ANADOLU ÜNİVERSİTESİ BİLİM VE TEKNOLOJİ DERGİSİ –B
Teorik Bilimler

ANADOLU UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY –B
Theoretical Sciences
Cilt/Vol.:1-Sayı/No: 1 : 49-56 (2011)

DOĞRUSAL REGRESYONDA SAĞLAM GÜVEN ARALIKLARI

Meral ÇETIN [1], Tuba KAVRUK

ÖZ

Sağlam regresyon yöntemlerine ilişkin çok sayıda çalışma olmasına rağmen, regresyon parametrelerinin sağlam güven aralığına ve testlerine ilişkin çalışmalar az sayıdadır. Bu çalışmaların çoğu da konum parametresinin güven aralığı üzerinedir. Bu çalışmada, doğrusal regresyon analizinde normallikten sapmalar ve aykırı değer varlığında, parametre tahminlerinin ve varyanslarının sağlam(robust) versiyonları kullanılarak β-parametrelerinin sağlam güven aralıkları hesaplanmış ve bu aralıklar En Küçük Kareler (EKK) yöntemine dayalı klasik güven aralıkları ile bir benzetim çalışması ile karşılaştırılmıştır.

Anahtar Kelimeler : Güven aralıkları, Sağlam regresyon, Sağlam kestiriciler, M-kestiriciler.

ROBUST CONFIDENCE INTERVALS IN LINEAR REGRESSION

ABSTRACT

Although it is many studies with respect to robust regression methods, there is a few studies with robust confidence interval and tests in the literature. Many of these studies are respect to confidence interval of location parameter. In this study, the aim is compute to confidence intervals for regression parameter and compare to classical confidence interval based on least squares estimator with simulation study, using robust version of parameter and variance, in case of non-normality and outlier.

Keywords: Confidence interval, Robust regression, Robust estimators, M-estimators.

1. GİRİŞ

Sağlam regresyon yöntemlerine ilişkin çok sayıda çalışma olmasına rağmen, regresyon parametrelerinin güven aralığına ve testlerine ilişkin çalışmalar az sayıdadır. Bu çalışmaların çoğu da konum parametresinin güven aralığı üzerinedir. Regresyon parametreleri için sağlam güven aralıklarına ilişkin ilk çalışma Gross (1977) tarafından yapılmıştır. Tiku vd. (1986) konum kestiricileri için sağlam güven aralıklarını tanımlamışlar ve ortalamanın sağlam kestirimi olarak kesilmiş (trimmed) ortalamayı kullanmışlardır. Du Mond ve Lenth(1987) çalışmalarında konum ölçüsü için sağlam güven aralığı vermişlerdir.

Field C.A(1997), küçük örneklemlerde regresyon parametresi için güven aralığının hesaplanmasında üç farklı alternatif yaklaşımda bulunmuştur. Androver vd. (2004), konum(location) ölçüsü ve basit doğrusal regresyon modelleri için sağlam güven aralıklarını tanımlamışlardır.

Bu çalışmada doğrusal regresyon analizinde normallikten sapmalar ve aykırı değer varlığında, regresyon parametresi için sağlam güven aralıkları incelenecektir. Sağlam parametre ve sağlam varyans

[1,] Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Beytepe ANKARA.

Geliş: 16 Eylül 2009; **Düzeltme:** 13 Nisan 2010; **Kabul:** 2 Haziran 2010

kestirimleri kullanılarak, normallik varsayım bozulumunda güven aralıklarının karşılaştırılması için bir benzetim çalışması yapılmıştır.

2. SAĞLAM REGRESYON KESTİRİCİLERİ

Sağlam bir regresyon tahmin edici bulmak için ilk adım Edgeworth tarafından yapılmıştır. Artık kare değerlerinin kullanılmasından dolayı aykırı değerlerin EKK üzerinde çok büyük etkisi olduğunu söylemiş ve aşağıdaki gibi tanımlanan en küçük mutlak değer regresyonunu önermiştir:

$$\underset{\hat{\beta}}{\text{enküçük}} \sum_{i=1}^{n} |e_i| \tag{1}$$

Bu konudaki bir sonraki adım M-tahmin edicilerinin kullanımıdır (Huber, 1981).

2.1 M-Kestiriciler

Huber, M-tahmin edicileri olarak bilinen bir tahmin edici sınıfı önermiştir. Bu tahmin yöntemi, artıkların kareleri (e_i) yerine, artıkların başka bir fonksiyonunu en küçükleme fikrine dayanmaktadır. M- tahmin edicisinin amaç fonksiyonu,

$$\min_{\hat{\beta}} \sum_{i=1}^{n} \rho(e_i) \tag{2}$$

biçiminde verilir ve artıkların simetrik bir fonksiyonudur. $\rho(-t) = \rho(t)$ ve sıfır noktasında en küçük değerini alır.

Bu amaç fonksiyonunun $\hat{\beta}_j$ regresyon parametresine göre türevi alındığında,

$$\sum_{i=1}^{n} \psi(e_i)\, x_i = 0 \tag{3}$$

denklem sistemi elde edilir. Burada ψ, ρ'nun birinci türevidir. Eşitlik (3)'deki denklem sistemi iteratif yöntemlerle çözülür. Literatürde ψ fonksiyonu için çeşitli öneriler verilmektedir. Bunlardan bazıları aşağıda verilmiştir:

Huber ψ fonksiyonu,

$$\psi_c(x) = \begin{cases} -c & x < -c \\ x & -c \le x \le c \\ c & x > c \end{cases} \tag{4}$$

biçimindedir ve burada c=1.345'dir.

Hampel (1974) ψ fonksiyonu,

$$\psi(x) = \operatorname{sgn}(t)\begin{cases} |x| & 0 \le |x| < a \\ -a & a \le |x| < b \\ a\left(\dfrac{c-|x|}{c-b}\right) & b \le |x| < c \\ 0 & c \le |x| \end{cases} \tag{5}$$

biçimindedir ve burada a= 1.7, b=3.4, c= 8.9'dır.

Andrews ψ fonksiyonu,

$$\psi(t) = \begin{cases} \sin(x/k) & |x| \le k\pi \\ 0 & |x| > k\pi \end{cases} \tag{6}$$

biçimindedir ve burada k=1.5 ya da k=2'dır.

Tukey(k) ψ fonksiyonu,

$$\psi(x) = \begin{cases} x(1-(x/k)^2)^2 & |x| \le k\pi \\ 0 & |x| > k\pi \end{cases} \tag{7}$$

biçimindedir ve k=5 ya da k=6'dır (Candan,1995).

Bu çalışmada Huber(1981) tarafından verilmiş regresyon parametresi sağlam kovaryans tahmini kullanılmıştır:

$$\operatorname{cov}(\hat{\beta}) = \frac{E(\psi)^2}{[E(\psi')]^2}(X'X)^{-1} \tag{8}$$

Eşitlik (8), aşağıdaki gibi de gösterilebilir:

$$\operatorname{cov}(\hat{\beta}) \approx \frac{(1/n)\sum \psi(r_i)^2}{\left[(1/n)\sum \psi'(r_i)\right]^2}(X'X)^{-1} \tag{9}$$

M-kestiriciler y-yönündeki aykırı değerlere karşı sağlamken, x-yönündeki aykırı değerlere karşı sağlam değildir ve bozulma noktası 1/n.

3. SAĞLAM GÜVEN ARALIKLARI

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ biçiminde verilen basit doğrusal regresyon denkleminde β_1 parametresinin klasik güven aralığı,

$$\hat{\beta}_1 \mp t_{\alpha/2,(n-2)} S(\hat{\beta}_1) \tag{10}$$

Çoklu doğrusal regresyon modeli için ise (j=1,2,...,k)

$$\hat{\beta}_j \mp t_{\alpha/2,(n-k-1)} S(\hat{\beta}_j) \tag{11}$$

şeklinde tanımlanır. Bu eşitliklerde t, tablo değerini, $S(\hat{\beta}_j)$ 'da parametrenin standart sapmasını göstermektedir.

En Küçük Kareler (EKK) yöntemi, artıkların normal dağıldığını, bağımsız değişkenlerin tüm düzeyleri için eşit varyanslı olduğunu ve bağımsız değişken ile artıkların ilişkisiz olduğunu varsayar. Bu varsayımlar sağlanmadığında, parametre tahminleri yanlı olabilir, EKK etkin olmaz, büyük artık değerleri hata varyansını ve standart sapma değerlerini şişirir. Böylece parametrelerin güven aralıkları uzamaya, yayılmaya başlar ve tahminler asimptotik olarak tutarlı olmayabilir. Aykırı değerler hata varyansını şişirdiğinde, modelin aykırı değeri tespit etme gücünü de zayıflatırlar. Böylece EKK yaklaşımı güvenilir sonuçlar vermez (Ryan,1997).

Bu nedenle bu çalışmada sağlam güven aralıkları elde etmek için parametre ve varyans tahminlerinin sağlam versiyonları kullanılmıştır. Sağlam regresyon yöntemlerine ilişkin çok sayıda çalışma olmasına rağmen sağlam kestiricilere dayalı güven aralıkları ve testlerine ilişkin çalışmalar az sayıdadır. Regresyon parametrelerinin güven aralığına ilişkin çalışmaya fazla rastlanılmamaktadır.

Regresyon parametreleri için sağlam güven aralıklarına ilişkin ilk çalışma Gross (1977) tarafından yapılmıştır. Gross bu çalışmasında, basit doğrusal regresyon modelinde bilinmeyen parametrelerin doğrusal kombinasyonuna ilişkin sağlam güven aralığı tahminini, ele alınan farklı bir X-dizayn matrisi için tanımlamıştır.

Gross tarafından önerilen güven aralığı,

$$T \pm t_{\alpha}^{*} S_g \tag{12}$$

biçimindedir.

Burada T, β_j 'lerin kombinasyon değerleri L, ile elde edilen bisquare nokta tahmin değeridir. t_{α}^{*} güven aralığının $(1-\alpha)$ olasılığıyla tahmin edilen parametreyi içermesini sağlamak ama- cıyla önerilen sabit değerdir ve $\alpha = 0.05$ alınması önerilir. Sg=D, g dizayn matrisinin bazı fonksiyonları olmak üzere T'nin standart hata tahminidir. Field ve Welsh (1998) küçük örneklemlerde regresyon parametreleri için koşullu sağlam güven aralığı tanımlamışlardır. Koşullu güven aralığı,

$$\hat{\beta}_p - \hat{\sigma}\, l(\alpha, f, a), \hat{\beta}_p - \hat{\sigma}\, u(\alpha, f, a)] \tag{13}$$

biçiminde tanımlanır. Aralığın alt sınırı ,

$$I(\alpha, f, a) = K_p^{-1}(1 - \frac{1}{2}\alpha / a) \tag{14}$$

üst sınırı,

$$u(\alpha, f, a) = K_p^{-1}(\frac{1}{2}\alpha / a) \tag{15}$$

biçimindedir. Bu eşitliklerdeki K değeri,

$K_p(x/a) = K(\infty,....,\infty, x, \infty/a)$ koşullu dağılım fonksiyonudur ve burada

$$K_p[I(\alpha, f, a/a) = 1 - \frac{1}{2}\alpha$$

$$K_p[u(\alpha, f, a/a) = \frac{1}{2}\alpha$$

biçiminde tanımlanır (Field and Welsch,1998).

Field (1997), sağlam regresyon için güven aralıklarını oluşturmada küçük örneklem tekniklerini kullanılmıştır. Bunun için de güven aralıklarının oluşturmada üç alternatif yaklaşımı ele almıştır.

Sağlam güven aralığına ilişkin bir diğer çalışma Field ve Zhou (2003) tarafından verilmiştir. Ancak bu çalışma da, regresyon parametreleri için M-tahmin yöntemi kullanılarak hatalar arasında zaman serisi türünde bir ilişkinin varlığına izin veren, güvenilir güven aralığı oluşturulmuştur. Adrover vd. (2004) ise konum ve basit doğrusal regresyon modelleri için aykırı değer varlığında ve diğer parametrik modelden sapma olduğu durumlarda bile gerçek sınırlara yakın ve sabit sağlam güven aralıklarını tanımlamışlardır. Bu güven aralığı, potansiyel asimptotik yanı dikkate almayan önceki sağlam aralıkların geliştirilmiş halidir.

4. UYGULAMA

Bu çalışmada bilinmeyen β parametrelerine ilişkin klasik güven aralığının sağlamlığını elde etmek için veri kümesinde aykırı değer olduğunda β parametrelerinin ve varyanslarının sağlam versiyonları dikkate alınmıştır.

Bunun için, bir benzetim çalışması planlanmıştır. Bu benzetim çalışması S-Plus paket programında hazırlanmıştır. Basit doğrusal bağlanım denkleminde $\hat{\beta}$ 'nın sağlam tahmin edicisini bulmak için yukarıda bahsettiğimiz M-tahmin edicileri kullanılmıştır, yine $\hat{\beta}$' nın varyans tahmini de Huber tarafından önerilen varyans-kovaryans tahmini(Eşitlik.9) kullanılmıştır. EKK, Huber, Hampel ve Andrews tahmin edicileri için $\hat{\beta}$' lara ilişkin güven sınırları hesaplatılmıştır. Bu hesaplamalarda aykırı değerin olmadığı ve aykırı değerin olduğu durumlar göz önüne alınmıştır. Hatta tek aykırı değer ve çok aykırı değer yaratılarak karşılaştırmalar yapılmıştır. Ayrıca, σ^2'nin tahminler üzerindeki etkisini görebilmek amacıyla 0.01, 0.1, 1 ve 10 olmak üzere dört düzeyi alınmıştır. Benzetim çalışmasında kullanılan etken olarak aykırı değer ve hata varyansı göz önüne alınmış ve 1000 tekrar üzerinden sonuçlar elde edilmiştir. Çalışma da ayrıca EKK, Huber, Hampel ve Andrews tahmin yöntemleri için ortalama güven aralığı uzunlukları, güven aralıklarının gerçek parametre değerini kapsama olasılıkları da hesaplanmış ve ayrıca parametreler için $H_0 : \beta_j = 0$ hipotez testi önem kontrolü yapılmıştır. Ortalama güven aralığı uzunluğu, her bir tekrar için bulunan aralık alt sınır ve üst sınır farkları toplamının tekrar sayısına bölünmesi ile elde edilmiştir. Kapsama olasılıkları, her bir tekrar için gerçek parametre değerlerinin hesaplanan aralık alt ve üst sınır değerleri arasında yer alıp almadığına bakılarak hesaplanmıştır. Ayrıca parametre önem kontrolleri yapılmıştır. Güven aralıkları %95 güven düzeyi için elde edilmiştir.

Tablo 1. Aykırıdeğer yokken kestiricilere ilişkin ortalama güven aralık (g.a) sonuçları

Tahmin Ediciler	Ortalama G.A. uzunlukları (σ^2=0.01)		Ortalama G.A. uzunlukları (σ^2=0.1)		Ortalama G.A. uzunlukları (σ^2=1)		Ortalama G.A. uzunlukları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	0.022	0.008	0.068	0.026	0.216	0.082	0.684	0.259
Huber	0.298	0.113	0.298	0.113	0.298	0.113	0.298	0.113
Hampel	0.379	0.143	0.379	0.143	0.379	0.143	0.379	0.143
Andrews	0.022	0.008	0.069	0.026	0.219	0.083	0.692	0.262

Tablo 2. Tek aykırıdeğer durumunda kestiricilere ilişkin ortalama güven aralık sonuçları

Tahmin Ediciler	Ortalama G.A. uzunlukları (σ^2=0.01)		Ortalama G.A. uzunlukları (σ^2=0.1)		Ortalama G.A. uzunlukları (σ^2=1)		Ortalama G.A. uzunlukları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	1.500	0.567	1.501	0.567	0.513	0.572	1.635	0.618
Huber	0.305	0.115	0.303	0.115	0.303	0.115	0.303	0.115
Hampel	0.305	0.115	0.319	0.121	0.315	0.119	0.461	0.174
Andrews	0.022	0.008	0.071	0.027	0.224	0.085	0.718	0.271

Tablo 3. Çok aykırıdeğer durumunda kestiricilere ilişkin ortalama güven aralık sonuçları

Tahmin Ediciler	Ortalama G.A. uzunlukları (σ^2=0.01)		Ortalama G.A. uzunlukları (σ^2=0.1)		Ortalama G.A. uzunlukları (σ^2=0.1)		Ortalama G.A. uzunlukları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	2.444	0.924	2.444	0.924	2.452	0.927	2.524	0.954
Huber	0.290	0.110	0.310	0.117	0.314	0.119	0.317	0.120
Hampel	1.069	0.404	0.885	0.335	0.635	0.240	0.536	0.203
Andrews	0.023	0.009	0.073	0.027	0.230	0.087	0.784	0.296

Tablo 4. Aykırıdeğer yokken kestiricilere ilişkin kapsama olasılıkları

Tahmin Ediciler	Kapsama olasılıkları (σ^2=0.01)		Kapsama olasılıkları (σ^2=0.1)		Kapsama olasılıkları (σ^2=1)		Kapsama olasılıkları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	0.956	0.957	0.956	0.957	0.956	0.957	0.956	0.957
Huber	1.000	1.000	1.000	1.000	0.980	0.984	0.579	0.616
Hampel	1.000	1.000	1.000	1.000	0.976	0.979	0.619	0.638
Andrews	0.946	0.950	0.946	0.950	0.946	0.950	0.942	0.954

Tablo 5. Tek aykırı değer durumunda kestiricilere ilişkin kapsama olasılıkları

Tahmin Ediciler	Kapsama olasılıkları (σ^2=0.01)		Kapsama olasılıkları (σ^2=0.1)		Kapsama olasılıkları (σ^2=1)		Kapsama olasılıkları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	0.000	1.000	0.000	1.000	0.000	1.000	0.086	0.999
Huber	1.000	1.000	0.998	1.000	0.599	0.964	0.164	0.530
Hampel	1.000	1.000	0.975	0.996	0.859	0.936	0.303	0.561
Andrews	0.848	0.951	0.848	0.951	0.846	0.953	0.837	0.952

Tablo 6. Çok aykırı değer durumunda kestiricilere ilişkin kapsama olasılıkları

Tahmin Ediciler	Kapsama olasılıkları (σ^2=0.01)		Kapsama olasılıkları (σ^2=0.1)		Kapsama olasılıkları (σ^2=1)		Kapsama olasılıkları (σ^2=10)	
	b1	b2	b1	b2	b1	b2	b1	b2
EKK	0.000	1.000	0.170	1.000	0.405	1.000	0.542	1.000
Huber	1.000	1.000	0.977	1.000	0.392	0.990	0.089	0.632
Hampel	0.000	1.000	0.008	1.000	0.064	0.977	0.076	0.672
Andrews	0.836	0.946	0.845	0.942	0.846	0.941	0.811	0.944

Yukarıdaki çizelgeler incelendiğinde, $\sigma^2 = 0.01$ ve $\sigma^2 = 0.1$ için en kısa güven aralıklarının EKK ve Andrews yöntemleri ile elde edildiği, ancak kapsama olasılıklarının Huber ve Hampel yöntemlerinden daha küçük bulunduğu görülmüştür. Huber ve Hampel tahminlerinde kapsama olasılıkları varyans küçükken 1'e yakın olmakla birlikte, varyans büyüdüğünde ($\sigma^2 = 10$), EKK ve Andrews yöntemlerine göre daha kısa aralık uzunlukları vermelerine rağmen kapsama olasılıkları oldukça düşmüştür ve güvenilirlikleri bozulmuştur. Huber ve Hampel tahmin edicileri tüm etkenler için, EKK ve Andrews tahmin edicileri ile benzer sonuçlar vermiştir. EKK ve Andrews yöntemlerinde tüm tekrarlar için parametrelerin anlamlı bulunduğu, Huber ve Hampel'da ise bir iki örneklem için parametrelerin anlamsız çıktığı görülmektedir.

Veri kümesi tek aykırı değer içerdiğinde, EKK yönteminde aralık uzunlukları sağlam yöntemlere göre oldukça büyük çıkmıştır ve β_1 parametre değerinin kapsama olasılıkları genelde sıfır bulunmuştur. Huber ve Hampel tahmin edicileri varyans küçük olduğunda iyi sonuçlar verirken, ancak varyans $\sigma^2 = 10$ olduğunda kapsama olasılıkları düşmüştür. Andrews yöntemi varyans büyük olduğunda, Huber ve Hampel yöntemlerinden daha uzun aralıklar vermiştir ancak kapsama olasılığı çok daha iyi bulunmuştur. Sonuç olarak, küçük varyans değerleri için en iyi sonucun Huber ve Hampel yöntemleri ile, varyans büyük olduğunda ise Andrews yöntemi ile elde edildiği görülmüştür. Veri kümesinde iki aykırı değer oluşturulduğunda, farklı varyans düzeyleri için EKK ve sağlam M-yöntemlerinden elde edilen ortalama güven aralıkları uzunlukları ve bu aralıkların gerçek parametreyi kapsama olasılıklarına ilişkin sonuçlar incelendiğinde de tek aykırı değer sonuçları ile benzer sonuçlar gözlenmiştir. EKK yönteminden elde edilen güven aralıkları sağlam yöntemlere göre oldukça uzun ve β_1 parametresini kapsama olasılıkları çok düşük bulunmuştur. Huber için bulunan aralık uzunluklarının yine

birbirine yakın ve varyanstan etkilenmediği ancak varyans arttığında kapsama olasılığının çok fazla düştüğü görülmektedir. Hampel yönteminde ise varyans arttıkça aralık uzunluğunun daha kısaldığı görülmektedir ve kapsama olasılıkları EKK yönteminden daha kötü sonuçlar vermiştir. Andrews yönteminde kapsama olasılıkları varyans küçük olduğunda Huber'den daha kötü çıkmıştır. Ancak varyans büyüdüğünde Huber'den daha iyi sonuçlar verdiği görülmüştür. Sonuç olarak; Hampel yöntemi iki aykırı değer olduğunda güvenilir değildir. Huber yöntemi varyans küçük olduğunda (σ^2=0.01 ve σ^2=0.1 için) iyi sonuçlar vermiştir. Andrews yöntemi ise daha kısa aralıklar verdiği için küçük varyansta Huber kadar iyi kapsama olasılıkları vermemiştir, ancak varyans arttığında aralık genişliğide büyüdüğünden en güvenilir sonuçları vermiştir. Ayrıca her yöntemde tüm tekrarlar için parametreler önemli çıkmıştır.

KAYNAKLAR

Adrover, J., Sal bian-Barrera, M. and Zamar, R. (2004). Globally Robust Inference for the Location and Simple Linear Regression Models. *Journal of Statistical Planning and Inference* 119, 353-375.

Candan, M. (1995). Doğrusal Regresyon Çözümlemesinde Sağlam Kestiriciler, Bilim Uzmanlığı Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. Ankara, 93s.

Du Mond, C. ve Lenth, R.V. (1987). A Robust Confidence Interval for Location. *Technometrics* 29,211-219.

Field, C.A. (1997). Robust regression and small sample confidence intervals. *Journal of Statistical Planning and Inference* 57,39-48.

Field, C. ve Welsh, A.H. (1998). Robust Confidence Intervals for Regression Parameters. *Australian &New Zealand Journal of Statistic* 40, 53-65.

Field, C.A. ve Zhou, J. (2003). Confidence Intervals Based On Robust Regression. *Journal of Statistical Planning and Inference* 115, 425-439.

Gross, A.M. (1977). Confidence Intervals for Bisquare Regression Estimates. *Jour. Amer. Statics. Assoc.* 72, 341-354.

Huber, P.J. (1981). Robust Statistics, John Wiley and Sons, New York, 308p.

Rouseeuw, P.J. ve Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York,329p.

Ryan, T.P. (1997). *Modern Regression Methods*. John Wiley and Sons, New York, 515p.

Tiku, M.L., Tan, W.Y. ve Balakrishnan, N. (1986). *Robust Inference*. Marcel Dekker, New York and Basel, 295p.