DAYANIKLI REGRESYON YÖNTEMİ VE ÇEŞİTLİ SOSYAL VERİLER ÜZERİNDE AYKIRI GÖZLEMLERİN TEŞHİSİ

Robust Regression Method and Diagnose Of Outliers on Several Social Data

Özlem YORULMAZ [*]

ÖZ

Araştırmanın Temelleri: Veride aykırı gözlem(ler)in bulunması durumunda En Küçük Kareler Tahmin (EKK) Yöntemi direncini yitirerek yanıltıcı sonuçlar verebilir. Aykırı gözlemlerin bir çok veride karşılaşılabilir olması araştırmacının her türlü istatistiksel çalışmada aykırı gözlemlere karşı dirençli tahminci kullanım eğiliminde olmasını gerektirir.

Araştırmanın Amacı: Aykırı gözlem varlığında kullanılabilecek EKK tahmin yöntemine alternatif dayanıklı tahmin yöntemlerinin bir kısmını tanıtmak, çeşitli veriler üzerinde aykırı gözlemlerin EKK ile dayanıklı tahminciler üzerindeki etkisini incelemek ve aykırı gözlemleri teşhis etmeye çalışmaktır.

Ana Tartışma ve Sonuçlar: Veride genel karakteristik yapıya uymayan gözlem(ler)in varlığı halinde, EKK tahmin sonuçları, bu tahminlerin standart hataları ve güven aralıkları olumsuz yönde etkilenir. Dayanıklı tahmin yöntemlerinin kullanımı bu tür gözlemlerin belirlenmesini sağlamakla beraber onların verideki etkinliğini azaltarak daha güvenilir sonuçların elde edilmesini sağlar. Demografik yapıya sahip çeşitli veriler üzerinde dirençli ve duyarlı tahmin yöntemlerinin sonuçları konu kapsamında değerlendirilerek doğrulanmış, iki boyutlu çizime olanak sağlayan verilerde de görsel algı ile desteklenmiştir. Ayrıca belirlenen aykırı gözlemler demografik istatistiklerle örtüşmektedir.

Anahtar Kelimeler: Aykırı Gözlemler, En Küçük Kareler, En Küçük Kırpılmış Kareler, En Küçük Varyans Kovaryans Determinantı.

ABSTRACT

Fundamentals of the Research: In the presence of outlier(s), the method of Ordinary Least Squares (OLS) can be affected and gives misleading results. The common applications of OLS and the frequent existence of outliers in researches require to be tended to use resistant estimators in every statistical work.

Purpose of the Research: Hence the purposes of this study are to introduce some high breakdown robust estimation techiques that are alternative to OLS, to investigate the effects of outliers on OLS and robust estimators and lastly diagnose of outliers.

Main Discussion and Results: In case of having atypical and infrequent observations, OLS estimation results, standard errors and confidence intervals are affected badly. Whereas reliable results and outlier diagnosis can be attained with robust estimators. The consequences of robust and sensitive estimators on demographically structured data are analyzed within the context of this subject. These consequences are also ascertained in some of the data set which provide two-dimensional plots with a visual perception. Furthermore, diagnosed outliers in data verify the past demographic statistics' summaries.

Key Words: Outliers, OLS, Least Trimmed Squares (LTS), Minimum Covariance Determinant (MCD).

1.GİRİŞ

Regresyon analizinde En Küçük Kareler (EKK) yöntemi sıklıkla tercih edilir, bunun en önemli nedenlerinden biri yöntemin hesaplama kolaylığı sağlamasıdır. Fakat bu yöntem veri kümesinin genel yapısının dışında olan gözlemlere (aykırı gözlem) karşı oldukça duyarlıdır. Bu tür gözlemler, EKK yönteminin hataların özdeş ve bağımsız dağılması varsayımının sağlanmamasına, tahminlerin yanlı olmasına ve etkin olmamasına neden olabilir. Veride söz konusu gözlemlerin varlığı durumunda onların etkisini sınırlandıran dayanıklı regresyon yöntemlerini kullanmak gerekir. Dayanıklı regresyon yöntemleri dirençli sonuçlar vereceği gibi aykırı gözlemlerin teşhisini de olanak sağlar. İlerleyen bölümde değinileceği gibi En Küçük Mutlak Sapmalar (LAD) Regresyon yöntemi ilk dayanıklı yöntem olarak

[*] Araş. Gör., İstanbul Üniversitesi İktisat Fakültesi Ekonometri Bölümü

kabul edilir. Geliştirilen birçok dayanıklı regresyon yöntemi olmasına karşın burada sadece sıklıkla kullanılan ve en çok bilinen iki yönteme değinilmiştir. Bunlar Rousseeuw' un önerdiği En Küçük Medyan Kareler (LMS) yöntemi ile yine Rousseeuw' un önerdiği En Küçük Kırpılmış Kareler (LTS) yöntemidir.

Çalışmanın amacı, çeşitli demografik yapıdaki veriler üzerinde EKK ve LTS yaklaşımlarının sonuçlarını göstermek, aykırı gözlem varlığının tahminci üzerinde neden olduğu etkiye vurgu yapmak ve bu aykırı gözlemleri teşhis etmektir. Uygulama kısmında özellikle basit regresyon yönteminin kullanımını gerektiren ilk iki veri kümesinde, aykırı gözlemlerin regresyon katsayılarının değişimine neden olduğu durum grafikler üzerinde de görsel olarak algılanabilmektedir. Bunun yanısıra uygulama kısmında dayanıklı yöntem kullanımıyla belirlenen aykırı gözlemler, demografik istatistikleri doğrulamaktadır.

Çalışmanın birinci bölümünde EKK tahmincisi ile çeşitli dayanıklı tahmincilerin amaç fonksiyonları tanımlanmış ve tahmincilerin dayanıklılığının ölçütlerinden biri olan kırılma noktası üzerinde durulmuştur.

İkinci bölümde çeşitli aykırı gözlem tanımları ve aykırı gözlem teşhis yöntemlerinden kısaca bahsedilmiştir.

Üçüncü bölümde ise üç farklı sosyal veriye regresyon analizi uygulanmış, dayanıklı ve dayanıklı olmayan EKK tahmincisi ile parametre tahminleri yapılarak katsayılardaki farklılıklar gösterilmiş ve bu katsayı farklılığına neden olabilecek aykırı gözlemler çeşitli teşhis yöntemleri ile tespit edilmiştir.

2. TEMEL KAVRAMLAR

İstatistiksel çalışmalarda, raslantısal olarak çekilen örneklemi oluşturan gözlemlerin birbirinden bağımsız ve özdeş dağıldıkları varsayılır. Özellikle gerçek verilerle yapılan çalışmalarda, bazı gözlemlerin diğerlerine göre aşırı büyük ya da küçük olduğu (aykırı gözlem) durumda, bu gözlemlerin verinin çoğunluğuyla özdeş dağılmaları beklenemez. Bu tür gözlemler, örnekleme ilişkin bilgiyi özetleyen tahmincileri etkileyebilir. Bir tahminci veride bulunan aykırı gözlem(lerin) varlığından etkilenmiyorsa o tahminci dayanıklı (robust), etkileniyorsa dayanıklı olmayan tahmincidir. Benzer bir yaklaşımla tahmin yöntemi de dayanıklı ve dayanıklı olmayan yöntem şeklinde isimlendirilebilir.

Aykırı gözlem(ler) verinin çoğunluğundan uzakta bulunan gözlem(ler)dir, verinin çoğunluğunun sahip olduğu dağılımdan farklı bir dağılıma ya da aynı dağılıma fakat farklı parametrelere sahip oldukları düşünülür. Rousseeuw ve Zomeren'a (1990) göre aykırı gözlemler, verideki toplam gözlem sayısının yarısından daha az sayıda olmasına rağmen, o verideki gözlemlerin çoğunun vermek istediği bilgiye engel olan ve sonuçlar üzerinde yanıltıcı bir etki yaratabilen gözlemlerdir.

Veride bulunan aykırı gözlemlerin varlığı nedeniyle, dayanıklı olmayan tahminciler ile bunlara bağlı olarak elde edilen tahminler, hipotez testleri ve güven aralıkları da ister istemez olumsuz yönde etkilenir.

Tahmincilerin dayanıklılık özelliği ve bu dayanıklılığın derecesi kırılma noktası (breakdown point), etki fonksiyonu (influence function) ve hassasiyet eğrisi (sensivitiy curve) gibi çeşitli büyüklüklerle ifade edilmektedir. Bu çalışmada sadece kırılma noktası üzerinde durulacaktır.

Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Cilt 12 Sayı 21
Haziran 2009
ss.76-88

Kırılma noktası, tahmincinin parametre değerinden uzaklaşmasına neden olacak aykırı gözlem sayısının toplam gözlem sayısına oranının supremumu olarak bilinir. Staudte ve Sheather (1990) tahmincilerin kırılma noktasının, tahmincideki değişim için bir sınır olarak düşünüldüğünde bu sınırın aşılmasına neden olan en yüksek aykırı gözlem oranının kırılma noktasını vereceğini ifade etmişlerdir.

$T(Z) = T(X, y)$ regresyon parametre tahmincisini göstermek üzere, bu tahmincinin kırılma noktası sonlu örneklem için aşağıdaki gibi ifade edilir.[1]

$$\varepsilon_n^*(T, Z) = \min\left\{\frac{m}{n}; \sup\left\|T(Z')\right\| = \infty\right\}$$

Burada $Z' = (X', y')$, $Z = (X, y)$ verisinde "m" tane gözlemin keyfi olarak başka gözlemlerle değiştirilmesiyle elde edilir, "m" aykırı gözlem sayısıdır.

Tahmincinin kırılma noktası 0'dan büyükse o tahminci dirençlidir. Aritmetik ortalamanın kırılma noktası 0'dır. Çünkü tek bir aykırı gözlemin varlığı duyarlı tahminci olan aritmetik ortalamayı tamamen değiştirebilir. Medyanın kırılma noktası bir tahminci için mümkün olabilecek en büyük sınır olan 0,5'tir. Kırılma noktası toplam gözlem sayısının % 50 'sinden fazlasına karşılık gelemez. Çünkü aykırı gözlem sayısı toplam gözlem sayısının yarısından fazlasına karşılık geliyorsa veriyi temsil eden dağılım hakkında anlamlı bir sonuç çıkarılamaz. Aykırı bir gözlem "düzenli" ve "düzenli" bir gözlem aykırı olarak algılanabilir.

2.1. En Küçük Kareler (EKK) Yöntemi

Yaygın olarak kullanılan ve işlem kolaylığı sağlayan EKK tahmin yöntemi aykırı gözlemlere karşı duyarlı olması, dayanıklı olmaması yüzünden eleştirilen bir tahmin yöntemidir. X açıklayıcı değişkenler matrisi, y bağımlı değişken vektörü, β parametre vektörü ve ε hata vektörü olmak üzere doğrusal regresyon modeli aşağıdaki biçimde matrisel olarak ifade edilebilir.

$$y_{nx1} = X_{nxp}\beta_{px1} + \varepsilon_{nx1}$$

β vektörünün En Küçük Kareler (EKK) tahmini şu şekilde elde edilir :

$$\hat{\beta} = (X'X)^{-1}X'y$$

$\hat{\beta}$ vektörünün elemanları olan $\hat{\beta}_1, \hat{\beta}_2 .... \hat{\beta}_p$ skaler biçimde yazılarak da ifade edilen regresyon modelinin katsayıları olarak adlandırılır.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p$$

Tahmin vektörü $\hat{y} = X\hat{\beta}$ olmak üzere y ile $\hat{y}$ vektörleri arasındaki farklar ise ε, un tahmini olarak düşünülen kalıntılar vektörü e'yi verir.

$$e = y - \hat{y}$$

$\hat{\beta}$ vektörünü elde etmek için kullanılan EKK yönteminin amaç fonksiyonu şöyledir:

$$\underset{\hat{\beta}}{Min} \sum_{i=1}^{n} e_i^2$$

EKK yöntemi verideki tüm gözlemlere eşit ağırlık vererek kalıntı kareler toplamını indirgeyen bir yöntemdir. Aykırı bir gözlemin vereceği büyük kalıntı, kalıntı kareler toplamını ve dolayısıyla paremetre tahminlerini de etkileyecektir. Bu yüzden amaç fonksiyonu daha güçlü olan, kırılma noktası daha yüksek tahminci seçilmesi önerilir. Bu tahmincilerden birkaçı şöyle sıralanabilir:

[1] Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection,* .s:9-10

2.2 En Küçük Mutlak Sapmalar Yöntemi (Least Absolute Deviations-LAD)

En küçük mutlak sapmalar (LAD) yönteminde $\hat{\beta}_1$, $\hat{\beta}_2$ .... $\hat{\beta}_p$ elemanlarından oluşan $\hat{\beta}$ vektörünün seçimi amaç fonksiyonu $\underset{\hat{\beta}}{Min} \sum_{i=1}^{n} |e_i|$, yi en küçük yapmayı amaçlar.

LAD tahmincisinin bulunuşunda EKK tahmincisindeki gibi analitik çözüm olmadığından çeşitli iteratif yaklaşımlar mevcuttur. Bu yaklaşımların ayrıntısına bu çalışmada değinilmeyecektir. LAD tahmincisi EKK tahmincisine göre daha dayanıklıdır.

2.3 En Küçük Medyan Kareler (Least Median Squares-LMS)

Rousseeuw (1984), bir çok yazarın EKK amaç fonksiyonunda tanımlı "kare" yerine değişik değerler kullanarak dayanıklı sonuçlar elde etmeye çalıştığını, oysa "$\sum$ " yerine "medyan" $\underset{\beta}{Min}\ medyan\ e_i^2$ yazılarak diğerlerinden daha dirençli, çok daha az duyarlı, kırılma noktası %50 olan tahminci elde edilebileceğini ifade etmiştir. Algoritma kısaca şöyle tanımlanabilir: p parametre sayısını göstermek üzere eleman sayısı en az p+1 olan tüm mümkün altkümeler elde edilir, her bir altküme için EKK tahmincileri bulunur ve bu tahmincilerle tüm gözlemler üzerinden elde edilen artıkların karelerinin ya da mutlak değerlerinin medyanı alınarak LMS amaç fonksiyonu elde edilir. Bu işlem altküme sayısı kadar tekrarlanarak amaç fonksiyonunu minimum yapan LMS parametre tahminlerine ulaşılır. LMS, 'nin kırılma noktası % 50' dir, EKK ve LAD tahmincilerine göre oldukça yüksek bir kırılma noktasına sahiptir.

2.4. En Küçük Kırpılmış Kareler (Least Trimmed Sum of Squares-LTS)

Rousseeuw (1984) bu tahminci için amaç fonksiyonunu şöyle tanımlamıştır:

$$\underset{\hat{\beta}}{Min} \sum_{i=1}^{h} (e^2(\hat{\beta}))_{i:n}$$

Yine verideki gözlem ve değişken sayıları dikkate alınarak çeşitli altkümeler elde edilir. Her bir altküme için h gözlem üzerinden EKK katsayıları bulunur ve bu katsayılarla kalıntıların kareleri sıralanır, hedef alınan belirli bir sıradaki kalıntıdan küçük bütün kalıntılar toplanır ve bu toplam başka tahminler sonucu elde edilen toplamdan daha düşük yapılmaya çalışılır. Rousseeuw h değerini (n+p+1)/2 olarak önermiş ve ayrıca çalışmasında gerek LTS'nin amaç fonksiyonunun LMS'ye göre daha düzgün olmasından gerekse LTS'nin istatistiksel etkinliğinin LMS'ye göre daha fazla olmasından dolayı LTS metodunun LMS ile yer değiştirebilir nitelikte olduğunu belirtmiştir.

2.5 En Küçük Varyans-Kovaryans Determinantı (Minimum Covariance Determinant-MCD)

Bu yöntem Rousseeuw (1984) tarafından önerilmiştir. Amaç varyans-kovaryans matrisinin determinantını en küçük yapacak olan h tane gözlemden oluşan veriyi bulmaktır. Elde edilecek olan h gözlemden hesaplanan ortalama MCD yer tahmincisi, aynı gözlemlerden elde edilecek varyans-kovaryans matrisi de yayılım tahmincisi olacaktır. h, (n+p+1)/2 'nin tam kısmı olarak belirlenebilir.

LTS ve MCD kırılma noktaları en yüksek olan tahmincilerdendir.

3. AYKIRI GÖZLEMLER

Regresyon analiziyle gözlemleri verinin çoğunluğunun sahip olduğu yapıyı (pattern) takip etmeyen gözlemler yani aykırı gözlemler ve düzenli gözlemler şeklinde iki alt başlığa ayırmak mümkündür.[2]

x ekseninde sapma göstermeyen ve elde edilen regresyon doğrusundan uzakta olmayan y değerleri düzenli gözlemler olarak nitelenmektedir.

Aykırı gözlemler ise genel olarak çalışmalarda üç başlık altında incelenmektedir. Bunlar;

- Dikey aykırı gözlemler, x ekseninde sapma göstermeyen fakat tahmin edilen regresyon doğrusuna uzakta olan y değerleri.
- İyi kaldıraç noktaları, x ekseninde sapma gösteren ve tahmin edilen regresyon doğrusuna uzakta olmayan y değerleri.
- Kötü kaldıraç noktaları, x ekseninde sapma gösteren ve tahmin edilen regresyon doğrusuna uzakta olan y değerleri.

EKK tahmincileri, regresyon doğrusunun eğiminde ciddi değişmelere neden olacak olan dikey aykırı gözlemlere ve kötü kaldıraç noktalarına karşı oldukça duyarlıdır. Aykırı gözlemler basit regresyonla çalışılırken serpilme diyagramından kolayca fark edilebilirken, çok değişkenli nokta kümelerinin içinde zor farkedilir. Değişken sayısı 2'yi aştığı zaman görsel kavrayış ortadan kalkar. Bu tür gözlemlerin teşhisi için değişik metodlar geliştirilmiştir. Gözlem(ler)in regresyon doğrusu üzerindeki etkisini ve bu etkinin önemli olup olmadığını inceleyen istatistiklere aykırı gözlem teşhisçileri denir (outlier diagnostics). Klasik analizlerde kullanılan bir çok teşhisçi EKK sonuçlarına dayanır, bu yüzden bunlar dayanıklı değildir. Bazı teşhisçilerden kısaca bahsetmek gerekirse şöyle sıralanabilir:

3.1 Standart-Studentize Artıklar

e, EKK regresyonunda elde edilen kalıntıları, n gözlem sayısı, p parametre sayısı ve s^2 hata terimleri varyansının yansız tahmincisi olmak üzere

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^{\,2}$$ iken

standart kalıntılar şöyle tanımlanabilir:

$$t_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

h_{ii} , H şapka matrisinin köşegen elemanlarından elde edilir. H matrisi ise tahmin vektörü $\hat{y} = X\hat{\beta}$ ifadesine $\hat{\beta}$ vektörü yerleştirilerek bulunur:

$$\hat{y} = X(X'X)^{-1}X'y$$

$$H = X(X'X)^{-1}X'$$

H matrisinin köşegen elemanları , h_{ii}, artıkları ağırlıklandırılmak için kullanılır.

$s(i)$, i. gözlemin yer almadığı veriden hesaplanan regresyonun hata terimlerinin varyansının tahmincisi iken, hesaplanan

[2] Hubert M.,(2004), *Regression Techniques,* ACCO, s:175

$$t(i) = \frac{e_i}{s_i\sqrt{1-h_{ii}}}$$

değeri ise student kalıntı (jackknife kalıntı) olarak isimlendirilir.

Teşhisçilerin bir kısmı gözlem silme metoduna dayanır. Bu metodla i. gözlemin olmadığı $\hat{\theta}_{(i)}$ tahmincisiyle, i. gözlemin olduğu durumdaki $\hat{\theta}$ tahmincisi arasındaki farka bakılır. Fakat hangi gözlemlerin silineceği açık olmadığından, bazı gözlemlerin grupça etkinlik gösterirken, tek başına etkin olmamaları yüzünden yöntem etkin olmayabilir. Bunun da ötesinde gözlem sayısı arttıkça altkümeleri incelemek mümkün olmayabilir. Bu sınıftaki teşhisçilere, i. gözlem, veriden çıkarıldığında hesaplanacak olan yeni regresyon denkleminin parametrelerinde meydana gelecek olan değişimi hesaplamada kullanılan DFBETA ve DFBETAS ölçüleri ile i. gözlem, veriden çıkarıldığında hesaplanacak olan yeni regresyondaki tahmini y değerlerinin değişimini hesaplamak için kullanılan DFFIT ve DFFITS ölçüleri örnek olarak gösterilebilir.

3.2 Mahalanobis Uzaklığı

Buraya kadar anlatılan aykırı gözlem teşhiscileri regresyon denkleminden elde edilen kalıntılar üzerinden hesaplanırken, Mahalanobis uzaklığı diğerlerinden farklı olarak çok değişkenli veri üzerinden hesaplanır. **X** , p boyutlu gözlemlerden oluşan veri matrisi, $\overline{x}$, veriden hesaplanan ortalama vektörü ve **S** de aynı veriden hesaplanan örneklem varyans-kovaryans matrisi olmak üzere Mahalanobis uzaklığı i. gözlem için şöyledir:

$$D_i = \sqrt{(x_i - \overline{x})' S^{-1} (x_i - \overline{x})}$$

Birden fazla aykırı gözlem olması durumunda, maskeleme (aykırı gözlemlerin düzenli gözlem gibi görünmesi) ve süpürme (düzenli gözlemlerin de aykırı gözlem gibi görünmesi) etkileri söz konusu olabilir ve Mahalanobis uzaklığı bu durumda doğru sonuç vermeyecektir.

Bu açıdan MCD varyans kovaryans matrisi ve ortalama vektörü kullanılarak, Mahalanobis uzaklığına benzer, dirençli hale getirilir. Bu uzaklığa dayanıklı uzaklık denilir. Benzer şekilde standart artıkların hesaplanmasında EKK'dan elde edilen artıklar yerine LMS ya da LTS artıkları kullanılarak dirençli hale getirilir.[3]

Rousseeuw ve Zomeren (1990) gözlem türlerinin Şekil 1 yardımıyla sağlıklı ve kolay biçimde sınıflanabileceğini açıklamışlardır.

Şekil 1. Aykırı Gözlem Teşhis Çizimi

Standartlaştırılmış EKK kalıntıları yada tandartlaştırılmış LTS kalıntıları

Dikey aykırı gözlemler	**Kötü kaldıraç noktası**
Düzenli gözlemler	**İyi kaldıraç noktası**
Dikey aykırı gözlemler	**Kötü kaldıraç noktası**

2.5

-2.5

Mahalanobis Uzaklığı Ya da Dayanıklı Uzaklık

Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Cilt 12 Sayı 21
Haziran 2009
ss.76-88

[3] Hubert M.,(2004), *Regression Techniques,* ACCO, s:178,188.

4. GERÇEK VERİ ÜZERİNDE ÇALIŞMA

Üç farklı sosyal veri grubuna EKK ve LTS tahmin yöntemleri uygulanmış, Mahalanobis Uzaklığı ve standardize edilmiş artıklar ile dayanıklı tahminciler kullanılarak dirençli hale getirilmiş artıklar ve uzaklıklar kullanılarak aykırı gözlemler teşhis edilmiştir. Her iki yöntemle bulunan aykırı gözlemler arasındaki farklılıklar teşhis çizimleriyle gösterilmiş ve bu aykırı gözlemlerin basit regresyon katsayılarında oluşturduğu etki görsel olarak sunulmuştur. Çalışmada kullanılan veriler çeşitli sosyal olgular üzerinedir. İlk uygulamada, Türkiye'de iç göç üzerine veriler alınarak modellenmiştir. İkinci uygulamada, Sağlıklı Yaşam Beklenti Süresinin ülkelere göre verileri alınarak, modelleme yapılmıştır. Üçüncü uygulamada ise, 5 Yaş altı Çocuk Ölüm Oranına ilişkin yine ülkeler düzeyinde kesit verilerle çalışılmıştır.

4.1 İç Göç

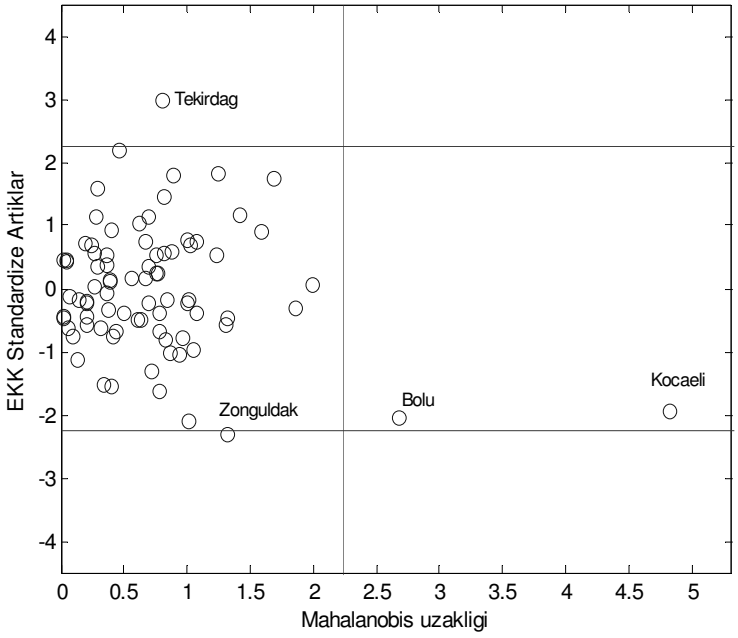
2000 yılı için Türkiye' nin 81 ilinin kişi başına düşen gelir rakamları ile illerin net göç hızı arasındaki ilişki incelenmiştir. Bağımlı değişken "Net Göç Hızı" ve bağımsız değişken "İllere göre kişi başına düşen GSMH – Türkiye'nin ortalama kişi başına düşen GSMH" değeri olarak ele alınmıştır.

EKK ve LTS yöntemi ile bulunan regresyon denklemlerine göre illerdeki kişi başına düşen ortalama gelir Türkiye ortalamasından uzaklaştıkça o ilin net göç hızı artmaktadır. EKK regresyon denklemi ve determinasyon katsıyısı şöyle bulunmuştur:

$$\text{EKK}: \ \hat{y} = -18.4302 + 0.019x$$

$$R^2 = 0.22$$

Şekil 2. İç göç verisi Aykırı Gözlem Teşhis Çizimi



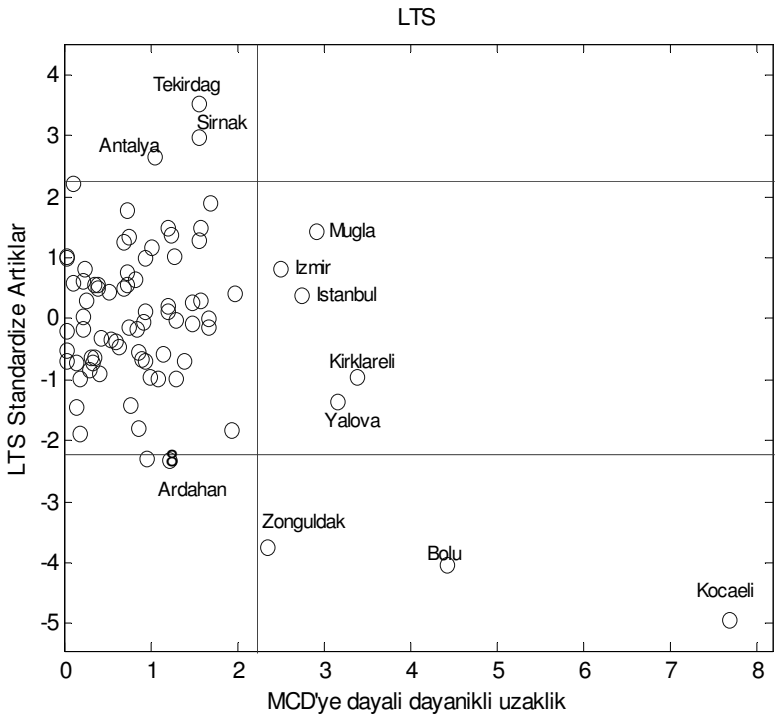
EKK regresyon denkleminden elde edilen artıklarlara ve Mahalanobis uzaklığına göre yapılan Şekil 2 çiziminde görüleceği gibi sadece Tekirdağ dikey aykırı gözlem olarak belirlenmiştir.

LTS regresyon denklemi şu şekilde tahmin edilmiştir:

$$\text{LTS}: \quad \hat{y} = -17.2906 + 0.0328x$$

$$R^2 = 0.45$$

Şekil 3. İç göç verisi Aykırı Gözlem Teşhis Çizimi (Dayanıklı)



LTS regresyon denklemine dayalı Şekil 3'e göre Tekirdağ, Şırnak, Antalya dikey yönlü aykırı gözlem ve Zonguldak, Bolu, Kocaeli ise kötü kaldıraç noktasıdır. Muğla, İzmir, İstanbul, Kırklareli, Yalova ise iyi kaldıraç noktasıdır. Her iki regresyon denkleminin sonucuna bağlı olarak elde edilen aykırı gözlemlerin genel yapısı şöyledir:

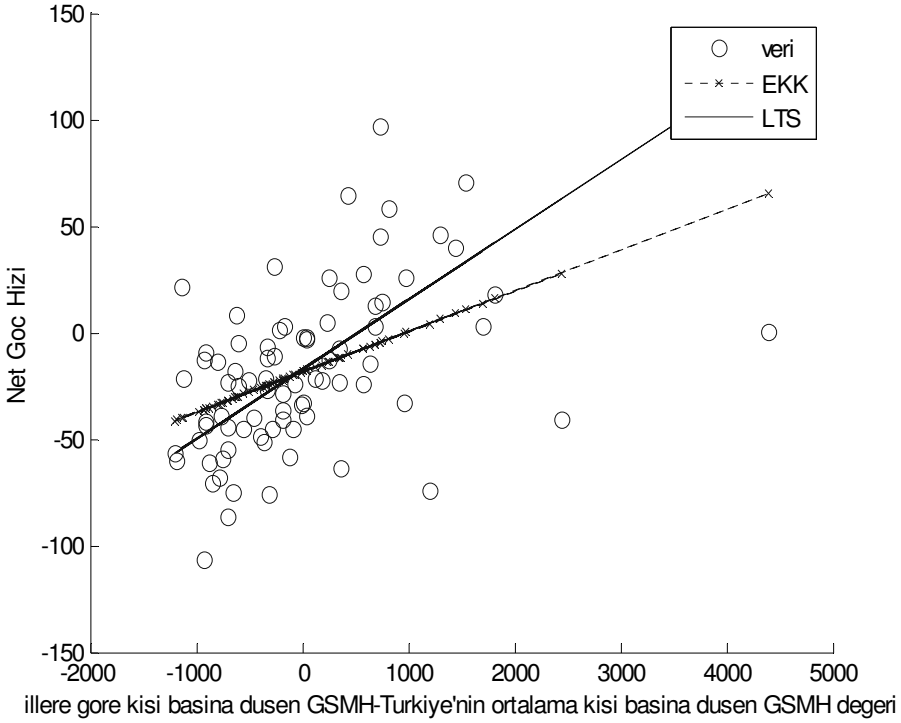
İzmir, Antalya, İstanbul, Tekirdağ, Kocaeli, Yalova, Muğla ve Kırklareli göç alan ve kişi başına düşen ortalama gelirleri Türkiye' nin ortalama gelirinin üzerinde olan illerdir.

Zonguldak, Bolu göç veren fakat kişi başına düşen ortalama gelirleri Türkiye' nin ortalama gelirinin üzerinde olan illerdir.

Şırnak ise kişi başına düşen ortalama geliri Türkiye' nin ortalama gelirinin altında olan ve göç alan (DİE, 2000) ildir.

Regresyon doğruları arasındaki eğim farklılığı ve aykırı gözlemlerin EKK doğrusunun eğimi üzerindeki etkisi, Şekil 4' deki serpileme grafiğinde görülmektedir. EKK doğrusu aykırı gözlemlere doğru çekilirken, LTS doğrusu gözlemlerin çoğunluğunu temsil etmekte, aykırı gözlemler karşısında dirençli kalmaktadır.

Şekil 4. LTS ve EKK Regresyon Doğruları Serpilme Grafiği Üzerinde Gösterimi



4.2 Ülkelere göre Sağlıklı Yaşam Beklenti Süresi

Sağlıklı Yaşam Beklenti Süresi (HALE) ve 5 yaş altı çocuk ölüm oranları ülkelerin gelişmişlik düzeylerinin birer ölçütleri olarak kabul edilmektedir. 5 yaş altı ölüm oranlarının sağlıklı yaşam beklentisi üzerindeki etkisi, 52 ülke için yapılan regresyon analizi ile incelenmiştir. Regresyon analizi öncesinde her iki değişkenin logaritmaları alınarak gerekli düzeltme işlemi gerçekleştirilmiştir. Bu çalışmada kullanılan 52 ülke şöyledir: Arnavutluk, Andora, Ermenistan, Rusya, Azerbaycan, Avusturya, Belarus, Bosna Hersek, Belçika, Bulgaristan, Hırvatistan, Kıbrıs, Çek Cumhuriyeti, Danimarka, Estonya, Finlandiya, Fransa, Almanya, Yunanistan, Macaristan, İzlanda, İrlanda, İtalya, Litvanya, Letonya, Lüksemburg, Malta, Monako, Hollanda, Norveç, Polonya, Portekiz, Moldova, Romanya, Sırbistan, Slovakya, Slovenya, İspanya, İsveç, İsviçre, Türkiye, Ukrayna, Irak, Japonya, Kanada, Timor, Senegal, Endonezya, Hindistan, ABD, Küba.

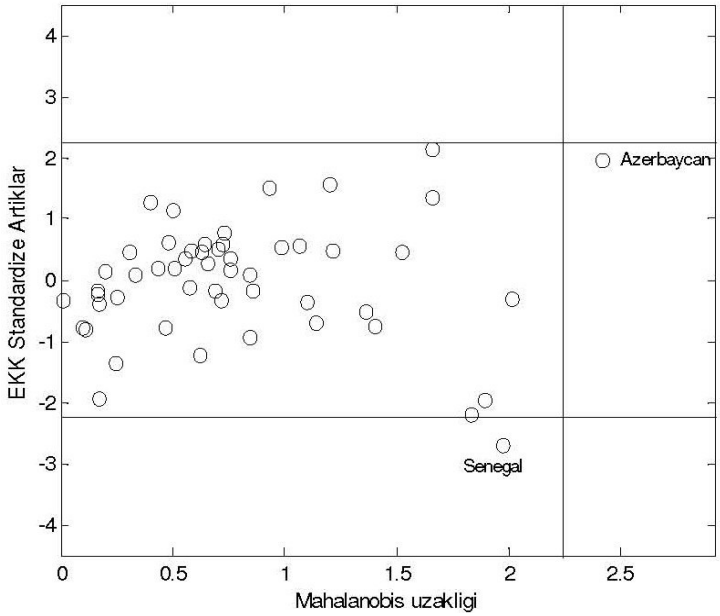
Şekil 5'in temsil ettiği EKK tahmin yöntemi sadece Azerbaycan ve Senegal'i aykırı gözlem olarak belirlerken; Şekil 6'nın temsil ettiği LTS; Türkiye, Hindistan, Senegal, Endonezya, Rusya, Ukrayna, Romanya, Timor, Irak, Azerbaycan ve Arnavutluk ülkelerini aykırı gözlem olarak belirlemiştir. Düzenli gözlemlerin büyük çoğunluğunu Avrupa Birliği'ne üye ülkelerle, ABD, Kanada, Japonya, Küba oluşturmaktadır.

EKK:

$\hat{y} = 4.4108\text{-}0.1036x$

$R^2 = 0.8130$

Şekil 5. Ülkelere göre Sağlıklı Yaşam Beklentisi Verisi
Aykırı Gözlem Teşhis Çizimi



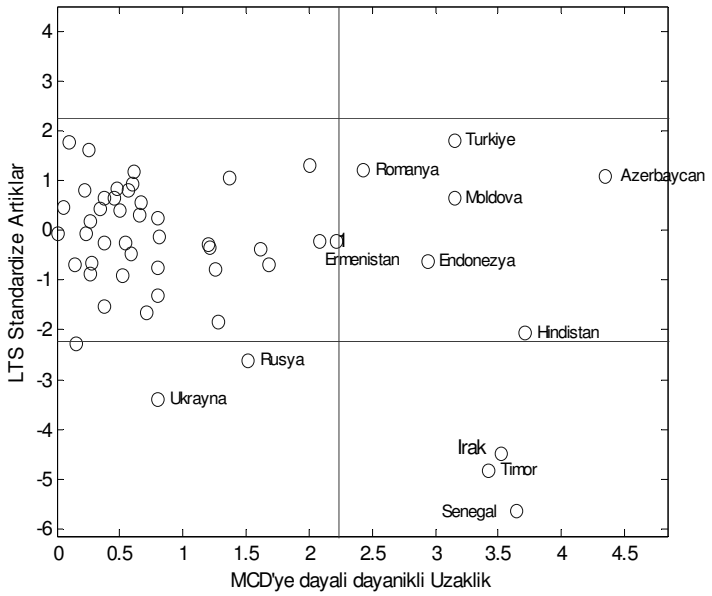
Mahalanobis uzaklığı ve standardize edilmiş EKK artıklarına göre yapılan çizim, önsel beklentinin oldukça dışında sonuçlar vermiştir. Sadece Senegal dikey yönlü aykırı gözlem ve Azerbaycan iyi kaldıraç noktası olarak belirlenmiştir. Oysa Şekil 6' da görülebileceği gibi LTS artıklarına ve dayanıklı uzaklıklara dayalı çizim gelişmiş ülkeleri, az gelişmiş ve gelişmekte olan ülkelerden keskin bir biçimde ayırmıştır; bu sonuç çok daha çarpıcı ve gerçeğe uygundur.

EKK regresyon denklemi katsayılarından farklılık gösteren LTS denklemi aşağıdaki gibidir.

LTS: $\hat{y} = 4.3801\text{-}0.0836x$

$R^2 = 0.8405$

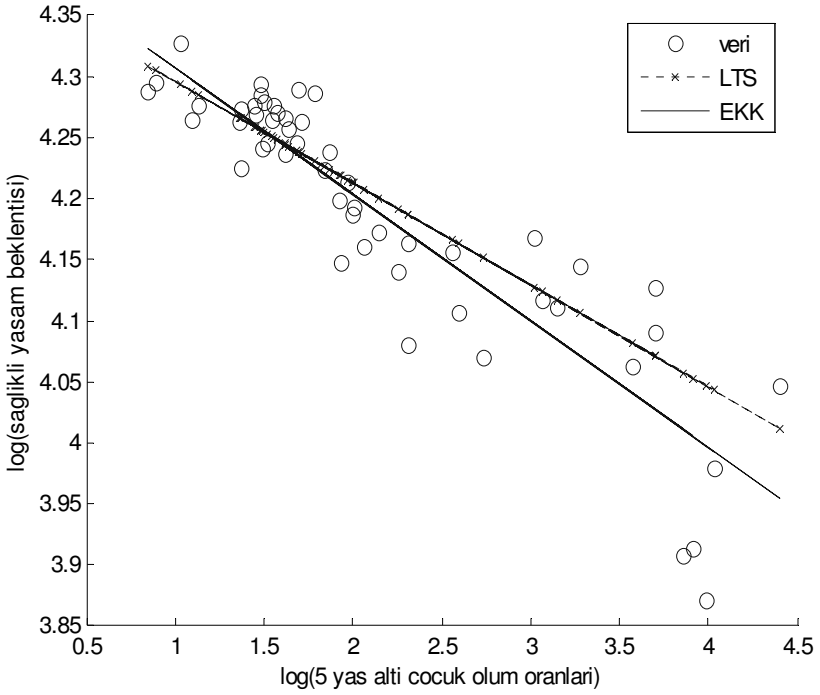
Şekil 6. Ülkelere Göre Sağlıklı Yaşam Beklentisi Verisi
Aykırı Gözlem Teşhis Çizimi (Dayanıklı)



Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Cilt 12 Sayı 21
Haziran 2009
ss.76-88

Şekil 7'den görülebileceği gibi aykırı gözlemler EKK doğrusunu kendilerine çekmekteyken, LTS daha dirençli bir yapı göstermektedir.

Şekil 7. LTS ve EKK Regresyon Doğruları Serpilme Grafiği Üzerinde Gösterimi



5.3 Ülkelere göre 5 Yaş altı Çocuk Ölüm Oranları

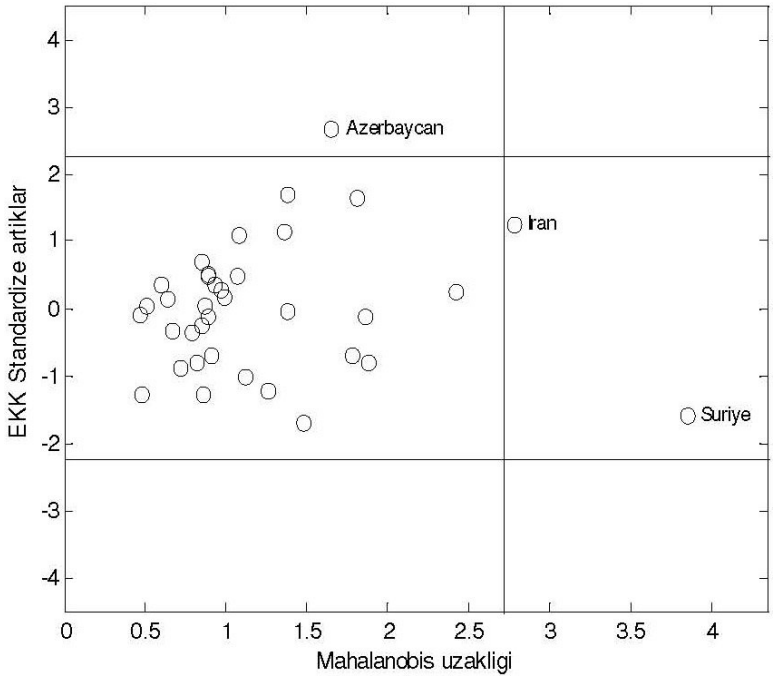
36 Farklı ülkeye ait 5 yaş altı çocuk ölüm oranları, kişi başına düsen milli gelir ve ülkelerdeki kadın nüfusun okur-yazarlık oranı ile açıklanmaya çalışılmıştır. Ülkeler aşağıdaki gibi seçilmiştir:

İsveç, Norveç, İngiltere, Belarus, Litvanya, İrlanda, Danimarka, İspanya, İtalya, Portekiz, Almanya, Fransa, Belçika, İsviçre, Macaristan, Arnavutluk, Yunanistan, Polanya, Türkiye, Romanya, Rusya, Estonya, Ermenistan, Ukrayna, Hollanda, Kıbrıs, Malta, Finlandiya, İzlanda, İran, Suriye, Azerbaycan, Bosna-Hersek, Avusturya, Hırvatistan.

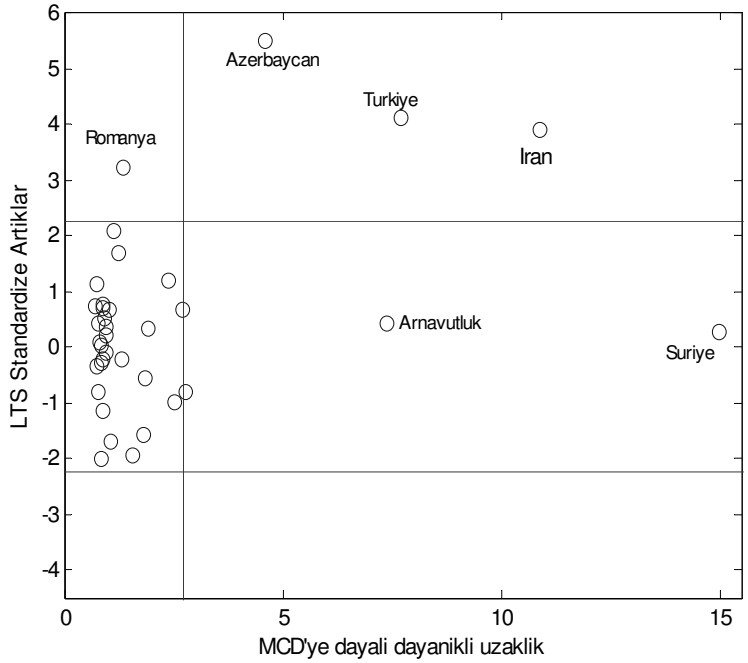
Beklendiği gibi bağımsız değişken, bağımlı değişkeni ters yönde etkilemektedir. Kişi başına düşen GSMH ile okuryazarlık oranındaki artış, bir gelişmişlik ölçütü olarak kullanılabilecek 5 yaş altı çoçuk ölüm oranında azalmaya neden olur.

Şekil 8'den görülebileceği gibi EKK tahmin yöntemi ile sadece Azerbaycan gözlemi regresyon doğrusunun eğiminde değişime neden olabilecek aykırı gözlem olarak görünmektedir. Oysa LTS yönteminde (Şekil 9) dikey aykırı gözlem değişmiş (Romanya) ve kötü huylu kaldıraç noktaları belirmiştir (Azerbaycan, Türkiye, İran). Yine EKK yöntemiyle düzenli birer gözlem olarak kabul edilen Türkiye ve Arnavutluk, LTS yöntemiyle aykırı gözlem olarak belirlenmiştir.

Şekil 8. Ülkelere göre 5 Yaş altı Çocuk Ölüm Oranları verisi Aykırı Gözlem Teşhis Çizimi



Şekil 9. Ülkelere Göre Sağlıklı Yaşam Beklentisi Verisi Aykırı Gözlem Teşhis Çizimi (Dayanıklı)



EKK ve LTS regresyon denklemleri şöyle bulunmuştur:

EKK : $\hat{y} = 13.3179 - 0.8261x_1 - 1.9719x_2$ LTS : $\hat{y} = 7.7008 - 0.7209x_1 - 0.8286x_2$

$R^2 = 0.7649$ $R^2 = 0.8054$

Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi
Cilt 12 Sayı 21
Haziran 2009
ss.76-88

6. SONUÇ

Üç farklı veri üzerinde yapılan bu çalışmada, dayanıklı tahminci ve dayanıklı uzaklıklarla elde edilen çizimin, önsel beklentiler doğrultusunda aykırı gözlem olarak düşünülebilecek gözlem(ler)i son derece etkin bir biçimde belirlediği anlaşılmıştır. Standartlaştırılmış LTS kalıntı değerleri ile dayanıklı uzaklıklara dayalı grafik, standartlaştırılmış EKK kalıntı değerleri ile Mahalanobis Uzaklığına dayalı grafiğin kaçırdığı gözlemleri kolaylıkla yakalamıştır. Aykırı gözlemlerin EKK regresyon doğrusunun eğiminde meydana getirdiği değişim, gerek katsayılarda gerekse çizilen serpilme diyagramlarında görülmüştür. Veride aykırı gözlem bulunması halinde, gözlemlerin homojen dağılmaması durumunda, dayanıklı tahmincileri kullanmak sonuçların güvenilirliğini artıracaktır.

KAYNAKÇA

Belsey, D., Kuh, Edwin., Welsch, R. (1980). *Regression diagnostics***.** USA, John Wiley& Sons. Inc.

Birkes, D., Dodge, Y. (1990). *Alternative methods of regression***.** New York, John Wiley&Sons, Inc.

Cook, D., Weisberg, S. (1992). *Residuals and influence in regression*. Great Britain, Chapman&Hall.

Hubert M.,(2004). *Regression techniques.* ACCO, Leuven, Belgium.

Rousseeuw, P.J. ve Leroy, A.M. (1987). *Robust regression and outlier detection*. John Wiley, New York.

Rousseeuw, P.J. ve Zomeren, B.C. (1990). Unmasking outliers and leverage points. *Journal of the American Statistical Association*, 85, 411, s.633-639.

Rousseeuw, P.,Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics***,** Vol.41, s.212-223.

Rousseeuw, P.J. and Van Driessen, K. (2002). Computing LTS regression for large data sets. *Estadistica*, 54, s.163-190.

Staudte, R., Sheather, S. (1989). *Robust estimation and testing*. USA, John Wiley& Sons. Inc.

Verboven, S. and Hubert, M. (2005). LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75, s.127-136.

Yorulmaz, Ö. (2003). *Robust regresyon ve mathematica uygulamaları*. Yayınlanmamış Yüksek Lisans Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

DİE, (2001). http://www.tuik.gov.tr/VeriBilgi.do.

CIA, https://www.cia.gov/cia/publications/factbook/index.html.

WHO, (2003). The World health report.http://www.who.int/whr/2003/en/Annex4-en.pdf.

Araş. Gör. Özlem Yorulmaz

Mimar Sinan Üniversitesi, Fen-Edebiyat Fakültesi İstatistik Bölümü'nden 2001 tarihinde mezun oldu. 2001–2003 tarihleri arasında Marmara Üniversitesi Sosyal Bilimler Enstitüsü'nde İstatistik yüksek lisans programını tamamladı. 2001 tarihinden itibaren İstanbul Üniversitesi İktisat Fakültesinde araştırma görevlisi olarak çalışmakta ve halen aynı üniversitede Ekonometri doktora programına devam etmektedir.