İstatistiksel Regresyon Yöntemlerinin Farklı Veri Gruplarına Uygulanması Üzerine Bir Analiz

Serkan ÖZTÜRK[1,*]

[1]Gümüşhane Üniversitesi, Müh. Fak. Jeofizik Müh. Böl., TR-29100, Bağlarbaşı, Gümüşhane.

Geliş tarihi/*Received* 14.02.2012
Düzeltilerek geliş tarihi/*Received in revised form* 10.04.2012
Kabul tarihi/*Accepted* 17.04.2012

Özet

Bu çalışmada, farklı veri grupları kullanarak deneysel ilişkiler için teorik, pratik ve doğru tahmin aracı olarak farklı regresyon yöntemlerinin uygulanması üzerine bir değerlendirme yapılmıştır. Sayısal merkezli disiplinlerde karşılaşılan birçok pratik problem, doğrusal denklem sistemlerinin hesabı için en uygun çözümün bulunmasını gerektirir. Bu amaçla, dört farklı regresyon normu arasında detaylı bir karşılaştırma yapılmıştır. Tahmin yöntemleri olarak (1) L_2 Norm veya En Küçük Kareler Yöntemi, (2) L_1 Norm veya En Küçük Toplamlı Mutlak Sapma, (3) Toplam En Küçük Kareler veya Ortogonal Regresyon ve (4) Robust Regresyon yöntemleri kullanılmıştır. Veri setleri için doğrusal bir regresyondaki uyumun kalitesini değerlendirmek ve en iyi deneysel ilişkiyi elde edebilmek için, oldukça kullanışlı ve pratik bir uygulama olarak tüm regresyon modelleri için ilişki katsayıları hesaplanmıştır. Detaylı istatistiksel analizler için, literatürde mevcut olan çalışmalardan derlenmiş üç farklı veri seti kullanılmıştır. Sonuçlar, kümelenmiş veri grupları için deneysel ilişkilerin temsilinin En Küçük Toplamlı Mutlak Sapma veya Robust Regresyon yöntemleri ile buna karşın dağınık veri grupları içinse En Küçük Kareler veya Ortogonal Regresyon yöntemleri ile daha uygun ve güvenilir olarak yapılabileceğini göstermektedir.

Anahtar Kelimeler: *L_1, L_2, Robust, Ortogonal regresyon, İlişki katsayısı.*

*Serkan ÖZTÜRK, serkanozturk@gumushane.edu.tr, Tel: (456)233 74 25

An analysis on the application of statistical regression methods for different data sets

Abstract

In this study, an assessment on the application of different regression methods is made by using the different data sets as a theoretical, practical and correct estimation tool for empirical relationships. Many practical problems encountered in quantitative oriented disciplines entail finding a best approximate solution to an over determined system of linear equations. For this purpose, a detailed comparison is made among four different regression norms. The estimation procedures are considered as (1) L_2 or Least Squares Regression, (2) L_1 or Least Sum of Absolute Deviations Regression, (3) Total Least Squares or Orthogonal Regression and, (4) Robust Regression. In order to assess the quality of the fit in a linear regression and in order to select the best empirical relationship for data sets, the correlation coefficients are calculated for all regression models as a quite simple and very practicable tool. For the detailed statistical analyses, three data sets compiled from different examples in the literature are used. The results show that the representation of empirical relationships will be made as more suitable and reliable by Least Sum of Absolute Deviations or Robust regressions for clustered samples whereas by Least Squares or Orthogonal regressions for scattered data.

Key Words: *L_1, L_2, Robust, Orthogonal regression, correlation coefficient.*

1. Giriş

Veri noktalarının parametrik eğriler ve yüzeylerle olan uyumu birçok bilim dalının çalışma konusudur. Çok sayıda standart istatistiksel yazılım, bazı bağımlı değişken ile tepki arasında ve bağımsız değişken ile varsayılan tahmin edici arasında bir ilişkinin var olup olmadığını hesaplamak için veri seti ile kolayca uyum sağlayan regresyon tahminlerinin kullanılmasına olanak sağlar. Dolayısıyla, araştırmacılar varsayılan bazı ilişkilere dayalı olarak tepki değişkeninin değerini hesaplamaya ihtiyaç duyarlar ve varsayılan ilişkiden somut bir model geliştirmek isteyebilirler. Bununla birlikte, birçok olayda bir ilişkinin varsayılan veya kabul edilen matematiksel model ile geçerli olduğunu söylemek mümkün olmayabilir. Ayrıca, regresyon yöntemlerinin bu değişken seçimi, yaygın bir tahmin yöntemi kullanan karmaşık veri gruplarına doğrudan uygulanamayabilir. Bunun yerine, veriler bu değişkenlerin mümkün veri gruplarından elde edilmelidir ve söz konusu olan bağımlı ve bağımsız değişkenler arasındaki deneysel bir ilişki veriye dayalı olarak belirlenmelidir [1].

Uygulamalı istatistikte önemli bir problem, gözlenmiş veriye uyum sağlayacak parametrik regresyon modellerinin yeterliliğinin veya uygunluğunun araştırılmasıdır. Etkili ve doğru bir regresyon yöntemi, sonuçların doğruluğu açısından oldukça önemlidir ve birçok bilimsel alanda ve mühendislik alanında temel bir araç olarak kullanılır. Regresyon, uygulamalı istatistikte bir tepki ile açıklayıcı

değişken arasındaki ilişkiyi belirlemede en yaygın olarak kullanılan yöntemlerden biridir [2]. Dolayısıyla, bir tepki değişkeni ile bazı değişkenler arasındaki ilişkiyi ortaya koyma problemi önemli bir sorun olabilmektedir. Regresyon yöntemlerinin kullanıldığı birçok alanda üzerinde durulan nokta yöntem ve değişken seçimi ile ilişkilidir. Modelin seçimi genellikle, önem veya tamamıyla doğruluk sırasına göre değişkenleri iyi tahmin edebilme özelliği üzerinde odaklanır. Değişken veya özellik seçimi, doğru değişkenlerin bulunmasında daha önemli bir yere sahiptir. Elbette değişken seçimi, model seçimini doğru olarak yapabilmenin bir yoludur.

Regresyon problemi ile ilişkili olarak matematikte, istatistikte ve bilgisayar bilimlerinde büyük eksiklikler mevcuttur. Değişik içeriklere sahip mevcut yöntemlerdeki farklılıklara rağmen çoğu regresyon yönteminin temeli, klasik optimizasyon (en iyi değer bulma) teorisi ve optimizasyon tekniklerdir. Bununla birlikte, sayısal sinyal işleme ve diğer sayısal merkezli disiplinlerde karşılaşılan çoğu pratik regresyon problemi, doğrusal denklem sistemlerinin hesabı için en uygun çözümün bulunmasını gerektirir. Dolayısıyla, araştırmacılar doğrusal bir denklem sisteminin veri özelliklerinin tanımlanması için doğrusal regresyon yöntemlerinin teorik kısımlarını kullanırlar [3]. Herhangi bir veri seti için regresyon tekniklerinin kullanımı ile ilgili temel problemlerden bir tanesi de, değişkenin önemli bir bileşeninin ve denklem hatasının ihmal edilmesidir. Bu problem, kabul edilen değişimin boyutuna bağlı olarak ölçüm hatasının ya üzerinde ya da altında bir düzeltmeye sebep olacaktır. Ayrıca, regresyon modelleri için ilişki katsayısının hesaplanması oldukça kullanışlı ve kabul edilebilir bir değerlendirme yöntemidir ve model uyumunun belirlenmesi, regresyon yöntemi için katsayının hesaplanması gibi bazı kriterlerin hesaplanması (tahmin edilmesi) yoluyla da yapılabilir. Aslında temel bir yöntemin seçim kriteri olarak kullanılması önerilmemekle birlikte çoğu istatistiksel analizde kullanılmaktadır ve regresyon yönteminin tepkisini tahmin etmede seçilen açıklayıcı değişkenin doğruluğu için bir ipucu sağlamaktadır [4].

Çoğu veri işlem uygulamalarında, ele alınan doğrusal denklem sistemleri sabit değildir. Bu tür hesaplamalarda arzu edilen, en iyi çözümü bulmaktır. Literatürde birçok regresyon yöntemi mevcuttur: *En Küçük Kareler Yöntemi* veya L_2 *Norm* [3], *En Küçük Toplamlı Mutlak Sapma* veya L_1 *Norm* [1], *Toplam En Küçük Kareler* veya *Ortogonal Regresyon* [5], *Robust Regresyon* [6], *Temel Bileşenler Regresyonu*, PCR [7], *Geometrik Ortalama Regresyonu,* GMR [8], *Traşlanmış En Küçük Kareler Yöntemi*, LTS [9] ve *Kovaryant Düzeltilmiş Regresyon*, CAR [10] yöntemleri örnek olarak verilebilir. Bu çalışmada ki temel amaç, farklı veri grupları için parametrik regresyon modellerinde teorik ve pratik bir araç olarak doğrusal denklemlerin en iyi regresyon çözümünün kullanımı üzerine bir araştırma yapmaktır. Bu amaçla, üç farklı veri seti kullanılarak en uygun standart istatistiksel modeli ortaya koyabilmek için yukarıda verilen ilk dört regresyon yöntemi arasında bir karşılaştırma yapılmıştır. PCR, GMR, LTS veya CAR gidi diğer yöntemler bu çalışma kapsamında tartışılmayacaktır. Çünkü bu tür yöntemler daha özel alanlar içindir ve jeofizik uygulamalarda kullanımı pek yaygın olmayan yöntemlerdir. Yukarıda bahsedilen tüm regresyon yöntemleri hakkında literatürde pek çok detay bulunabilir [örneğin, 11, 12, 13].

2. Regresyon yöntemleri için algoritma tanımlamalarına genel bir bakış

Bu bölümde, analizler için kullanılacak olan matematiksel algoritmaların ayrıntılı, matematiksel ifadelerin karmaşık olması ve ayrıca bu çalışmanın amacın mevcut yöntemlerin doğruluğunu tartışmak veya yeni bir model üretmek değil; mevcut yöntemlere ait algoritmaları modelleyerek farklı veri grupları üzerinde kullanılabilirliğini tartışmak olduğundan, matematiksel işlemleri ayrıntılı olarak vermek yerine yöntemlere genel bir giriş yapılarak kullanılacak regresyon yöntemlerine ait temel tanımlamalar verilecektir.

Doğrusal regresyon problemi, en önemli veri analizi işlemlerinden biridir. Bu tür problemler, Öklid (Euclidean) geometrisinde ki uzaklıklarla doğal bir ilişki içerisindedir ve çözümler doğrusal cebir işlemleri kullanılarak analitik olarak hesaplanabilir. Doğrusal regresyon modellerini formüle etmek için, bağımlı değişken y üzerinde yapılan n tane ölçüm veya gözlem ile n değerleri bilinen her bir bağımsız değişkenler $\chi_1,\ldots\ldots, \chi_p$ için birkaç tane $p\geq 1$ değerinin olduğu varsayılır. Dolayısıyla denklem aşağıdaki gibi verilir [14]:

$$y=\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix}, \qquad X=\begin{pmatrix} \chi_1^1 \ldots\ldots\ldots \chi_p^1 \\ . \qquad\qquad . \\ . \qquad\qquad . \\ . \qquad\qquad . \\ \chi_1^n \ldots\ldots\ldots \chi_p^n \end{pmatrix}=\begin{pmatrix} x^1 \\ . \\ . \\ . \\ x^n \end{pmatrix}=(x_1,\ldots\ldots,x_p) \tag{1}$$

Burada $y \in R$, n tane gözlemin bir vektörüdür ve X, tasarım matrisi olarak tanımlanan gerçek frekanslı bir n x p matrisidir. Üstelik, $\chi_1,\ldots\ldots, \chi_p$, n bileşenli bir kolon vektörüdür ve $x^1,\ldots\ldots,x^n$, sırasıyla X dizileri ve kolonlarla ilişkili p bileşenlere sahip dizi vektörleridir. İstatistiksel (veya varsayılan) doğrusal regresyon modeli:

$$y = X\beta + \varepsilon \tag{2}$$

denklemi ile verilir. Burada, $\beta^T=(\beta_1,\ldots\ldots,\beta_p)$ doğrusal modelin parametrelerinin bir vektörüdür ve $\varepsilon^T=(\varepsilon_1,\ldots\ldots,\varepsilon_n)$ ise varsayılan ilişkideki hata terimleriyle ilişkili n rasgele değişkenlerinin bir vektörüdür. T üst indeksi, bir vektörün veya bir matrisin transpozesini ifade eder. İstatistiksel modelde sonuç olarak, bağımlı değişken y, hata terimleri ε içerisinde mevcut olan ölçüm hataları veya bazı gürültüleri içeren gözlemler veya ölçümler için elde edilen rasgele bir değişkendir. Başka bir deyişle, karşılaşılan sayısal problem için:

$$y = X\beta + r \tag{3}$$

denklemi yazılabilir. Burada, keyfi olarak verilen bazı sabit parametre vektörü β ve $r^T = (r_1, ..., r_n)$ vektörünün r_i bileşenleri, verilen y gözlemleri, sabit bir tasarım matrisi X ve seçilen vektör $\beta \in R^p$ ile sonuçlanan rezidüellerdir. r rezidüelleri ise istatistiksel model ile, belirli y gözlemlerinin rasgele hata terimleri ε ve parametre düzenlemeleri β ile ilişkilidir. y ve X ile ilişkili olarak doğrusal regresyonda genel kabul $\beta \in R^p$ parametre düzenlemelerini bulmaktır. Yani $r \in R^n$ için sonuç rezidüellerinin dispersiyonun bazı uygun ölçümünü mümkün olduğunca küçük bir olasılıktır [1].

Giloni ve Padberg [1], tasarım matrisi X'teki her $j \in \{1, ..., n\}$ için $\chi_1^j = 1$ durumunun tümüyle olası olduğunu ifade etmiştir. Bu durumda, örneğin p=2 olduğunda iki parametre olması durumuyla ilişkili sınırlı terim β_1 tanımlanır. Eğer her $j \in \{1, ..., n\}$ ve p=1 için $\chi_1^j = 1$ ise iyi bir uyum ölçeği β_1 bulma problemi, y gözlemlerinin bazı iyi merkezcil ölçümlerinin bulunması anlamına gelir.

Ek Küçük Kareler Regresyonu (***L_2 Norm***), en iyi bilinen en eski ve en çok kullanılan faydalı bir eğri uydurma tekniğidir. *L_2 Norm*, en küçük kareler optimizasyonunun en temel şeklidir ve *L_2 Norm*unun temel doğrusal-cebirsel problemleri için örnekleme algoritmaları, en temel regresyon problemlerinden bir tanesidir. Dolayısıyla, birçok farklı bilimsel alan yanında matematik ve istatistiksel veri analizinde çok sayıda uygulamaları mevcuttur. Bu istatistiksel doğrusal regresyon modeli yaklaşık 200 yılı aşkın bir süredir yaygın bir şekilde kullanılmaktadır. Hata terimi ε'nin normal (Gauss veya üstel) dağılım gösterdiği varsayımı altında etkili bir istatistiksel yaklaşımdır. Doğrusal regresyon modelinin istatistiksel özellikleri yanında, verinin uyumu, regresyon katsayılarının alt setinin ve/veya özgün bir kalitenin değerlendirilmesi için oluşturulur. Sonuç olarak bu yaklaşım, sabit sayıdaki dış değerlere sahip büyük örnekleri içeren çok büyük veri setlerinin çalışılmasını içeren durumlarda kısmen faydalı bir değerlendirme yöntemi olarak kullanılabilir [1, 3, 4].

En küçük kareler doğrusal eğri uydurma tahminlerinin verideki anormal gözlemlere karşı oldukça hassas olduğu bilinir ve bunun bir sonucu olarak ta çok daha güçlü tahminler alternatif modeller olarak üretilmiştir. İlk üretilen yöntemlerden bir tanesi de ***En Küçük Toplamlı Mutlak Sapma*** (***L_1 Norm***) regresyonudur. Burada regresyon katsayısı, rezidüellerin tüm değerlerinin toplamının minimum yapılması ile tahmin edilir. *L_1* regresyonu, birçok araştırmacı tarafından 1960'lı yıllardan sonra yeni bir çözüm yöntemi olarak önerilmiştir [15, 16]. *L_1* regresyonu, en küçük karelere daha güçlü bir alternatif olarak büyük oranda kullanılmaz. Çünkü tek bir gözlemden bile güçlü bir şekilde etkilenir. *L_1* regresyonu için asimptotik teori *L_2* regresyonu kadar iyi gelişmemiştir. Bu bir dereceye kadar doğru olmakla birlikte yüksek analizli regresyon tahminleri içinde doğrudur. Ayrıca, *L_1* regresyon tahmini, anormal tahminli gözlemler için her zaman güçlü bir analiz yöntemi değildir. Yani, düşük bir analiz noktasına sahiptir [1, 3].

Basit doğrusal regresyon modellerinde değişken tahminlerindeki hatalar için en yaygın olarak bilinen tekniklerden bir tanesi de ***Ortogonal Regresyon*** (***Toplam En Küçük Kareler***) yöntemidir. Bazen bilinen hata değişim oranının sınırlı olması durumunda ***fonksiyonel maksimum olasılık tahmini*** olarak ta isimlendirilir. Olağan doğrusal regresyon analizlerinde amaç, uyumlu eğri üzerindeki ilişkili y değerleri ile y veri değerleri arasındaki düşey uzunlukların karelerinin toplamını minimum yapmaktır. Ortogonal regresyon analizinde ise amaç, veri noktalarından uyumlu eğriye olan ortogonal (dik) uzaklıkları minimum yapmaktır. Dolayısıyla, varsayım geçerli ise, ortogonal regresyon mükemmel olarak kabul edilebilecek bir tahmin değerlendirme yöntemidir. Bununla birlikte bu yöntem hesaplamalardaki denklem hatalarını dikkate almaz. Bu iyi bilenen ortogonal regresyon tahmini eski bir yöntemdir ve birçok çalışmada kullanılmıştır [5, 8, 17, 18]. Ortogonal regresyon, sadece ölçüm hata değişim oranının olağan tahmini değildir ve bu kullanımı dikkatli bir denklem hata değerlendirmesini içermelidir.

En küçük kareler regresyonundaki en ciddi problem dış değerlerin çok güçlü olmamasından kaynaklanır. Eğer, kötü veri noktası sadece bir değer bile olsa bu değer çözüm üzerinde güçlü bir etkiye sebep olacaktır çünkü dış değerler regresyon parametreleri üzerinde güçlü bir etkiye sahiptir. Basit bir çözüm, kötü uyumlu veri noktasını tekrarlı olarak hesap dışı bırakmak ve kalan veriyi kullanarak en küçük kareler uyumunu yeniden hesaplamaktır. Diğer bir yaklaşım ise ***Robust Regresyon*** olarak isimlendirilen ve anormal veri için en küçük kareler kadar kullanışlı olmayan bir uyum kriterini kullanmaktır. Robust regresyon için en yaygın genel yöntem Huber [6] tarafından tanımlanan M-tahminidir. Doğrusal olmayan regresyon modelleri birçok alanda önemli bir rol oynar. Doğrusal olmayan bir modelin parametrelerin tahmini için klasik en küçük kareler (veya maksimum olasılık) yöntemi birçok durumda yaygın olarak kullanılır. Bununla birlikte, bu klasik yöntemlerin dış değerlere ve belli başlı dağılımlardan olan diğer uzaklıklara çok hassas olduğu bilinir. Regresyon modellerinin tahmininde çoğu güçlü gelişmeler, maksimum olasılık yöntemleri veya en küçük karelerin genelleştirilmesine dayalıdır [14]. Robust regresyon yöntemi bu uç değerlerden çok az etkilenir. Bununla birlikte, Robust regresyon tahminlerinin davranış değerlendirmesinde küçük örnekli asimptot teknikleri çok faydalıdır. Robust regresyon tahmininin kullanımı Huber [6]'dan başlar. Bununla birlikte birçok Robust regresyon tekniği farklı kaynaklarda mevcuttur [19, 20, 21].

3. Regresyon yöntemleri için ilişki katsayılarının hesaplanması

Regresyon analizi, gözlenmiş veriye matematiksel modelin uyumunu sağlamak için farklı disiplinlerde araştırmacılar tarafından yaygın olarak kullanılmaktadır. Eğer hata terimleri eğrilerden bağımsız ise, benzer ise ve normal olarak bağımsız bir dağılım gösteriyorsa En küçük kareler için olağan tahmin teknikleri etkilidir. Bir regresyon modelinde gözlenmemiş rasgele bozukluklar sıkça normal dağılmış olarak kabul ediliyorsa, gerçek veride sıkça dış değerlerin çokluğu söz konusudur. Bu hatalar yanlış ölçümlerin veya insan kaynaklı kayıt hatalarının sonucu olmasına rağmen birçok dış değer asimetrik hata dağılımından kaynaklanır. Bu tür olaylarda, dış değerleri elemek uygun olmaz çünkü bu değerler süreci devam ettiren gerçek veriyi temsil ederler [22].

Regresyon analizlerinde karşılaşılan en önemli problemlerden bir tanesi de, verilen bir veri seti için uygun olasılık dağılımının seçimidir. Çünkü bu dağılım güçlü ve doğru bir yaklaşım sunabilir. Literatürde verildiği gibi uygun dağılımın seçimi için kesin bir kural veya parametre tahmin tekniği yoktur ve farklı dağılımlar uygulanarak en iyi model seçilmelidir. Sonuç olarak, en iyi uyumu sağlayan modelin seçimi frekans analizlerinde oldukça önemli bir kuraldır. Birçok olayda, uygun dağılımın seçimi uyum kalitesinin değerlendirilmesine dayalı olarak yapılır. Uyum kalitesi tekniği, varsayılan bir olasılık dağılımı ile örnek verinin nasıl iyi bir uyum sağlayacağının tanımlanmasını gerektirir. Mühendislik çalışmalarında kullanılan birkaç uyum kalitesi tekniği geliştirilmiştir. Bu yöntemler arasında seçim kriteri olarak, ilişki katsayısı (R^2 veya bazen r kullanılır) güçlü ve kabul edilebilir bir yöntem olarak bilinmektedir. R^2 yalnızca kovaryans (öz ilişki) hatasına bağlı olmasına rağmen model uyum değerlendirmesinde önemli bir rol oynar. Tek başına bir uyum aracı olarak kullanılmamalıdır. Fakat uyum kalitesi için hızlı ve geçerli bir modeldir [23].

R^2 genellikle, açıklayıcı değişkenlerle birlikte doğrusal ilişkinin ortaya koyduğu tepki değişkeninin değişim yüzdesini tahmin eden nitelik olarak ifade edilir. Oransal olarak şu şekilde verilir:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

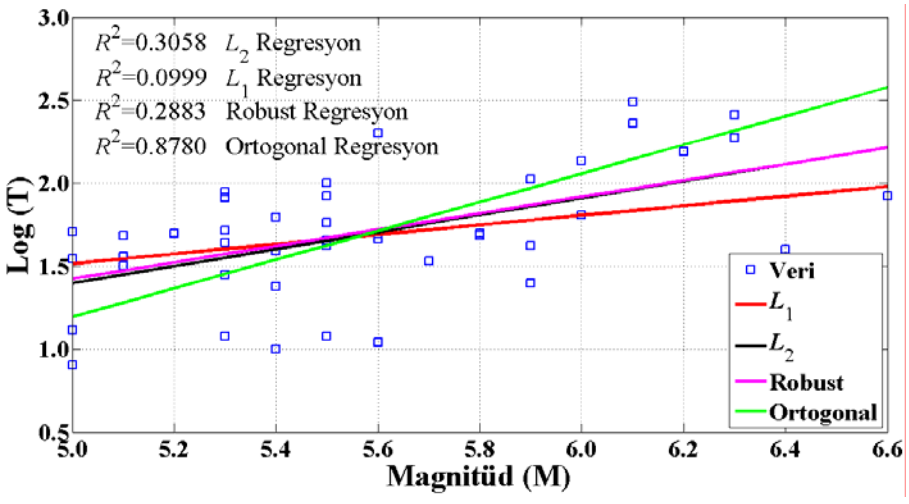
Burada, ESS, TSS ve RSS sırasıyla hesaplanan, toplam ve rezidüel kare toplamlarıdır. Doğrusal modelde sınırları belirli bir terim varsa, bu ilişki katsayısının hesabı genellikle y_i ve $\hat{y}_i$ arasındaki ilişki katsayısının karesine eşittir [24]:

$$R^2 = \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \right)^2 \tag{5}$$

Burada $\bar{y}$ ve $\bar{\hat{y}}$, sırasıyla y_i gözlemlerinin ortalama değerlerini ve uyum sağlayan $\hat{y}_i$ değerinin ortalamasını gösterir. (5) denklemi iyi bir yorumdur yani R^2 regresyon modelinin uyum kalitesini ölçer. İlişki katsayısı konum ve ölçekle değişmez ve istatistiksel olarak standart sapma ve $\bar{y}$ ortalamasından bağımsızdır. Netice olarak, R^2 olasılık çiziminin doğrusallığını ölçer ve uyumun nitel bir değerlendirmesine olanak sağlar. Eğer R^2, 1'e yakın ise gözlemlerin uyumlu dağılım gösterdiği kabul edilir [23].

4. Farklı veri grupları üzerinde istatistiksel regresyon yöntemlerinin uygulamaları

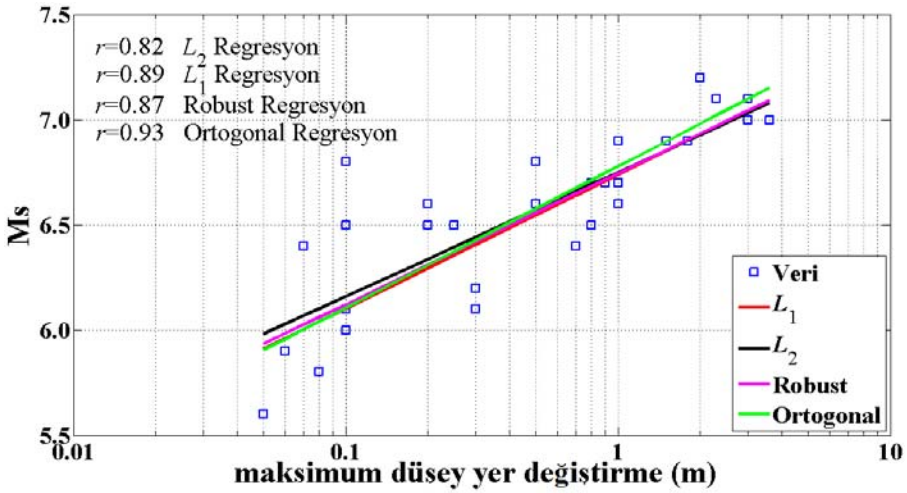
Bu çalışmanın temel amacı, farklı veri grupları için istatistiksel regresyon yöntemlerinin kullanımını ortaya koymaktır. Belirli veri setlerinin regresyon yöntemleriyle iyi temsil edilemediği ve sonuç olarak en yüksek ilişki katsayısına sahip en iyi regresyon uyumunun seçilebileceği anlamına gelir. Bu çalışma kapsamında analiz edilen ilk veri seti Drakatos ve Latoussakis [25]'ten derlenmiştir. Bu veri seti, Yunanistan ve civarında son 25 yılda meydana gelmiş $M_L \geq 5.0$ olan depremlere ait artçı şok dizilerinin tam bir kataloğuna aittir. Drakatos ve Latoussakis [25], ana şok magnitüdü ile yüzey kırık uzunluğu arasındaki ilişkiyi, ana şok magnitüdü ile artçı şok dizisinin devam etme süresi, artçı şokların sayısı, en büyük artçı şokun magnitüdü ve ana şoktan sonraki zaman arasındaki ilişkileri hesaplamışlardır. Bu veri grubu için farklı regresyon yöntemlerinin kullanımını göstermek amacıyla ana şok magnitüdü ile artçı şok dizisinin devam etme süresi arasındaki ilişki irdelenmiştir. Drakatos ve Latoussakis [25], bu iki veri arsındaki ilişkiyi R^2=0.3058 gibi oldukça düşük bir ilişki katsayısına sahip olan $\log(T) = 0.51 * M - 1.15$ bağıntısıyla vermiştir. Şekil 1'deki regresyon sonuçlarından görüldüğü gibi, bu veri grubu için en iyi regresyon modeli Ortogonal regresyon olarak verilebilir. Çünkü bu regresyon için ilişki katsayısı (R^2=0.878) diğer modellere göre en yüksektir. Şekil 1'de görüldüğü gibi Drakatos ve Latoussakis [25], yaptıkları hesaplamalarda L_2 normunu (En Küçük Kareler Regresyonu) kullanmışlardır. Bununla birlikte L_1 normu için elde edilen ilişki katsayısı (R^2=0.0999) ve Robust regresyon için elde edilen ilişki katsayısı (R^2=0.2883), L_2 ve Ortogonal regresyondan elde edilenden küçüktür. Ayrıca, bu veri seti için farklı regresyon yöntemlerinden elde edilen ilişkiler, L_1 regresyonu için $\log(T) = 0.29 * M + 0.07$ olarak, Robust regresyon için $\log(T) = 0.49 * M - 1.04$ olarak ve Ortogonal regresyon için $\log(T) = 0.86 * M - 3.13$ olarak hesaplanmıştır. Kriter olarak ilişki katsayısı dikkate alındığında, bu veri seti için Ortogonal regresyon uyumunun diğer regresyon yöntemlerine kıyasla daha uygun ve kullanılabilir olduğu söylenebilir.



Şekil 1. Artçı şok süresi log(T) ile ana şok magnitüdü (M) arasındaki ilişki. Kullanılan veri Drakatos ve Latoussakis [25]'ten derlenmiştir.

Bu çalışmada kullanılan ikinci veri seti Pavlides ve Caputo [26]'dan derlenmiştir. Bu araştırmacılar, Ege bölgesindeki deprem tehlikesi analizlerini değerlendirmek için, ortalama yer değiştirme veya maksimum yer değiştirme ile yüzey kırık uzunluğu değerlerinin mevcut olduğu 36 depremin bir listesini hazırlamışlardır. Şekil 2'de görüldüğü gibi, veri seti üzerinde farklı regresyon

yöntemlerinin değişimini değerlendirmek için bu çalışma kapsamında magnitüd ile maksimum düşey yer değiştirme arasındaki ilişki dikkate alınmıştır. Bu ilişki Pavlides ve Caputo [26] tarafından r=0.82 gibi nispeten iyi bir ilişki katsayısına sahip olarak $Ms = 0.59 * Log(MVD) + 6.75$ ilişkisi ile verilmiştir. Şekil 2'de ise, bu çalışmada hesaplanan ilişki katsayıları L_1 norm (r=0.89) ve Robust regresyon (r=0.87) için hesaplananlara oldukça yakındır. Bununla birlikte, Ortogonal regresyon için hesaplanan ilişki katsayısı (r=0.93) oldukça iyidir Pavlides ve Caputo [26], hesaplamalarında L_2 normunu kullanmışlardır. Bu çalışmada, bu veri seti kullanılarak L_1 regresyonu için $Ms = 0.63 * Log(MVD) + 6.74$ ilişkisi, Robust regresyon için $Ms = 0.62 * Log(MVD) + 6.74$ ilişkisi ve Ortogonal regresyon için $Ms = 0.67 * Log(MVD) + 6.78$ ilişkisi elde edilmiştir.

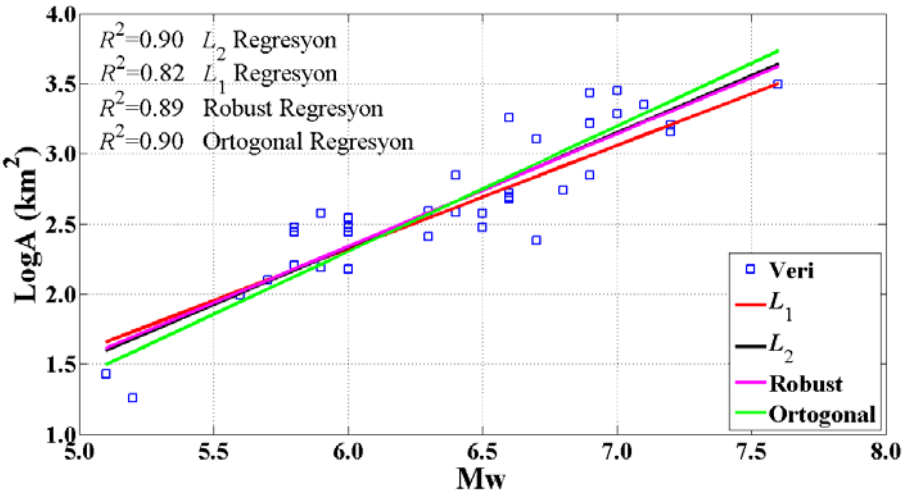


Şekil 2. Magnitüd (Ms) ile maksimum düşey yer değiştirme (MVD) arasındaki ilişki. Kullanılan veri Pavlides ve Caputo [26]'dan derlenmiştir.

Tüm regresyon yöntemlerinde elde dilen ilişkiler Pavlides ve Caputo [26] tarafından elde edilen ilişkilere çok benzerdir. Çünkü farklı regresyonlardan elde edilen ilişki katsayıları birbirlerine çok yakındır. Bununla birlikte, en yüksek ilişki katsayısı dikkate alındığında Ortogonal regresyonun bu veri grubunu diğerlerine oranla daha iyi temsil ettiği söylenebilir.

Çalışma kapsamında analiz edilen son veri seti ise Konstantinou vd., [27]'den derlenmiştir. Araştırmacılar, Akdeniz bölgesinde meydana gelmiş olan 36 depreme ait artçı şok dizisi için, hem yüzey dalgası magnitüdü (Ms) hem de moment magnitüdünün (Mw) bir fonksiyonu olarak artçı şok bölgesinin uzunluğu (L), genişliği (W) ve alanı (A) arasındaki deneysel ilişkileri ortaya koyabilmek amacı ile istatistiksel bir regresyon analizi yapmışlardır. Fay boyu ve magnitüd bağımlı tüm regresyonlarda genel ilişkiler, Ms veya Mw bağımlı uzunluk, genişlik ve alanın bir fonksiyonu olarak $Log(L,W,A)=a+b(Ms,Mw)$ denklemiyle verilmiştir. Bu çalışma kapsamında ise Konstantinou vd., [22]'de tanımlanan yüzey kırık uzunluğu ile moment magnitüdü arasındaki ilişki istatistiksel regresyon yöntemleriyle analiz edilmiştir. Konstantinou vd., [27], bu iki parametre arasındaki ilişkiyi L_2 normunu kullanarak oldukça yüksek bir ilişki katsayısına (r=0.9) sahip $LogA = 0.81 * Mw - 2.57$ denklemiyle vermişlerdir. Şekil 3, bu çalışmada kullanılan farklı yöntemlerin regresyon uyumunu ve ilişki katsayılarını göstermektedir. İlişki katsayıları, L_1 regresyon için r=0.82, Robust regresyon için

r=0.89 ve Ortogonal regresyon için r=0.9 olarak hesaplanmıştır. Bu veri seti için hesaplanan deneysel ilişkiler ise L_1 regresyon için $LogA = 0.74 * Mw - 2.10$, Robust regresyon için $LogA = 0.80 * Mw - 2.49$ ve Ortogonal regresyon için $LogA = 0.89 * Mw - 3.06$ eşitlikleri ile verilmiştir. Farklı regresyon teknikleri kullanılarak hesaplanan bu ilişkiler Konstantinou vd., [27] tarafından elde dilen sonuçlara çok yakındır. Ayrıca, bu veri seti için Konstantinou vd., [27] tarafından en küçük kareler yöntemi kullanılarak elde edilen ilişki katsayısı, bu çalışmada Ortogonal regresyon kullanılarak elde edilen ilişki katsayısıyla aynıdır. Sonuç olarak, başta Ortogonal regresyon yöntemi olmak üzere tüm regresyon yöntemleri bu veri seti için daha uygundur ve en iyi sonuçları vermektedir.



Şekil 3. Moment magnitüdü (Mw) ile yüzey kırık alanı (A) arasındaki logaritmik ilişki. Kullanılan veri Konstantinou vd., [27]'den derlenmiştir.

5. Tartışma ve sonuçlar

Bu çalışmada, farklı istatistiksel regresyon analizlerinin uygulamaları tanımlanmış ve farklı bölgelerden derlenen veri grupları için deneysel ilişkilerin bir karşılaştırılması yapılmıştır. Bu amaçla, üç farklı veri seti kullanılmıştır: (*i*) ana şok magnitüdü (M) ile artçı şokları devam etme süresi arasındaki ilişki için Drakatos ve Latoussakis [25]'ten, (*ii*) yüzey dalgası magnitüd (Ms) ile maksimum düşey yer değiştirme (MVD) arasındaki ilişki için Pavlides ve Caputo [26]'dan ve (*iii*) yüzey kırık alanı (A) ile moment magnitüdü (Mw) arasındaki ilişki için Konstantinou vd., [27]'den derlenen veri. Bu hesaplamalar sonucunda, farklı araştırmacılara ait bu örnek gruplarını kullanarak verilen bir veri seti için en iyi istatistiksel regresyon yönteminin seçimine nasıl karar verilebileceği konusunda bazı önemli yaklaşımlar ortaya konulmuştur.

Drakatos ve Latoussakis [25], 34°-42°K ve 19°-30°D arasında 1971-1997 yılları arasında meydana gelmiş $M_L \geq 5.0$ olan tüm depremler için ana şok magnitüdü ile artçı şokların devam etme süreleri arasında bir ilişki önermişlerdir. En Küçük Kareler yöntemini kullanmışlar ve oldukça küçük bir ilişki katsayısına (0.3058) sahip olan $\log(T) = 0.51 * M - 1.15$ ilişkisini elde etmişlerdir. Bununla birlikte, bu çalışmada kullanılan Ortogonal regresyon, bu veri seti için en iyi istatistiksel regresyon yöntemi olarak gözükmektedir. Çünkü Şekil 1'de görüldüğü gibi, en yüksek ilişki katsayısı (0.878) bu ilişkide elde edilmiştir. Dolayısıyla, Drakatos ve Latoussakis [25] tarafından ana şok magnitüdü ile artçı şokların devam etme süreleri arasındaki bu ilişki Ortogonal regresyon yöntemiyle $\log(T) = 0.86 * M - 3.13$ olarak verilebilir.

Pavlides ve Caputo [26], Ege bölgesinde deprem tehlikesi analizi yapabilmek için 36 depremi içeren bir katalog kullanmışlar ve ana şok magnitüdü ile maksimum düşey yer değiştirme arasında bir ilişki önermişleridir. En Küçük Kareler yöntemini kullanarak, nispeten iyi bir ilişki katsayısına (0.82) sahip olan $Ms = 0.59 * Log(MVD) + 6.75$ ilişkisini vermişlerdir (Şekil 2). Bu çalışma kapsamında ise, Ortogonal regresyon kullanılarak oldukça yüksek bir ilişki katsayısına (0.93) sahip olan $Ms = 0.67 * Log(MVD) + 6.78$ ilişkisi elde edilmiştir. Farklı istatistiksel regresyon yöntemlerinden elde edilen ilişki katsayıları birbirine çok yakın olmasına rağmen, Ortogonal regresyon yönteminin bu veri setini diğerlerine kıyasla daha iyi temsil ettiği söylenebilir.

Konstantinou vd., [27], hem yüzey dalgası magnitüdü hem de moment magnitüdünün bir fonksiyonu olarak artçı şok alanının boyutları arasında deneysel ilişkiyi ortaya koyabilmek için Akdeniz civarında meydana gelmiş 36 depremi içeren bir katalog kullanmışlar ve En Küçük Kareler yöntemi ile oldukça iyi bir ilişki katsayısına (0.9) sahip $LogA = 0.81 * Mw - 2.57$ ilişkisini elde etmişlerdir. Bu çalışma kapsamında, tüm regresyon yöntemleriyle birlikte özellikle Ortogonal regresyon yöntemi kullanılarak elde edilen ilişki katsayısı (0.9), Konstantinou vd., [27] tarafından önerilen ilişki katsayısıyla çok yakın değerlere sahiptir. Sonuç olarak, çalışma kapsamında Ortogonal regresyon yöntemi ile elde edilen $LogA = 0.89 * Mw - 3.06$ ilişkisi bu veri seti için daha uygundur denilebilir.

Farklı veri grupları üzerinde farklı istatistiksel regresyon yöntemleri ile yapılan bu değerlendirmeler dikkate alındığında, ayrıntılı analizlerden genel olarak şöyle bir sonuca varılabilir: "E***n güvenilir regresyon analizleri, kümelenme gösteren veri grupları için En Küçük Toplamlı Mutlak Sapma (L_1 Norm) veya Robust regresyon yöntemleri ile, dağınıklık gösteren veri grupları için ise En Küçük Kareler Yöntemi (L_2 Norm) veya Ortogonal regresyon yöntemleri ile yapılabilir***".

Teşekkür

İstatistiksel regresyon yöntemlerinin algoritmalarının modellenmesinde yardımlarını esirgemeyen Doç. Dr. Hakan Karslı (KTÜ)'ya, yapıcı tavsiyelerde bulunan hakem kuruluna ve editöre teşekkür ederim. Bu çalışma Gümüşhane Üniversitesi Bilimsel Araştırma Projesi (GÜBAP) tarafından desteklenmektedir (Proje no: 2012.02.1717.2).

Kaynaklar

[1] Giloni, A., Padberg, M., 2002. Alternative Methods of Linear Regression, *Mathematical and Computer Modelling*, 35: 361-374.

[2] Durio, A., Isaia, E.D., 2003. Parametric Regression Models by Minimum L_2 Criterion. A Study on the Risks of Fire and Electric Shocks of Electronic Transformers, *Developments in Applied Statistics*, 19: 69-83.

[3] Cadzow, J.A., 2002. Minimum ℓ_1, ℓ_2 and ℓ_∞ Norm Approximate Solutions to an Overdetermined System of Linear Equations, *Digital Signal Processing*, 12: 524-560.

[4] Renaud, O., Victoria-Feser, M.P.,2010. A robust coefficient of determination for regression, *Journal of Statistical Planning and Inference*, doi:10.1016/j.jspi.2010.01.008.

[5] Carrol, R.J., Ruppert, D., 1996. The use and misuse of orthogonal regression estimation in linear errors-in-variables models, *The American Statistician*, 50: 1-6.

[6] Huber, P.J., 1964. Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35: 73–101.

[7] Maronna, R., 2005. Principal components and orthogonal regression based on robust scales, *Technometrics*, 47: 264-273.

[8] Leng, L., Zhang, T., Kleinman, L., Zhu, W., 2007. Ordinary Least Square Regression, Orthogonal Regression, Geometric Mean Regression and their Applications in Aerosol Science, *Journal of Physics*: Conference Series 78, doi:10.1088/1742-6596/78/1/012084.

[9] Rousseeuw, R.J., Leroy, A.M., 1987. *Robust Regression and Outlier Detection*, New York: Wiley.

[10] Şentürk, D., Nguyen, D.V., 2006. Estimation in covariate-adjusted regression, *Computational Statistics & Data Analysis,* 50: 3294-3310.

[11] Branham, Jr, R.L., 1982. Alternatives to least-squares, *Astr. J.*, 87: 928-937.

[12] Spiess, M., Hamerle, A., 2000. A comparison of different methods for the estimation of regression models with correlated binary responses, *Computational Statistics & Data Analysis*, 33: 439-455.

[13] Sen, A., Srivastava, M., 1990. *Regression Analysis: Theory, Methods, and Applications*, Springer-Verlag, NewYork.

[14] Giloni, A., Simonoff, J.S., Sengupta, B., 2006. Robust weighted LAD regression, *Computational Statistics & Data Analysis*, 50: 3124-3140.

[15] Blattberg, R.C., Sargent, T., 1971. Regression with non-Gaussian stable disturbances: Some sampling results, *Econometrica*, 39: 501-510.

[16] Huber, P.J.1987. *The place of the L_1 norm in robust estimation. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1 norm and Related Methods*, North-Holland, Amsterdam.

[17] Kendal, M.G., Stuart, A., 1979. *The Advanced Theory of Statistics*, vol 2, 4th edition. Hafner, New York.

[18] Weisberg, S. 1985. *Applied Linear Regression, second edition*, John Wiley & Sons. New York.

[19] Huber, P.J., 1981. *Robust Statistics*, Wiley, New York.

[20] Field, C.A., 1997. Robust regression and small sample confidence intervals, *Journal of Statistical Planning and Inference*, 57: 39-48.

[21] Sinha, S.K., Field, C.A., Smith, B., 2003. Robust estimation of nonlinear regression with autoregressive errors, *Statistics & Probability Letters*, 63: 49-59.

[22] Boyer, B.H., McDonald, J.B., Newey, W.K., 2003. A comparison of partially adaptive and reweighted least squares estimation, *Econometric Reviews*, 115-134.

[23] Heo, J.H., Kho, Y.W., Shin, H., Kim, S., Kim, T., 2008. Regression equations of probability plot correlation coefficient test statistics from several probability distributions, *Journal of Hydrology*, 355: 1-15.

[24] Greene, W., 1997. *Econometric Analysis, third ed*, Prentice-Hall, Englewood Cliffs, NJ.

[25] Drakatos, G., Latoussakis, J., 2001. A catalog of aftershock sequences in Greece (1971-1997): Their spatial and temporal characteristics, *Journal of Seismology,* 5: 137-145.

[26] Pavlides, S., Caputo, R., 2004. Magnitude versus faults' surface parameters: quantitative relationships from the Aegean region, *Tectonophysics*, 380: 159-188.

[27] Konstantinou, K.I., Papadopoulos, G.A., Fokaefs, A., Orphanogiannaki, K. 2005. Empirical relationships between aftershock area dimensions and magnitude for earthquakes in the Mediterranean Sea region, *Tectonophysics*, 403: 95-115.