde sapan değer bulunmamaktadır.

Eğer iterasyona EKK doğrusu ile başlarsak Tablo 3'te verilen kestirimlere ulaşırız. p=2.5 ve q=1 için GT kestiricisi 4. ve 21. gözleme iterasyonun son adımında en küçük ağırlıklar olarak sırasıyla 0.4 ve 0.2 vermektedir. Dağılımın şekil parametrelerini biraz daha küçük seçtiğimizde nispeten daha küçük ağırlıklar alan gözlemlerin sayısı artar. Örneğin, p=1.5 ve q=0.2 için GT kestiricisi 1., 3., 4. ve 21. gözlemlerin hepsine son iterasyonda sıfır ağırlık vermektedir. Bu gözlemlerin birer sapan değer olduğunu $GT(p=1.5,q=.2)$ kestiricisine ait endeks grafiğinden de anlayabiliriz. Bu grafiğe göre 1., 3., 4. ve 21. gözlemler (-2.5;2.5) yatay bandının dışında kalmaktadır (Şekil.6.)

Tablo 3. Stackloss verisi için regresyon kestirimleri.

Yöntem	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\sigma}$	Loglik
EKK	-39.92	.72	1.30	-.15	2.97	-52.65
LMS	-34.25	.75	.50	0.00	1.21	
Huber	-40.75	.76	1.17	-.14		
Tukey	-41.33	.83	.95	-.13		
t_1	-38.63	.85	.49	-.07	.76	-49.58
t_2	-38.12	.85	.56	-.09	1.80	-50.31
$GT(p=2.5,q=1)$	-39.96	.86	.69	-.11	2.52	-14.56
$GT(p=1.5,q=.5)$	-40.47	.84	.55	-.05	.80	-33.28
$GT(p=1.5,q=.2)$	-39.94	.83	.57	-.06	.09	-49.23
$GT(p=1.3,q=.5)$	-40.09	.83	.56	-.06	.58	-34.34





Şekil 5. EKK endeks grafiği.
Şekil 6. *GT*(p=1.5,q=.2) kestiricisine ait endeks grafiği.

Eğer bu 4 sapan değeri veri kümesinden atıp kalan gözlemler üzerine modeli kurarsak EKK* doğrusu $\hat{y} = -37.65 + .80x_1 + .58x_2 - .07x_3$ olarak elde edilir. Bu

doğrunun yeteri kadar küçük p ve q için bulunan ve Tablo 3'te verilen *GT* doğrularıyla benzer olduğunu görmekteyiz.

5. SONUÇ

Regresyonda hata teriminin normal dağılımlı olduğu yaygın olarak varsayılır. Fakat gerçekte hata dağılımı, özellikle de sapan değerlerin etkisiyle, normalden daha kalın kuyruklu olabilir. Bu durumda EKK analizi doğru sonuç vermeyecektir.

Biz bu çalışmada normal dağılıma alternatif olarak regresyonda hataları *GT* dağılımıyla modelledik. Dağılıma ait şekil parametrelerinin bilindiği varsayımı altında regresyon parametrelerini kestirdik. Elde ettiğimiz kestiriciler sıfıra azalan M-kestiricileri olup bunlar yinelemeli tekrar ağırlıklandırmalı en küçük kareler formunda bulunmuştur. Dağılımın şekil parametreleri aynı zamanda dayanıklılık ayar sabitleri olup bunlar ağırlıkların sapan değerlere karşı hassalığını kontrol etmektedir. Bulduğumuz *GT* kestiricilerinin performansını görmek ve diğer dayanıklı kestiricilerle karşılaştırmak için *GT* kestiricilerini problemli veri kümelerinde kullandık. p ve q nun küçük olması durumunda ve iterasyona uygun bir başlangıç değerin seçilmesiyle *GT* kestiricilerinin iyi bir alternatif olabileceğini gördük. Ayrıca Örnek 1'de sapan değerlerin bir alt grup oluşturması durumunda da *GT* kestiricilerinin bozulmayacağını, dayanıklı kalabileceğini gördük.

KAYNAKLAR

ANDREWS, D. F. (1974). *A robust method for multiple linear regression.* Technometrics, 16, 4, 523-531.

ARSLAN, O. (1992). *Multivariate robust analysis based on the t distribution and the EM algorithm.* Unpublished PhD thesis, Leeds University, Leeds, U.K.

ARSLAN, O., GENÇ, A. İ. (2002). *Robust location and scale estimation based on the univariate generalized t (GT) distribution.* (Submitted.)

ARSLAN (2002). *A simple test to identify good solutions to redescending M-estimating equations for regression.* In Development in Robust Statistics, Proceedings of ICORS 2001. Edited by R. Dutter, U. Gather, P.J. Rousseeuw and P. Filzmoser, pp. 50-61.

BLOOMFIELD, P., STEIGER, W. L. (1983). *Least Absolute Deviations Theory: Applications and Algorithms.* Boston: Birkhauser.

BUTLER, R. J., McDONALD, J. B., NELSON, R. D., WHITE, S. B. (1990). *Robust and partially adaptive estimation of regression models.* The Review of Economics and Statistics, 72, 321-327.

KLEIN, R., SPADY, R. (1984). *Quasi-maximum likelihood as a parametric approach to robust estimation.* working paper, Bell Communication Research.

LANGE, K. L., LITTLE, J. A., TAYLOR, J. M. G. (1989). *Robust statistical modeling using the t distribution.* Journal of the American Statistical Association, 84, 881-896.