T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

ÇOKLU DOĞRUSAL REGRESYON ANALİZİNDE ÇAKI TAHMİN YÖNTEMİ VE TEST PROBLEMLERİNE KATKILAR

Tolga ZAMAN

İSTATİSTİK ANABİLİM DALI

SAMSUN
2017

TEZ ONAYI

Tolga Zaman tarafından hazırlanan "Çoklu Doğrusal Regresyon Analizinde Çakı Tahmin Yöntemi ve Test Problemlerine Katkılar" adlı tez çalışması 25/08/2017 tarihinde aşağıdaki jüri tarafından Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda **Doktora Tezi** olarak kabul edilmiştir.

Danışman Doç. Dr. Kamil ALAKUŞ
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Jüri Üyeleri

Başkan Prof. Dr. Vedat Suat ERTÜRK
Ondokuz Mayıs Üniversitesi
Matematik Anabilim Dalı

Üye Doç. Dr. Kamil ALAKUŞ
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Yüksel ÖNER
Ondokuz Mayıs Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Mehmet GÜRCAN
Fırat Üniversitesi
İstatistik Anabilim Dalı

Üye Doç. Dr. Necati Alp ERİLLİ
Cumhuriyet Üniversitesi
Ekonometri Anabilim Dalı

Yukarıdaki sonucu onaylıyorum. …/…/20…

Prof. Dr. Bahtiyar ÖZTÜRK
Enstitü Müdürü

ETİK BEYAN

Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez içindeki bütün bilgilerin doğru ve tam olduğunu, bilgilerin üretilmesi aşamasında bilimsel etiğe uygun davrandığımı, yararlandığım bütün kaynakları atıf yaparak belirttiğimi beyan ederim.

25.08.2017

Tolga ZAMAN

ÖZET

Doktora Tezi

ÇOKLU DOĞRUSAL REGRESYON ANALİZİNDE ÇAKI TAHMİN YÖNTEMİ VE TEST PROBLEMLERİNE KATKILAR

Tolga Zaman

Ondokuz Mayıs Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Doç.Dr. Kamil Alakuş

Regresyon analizi, iki ya da daha fazla değişken arasındaki ilişkinin modellenmesi amacı ile çok sık kullanılmaktadır. Günümüzde değişkenler arasındaki ilişkilerin belirlenmesinde önemli bir yeri olan parametrik regresyon analizi, ilişkinin doğrusal olduğu temeline dayanmakta ve güçlü varsayımların sağlanmasını gerektirmektedir. Bu varsayımlar sağlanmadığında ilgilenilen ortaya çıkarılması gerçeğe uymamakta ve yanlış sonuçlara neden olabilmektedir. Gerçek hayatta da eldeki veriler her zaman parametrik regresyon analizindeki varsayımları sağlamadığı için uygulama alanında zorluklarla karşılaşılmaktadır. Parametrik olmayan regresyon analizi ise hiçbir varsayıma yer vermediği için, değişkenler arasındaki ilişkinin fonksiyonel yapısının ortaya çıkarılmasında iyi bir alternatiftir. Çakı yeniden örnekleme yöntemi parametrelerin güven aralıklarının tahmininde ve istatistiksel testlerin performanslarının değerlendirilmesinde kullanılan bir tekniktir. Çakı yöntemi veride aşırı uç değerlerin varlığında kullanılmaktadır. Yöntemin uygulanması aşamasında veri kümesindeki her bir gözlem değeri bir kez dışarıda bırakılarak geride kalan gözlemlerden örneklem istatistikleri hesaplamaya dayanmaktadır.

Bu çalışmada, çoklu doğrusal regresyon analizinde Theil-Sen ve çakı yöntemlerine ait teorik esaslar tanıtılarak, kullanılma neden ve sonuçları açıklanmıştır. Çoklu doğrusal regresyon analizinde Theil-Sen tahminine ait model parametrelerinin çok değişkenli bir medyana bağlı olarak elde edilmesi ve buradan hareketle regresyon katsayılarının çakı tahminleri incelenmiştir. Çakı yöntemi veriden her defasında bir gözlem değerini silmesi bakımından uç değerlerin etkisini azaltan bir yöntemdir. Bu bağlamda çoklu doğrusal regresyon analizinde Theil-Sen yöntemine, çakı yönteminin ikame edilmesiyle Çakı-Theil adında yeni bir tahmin yöntemi önerilmiştir. Önerilen bu yöntemin veride aykırı değerlerin etkisini daha da azaltıp daha güvenilir sonuçlar verdiği gözlenmiştir. Önerilen Çakı-Theil tahmin yönteminin asimptotik özellikleri çeşitli simülasyon çalışmaları ve orijinal veri yapılarıyla incelenmiştir ve önerilen Çakı-Theil tahmin yönteminin ilgilenilen diğer tahmin yöntemlerine nazaran avantajları belirlenmiştir.

Ağustos 2017, 96 sayfa

Anahtar Kelimeler: Theil-Sen tahmini, Çakı tahmini, Önerilen Çakı-Theil tahmini, Spatial medyan, Çoklu doğrusal regresyon, Sağlamlık, Etkinlik.

ABSTRACT

Doctoral Thesis

CONTRIBUTIONS TO THE JACKKNIFE ESTIMATION AND TEST PROBLEMS IN MULTIPLE LINEAR REGRESSION ANALYSIS

Tolga Zaman

Ondokuz Mayis University
Graduate School of Sciences
Departmant of Statistics

Supervisor: Assoc. Prof. Dr. Kamil Alakuş

Regression analysis is often used to determine the relationship between two or more variables. Today, parametric regression analysis has a significant role in determining the relationships between variables, is based on the fact that the relationship is linear, and requires strong assumptions. When these assumptions are not met, interest does not reflect reality and can lead to inaccurate results. In real life, it is possible to meet with difficulties in the field of application because the data at hand does not always meet with the assumptions in parametric regression analysis. As for nonparametric regression analysis, there are no assumptions, therefore, it is a good alternative in discovering the functional structure of the relationship between variables. Jackknife resampling method is a method which is used to estimate the confidence intervals of parameters and determine the performances of the statistical tests. Jackknife method is used when there are extreme peak values in data. At the stage of application of the method, comes from calculating sampling statistics from remaining observations through excluding an observation one time in data cluster.

In this study, the principals regarding Theil-Sen and Jackknife methods in multiple linear regression analysis are introduced and results are explained. In multiple linear regression analysis, the process of obtaining the parameters of Theil-Sen estimation based on a multivariate median and Jackknife estimations of regression coefficients from this point are explored. Jackknife method is a method which reduces the effect of peak values since it deletes an observation value each time. In this regard, a new estimation method named Jackknife-Theil is proposed by substituting Jackknife method to Theil-Sen method in multiple linear regression analysis. It is observed that suggested method reduces the effects of outliers even more and gives more reliable results. The asymptotic features of the suggested Jackknife-Theil estimation method are investigated by using various simulation studies and it is found that the suggested Jackknife-Theil method has advantages over other estimation methods.

August 2017, 96 pages

Key Words: Theil-Sen estimation, Jackknife estimation, Proposed Jackknife-Theil estimation, Spatial median, Multiple linear regression, Robustness, Efficiency.

ÖNSÖZ VE TEŞEKKÜR

Akademik eğitim sürecimin bir üst noktası olan doktora tez çalışmalarım boyunca yardım ve desteğini benden esirgemeyen danışman hocam Sayın Doç. Dr. Kamil Alakuş 'a teşekkürü bir borç bilirim.

Beni bugünlere getirmek için hiçbir fedakârlıktan kaçınmayan aileme sonsuz teşekkür ederim.

Ayrıca, lisans, yüksek lisans ve doktora sürecinde bana yardımcı olan bölümümüz tüm hocalarına teşekkürlerimi bir borç bilirim.

Doktora öğrenimim boyunca, gerek ders aşaması, gerekse de tez aşamasında verdikleri destek ve katkıdan dolayı TUBİTAK ve tüm çalışanlarına sonsuz teşekkürlerimi sunmak isterim.

Ağustos 2017, Samsun

Tolga Zaman

İÇİNDEKİLER DİZİNİ

Sayfa

SİMGELER VE KISALTMALAR

ARE	: Asimptotik Göreli Etkinlik
ÇÇTSE	: Çakı Çoklu Theil-Sen Tahmini
Çmed	: Çok Değişkenli Medyan
ÇTSE	: Çoklu Theil-Sen Tahmini
DHKO	: Deneysel Hata Kareler Ortalaması
EKK	: En Küçük Kareler
HKO	: Hata Kareler Ortalaması
IF	: Etki Fonksiyonu
Med	: Medyan
OMS	: Ortalama Mutlak Sapma
S(t)	: İşaret Fonksiyonu
T(F)	: Medyan Fonksiyonu
TSE	: Theil-Sen Tahmini
$\hat{R}(t)$	: Sıra Fonksiyonu

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

1. GİRİŞ

İstatistiksel çalışmalarda regresyon analizi değişkenler arasındaki ilişkinin belirlenmesinde yaygın olarak kullanılan önemli bir yöntemdir. Bu analiz yöntemi ile değişkenler arasındaki ilişki belirlenerek parametre tahmini ve ileriye dönük tahmin yapılabilmektedir. Regresyon analizinin temelinde, gözlenen bir olayın fonksiyonel yapısının belirlenmesi yatmaktadır. Regresyon analizi yapılırken gözlem değerlerinin ve etkilenilen olayların matematiksel bir gösterimle ifade edilmesi gerekmektedir. Tahmin edilen bu model regresyon modeli olarak anılır. Regresyon modeli sosyal bilimler, tıp, mühendislik, fen bilimleri gibi bir çok alanda çeşitli değişkenler arasındaki ilişkilerin belirlenmesi ve tahmin edilmesi amacıyla uygulanmaktadır.

Regresyon analizi, parametrik, yarı parametrik (Cox regresyon) ve parametrik olmayan regresyon analizi şeklinde ayrılmaktadır. Parametrik regresyon analizinin en karakteristik özelliği bağımlı ve bağımsız değişkenler arasındaki ilişkinin matematiksel formunun önceden biliniyor olmasıdır. Bu regresyon modeli elde edilirken yaygın olarak En Küçük Kareler (EKK) tekniği kullanılmaktadır. Regresyon analizinde EKK yöntemi ile parametre tahminlerinin yapılabilmesi için temel bazı varsayımlar vardır. Bu varsayımlar genellikle hata terimleri ile ilgilidir. Bu varsayımları özetlersek; bağımsız değişken değerlerinin tümü için hata varyanslarının sabit olması, farklı gözlemlerin tesadüfi hata terimleri $(\varepsilon_i, \varepsilon_j)$ birbirinden bağımsızdır. Yani herhangi bir ε_i teriminin başka bir ε_j ile olan ortak varyansı sıfıra eşittir. Hata terimleri arasında otokorelasyon olmaması, hata terimlerinin normal dağılıma uyması ve eğer çoklu regresyon analizi söz konusu ise, bağımsız değişkenler arasında çoklu doğrusal bağlantı probleminin olmaması şeklindeki varsayımların sağlanması gerekmektedir. Varsayımların sağlanmaması durumunda ise, regresyon modeli hakkında yapılan çıkarımlar gerçeği yansıtmamakta ve sonuçların yanlış yorumlanmasına sebep olmaktadır. Yani iyi bir tahmin niteliği taşımazlar. Bu durumda daha iyi tahmin yapılabilmesi için parametrik regresyondaki doğrusallık varsayımının esnetilmesine olanak sağlayan regresyon yöntemlerine ihtiyaç duyulmaktadır. Bu yöntemler parametrik olmayan yöntemler olarak bilinir (Erilli ve Alakuş, 2016). Parametrik olmayan regresyon analizinde ise, modelin şekli önceden belli değildir ve parametrik regresyon analizinde olduğu gibi önemli varsayımlar bulunmamaktadır. Yapılan tek varsayım hata terimlerinin ortalamasının sıfır, varyansınında sonlu bir sayı olmasıdır. Böylece değişkenler arasındaki ilişkinin ortaya çıkarılmasında esneklik sağlanmaktadır. Bu avantajlarının yanısıra parametrik olmayan regresyon yönteminin uygulanmasında bazı zorluklarda mevcuttur. Örneğin; işlemlerin matematik olarak alt yapısının karmaşıklığı ve yoğun bilgisayar kullanımı gerektirmesi bu zorluklardan bazılarıdır. Etkinlik bakımında karşılaştırıldığında ise; parametrik olmayan regresyonun, parametrik regresyon analizine göre daha az etkin olduğu unutulmamalıdır (Eubank, 1988).

Bu tez kapsamında parametrik olmayan regresyon yöntemlerinden biri olan Theil-Sen tahmini (TSE) üzerinde durulacaktır. Theil-Sen yöntemi literatürde Theil-Kendall ya da Theil yöntemi olarak da ifade edilmektedir. Eğimi bulmaya yönelik olarak önerilen Mood-Brown yöntemi hızlı; ancak, çok güvenilir bir yöntem değildir (Zaman ve Alakuş, 2015). Özellikle eğim katsayısını bulmak için önerilen Theil yöntemi daha kullanışlıdır. Bu yöntemde basit doğrusal regresyon modeli:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, 2, \dots, n \qquad \mathbf{(1.1)}$$

şeklinde ifade edilmektedir. Burada β_0 kesim parametresi, β_1 ise eğim parametresi olup, bu parametreler tahmin edilmeye çalışılmaktadır. Basit doğrusal regresyona ait bu parametrelerin tahmin edilmesi için bazı varsayımlar vardır. Bu varsayımlar:

a. Her bir X_i değeri için Y'lerin bir alt kitlesi ve ε_i'ler karşılıklı bağımsızdır.
b. X_i'ler tekrarsız olup, $X_1 < X_2 < \cdots < X_n$ sırasındadır.
c. Veri kümesi $(X_1, Y_1), \dots, (X_n, Y_n)$ şeklinde n gözlem ikililerinden oluşmaktadır.

Bu varsayımların doğruluğu altında, β tahminine ulaşmak için tüm olası $S_{ij} = \frac{(Y_j - Y_i)}{(X_j - X_i)}$ eğim değerleri $(i < j\ için)$ hesaplanır. $N = \binom{n}{2}$ tane S_{ij} eğimi elde edilir. β tahmini S_{ij} değerlerinin medyanı olarak hesaplanır. Şöyle ki;

$$\widehat{\beta_1} = Ortanca(S_{ij}) \qquad \mathbf{(1.2)}$$

ve sabit terim ise;

$$\widehat{\beta_0} = Ortanca(Y) - \widehat{\beta_1} Ortanca(X) \qquad \mathbf{(1.3)}$$

şeklinde elde edilir (Kıroğlu, 2001). Ayrıca kesim parametresinin hesaplanılması konusunda çeşitli çalışmalar mevcuttur (Wilcox, 2013; Granato, 2006; Erilli ve Alakus, 2014; 2016).

Uygulama alanının kısıtlı olması nedeniyle parametrik olmayan regresyon analizi literatürde çok fazla ilgi görmemektedir. Yine de parametrik olmayan yöntemler ile ilgili çeşitli çalışmalar yapılmış ve yöntemler geliştirilmiştir. Theil (1950), basit bir doğrusal modelde, eğim parametresinin bir tahmini olarak ikili eğimlerin medyanını önermiştir. Daha sonra Sen (1968) ‘de bu tahmini genişletmiştir. Theil-Sen tahmini (TSE) % 29.3 ‘lük bir yüksek bozulma noktası ile sağlam, sınırlı bir etki fonksiyonuna ve yüksek asimptotik etkinliğe sahip olduğu belirlenmiştir. Bu nedenle, diğer eğim tahminlerine iyi bir rakiptir (örneğin, en küçük kareler tahminine) (Sen, 1968; Dietz, 1989; Wilcox, 1998).

TSE, sağlam bir tahmindir. Alternatif eğim tahminlerine nazaran hata kareler ortalama hesabı kolaydır (Dietz, 1989). $x_i, i = 1,2, \dots, n$ bilinen özdeş olmayan sabitler için, Sen (1968) ve Wilcox (1998) tarafından en küçük kareler tahminine göre TSE'nin asimptotik göreli etkinliği araştırılmış ve özellikle en iyi ya da asimptotik olarak en iyi tasarımlar için şu sonuçlar bulunmuştur;

1. Dağılım fonkisyonu F normalse, asimptotik göreli etkinlik $= 3/\pi \approx$ % 95.5 ‘dir.
2. Dağılım fonksiyonu F lojistik veya çift üstel dağılımdan geldiğinde, asimptotik göreli etkinlik >1'dir.
3. Ağır kuyruklu dağılımlar (örneğin, Cauchy dağılımı) için, asimptotik göreli etkinlik sonsuz sayıda büyük olabilir.
4. Her bir sürekli F için , asimptotik göreli etkinlik ≥ 0.864'dür (Wang ve Yu, 2005; Wang, 2005).

Sen (1968) ‘de $x_i, i = 1,2, \dots, n$ bilinen özdeş olmayan sabitler için, TSE tahmini yansız ve asimptotik olarak normaldir (Wang, 2005). Bütün bu önemli özelliklerinden dolayı TSE, parametrik olmayan ve sağlam istatistik üzerine yazılmış birçok değerli ders kitabında yer almıştır. Örneğin, Sprent (1993); Hollander ve Wolfe (1973, 1999) ve Rousseeuw ve Leroy (2005) bunlardan bazılarıdır. Yine bu yöntemin önemli uygulama alanları vardır. Örneğin astronomi Akritas vd (1995), Jones (1997), Mount ve Netanyahu (2001) ve sansürlü veriler ve uzaktan algılamada (Fernandes ve Leblanc, 2005) gibi. Fernandes ve Leblanc (2005), Theil-Sen regresyonu ile ölçme hataları altında regresyon tahminini incelemişlerdir. Ayrıca, Sen (1968) mutlak sürekli hata dağılımı ve bir özdeş olmayan eş değişken için tahminin yansızlık ve asimptotik normalliğini araştırmıştır. Genelleştirilmiş L-istatistiği olarak bakıldığında, TSE'nin asimptotikliği Serfling (1984)'den elde edilebilir. Lavagnini vd. (2011) Theil-Sen regresyon yardımıyla ters regresyon tahminini incelemişlerdir. Wang (2005), eş değişkenin rastgele olduğu durumlarda, TSE'nin asimptotik davranışlarını incelemiştir. Topal vd. (2004), Theil regresyon ve EKK yöntemi ile elde edilen sonuçları karşılaştırarak, Theil regresyonun uygulamadaki üstünlüklerini araştırmışlardır. Peng vd. (2008) ise hata dağılımının keyfi olduğu ve Sen (1968) tarafından elde edilen asimptotik normalliğin özel bir durumu takip ettiği durumlarda TSE'nin tutarlılık ve asimptotik dağılımını elde etmişlerdir. Ayrıca, hata dağılımının bir noktada süreksiz olduğu durumlarda TSE'nin oldukça etkin olduğunu göstermişlerdir. Erilli ve Alakuş (2014), eşit bağımsız değerlere sahip verilerin parametrik olmayan hesaplamaları için (Theil, Mood-Brown ve Hodges-Lehmann gibi) düzeltilmiş ortalama kullanımını önermişlerdir. Peng vd. (2008), keyfi hata dağılımlı doğrusal regresyon modellerinde Theil-Sen tahmininin asimptotik dağılımını ve güçlü tutarlılığını araştırmışlardır. Ayrıca hata dağılımı süreksiz ve sürekli olduğunda, TSE'nin asimptotik dağılımı normal olsa da olmasa da Theil-Sen tahmininin süper-etkin olduğunu göstermişlerlerdir. Lavagnini vd. (2011) 'de analitik problemlerin çözümü için doğrusal regresyon parametrelerine ilişkin parametrik olmayan Theil-Sen regresyon tekniğinin ve Lancaster-Quada (LQ) istatistiğinin birleştirilmesiyle yeni bir yaklaşım sunmuşlardır. Lia vd. (2009), Mahalanobis derinliğinin, çoklu doğrusal bir modeldeki parametrelerinin tamamen afin eşdeğer Theil-Sen tahmincileri oluşturmak için kullanılmasını önermişlerdir. Önerilen bu yapı, derinlik tabanlı Theil-Sen tahmin edicilerinin özel bir durumu olarak sunulmuştur. Önerilen tahmin edicilerin EKK tahmin edicilerine göre etkinliğinin, bağımsız ve aynı dağılımlı normal hatalar altındaki parametrelerin sayısı ile arttığını sergilemişler ve ayrıca tahminlerin olası en büyük kırılma noktası ile sağlam olduğuna ve sınırlı etkilere sahip olduğuna dikkat çekmişlerdir. Kısaca uygun koşullar altında hem deterministik hem de rastgele değişkenler için tahmin edicilerin tutarlı ve asimptotik normal olduğunu göstermişlerdir.

Her ne kadar TSE tahmininin birçok iyi özellikleri açık bir şekilde geometrik olarak yorumlanmış ve birçok istatistikçi tarafından bu özelliklerin genişletilebilmesi için çokça çaba harcanmış olsa da, (bakınız örneğin, (Oja ve Niinimaa, 1984; Zhou ve Serfling, 2006), TSE tahmini sadece basit bir lineer model için formüle edilmiş olduğundan az gelişmiş ve çok az değerlendirilmiştir. TSE'nin çoklu doğrusal bir modele genişletilmesi geometrik olarak uygun ve kullanışlı iken, teknik açıdan zorlayıcı özelliklerin olması nedeni ile genellenme ve incelenmesini geciktiren bir durum olmuştur. Dang vd. (2008), basit doğrusal bir modeldeki parametrenin Theil-Sen tahminini çok değişkenli lineer modele genişletmede çok değişkenli medyan kullanımının birkaç farklı tekniğini önermişlerdir. Çok değişkenli medyan (bazı yazarlar tarafından da kullanıldığı gibi çok boyutlu medyan) tek değişkenli medyanın genelleştirilmesidir ve literatürde yer alan köklü bir kavramdır (Bakınız örneğin, Small, 1990). Dang vd. (2008) çalışması derinlik fonksiyonları yardımı ile tanımlanan herhangi bir çok değişkenli medyan için de geçerlidir. Özellikle, derinlik fonksiyonuna dayalı çok değişkenli medyan, derinlik fonksiyonunun bir arttırıcısıdır. Matematikte derinlik fonksiyonu teorisi nispeten çok yenidir ve hala geliştirilme aşamasındadır. Bir bakıma doğrusal sıralamaya benzerlik göstermektedir. İstatistiksel derinlik fonksiyonları çok boyutlu verilerin merkezden dışa doğru sıralanmasını sağlamaktadır. Tukey (1975) ilk olarak yarı uzay derinliğini ortaya koymuştur. Oja (1983) Oja derinliğini tanımlamıştır. Liu (1990) simplicial derinliği önermiştir. Zou ve Serfling (2000) projeksiyon derinliğini değerlendirmişlerdir. Diğer kavramlar ise şunlardır; Zonoid derinliği (Koshevoy ve Mosler, 1997), genelleştirilmiş Tukey derinliği (Zhang, 2002) ve spatial derinlik (Chaudhuri, 1992). Bu çeşitli derinlikler arasından, hesaplama kolaylığı ve matematiksel çözülebilirliğinden ötürü özellikle spatial derinlik en kullanışlı ve tercih edilenidir (Dang vd., 2008). Böylece daha çok spatial derinlik tabanlı Theil-Sen tahminine bir diğer ifadeyle ÇTSE'ye odaklanılmıştır. Dang vd. (2008) bunun nispeten yüksek bozulma noktası ile sağlam olduğunu ve sınırlı bir etki fonksiyonuna sahip olduğunu göstermişlerdir. Normal koşullar altında güçlü tutarlılığı olduğunu, süreksiz hata dağılımı için süper etkili ve asimptotik normallikte olduğu ispat edilmiştir. Daha geniş bir örneklem büyüklüğü için, alt kitlesinin olasılıksal örneklemesinin kullanılması önerilmektedir (Wang vd., 2009).

Sağlam tahminler sayesinde, Theil tahmin edicisinin arkasındaki fikrin genişletilmesi için çaba sarfedilmektedir. Oja ve Niinimaa (1984), bütün n gözlemlerinin m öğelerinin her bir alt kümesi için, m veri noktalarından geçen hiperdüzleme (aşırı düzlem) tekabül eden parametre vektörünü tanımlamışlardır. Bunların m seçimlerinin sayısı n'dir. n kadar m seçimi vardır ki bunları sözde gözlemler olarak adlandırmışlardır. Konum parametresinin sağlam tahmini sözde gözlemlerin çok değişkenli medyanıdır. Shen (2009)'da Theil-Sen regresyon yardımıyla Oja medyanına bağlı asimptotik çoklu doğrusal regresyon tahminini incelemiş ve küçük örnek hacmindeki afin değişmezliğe ve dayanıklılığına ilaveten, mutlak sapma tahmininden asimptotik olarak daha etkili olduğu sonucuna varmıştır. Bu çalışmalar ışığında, Theil-Sen tahmin edicisinin çoklu doğrusal bir modele genişletilmesi konusu üzerinde düşünülmeye başlanmıştır ve Zhou ve Serfling (2006) tarafından bu çalışmaların bir alt yapısı oluşturulmuştur. Daha sonra TSE'nin çoklu doğrusal modele genelleştirilmesi (ÇTSE) için kapsamlı bir çalışma yapılmış ve mevcut sonuçlar iyileştirilmiştir (Dang vd, 2008). Buradan hareketle tez kapsamında; sonuçları iyileştilen ÇTSE'nin, çakı yöntemiyle etkileşmesi sonucunda mevcut sonuçlar daha da iyileştirilip, Çakı-Theil (ÇÇTSE) isminde yeni bir tahmin yöntemi önerilmiştir.

Çakı-Theil tahmini, çoklu doğrusal regresyon analizinde kullanılan Theil-Sen yöntemi ve çakı yöntemlerinin birleşmesinden meydana gelen regesyon katsayılarının tahmininde kullanılması önerilen yeni bir yöntemdir. Bu tezin temel amacı, çoklu doğrusal regresyon analizinde Çakı-Theil tahminine ait model parametrelerinin çok değişkenli bir medyana dayalı olarak elde edilmesi ve buradan hareketle parametrelerin çakı tahminlerini incelemektir. Literatürde çoklu doğrusal regresyon analizinde Theil-Sen tahmin yönteminin kullanılması matematiksel olarak oldukça zordur. Dolayısıyla bilgisayar programlarının etkin şekilde kullanılması gerekmektedir. Bu sebeple tez çalışması kapsamında önerilen tahmin yönteminin rakiplerine göre üstünlüğünün araştırılması için yapılan simülasyon ve orijinal veri uygulamaları R paket programında hazırlanan kodlar yardımıyla incelenecektir. Sonuçta, önerilen Çakı-Theil yönteminin asimptotik özellikleri araştırılacak, ilgilenilen diğer tahmin yöntemlerine göre avantajları belirlenmiş ve literatüre yeni bir yöntem olarak sunulacaktır.

Analizler sonucunda çoklu doğrusal regresyon analizinde model parametrelerinin tahmininde önerilen Çakı-Theil yönteminin kullanılmasının doğruluğu ve asimptotik özellikleri, orijinal veri kümeleri ve simülasyon çalışmaları yardımıyla değerlendirilmiştir. Çalışma sonucunda önerilen Çakı-Theil tahmin yönteminin daha doğru sonuçlar verdiği durumlar incelenmiştir. Önerilen yöntemin güvenilirliği, sağlamlık, etkinlik, süper etkinlik ve çoklu doğrusal regresyon modelindeki hata terimine ait farklı dağılımlardan örneklemler üretilip, elde edilen sonuçlara göre değerlendirilmiştir. Bu sonuçlar ışığında gelecek bilimsel çalışmalar için çoklu doğrusal regresyon analizi kapsamında spatil medyana bağlı olarak önerilen Çakı-Theil tahmin yöntemi elde edilen çeşitli koşullar altında daha etkin sonuçlar verdiği belirlenmiştir ve böylece bu yaklaşımın ilgili çeşitli alanlarda da etkin şekilde kullanılması beklenmektedir.

Literatürde çoklu doğrusal regresyon analizinde Theil-Sen yönteminin kullanılması matematiksel olarak oldukça zordur. Bu zorluğundan dolayı bu konuyla ilgili çok az çalışma vardır. Yine de çoklu doğrusal regresyon analizinde çok değişkenli medyan yaklaşımı kullanılarak model parametrelerinin tahmini yeni bir problem değildir; ancak çoklu doğrusal regresyon analizinde model parametrelerinin önerilen Çakı-Theil tahmin yöntemiyle elde edilmesi literatür de hiç çalışılmamıştır. Ayrıca tahmin yaparken ele alınan veri yapısı olumlu ya da olumsuz anlamda genel eğilimin dışında bulunabilir. Bu durum aykırı değer sorununu ortaya çıkarır. Önerilen Çakı-Theil tahmin yöntemi daha dar güven aralığı ve aşırı uç değerlerin etkisini daha da azaltan bir tekniktir. Bunun yanında önerilen Çakı-Theil tahmin yönteminin, regresyon modelindeki hata terimine ait dağılımın ağır kuyruklu dağılımlardan geldiğinde belirgin bir üstünlüğü vardır. Yani önerilen Çakı-Theil yöntemi ile parametre tahmini çoklu doğrusal regresyon analizinde henüz çalışılmamış bir problem olduğu için bu tez kapsamında incelenerek yeni bir tahmin yöntemi olarak literatüre sunulmuştur. Önerilen bu yaklaşım disiplinler arası çeşitli araştırmaların modellemesi aşamasında kullanılabilecektir.

2. REGRESYON PARAMETRELERİ İÇİN KLASİK TAHMİN METODU

Regresyon çözümlemesi, bağımsız değişken(ler) ile bağımlı değişken arasındaki ilişkiyi matematiksel modellerle açıklayarak regresyon denklem(ler) bulmak olarak ifade edilebilir. Bulunan regresyon denklemleri değişik amaçlar için kullanılabilir. Bu amaçlar içerisinde en önemli olanı tahmin yapmaktır. Doğrusal regresyon çözümlemesi genel olarak iki ana başlıkta incelenir. Bir bağımlı ve bir bağımısz değişkenin olduğu doğrusal regresyon çözümlemesine basit doğrusal regresyon, bir bağımlı ve birden fazla bağımsız değişkenin olduğu doğrusal regresyon çözümlemesine ise çoklu doğrusal regresyon çözümlemesi denir. Tez kapsamında çoklu doğrusal regresyona bağlı olarak hibrit bir yaklaşım önerileceği için bu bölümde çoklu doğrusal regresyon konusu incelenecektir. Yine de basit doğrusal regresyon analizinden biraz bahsedecek olursak, örneğin elimizde Y bağımlı ve X bağımsız değişken arasındaki nedensellik ilişkisinin $Y = \beta_0 + \beta_1 X + \varepsilon$ ile temsil edildiği düşünülsün. Bu ifadeye göre herhangibir x değerine karşılık gelen y gözlem değeri $\beta_0 + \beta_1 x$ ile bir ε miktarının toplamından ibarettir. Bu sebeple her bir y gözlem değeri regresyon doğrusu üzerinde yer almaz. Bu model değişkenler arasında olabileceğine inanılan bir ilişkiyi ifade etmektedir. Regresyon analizine başlarken bu modelin incelenen veriye uygun bir model olduğu varsayımı yapılmaktadır. Ancak bu varsayımın geçerliliği regresyon analizinin ileriki aşamalarında mutlak suretle kontrol edilmelidir. Bu modeldeki β_0 ve β_1 model parametreleridir. Bir modelin doğrusal veya doğrusal olmayan şeklindeki ifadesi parametrelere göre yapılır. Yani parametrelerin doğrusal olup olmaması durumuna göre model doğrusaldır veya değildir denilmektedir. Örneğin $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + C$ modeli parametrelerine göre doğrusal bir modeldir. X değişkenine göre doğrusal olmayan bir modeldir. Regresyon analizinde zaman zaman modelin doğrusal olmadığı belirtilmediği sürece doğrusal ifadesi kullanılmamaktadır. Yani doğrusal modelin sadece model olarak ifade edildiği unutulmamalıdır. Doğrusal modeller her regresyon parametresinin (β) üssünün 1 olduğu, regresyon katsayılarının modellerde üs olarak yer almadığı ve parametrelerin diğer parametrelerle bölünüp çarpılmadığı modellerdir (Alpar, 2013). İki değişkenden birindeki bir birim artışa karşılık, diğerinde sabit bir değişiklik meydana geliyorsa bu değişkenler arasında doğrusal ilişki vardır. Bu ilişki bir doğru denklemi ile ya da doğrusal regresyon grafiği ile gösterilebilir. Değişkenlerden birindeki bir birim artışa karşılık, diğerinde sabit olamayan bir değişiklik meydana geliyorsa, bu değişkenler arasında doğrusal olmayan bir ilişki vardır. Bu ilişki uygun bir eğri denklemiyle veya eğrisel regresyon grafiği ile gösterilebilir. İki değişken arasındaki doğrusal ve doğrusal olmayan ilişkilerden bazıları,

Doğrusal Denklem: $Y = \beta_0 + \beta_1 X + \varepsilon_i$

Kuadratik Denklem: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon_i$

Kübik Denklem: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon_i$

Bileşik Denklem: $Y = \beta_0 {\beta_1}^X \rightarrow logY = log\beta_0 + Xlog\beta_1 + \varepsilon_i$

Power Denklem: $Y = \beta_0 X^{\beta_1} \rightarrow logY = log\beta_0 + \beta_1 logX + \varepsilon_i$

Ters Denklem: $Y = \frac{1}{\beta_0 + \beta_1 X} \rightarrow \frac{1}{Y} = \beta_0 + \beta_1 X + \varepsilon_i$

Üstel Denklem: $Y = \beta_0 \exp(\beta_1 X) + \varepsilon_i$

Growth Denklem: $Y = \exp(\beta_0 + \beta_1 X) + \varepsilon_i$

S Denklem: $Y = \exp(\beta_0 + \frac{\beta_1}{X}) + \varepsilon_i$

Logaritmik Denklem: $Y = \beta_0 + \beta_1 logX + \varepsilon_i$

şeklindedir.

Yine bu modellerde β_0, β_1 ve ε bilinmeyenlerdir. Buradaki ε terimi her bir y bağımlı değişken değeri için değişmektedir. Ancak β_0 ve β_1 sabit kalır. Bu bilinmeyenleri gözlemlerin tamamını bilmeksizin ortaya çıkartmak mümkün olmasa da örneklem verileri kullanılarak $\hat{\beta}_0$ ve $\hat{\beta}_1$ şeklinde tahmin edilebilir. O halde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ile veriler herhangi bir x bağımsız değişken değeri için y'nin tahmini değerini ifade eder. Daha doğru şekilde ifade edecek olursak, $\hat{y}_i$ verilen x_i değerine karşılık gelen y bağımlı değişkeninin alabileceği ortalama değeri tahmin eder.

2.1. Basit Doğrusal Regresyon Modeli

Tek değişkenli regresyon analizi bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi inceleyen analiz tekniğidir. Bu analizle bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişkiyi temsil eden bir doğru denklemi formüle edilmektedir (Bowerman vd., 2013). Korelasyon analizinde olduğu gibi, regresyon analizinde üzerinde durulan ilişki, değişkenler arasındaki doğrusal ilişkidir. Bu doğrunun hesaplanması ise EKK yöntemi yardımıyla yapılmaktadır. EKK yönteminin, en çok olabilirlik yöntemine göre kullanılma sebebi sezgisel olarak anlaşılabilmesi ve en yüksek olabilirlik yöntemine göre matematiğinin daha basit olmasıdır. Yine de doğrusal regresyon çerçevesinde iki yöntemde genellikle benzer sonuçlar verir (Gujarati ve Porter, 2012).

Regresyon analizi sonuçlarının yorumlanmasında ciddi hatalar yapılmaktadır. En yaygın hata, regresyon analizi sonuçlarının yorumlanmasında, X bağımsız değişkeninin y bağımlı değişkenine sebep olduğu şeklindeki yorumdur. Bağımsız değişkenlerin bağımlı değişkendeki değişimi açıklıyor olması nedenselliği gerekli kılmaz. Başka bir ifade ile, bağımlı ve bağımsız değişkenler arasında (pozitif ve negatif) bir ilişkinin olması her zaman bağımsız değişken(lerin) bağımlı değişkenin sebebi olduğu sonucunu doğurmayacaktır. İki değişken arasında bir ilişkinin olabilmesi için nedensellik şart değildir. İlişki, iki değişkenin üçüncü bir değişkenle olan ilişkilerinden kaynaklanıyor olabileceği gibi, tamamen tesadüfi olarak da ortaya çıkmış olabilir. Nedensellik ile ilişkiselliğin aynı şeyler olmadığı unutulmamalıdır. Regresyon analizi değişkenler arasındaki ilişkinin yapısı ve derecesi ile ilgilenmektedir. Basit doğrusal regresyon modeli,

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \textbf{(2.1)}$$

şeklinde ifade edilir. Burada,

Y: Bağımlı değişken olup, belli bir hataya sahip olduğu varsayılır.

X: Bağımsız değişken olup, hatasız ölçüldüğü varsayılır.

β_0: Sabit olup $X = 0$ olduğunda Y'nin aldığı değerdir. Yani doğrunun y-eksenini kestiği noktadır.

β_1: Doğrunun eğimi veya regresyon katsayısı olup, X'in kendi birimi cinsinden 1 birim değişmesine karşılık Y'de kendi birimi cinsinden meydana gelecek ortalama değişme miktarını ifade eder.

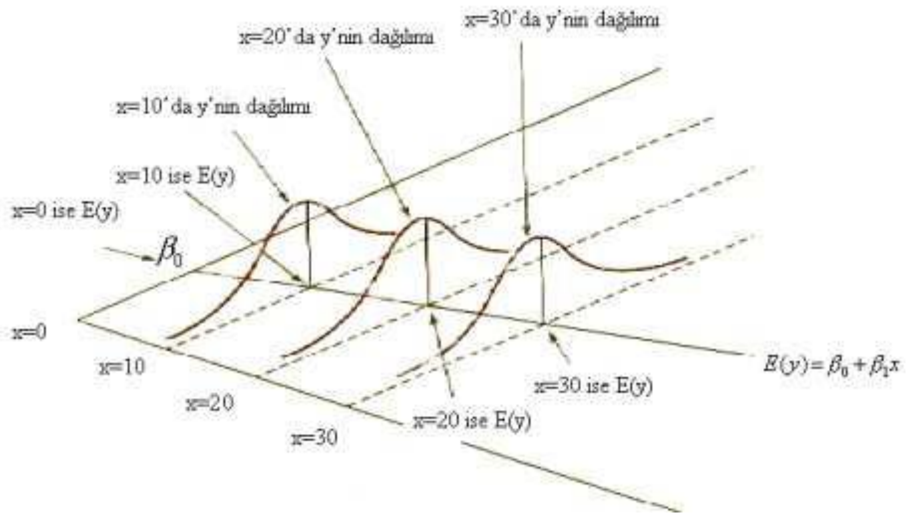
ε: Tesadüfi hata terimi olup, ortalaması sıfır varyansı bilinmeyen σ^2'ye sahip olan normal dağılış gösterdiği varsayılır. Bunun anlamı bir hata değerinin diğer bir hata değerinden etkilenmediğidir. Bu varsayım parametre tahminleri için değil, tahmin edilen regresyon katsayılarının önem kontrolleri için gereklidir.

Burada regresyon katsayıları olan β_0 ve β_1 değerleri tüm kitle verileri kullanılarak hesaplanan teorik değerlerdir. Ancak yine de dikkate alınmayan bağımsız değişkenler olabileceğinden, verilerin tesadüfi değişimlerini gösteren hata değeri ε modele eklenmiştir. Burada dikkat edilmelidir ki, Y bağımlı değişkeni bir tesadüfi değişkenidir. X bağımsız değişkeni ise bir tesadüfi değişken değildir. Deney tekrar edildiğinde X bağımsız değişkenine ait değerler sabit tutulabilir. Bağımsız değişkenlere ait gözlemlerin araştırmacı tarafından kontrol edilebildiği veya ihmal edilebilir en küçük bir hata ile ölçüldüğü varsayılır. Dolayısıyla X bağımsız değişkeninin her bir değeri için Y'nin bir olasılık dağılımı mevcuttur. Bu dağılımın ortalaması ve varyansı;

$$E(Y) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$$

$$Var(Y) = Var(\beta_0 + \beta_1 X_i + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2 \tag{2.2}$$

şeklindedir. O halde Y'nin varyansı X'e bağlı olmamasına karşılık ortalaması X'in bir doğrusal fonksiyonudur. Hata terimi ε'nun birbirinden bağımsız olduğu varsayımı yapıldığı için bağımlı değişken Y içinde bu varsayım geçerli olacaktır. Bu demektir ki Y'ye ait gözlemler arasında ilişki yoktur.



Şekil 2.1. Farklı X değerleri için Y ' nin dağılımı

Şekil 2.1'de Y dağılımlarının şekli her X noktasında aynıdır. Burada Y'nin ortalaması X'e bağlıdır, ancak varyans X değerinden etkilenmez. Yani hata değişkeninin dağılımı şekilde görüldüğü gibi farklı X değeri için aynıdır.

Basit doğrusal regresyon modelinin varsayımlarını aşağıdaki şekilde ifade edilir,

1. Bağımsız değişkenin değerleri sabit kabul edilir. Bağımlı değişkenin değerleri ise tesadüfidir. Fakat korelasyon analizinde her iki değişken de tesadüfidir.

2. Değişkenler hatasız ölçülmüştür.

3. Her bir x_i değeri için;

a. y_i değerleri birbirinden bağımsızdır,

b. y_i gözlemlerinin tüm dağılımları normaldir,

c. y_i gözlemlerinin tüm dağılımları aynı varyansa sahiptir.

4. Bu alt küme değerlerinin (Y'lerin) varyansları eşittir.

5. Bağımlı değişken ve bağımsız değişkenler arasındaki ilişki doğrusaldır.

6. Gözlem sayısı bağımsız değişken sayısından çok olmalıdır.

Şimdi de hata terimi (ε_i)'ye ait varsayımlar ise;

1. Tesadüfi hataların beklenen değeri sıfırdır. E(ε)=0

2. ε' ların olasılık dağılımının varyansı sabittir. $Var(\varepsilon_i) \ = \sigma^2$

3. Hata değerleri birbirinden bağımsızdır.

4. Tesadüfi hataların dağılımı normaldir.

5. Hatalar ile bağımsız değişkenler birbirinden ilişkisizdir. $Cov\ (\varepsilon,\ X_i) = 0$

olarak ifade edilir (Gujarati ve Porter, 2012).

Basit doğrusal regresyon parametrelerinin EKK yöntemi ile tahmininde esas, serpme diyagramında görülen tüm noktalar için doğruya uzaklıklarının bulunması ve bunların toplamının en küçüklenmesidir. Basit doğrusal regresyon modeli, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1,2, \dots, n$ idi. Bu modelin tahmin denklemi,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \mathbf{(2.3)}$$

Burada, $\hat{Y}_i$, Y_i'nin tahmin edilmiş koşullu ortalama değeridir (Gujarati ve Porter, 2012). Y_i ile $\hat{Y}_i$ arasındaki fark olarak tanımlanan hata teriminin tahmini ise

$$e_i = y_i - \hat{y}_i \qquad \mathbf{(2.4)}$$

ile verilir. Burada, e_i artık olarak ifade edilir ve açıklanan değişkenin gözlemlenen ve tahmin edilen y_i değerleri arasındaki farktır. Tahmin edilen hata terimlerinin toplamı sıfırdır. Bu durumda hataların kareleri toplamı bulunarak yeni bir fonksiyon tanımlanır.

$$\sum {\varepsilon_i}^2 = \sum (Y_i - \beta_0 + \beta_1 X_i)^2 = \sum (Y_i - \beta_0 + \beta_1 X_i)^2 \to min \quad \mathbf{(2.5)}$$

toplamı en küçük olur. (x_1, y_1), (x_2, y_2),…,(x_n, y_n) noktaları arasından sonsuz sayıda doğru geçebilir. Her doğru için y_i değerleri ile x_i değerleri arasında değişik farklar çıkacaktır. Bunlar içerisinde herhangi bir doğru için, bu farkların kareler toplamı en küçük ise o doğru, dağılımı en iyi temsil eden doğrudur. Eşitlik (2.5)' deki fonksiyonu en küçük yapacak β_0 ve β_1 parametrelerinin en iyi tahmin edicilerini bulmak için, fonksiyonun β_0 ve β_1 'e göre birinci kısmi türevlerini sıfır yapan değerler elde etmek gerekir. Buna göre,

$$\frac{\partial}{\partial\beta_0}\sum(Y_i-\beta_0+\beta_1X_i)^2=0 \qquad (\mathbf{2.6})$$

$$\frac{\partial}{\partial\beta_1}\sum(Y_i-\beta_0+\beta_1X_i)^2=0 \qquad (\mathbf{2.7})$$

eşitlikleri yazılabilir. Bu eşitlikler birlikte çözülerek aranan tahmin ediciler,

$$\hat{\beta}_0=\bar{Y}-\hat{\beta}_1\bar{X} \qquad (\mathbf{2.8})$$

$$\hat{\beta}_1=\frac{\sum X_iY_i-\frac{\sum X_i\sum Y_i}{n}}{\sum X_i^2-\frac{(\sum X_i)^2}{n}} \qquad (\mathbf{2.9})$$

olarak elde edilir. Bu yolla elde edilen tahmin ediciler, EKK tahmin ilkesinden türedikleri için EKK tahmin edicileri adını alırlar.

EKK tahmin edicileri, gözlem değerleri ölçüm hatalarından bağımsız olduğunda doğrusal regresyon çözümlemesindeki eğim parametresi için en iyi doğrusal ve sapmasız tahmini vermektedir (Srivastava ve Shalabh, 1997).

2.2. Çoklu Doğrusal Regresyon Modeli

Bir bağımlı değişken ve birden fazla bağımsız değişkenin yer aldığı regresyon modellerinin analizine çoklu regresyon analizi denir. Çoklu regresyon analizinde bağımlı değişken ve bağımsız değişkenler arasındaki ilişki matematiksel modelle açıklanır ve bağımsız değişkenler yardımı ile bağımlı değişkenin tahmini yapılır. Çoklu regresyon analizinde bağımsız değişkenler ile bağımlı değişkendeki değişim açıklanmaya çalışılır. Bağımsız değişkenleri seçerken kendi aralarında yüksek korelasyona sahip olan değişkenlerden sadece birisinin seçilmesi gerekir. Bunun içinde analize başlamadan bütün değişkenlerin korelasyon matrisine bakıp, aralarında yüksek korelasyon olanlardan sadece biri seçilmelidir. Çoklu regresyonda bazen hangi bağımsız değişkenin daha önemli olduğunu ve bağımlı değişkeni daha çok etkilediğini bilmek gerekir. Bunun için önce korelasyonlara bakılır. Yüksek korelasyon, daha güçlü doğrusal ilişki gösterir. Başka bir deyişle, bağımsız değişkenin bağımlı değişkeni bir kez ikili değişken formülünde bir kez de çoklu doğrusal regresyon içinde nasıl etkilendiğine bakmaktır. Çoklu doğrusal regresyon modeli,

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon \qquad (\mathbf{2.10})$$

şeklinde ifade edilir (James vd, 2013). Eşitlik (2.10), matris notasyonu ile

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (\mathbf{2.11})$$

eşitliğindeki gibi yazılır. Burada,

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}_{n*1}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}_{n*(p+1)}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{bmatrix}_{(p+1)*1}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}_{n*1} \qquad \textbf{(2.12)}$$

şeklindedir.

Çoklu doğrusal regresyon modelinin varsayımları şu şekilde ifade edilir,

1) Parametrelerde doğrusallık: Çoklu doğrusal regresyon modeli $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$ gibi doğrusal olmalıdır.

2) Tesadüfi örnekleme: Kitleden çekilen örneklemler rastgele olmalıdır.

3) Sıfır koşullu ortalama: Bağımsız değişkenlerle hatalar ilişkisizdir. Yani $E\left[\varepsilon/x_1, x_2, \dots, x_p\right] = 0$ 'dir. Bağımsız değişkenler ile hata terimi (ε) arasındaki ilişki EKK tahmin edicilerinin sapmalı olmasına yol açar. Bu varsayımın sağlanmamasının sebepleri, regresyonun fonksiyonel biçiminin yanlış seçilmesi, önemli bir bağımsız değişken model dışında bırakılmış olması ve değişkenlerin ölçümündeki hatalardan kaynaklı olması olabilir.

4) Çoklu bağlantı olmaması: Bağımsız değişkenler arasında tam ilişki (yüksek korelasyon, doğrusal ilişki) olursa çoklu bağlantı problemi oluşur. Regresyon modeline aynı bağımsız değişkenin değişik doğrusal olmayan fonksiyonları sokulabilir. Bu durum çoklu bağlantıya sebep olmaz.

5) Sabit varyans: Sabit varyans hata terimlerinin varyanslarının tüm gözlemler için aynı olmasıdır. Yani her bağımsız değişkenin değerlerine ait olan bağımlı değişken değerlerinin alt setlerinin varyansları birbirine eşittir. Şöyle ki; $Var(\varepsilon_i/x_i) = \sigma^2$ ifadesi sabit varyansı, $Var(\varepsilon_i/x_i) = \sigma_i^2$ ifadesi ise değişen varyansı gösterir. Değişen varyans hata terimleri varyansının x_i değerlerine göre değiştiğini gösterir. Değişen varyans sapmaya ve tutarsızlığa sebep olmaz. Ancak; regresyon katsayılarının varyansı sapmalı olur (Kalaycı, 2016).

Yukarıdaki beş varsayımın sağlanması durumunda EKK tahmin edicileri olan $\hat{\beta}_0$, $\hat{\beta}_1$,…,$\hat{\beta}_p$, kitle parametrelerinin doğrusal en iyi sapmasız tahmin edicileridir. Bu ifade literatürde Gauss-Markov teoremi olarak bilinir (Terzi, 2015).

6) Tahmin hataları rastgele olup ortalaması sıfır $(E(\varepsilon) = 0)$ ve varyansı

$Var(\varepsilon) = \sigma^2 I$ olan normal dağılıma uyar. Hipotez testlerinin yapabilmek ve güven aralıklarını oluşturabilmek için hataların normal dağılıma sahip olduğu varsayılır. Ayrıca hatalar birbirinden bağımsızdır (otokorelasyon yoktur), yani birbirleriyle ilişki göstermezler $\left(Cov(\varepsilon_i, \varepsilon_j) = 0\right)$

Çoklu doğrusal regresyon modelinde parametrelerin tahminini bulmak için EKK metodu kullanılır. Amaç fonksiyonu olan artık kareler toplamını en küçük yapan β vektörü bulunur.

$S = \sum \boldsymbol{\varepsilon_i^2} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = [\varepsilon_1\ \varepsilon_2 \dots \varepsilon_n] \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$ ifadesini minimum yapacak $\boldsymbol{\beta}$ vektörü,

$\hat{\beta} = (X'X)^{-1}X'Y$ şeklinde elde edilir. Burada,

$$X'X = \begin{bmatrix} n & \sum x_{1i} & \dots & \sum x_{pi} \\ \sum x_{1i} & \sum x_{1i}^2 & \dots & \sum x_{1i}x_{pi} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \dots & \sum x_{2i}x_{pi} \\ \dots & \dots & \dots & \dots \\ \sum x_{pi} & \sum x_{1i}x_{pi} & \dots & \sum x_{pi}^2 \end{bmatrix}, \qquad X'Y = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \dots \\ \sum x_{pi}\, y_i \end{bmatrix}$$ şeklinde

gösterilir.

2.3. Regresyon Analizinde Aykırı Gözlemlerin İncelenmesi

Regresyon analizindeki model bozukluklarını araştırmak ve aykırı değerleri belirlemek için en basit ve kullanışlı yöntem artıkların incelenmesidir. Çünkü artıklar, hataların gerçekleşen ya da gözlenen değerleri olarak düşünülebilir. Bu sebeple, hatalar için söz konusu olan varsayımlara ilişkin bozulmalar, artıklar aracılığıyla araştırılabilir. Bu araştırmada farklı artık türlerinden yararlanılır. Bunlar; standartlaştırılmamış (ham) artıklar, standartlaştırılmış artıklar, student türü artıklar vb. değişik artık türlerine örnek oluştururlar. Verideki gözlem sayısı artıkça artıklarla yapılacak incelemelerin güvenilirliği de artacaktır (Alpar, 2013).

Aykırı değerler regesyon analizi uygulamalarında sıklıkla karşılaşılan bir durum olan gözlem değerlerinden herhangi birinin veya birkaçının ölçme, kopyalama ya da aktarma hataları sebebiyle hatalı bir gözlem olarak alınmasından da kaynaklanabilir. Sonuç olarak regresyon aykırı değerleri, verilerin çoğunluğu tarafından oluşturulan doğrusal yapıya uyum sağlamayan gözlemlerdir. Bu tür gözlemler, hem bağımlı değişken ve hem de bağımsız değişken(ler)in değerleri veya her ikisi ile birlikte de olabilir (Rousseeuw ve Aelst, 1999).

2.3.1. Standartlaştırılmamış (ham) artıklar

i. gözlem için ham artık,

$$e_i = y_i - \hat{y}_i \quad (i = 1, 2, \dots, n) \qquad \mathbf{(2.13)}$$

şeklinde tanımlanır. Eşitlik (2.13)' de gösterilen artıklar adından da anlaşılacağı gibi ölçeklenmemiş ya da ham artıklar olarak ifade edilir. Standartlaştırılmamış artıkların toplamı sıfır (dolayısıyla ortalaması da sıfır) olmakla birlikte, varyansı örneklemden örnekleme değişir. Varyansın değişkenlik göstermesi sorunu, artıkların standartlaştırılması yoluyla giderilmeye çalışılır.

2.3.2. Standartlaştırılmış artıklar

Eğer ε_i'ler 0 ortalama ve σ^2 varyansı ile normal dağılım gösteriyor ise ε_i/σ'nın standart normal dağılım göstermesi söz konusudur. Bu nedenle e_i artıkları $\hat{\sigma}$'ye bölünerek standartlaştırılmış artıklar (e_{is}) elde edilir ve

$$e_{is} = \frac{e_i}{\sqrt{AKO}} = \frac{e_i}{\hat{\sigma}} \quad (i = 1, 2, \dots, n) \qquad \mathbf{(2.14)}$$

şeklinde elde edilir. Burada $\hat{\sigma}$ modelin standart hata tahminidir. Gözlem sayısının fazla olması durumunda, e_{is}'lerin 0 ortalamalı ve 1 standart sapma ile yaklaşık standart normal dağıldıkları $\left(e_{is} \sim N(0,1)\right)$ ve artıkların birbiri ile ilişkili olmadıkları söylenir. e_{is} artıklarının % 95' inin çoğunlukla $[-2, +2]$ sınırları arasında değiştiği kabul edilir. Zaten biliniyor ki standart normal dağılımda gözlemlerin % 95'i -1.96 ile $+1.96$ arasında değişmektedir. Buradaki aralık ise daha esnektir. İşte bu sınırların dışına çıkan gözlemlerin aykırı değer kuşkusu ile incelenmesi gerekir.

2.3.3. Student türü artıklar

e_{is} artıkları her zaman 0 ortalama ve 1 standart sapma ile normal dağılım göstermemektedir. Çünkü paydadaki $\hat{\sigma}$ istatistiği e_i değerlerinin standart sapması değildir. Bu sebeple e_i artıkları kendi standart sapması olan $\hat{\sigma}\sqrt{1 - p_{ii}}$ ile standartlaştırılır ve student türü artıklar

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}} \quad (i = 1, 2, \dots, n) \qquad \mathbf{(2.15)}$$

şeklinde gösterilir. Burada p_{ii} gözlem uzaklığı olarak ifade edilir. X matrisi, ilk sütunu 1'lerden oluşan, diğer sütunları bağımsız değişken(ler) değerlerinden oluşan veri matrisi olmak üzere $n \times n$ boyutlu P matrisi,

$$P = X(X'X)^{-1}X' \qquad (\mathbf{2.16})$$

şeklinde yazılırsa, P matrisinin esas köşegen elemanları p_{ii}, i. gözlem uzaklığını verir. p_{ii} değerleri hat değerleri olarakta bilinir.

r_i'ler 0 ortalama ve 1 standart sapma ile standart normal dağılıma yakın bir dağılım gösterirler. r_i'ler az da olsa tahmini $\hat{y}_i$ değerleri ile ilişkilidirler. r_i değerleri model doğru olduğunda ve gözlem sayısı fazla olduğunda (p_{ii} değerleri sıfıra yaklaşırken r_i'lerde e_{is} artıklarına yaklaşacağı için) genellikle $[-2, +2]$ sınırları arasında ve sıfır etrafında bir dağılım gösterirler. Dolayısıyla böyle durumlarda bu sınırların dışına çıkan gözlemlerin aykırı gözlem olabileceği düşünülerek bu gözlemler incelenmelidir.

2.3.4. Silinmiş artıklar

Aykırı değerleri araştırmanın bir diğer yolu, problemli olduğu düşünülen gözlemlerin veriden çıkartılması durumunda, ilgili gözlemin tahmin modelinde yapacağı değişikliklerin incelenmesidir. Gözlemi veriden çıkartarak elde edilen artıkları $\left(e_{(i)}\right)$ bulabilmek için i. gözlem veriden silinerek $n-1$ gözlem için regresyon denklemi bulunur. Daha sonra i. gözlemin bağımsız değişken değerleri bu denklemde yerine konarak $\hat{y}_{(i)}$ kestirim değeri elde edilir. $\hat{y}_{(i)}$ kestirim değerinin i. gözlemin y_i değerinden çıkartılması ile $e_{(i)}$ hesaplanır $\left(e_{(i)} = y_i - \hat{y}_{(i)}\right)$. Bu işlemler tüm gözlemler için tekrarlanarak n adet silinmiş artık $\left(e_{(1)}, e_{(2)}, \dots, e_{(n)}\right)$ elde edilir. Bu artıklara PRESS artıkları da denir. Uygulamada, özellikle gözlem sayısı fazla olduğunda $n-1$ gözlemli n tane regresyon denklemi elde etmek zor ve pratik olmadığından, silinmiş artık değerleri;

$$e_{(i)} = \frac{e_i}{1 - p_{ii}} \qquad (\mathbf{2.17})$$

biçiminde elde edilir (Alpar, 2013).

3. THEİL-SEN TAHMİNİ (TSE)

İkinci bölümde de bahsedildiği gibi, regresyon analizi aralarında nedensellik bulunan iki ya da daha fazla değişken arasındaki ilişkiyi inceler. Modeldeki bağımlı değişkenin bir bağımsız değişken tarafından açıklanması basit doğrusal regresyon modeli olarak tanımlanır. Basit doğrusal regresyon modelinde varsayımlar sağlandığında elde edilen tahmin değerleri, Gauss-Markov teoremine göre parametrelerinin doğrusal, yansız ve en küçük varyanslı tahmin edicileridir. Ancak, varsayımlar sağlanmadığında elde edilen tahminler sağlaması gereken özellikleri kaybederler. Bu durumda parametrik olmayan ve sağlam regresyon yöntemleri kullanılarak gerçek verilere uyumluluk araştırılır.

Theil (1950), tarafından, β_1 eğim katsayılarının nokta tahminini bulmayı amaçlayan ve kendi anıyla anılan bir tahmin yöntemni geliştirdi. Bir parametrik olmayan regresyon yöntemi olarak bilinen Theil yöntemi, yine eğim parametresini tahmin etmek için önerilen Mood-Brown yöntemine göre daha güvenilirdir (Theil, 1950).

Sen (1968), iki veya daha fazla eğim parametresinin birbirine eşit olduğunu iddia eden sıfır hipotezlerini test eden bir sıra puanı yöntemini incelemiştir. Kendall'ın Tau' sundan esinlenerek β_1' in basit ve sağlam tahmincileri üzerinde çalışmıştır.

Kendall' ın Tau test ölçütü hesabının güç ve etkinliğini incelemiştir. Nokta tahmin edicisini, $x_j \neq x_i$ ile noktaların $(y_j - y_i)/(x_j - x_i)$ eğim çiftleri kitlesinin medyanı olarak tarif etmiştir. Sen, ileri sürdüğü tahmin edicilerin çeşitli özelliklerini incelemiş ve kendi ismini verdiği yöntemi EKK ve diğer parametrik olmayan tahmin ediciler ile karşılaştırmasını yapmıştır.

3.1.Basit Doğrusal Regresyon Analizinde Theil Tahmini

Basit doğrusal regresyon modeli,

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \mathbf{(3.1)}$$

olarak tanımlansın. Burada Y bağımlı değişkeni; X bağımsız değişkeni gösterir. ε ise regresyon modelinin hata terimidir. β_0 ve β_1 nicelikleri ise sırasıyla modelin kesim ve eğim parametreleridir. Regresyon modelinde yer alan katsayıların tahmini için EKK yöntemine alternatif olarak bazı parametrik olmayan ya da sağlam yöntemler kullanılır. Bu yöntemler hata terimi normal dağılmayan ve/ veya aykırı değerler modeli etkilediği zamanlarda EKK yöntemine alternatif olarak kullanılmaktadır (Candan, 1995).

Theil-Sen yöntemi literatürde Theil-Kendall ya da Theil yöntemi olarak da ifade edilmektedir (Zaman ve Alakuş, 2015). 1950 yılında Theil tarafından ileri sürülen yöntem araştırmacıların en çok başvurdukları eğim bulma yöntemlerindendir. Bir doğrunun eğimi tahmininde kullanılan tüm (x_i, y_i) ve (x_j, y_j) gözlem ikililerinden hesaplanan eğim değerlerinin medyan hesabına dayandırılmaktadır (Hussain ve Sprent, 1983). Sahip olunan veri $(x_1, y_1), \dots, (x_n, y_n)$ şeklinde n tane gözlem ikililerinden oluşmaktadır. Burada x_i değerleri birbirinden farklı bilinen bağımsız değişken değerleri olup $x_1 < x_2 < \cdots < x_n$ şeklinde sıralanmaktadır (Yıldız ve Topal, 2001). Eşitlik (3.1)'de ε_i'lerin varyansı σ_ε^2 ve medyanı sıfır olan simetrik bir sürekli dağılışa sahip bağımsız ve aynı dağılımdan meydana gelen rastgele hatalardan oluşurlar (Rao ve Gore, 1982). Theil yönteminde β_0 ve β_1 öyle tahmin edilmelidir ki hata teriminin medyanı sıfır olmalıdır (Maritz, 1979). β_1' in tahmin edicisi $\hat{\beta}_1$, $i < j$ ve $(x_i \neq x_j)$ olmak üzere elde edilen $N = \binom{n}{2}$ tane $S_{ij} = \frac{y_j - y_i}{x_j - x_i}$ eğimlerinin tümünün bir medyanı olur (Daniel, 2000; Wang ve Yu, 2005). Şöyle ki,

$$\hat{\beta}_1 = ortanca\{S_{ij}\} \qquad \mathbf{(3.2)}$$

ve

$$\hat{\beta}_0 = ortanca(Y) - \hat{\beta}_1 ortanca(X) \qquad \mathbf{(3.3)}$$

olarak elde edilir (Hussain ve Sprent, 1983).

Eğim ve kesim parametrelerinin ortak tahmini için iki yaklaşım vardır. Bunlar;

3.1.1. İşaret testine dayalı en iyi tahmin yöntemi

$d_i = y_i - \hat{\beta}_1 x_i$ değerleri hesaplanır ve bu değerlerin medyanı β_0'ın tahmini, $\hat{\beta}_0$ olur. Bu tür düzenlemeler ya da yaklaşımlar, simetrik dağılımlı d_i'lerin varsayımını gerektirmez. Özellikle uç değerli veriler için daha uygundur.

3.1.2. Hodges-Lehmann yöntemi

$d_i = y_i - \hat{\beta}_1 x_i$ tesadüfi değişken değerleri hesaplanır. Bu yaklaşım d_i'lerin β_0 etrafında simetrik dağıldığı varsayımı gerektirir. Hodges-Lehmann yöntemine göre $\hat{\beta}_0$, d_i'lerin aritmetik ortalamasıdır. Bu tür düzenleme, uç değerli veriler için uygun olmayabilir (Alakuş, 2015, Lehmann, 2006).

$w_{ij1} = (x_j - x_i)$ ve $w_{ij2} = (x_j - x_i)^2$ iki farklı ağırlık olmak üzere Theil yöntemindeki ağırlıklı eğim parametresi tahmin edicileri eşitlik (3.4) ve (3.5) ile verilir. Görüldüğü gibi tahmin ediciler, S_{ij}'lerin ağırlıklı ortalama ve medyanlarıdır.

$$\hat{\beta}_{1wij(1,2)(ort)} = \frac{\sum_{i<j} w_{ij(1,2)} S_{ij}}{\sum_{i<j} w_{ij(1,2)}} \qquad (\mathbf{3.4})$$

ve

$$\hat{\beta}_{1wij(1,2)(med)} = \frac{med_{i<j}\left(w_{ij(1,2)} S_{ij}\right)}{med_{i<j}\left(w_{ij(1,2)}\right)} \qquad (\mathbf{3.5})$$

olup, kesim parametresinin tahmin edicisi ise

$$\hat{\beta}_0 = ortanca(Y) - \hat{\beta}_{1wij(ort,med)} ortanca(x_i) \qquad (\mathbf{3.6})$$

eşitliği ile verilir (Sievers, 1978 ve Scholz, 1978). Ayrıca Randles ve Wolfe (1979), S_{ij}'nin ağırlıklandırılmamış ortalamalarını $\hat{\beta}_1$' in tahmin edicisi olarak önermişlerdir. Bu durumda eğim parametresinin tahmin edicisi ise

$$\hat{\beta}_{1sij(ort)} = \frac{\sum_{i<j} S_{ij}}{N} \qquad (\mathbf{3.7})$$

eşitliği ile verilir (Toka vd., 2011)

3.2.Basit Doğrusal Regresyonda TSE'nin Asimptotik Özellikleri

Bu alt bölümde, Theil-Sen tahmin edicisinin (TSE) bazı asimptotik özelliklerinden bahsedilecektir. Öncelikle Sen (1968)' de, x_i, $i = 1,2,\ldots,n$ bilinen özdeş olmayan sabitler ve dağılım fonksiyonu $F(.)$ sürekli ise TSE tahmininin asimptotik olarak normal dağılımlı olduğunu göstermiştir. Ayrıca bu tahmin edicinin 0.293'lük bir kırılma noktasına sahip olduğu belirlenmiştir (Peng vd, 2008).

Wang (2005), X rasgele değişken varsayımı altında TSE tahmininin asimptotik özelliklerini incelemiştir. ε ile X tesadüfi değişkenlerinin bağımsız ve dağılım fonksiyonu $F(.)$ keyfi (sürekli ya da süreksiz) varsayımı altında, TSE tahmin edicisinin güçlü tutarlılığını ve asimptotik dağılımını araştırmıştır. Şimdi TSE tahmininin güçlü tutarlılığını ve asimptotik dağılımını kısaca inceleyelim.

3.2.1. Güçlü tutarlılık

$\Omega_0 = \{\omega: \hat{\beta}_n(\omega) = \beta\}$, ($\forall\, n$ için) bir olay olsun. $n > n_\omega$ olduğunda, $\hat{\beta}_n(\omega) = \beta$ olur.

Teorem 1: $X_1, \dots, X_n$ rastgele olmayan ortak değişkenlerinin

$\bar{a}_n = \binom{n}{2}^{-1} \sum_{i<j} 1[x_i \neq x_j]$ iken

$$\frac{n^{-1}\log n}{\bar{a}_n^2} = o(1) \qquad (3.8)$$

olsun.

(i) Eğer dağılım fonksiyonu $F(.)$ sürekli değil ise,

$$P\left(\omega: \hat{\beta}_n \neq \beta \text{ sadece sınırlı kez gerçekleşir}\right) = 1$$

Dolayısıyla TSE güçlü bir şekilde tutarlıdır. Yani, $\mathrm{P}(\lim_{n\to\infty} \hat{\beta}_n = \beta) = 1$ şeklinde ifade edilir.

(ii) Eğer dağılım fonksiyonu $F(.)$ sürekli ve

$$\lim_{n\to\infty} \inf \left\{|x_i - x_j|: x_i \neq x_j: i < j\right\} > 0 \text{ ise}$$

o zaman TSE güçlü tutarlıdır (Peng vd, 2008).

3.2.2. Asimptotik dağılım

Teorem 2. Öncelikle Teorem 1 'deki durum 1 dikkate alınsın.

Eğer dağılım fonksiyonu $F(.)$ sürekli değilse,

$$P\left(n^v\left(\hat{\beta}_n - \beta\right) \to 0\right) = 1, \qquad v \geq 0 \qquad (3.9)$$

olur. Dolayısıyla $\hat{\beta}_n$ güçlü etkindir.

Durum 2 için, dağılım fonksiyonu $F(.)$ sürekli olduğundan,

$$c_i = \sum 1[x_j > x_i] - 1[x_j < x_i] \text{ iken } C_n^2 = \sum c_i^2 \text{ olsun.}$$

Eğer $k \to \infty, \max_{i,j}|x_i - x_j|/k_n \to 0, \lim\ \inf({}_nC_n/n^{\frac{3}{2}}) > 0$ ve

$$C_n^{-1} \sum_{i<j} [1 - 2F_2(t|x_i - x_j|/k_n)] \to m(t), \qquad F_2(t) = \mathbb{E}F(\varepsilon + t) \text{ ise}$$

ϕ standart normal dağılım fonksiyonu iken

$$\lim_{n\to\infty} P\big(k_n(\hat{\beta}_n - \beta_0) \le t\big) = \phi\left(-\sqrt{3}m(t)\right), \qquad t \in R \text{ olur.}$$

Asimptotik dağılım $\overset{\leftrightarrow}{\underset{\leftrightarrow}{}} m(t)$ t'de doğrusalsa normal dağılımlıdır.

Sonuç: F, $B(F) = \int f^2(t)dt < \infty$ gibi f olasılık dağılımı ile kesinlikle sürekli olduğu varsayılsın. Eğer $\lim\ \inf(C_n/n^{3/2}) > 0$ ise $d_i = \sum_j |x_i - x_j|$ ile $D_n = \sum_{i=1}^{n} d_i$ iken $(D_n/C_n)(\hat{\beta}_n - \beta_0) \Rightarrow \mathcal{N}(0,1/3B^2(F))$ olur (URL-1).

Özetle, genel hata dağılımı F için (hataların dağılım fonksiyonu) TSE tahmin edicisinin asimptotik dağılımını ve güçlü tutarlılığı gösterilmiştir. Burada hataların dağılım fonksiyonu keyfi ve hem sürekli hem de kesikli olabilir (Peng vd, 2008).

Ayrıca TSE tahmin edicisine ait güçlü tutarlılık ve asimptotik dağılım ile ilgili daha detaylı bilgi için Peng vd (2008) ve Wang (2005) çalışmaları incelenebilir.

Bu bölümde Theil tahmin yöntemi için kısa bir örnek verilmiştir.

3.3. Basit Doğrusal Regresyonda Theil Tahmin Yöntemi İçin Uygulama

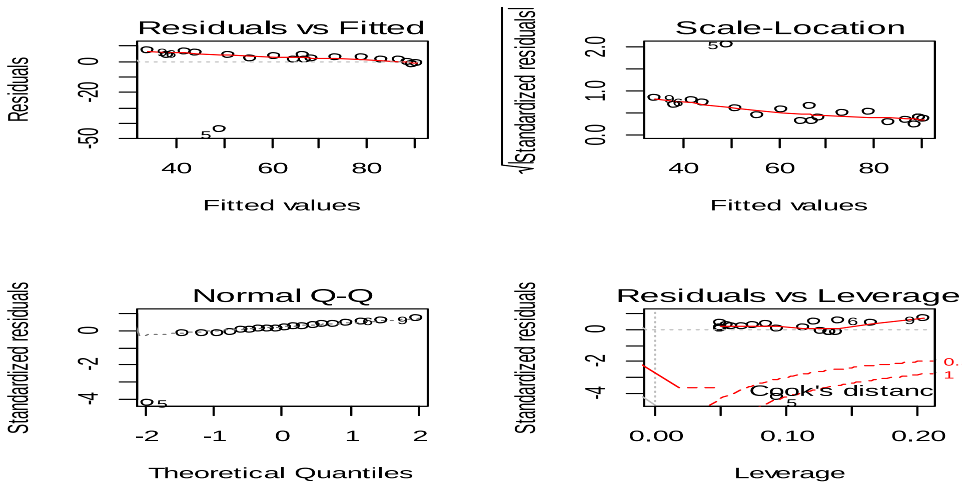
Çalışmada basit doğrusal regresyon modelinde parametrik olmayan analiz yöntemine olan ihtiyacı göstermek amacıyla Çizelge 2.1 de verilen Pilot-Plant (Daniel ve Wood, 1971) verisi çalışılmıştır. Burada bağımlı değişken titrasyonla belirlenen asit miktarı, bağımsız değişken ise çekme ve tartma ile belirlenen organik asit miktarıdır.

Veriler Çizelge 3.1'de verildiği gibidir. Ayrıca artık terimine ilişkin grafik Şekil 3.1'de sergilenmiştir.

Çizelge 3.1. Pilot-Plant verisi

Gözlem	Çekme	Titrasyon
1	123	76
2	109	70
3	62	55
4	104	71
5	57	55**(5.5)**
6	37	48
7	44	50
8	100	66
9	16	41
10	28	43
11	138	82
12	105	68
13	159	88
14	75	58
15	88	64
16	164	88
17	169	89
18	167	88
19	149	84
20	167	88

Bir değerin yanlış girildiğini kabul edelim. 5. sıradaki y gözleminde 55 yerine 5.5 olarak alınsın. Bu hata y yönünde bir sapan değere neden olur.



Şekil 3.1. Pilot-Plant verisinden elde edilen artık terimine ait grafikler

Şekil 3.1 incelendiğinde, sol üsteki grafik artıkların, tahmini $\hat{Y}$ değerlerine karşı grafiğini göstermektedir. Sıfırın etrafındaki hataları temsil eden yatay doğru etrafında rastgele dağılmış olmalı. Yani noktaların dağılımda açık bir trend olmamalı. Sol alttaki grafik, hataların normal dağılışlı olup olmadığını gösteren standart Q-Q grafiğidir. Sağ üstteki grafik standardize artıkların karekökü ile $\hat{Y}$ tahmin değerlerinin grafiğini göstermektedir. Yine bu noktalarında açık bir trende sahip olmaması gerekir. Son olarak sağ alttaki grafik, regresyon sonuçlarını değerlendirmede önemli bir ölçüm olan her bir nokta kaldıraç kuvvetini gösterir. Ayrıca regresyonda her bir gözlemin bir diğer önemli ölçümü olan Cook's uzaklığı da gösterilmiştir. Uzaklık, 1' den büyük ise şüpheli ve olası aykırı bir gözlem ya da zayıf model var demektir. 4 grafikte incelendiğinde 5. gözlemin regresyon doğrusundan sapmış olduğu söylenebilir.

EKK yöntemi için,

Çizelge 3.1'de verilen gözlem ikililerine dayalı tahmin denklemi,

$$\hat{y}_i = 28.193 + 0.368x_i$$

olarak bulunmuştur. EKK yönteminin uygulanabilmesi için hata terimleri ε_i'lerin bağımsız ve aynı dağılıma sahip ortalaması sıfır, varyansı σ^2 olan normal dağılım şartını sağlaması gerekir. Normal dağılım varsayımının sağlanıp sağlanmadığının kotrolü için Şekil 3.1'deki hata terimlerinin Q-Q grafiği incelenebilir. Şekil 3.1' de görüldüğü üzere hata terimleri normal dağılıma sahip değildir. Çünkü kutucuklar doğru üzerinde değildir. Dolayısıyla parametrik olmayan regresyon tekniklerinin kullanılması daha güvenilir sonuçlar elde edilmesini sağlayacaktır.

EKK yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 4.899$$

olarak elde edilir.

Theil yöntemi için,

Bu iki değişken arasındaki ilişkiyi tanımlamak için $\hat{\beta}_1$ eğim katsayısını hesaplayalım. Ancak ilgilenilen veri seti dikkate alındığında, 18. ve 20. gözlem değerlerine ait değerler birbirine eşittir. Eşit olan bağımsız değişkene ait değerlerden dolayı hesaplanan ikili eğim değerleri (S_{ij}) $\pm\infty$ gibi bir değer alır. Bu durumda yapılan çalışma bir belirsizliğe neden olur. Sonsuz çıkan ikili eğim değerleri (S_{ij}) analiz dışı bırakılırsa veri kaybı olacağından analize olan güven de azalacaktır. Bu problemden kurtulmak için, eşit olan bağımsız değişkenlerin sadece biri yazılır. Karşısına ise bağımsız değişkenlere karşılık olarak gelen bağımlı değişken değerlerinin medyanı yazılır. Buradan elde edilen eğim katsayısının tahmini tüm veriye uygulanarak $\hat{\beta}_0$ hesaplanır.

Bunun için, Theil yöntemi kullanıldığında $\binom{19}{2} = 171$ tane S_{ij} değeri hesaplanmıştır. Normal veride aynı x değerine sahip olan gözlem ve karşılık gelen y değerlerinin ortalaması alınarak Theil yöntemi için hesaplanan eğim bakımından $(x_i \neq x_j)$ varsayımı sağlanmıştır.

$$\hat{\beta}_1 = ortanca\{S_{ij}\} = 0.326087$$

ve

$$\hat{\beta}_0 = ortanca(Y) - \hat{\beta}_1 ortanca(X) = 69 - 0.326087 * 104.5 = 34.92391304$$

olarak hesaplanır. Theil-1 yöntemine ait regresyon tahmin denklemi,

$\hat{y}_i = 34.924 + 0.326x_i$ olarak elde edilir.

Theil yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 3.392$$

olarak elde edilir.

Optimum Tip Theil yöntemi için,

Bu yöntemle tahmin denklemi,

$$\hat{y}_i = 34.652 + 0.326x_i$$

olarak hesaplanır. Bu yönteme göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 3.378$$

olarak elde edilir.

Hodges-Lehmann Tip Theil yöntem için

Bu yöntemle elde edilen model denklemi,

$$\hat{y}_i = 32.522 + 0.326x_i$$

‘dir. Bu yönteme göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 4.560$$

olarak elde edilir.

Ağırlıklı Theil-1 yöntemi için

$w_{ij1} = (x_j - x_i)$ olarak alalım. Burada da S_{ij} değeri kadar w_{ij1} değeri vardır. Yani $\binom{19}{2} = 171$ tane w_{ij1} değeri hesaplanmıştır. Normal veride aynı x değerine sahip olan gözlem ve karşılık gelen y değerlerinin ortalaması alınarak hesaplanan eğim bakımından $(x_i \neq x_j)$ varsayımı sağlanmıştır. Bu ağırlık için hem ortalama hem de medyan değerine göre ayrı ayrı iki tane regresyon modeli tahmini yapılmıştır.

- Ortalamaya dayalı sonuçlar;

$$\hat{\beta}_{1wij1(ort)} = \frac{\sum_{i<j} w_{ij1}S_{ij}}{\sum_{i<j} w_{ij1}} = \frac{2095}{5182} = 0.40428$$

$$\hat{\beta}_{0wij1(ort)} = 69 - 0.40428 * 104.5 = 26.7523$$

olarak hesaplanır ve bu sonuçlara dayalı regresyon tahmin denklemi,

$$\hat{y}_i = 26.752 + 0.404x_i$$

şeklinde hesaplanır.

Ağırlıklı Theil-1(Ortalama) yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 5.532$$

olarak elde edilir.

- Medyana dayalı sonular;

Ağırlık $w_{ij1} = (x_j - x_i)$ ‘dir.

$$\hat{\beta}_{1wij1(med)} = \frac{ortanca_{i<j}\left(w_{ij1}S_{ij}\right)}{ortanca_{i<j}\left(w_{ij1}\right)} = \frac{13}{38} = 0.3421$$

$$\hat{\beta}_{0wij1(ort)} = 69 - 0.3421 * 104.5 = 33.25$$

dir. Böylece regresyon tahmin denklemi,

$$\hat{y}_i = 33.25 + 0.342x_i$$

şeklinde yazılır.

Ağırlıklı Theil-1(Medyan) yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1} |y_i - \hat{y}_i|}{n} = 3.594$$

olarak elde edilir.

Ağırlıklı Theil-2 yöntemi için

Şimdi de $w_{ij2} = (x_j - x_i)^2$ olarak alalım ve hem ortalamaya göre hem de medyana göre tahmin edilen regresyon modellerini inceleyelim.

- Önce ortalama için;

$$\hat{\beta}_{1wij2(ort)} = \frac{\sum_{i<j} w_{ij2} S_{ij}}{\sum_{i<j} w_{ij2}} = \frac{314950.5}{849834} = 0.3706$$

$$\hat{\beta}_{0wij2(ort)} = 69 - 0.3706 * 104.5 = 30.2720$$

‘dir. Ağırlıklı Theil-2 (Ortalama) yöntemine için regresyon tahmin denklemi,

$$\hat{y}_i = 30.272 + 0.371 x_i$$

şeklindedir.

Ağırlıklı Theil_2(Ortalama) yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1} |y_i - \hat{y}_i|}{n} = 4.374$$

olarak elde edilir.

- Şimdi de medyan için;

$$\hat{\beta}_{1wij2(med)} = \frac{ortanca_{i<j}\left(w_{ij2} S_{ij}\right)}{ortanca_{i<j}\left(w_{ij2}\right)} = \frac{1071}{3025} = 0.35404$$

$$\hat{\beta}_{0wij2(ort)} = 69 - 0.35404 * 104.5 = 32.0018$$

‘dir. Ağırlıklı Theil-2 (Medyan) yöntemine için regresyon tahmin denklemi,

$$\hat{y}_i = 32.002 + 0.354 x_i$$

şeklindedir.

Ağırlıklı Theil-2 (Medyan) yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 3.90$$

olarak elde edilir

- Theil-2 yöntemi için;

$$\hat{\beta}_{1sij(ort)} = \frac{\sum_{i<j} S_{ij}}{N} = 0.3759$$

$$\hat{\beta}_{0sij(ort)} = 69 - 0.3759 * 104.5 = 29.71549$$

şeklindedir. Model ise,

$$\hat{y}_i = 29.716 + 0.376x_i$$

'dir.

Theil-2 yöntemine göre bağımlı değişkenin tahmin değerlerinden mutlak sapmalarının aritmetik ortalaması,

$$OMS = \frac{\sum_{i=1}|y_i - \hat{y}_i|}{n} = 4.535$$

olarak elde edilir.

Elde edilen tüm tahmin yöntemlerine ait tahmin modeli ve deneysel ortalama mutlak sapma değerleri Çizelge 3.2'de özetlenmiştir.

Çizelge 3.2. İncelenen yöntemlere ait model tahmini ve ortalama mutlak sapma değerleri

Yöntem	Tahmin Denklemleri	OMS
EKK	$\hat{Y} = 28.193 + 0.368x_i$	4.899
Theil-1	$\hat{Y} = 34.924 + 0.326x_i$	3.392
Theil-Opt	$\hat{Y} = 34.652 + 0.326x_i$	**3.378***
Theil-Hod.	$\hat{Y} = 32.522 + 0.326x_i$	4.560
AğırlıklıTheil-1(Ort.)	$\hat{Y} = 26.752 + 0.404x_i$	5.532
AğırlıklıTheil-1(Med)	$\hat{Y} = 33.25 + 0.342x_i$	3.594
AğırlıklıTheil-2(Ort)	$\hat{Y} = 30.272 + 0.371x_i$	4.374
AğırlıklıTheil-2(Med)	$\hat{Y} = 32.002 + 0.354x_i$	3.900
Theil-2	$\hat{Y} = 29.716 + 0.376x_i$	4.535

EKK yönteminin gözlem değerleri içinde bulunan aykırı değerlerden etkilendiği ve hata terimlerinin dağılımının normal olmadığı durumlarda alternatif regresyon tekniklerinin kullanılması gerekliliği üzerinde durulmuştur. Bilindiği üzere gerçek verilerle çalışılırken en büyük sıkıntı verilerin dağılımının normal dağılıma uymamasıdır. Bu sebeple sağlam regresyon yöntemlerine ihtiyaç duyulmuştur. Çünkü EKK yönteminin varsayımlarından birisi de hata terimlerinin dağılımının normal olmasıdır.

Çalışmanın uygulama kısmında parametrik olmayan basit doğrusal yöntemler karşılaştırılmıştır. Sonuçlar elde edildikten sonra ortalama mutlak sapma değeri hesaplanmıştır. Çizelge 3.2 incelendiğinde ortalama mutlak sapma değeri, veride uç değerlerin varlığında iyi sonuç veren optimum tip Theil yöntemi olarak bulunmuştur. Daha sonra sırasıyla Theil-1, medyana bağlı Ağırlıklı Theil-1, medyana bağlı Ağırlıklı Theil-2, ortalamaya bağlı Ağırlıklı Theil-2, Theil-2, Hodges-Lehmann tip Theil, EKK ve ortalamaya bağlı Ağırlıklı Theil-1 yöntemleridir. Sonuçlar virgülden sonra alıanacak basamak sayısına göre değişkenlik gösterebilir.

Ayrıca sonraki çalışmalarda, incelenen bu yöntemlere yeniden örnekleme yöntemlerinden olan çakı ve bootstrap yöntemleri ikame edilerek istatistiksel çıkarsamalarda bulunulacaktır.

3.4. Çoklu Doğrusal Regresyon Analizinde Theil-Sen Tahmini (ÇTSE)

Matris gösterimi ile bir çoklu doğrusal regresyon modeli

$$Y = \beta_0 + X^T \beta_1 + \varepsilon \qquad (\mathbf{3.10})$$

ile verilir. Burada β_0 kesim parametresi ve β_1 ise p-boyutlu regresyon parametre vektörünü, X, ya tesadüfi ya da deterministik olan ortak değişkenlerin bir matrisi ve ε bağımsız aynı dağılımlı tesadüfi hataları göstermektedir. Bağımsız değişkenlerle hata terimlerleri ilişkisizdir (Lia vd, 2009).

Basit doğrusal regresyon $p = 1$ ile başlar. Geometrik olarak eğim parametresi olan β_1'i tahmin etmek için, sadece iki farklı noktaya $(X_i, Y_i), (X_j, Y_j)$ $(X_i \neq X_j$ iken) ve eğim parametresinin bir tahminine $b_{i,j} = (Y_i - Y_j)/(X_i - X_j)$ ihtiyaç duyulur.

Buna alternatif olarak, her iki ayrı nokta ile artık kareler toplamı $(Y_i - \beta_0 - \beta_1 X_i)^2 + (Y_j - \beta_0 - \beta_1 X_j)^2$' nı β_0 ve β_1 parametreleri, artık kareler toplamı denklemini sağladığı durumda en küçüklenmiş olur. Şöyle ki,

$$Y_i - \beta_0 - \beta_1 X_i = 0, Y_j - \beta_0 - \beta_1 X_j = 0 \qquad (\mathbf{3.11})$$

eşitliğini sağlar. Çözüm $\beta_{0_{i,j}} = Y_i - b_{i,j}X_i$ ve $b_{i,j} = (Y_i - Y_j)/(X_i - X_j)$ en küçük kareler tahminleridir. β_1 eğiminin sağlam bir tahmini $\hat{\beta}_n$, EKK tahmininin medyanı olur. Yani;

$$\hat{\beta}_n = Med\left\{b_{i,j} = \frac{Y_i - Y_j}{X_i - X_j} : (X_i \neq X_j, 1 \leq i < j \leq n\right\} \quad (\mathbf{3.12})$$

Burada $Med\,\{B_j : j \in J\}$; B_j sayılarının medyanını temsil etmektedir. Bu yüksek kırılma noktası ile sağlam olan Theil-Sen tahmini olarak bilinmektedir. Eğer sadece eğim (β_1) tahmini ele alınırsa, hatada hiçbir tanımlanabilirlik varsayımına ihtiyaç olmaz (Dang, 2008).

Kesim parametresini (β_0) belirleyebilmek için, hata dağılımı üzerindeki tanımlanabilirlik koşulunun bilinmesi gereklidir. Şimdi şöyle olduğunu düşünelim. Hatalar yaklaşık olarak sıfıra simetrik olan bir dağılıma sahiptir (ortalaması sıfır olan normal dağılım, sıfıra simetriktir) ve varyansı sabit ya da değişken olabilir. Örneğin standart normal dağılım sıfıra simetriktir. Bu yeterli bir durumdur. Ardından aynı şekilde, kesim parametresi EKK tahmininin medyanı

$$\widehat{\beta_{0}}_n = Med\left\{a_{i,j} = \frac{Y_jX_i - Y_iX_j}{X_i - X_j} : (X_i \neq X_j, 1 \leq i < j \leq n\right\} \quad (\mathbf{3.13})$$

eşitliği ile tahmin edilebilir.

Bu durum parametrenin (β_0, β_1) bileşen-odaklı medyan tahmini $(\widehat{\beta_0}_n, \widehat{\beta}_n)$ ile sonuçlanır.

Bu bileşen-odaklı medyan tahmininin çok zayıf bir tahmin olabileceği bilinmektedir (Wang vd, 2009). Bu zayıflığı aşmak için, kesim parametresi β_0 'ın sağlam bir tahminini oluşturmak amacıyla $\hat{\beta}_n$ sağlam tahmini kullanılabilir. Örneğin, $Med\,\{Y_i - \hat{\beta}_nX_i : 1 \leq i,j \leq n\}$. Buna alternatif olarak, çok değişkenli medyan ile *aynı anda* (β_0, β_1) tahmin edilebilir. Şöyle ki;

$$(\widehat{\beta_0}_n, \hat{\beta}_n) = Çmed\left\{\beta_{0_{i,j}}, \beta_{1_{i,j}}\right): X_i \neq X_j, \quad 1 \leq \mathrm{i} < j \leq \mathrm{n}\} \quad (\mathbf{3.14})$$

Burada $Çmed\{B_j : j \in J\}$; $\{B_j \in IR^d : j \varepsilon J\}$ vektörlerinin çok değişkenli medyanıdır (Dang vd, 2008). Çoklu doğrusal regresyon analizindeki parametrelerin Theil-Sen tahminlerini hesaplamak için çok değişkenli medyanların kullanılması gerekir.

Çok değişkenli medyan kullanımı ile ilgili çeşitli tartışmalar vardır. Çok değişkenli medyan konusu dördüncü bölümde daha kapsamlı şekilde incelenecektir.

Çoklu doğrusal regresyon analizinde TSE regresyon parametrelerinin tahmini aynı anda çok değişkenli medyanı alınarak tahmin edilebilir. Şimdi eş zamanlı olarak kesim ve normal vektörün tahmin edilmesi konusu üzerinde duralım.

$p > 1$ olan çoklu doğrusal bir regresyon modeli ele alalım. Bölüm 3.4'deki süreci izleyerek, ilk önce $\theta = (\beta_0, {\beta_1}^T)^T$ 'ün tahmini $(p + 1)$ adet eşitliğin çözümü olarak bulunabilir.

$$Y_i - \beta_0 - X_i^T \beta_1 = 0, i \in k_{p+1} = \{i_1, \ldots, i_{p+1}\} \qquad \mathbf{(3.15)}$$

ki burada k_{p+1}, $\{1, \ldots, n\}$' nin $(p+1)$ alt-kümesidir. Şöyle ki $(p + 1) \times (p + 1)$ matrisi $(X_k : k \in k_{p+1})$'se dönüştürülebilir. $(p + 1)$ gözlemlere bağımlılığı vurgulamak için, bu tahmin $\hat{\theta}_{k_{p+1}}$ ile ifade edilmiştir. Daha sonra ise Theil-Sen tahmininin basit doğrusal bir regresyon modelinden çoklu doğrusal bir regresyon modeline genişlemesi, çok değişkenli medyan yardım ile olmaktadır. Şöyle ki;

$$\hat{\theta}_n = Çmed\left\{\hat{\theta}_{\boldsymbol{k}_{p+1}} : \forall\, \boldsymbol{k}_{p+1}\right\} \qquad \mathbf{(3.16)}$$

Burada $\hat{\theta}_{\boldsymbol{k}_{p+1}}$ 'nın aynı zamanda $\{(X_i, Y_i) : i \in \boldsymbol{k}_{p+1}\}$ $p + 1$ gözlemlerine bağlı olarak θ' nın EKK tahmini olduğuna da dikkat edilmelidir. Bu açıdan bakıldığında, $p + 1 \leq m \leq n$ arasında olan $\{(X_i, Y_i) : i \in k_m\}$ gözlemlerinin m farklı keyfi bir kombinasyonu seçilebilir ve $\hat{\theta}_{k_m}$ tahminleri EKK tahminini oluşturur. Öyle ise θ parametre vektörünün çoklu Theil-Sen tahmini (ÇTSE) $\hat{\theta}_n$, tüm olası en küçük kareler tahminlerinin çok değişkenli medyanı olarak tanımlanır. Böylece θ'nın çok değişkenli medyan tahmin edicisi,

$$\hat{\theta}_n = Çmed\{\hat{\theta}_{k_m} : \forall\, \boldsymbol{k}_m\} \qquad \mathbf{(3.17)}$$

eşitliğiyle verilir. Burada olası en küçük kareler tahmin edicisi ise;

$$\hat{\theta}_k = (X_{\boldsymbol{k}}^T X_{\boldsymbol{k}})^{-1} X_{\boldsymbol{k}}^T Y_{\boldsymbol{k}} \qquad \mathbf{(3.18)}$$

ile ifade edilir.

$X_{\boldsymbol{k}}^T X_{\boldsymbol{k}}$, $(1, X_i^T) : i \in \boldsymbol{k}$ ve $Y_{\boldsymbol{k}} = (Y_i : i \in \boldsymbol{k})^T$ satırları ile $(1 + p) \times m$ boyutlu bir matris olarak düşünülsün. Kolaylık olsun diye, $\boldsymbol{k} = \boldsymbol{k_m}$ olarak alınmıştır. Burada dikkat edilmesi gereken durum m değerini seçerek herhangi bir önceden belirlenmiş olası sağlamlığın ve etkinliğin elde edebilir olmasıdır (Wang vd., 2009).

Normal vektörün tahmini için, eğer β normal parametrenin tahmin edilmesiyle ilgileniyorsak, o zaman kesim parametresinin sahip olduğu hata üzerindeki

tanımlanabilirlik koşulu (hata yaklaşık olarak sıfır etrafında simetrik olan bir dağılıma sahip olma varsayımı) tek değişkenli TSE'deki gibi gerekli değildir. Zhou ve Serfling (2006) spatial U-dağılım spatial teorisini geliştirmişler ve teorinin uygulaması olarak, gözlemlerin ikili farklılıklarına dayalı genelleştirilmiş TSE ve ÇTSE oluşturmuşlardır. Wang vd. (2009) çalışmalalarında, kısaca Zhou ve Serfling (2006) çalışmasından elde edilen sonucu gözden geçirmişlerdir (kendi yapılarında biraz daha genel, $m = p + 1$) ve onların TSE tahmini spatial derinliğe dayanır, fakat bununla birlikte farklı yeni bir çok değişkenli medyana da genişletilebilir.
Şimdi (3.10) denklemindeki ikili farklar

$$Y_j - Y_k = (X_j - X_k)^T\beta + \varepsilon_j - \varepsilon_k, \qquad j,k = 1,2,\dots,n \qquad (\mathbf{3.19})$$

olarak tanımlansın. Burada, $N = n(n-1)/2$ kadar farklı gözlem ikilisi vardır. $m \le N$ bir tamsayısı için, $K, \nabla \equiv \{(j,k) : j < k, j,\ k = 1,\dots,n\}$' dan (j,k) $\binom{N}{m}$ kombinasyonları olsun ve $\{(k_{1,i}, k_{2,i}) : i = 1,\dots,m\} \in K$ tarafından $j = 1,2$ için $\boldsymbol{k_j} = (k_{j,i} : i = 1,\dots,m)$ genel bir kombinasyonu yazılsın. Ya $\boldsymbol{k_1}$ ya da $\boldsymbol{k_2}$ için $\boldsymbol{k}$ yazılırsa ve (3.14) ifadesi,

$$Y_{k_1,k_2} = X_{k_1,k_2}\beta + \varepsilon_{k_1,k_2} \qquad (\mathbf{3.20})$$

şeklinde matris formunda yazılabilir. Burada $Y_{k_1,k_2} = Y_{k_1} - Y_{k_2}$, $X_{k_1,k_2} = X_{k_1} - X_{k_2}$ ve $\varepsilon_{k_1,k_2} = \varepsilon_{k_1} - \varepsilon_{k_2}$ $\varepsilon_k = (\varepsilon_k : k \in k)^T$ 'dır.
$\hat{\beta}_{k_1,k_2}$ gözlemlerin alt kümesine dayalı EKK tahminini göstermek üzere

$$\hat{\beta}_{k_1,k_2} = ({X_{k_1,k_2}}^T X_{k_1,k_2})^{-1} {X_{k_1,k_2}}^T Y_{k_1,k_2} \qquad (\mathbf{3.21})$$

eşitliği ile tahmin edilir. Böyelece TSE, ÇTSE'ye spatial medyan olarak genişletilmiş olur. Sonuç olarak bu tahmin edici

$$\hat{\beta}_n = Çmed\{\hat{\beta}_{k_1,k_2} : (\boldsymbol{k_1}, \boldsymbol{k_2}) \in K_0\} \qquad (\mathbf{3.22})$$

ile verilir. Burada, K_0, tüm EKK tahmin yönteminin mevcut olduğu K'nın alt uzayıdır. Yani aslında eldeki örneklem hacmine ve seçilen keyfi m sayısına bağlı olarak oluşturulan kombinasyonların tümünü almak yerine tesadüfen seçilen alt örneklem alınıp, spatial medyana bakılarak $\hat{\beta}_n$ tahmin edicisi hesaplanır. Çünkü seçilen tüm kombinasyonlar dikkate alınarak $\hat{\beta}_n$'in hesaplanması çok zaman kaybına neden olacağından belli bir alt uzayı seçilir. Ancak bu şekilde alt uzay seçilip işlemler yapıldığında her alt örneklem için sonuçlar değişebilir. Bu ise bu yöntemin eksik yönü olarak değerlendirilebilir. İleriki çalışmalar bu eksikliğin giderilmesine yönelik olacaktır.

Basit bir doğrusal regresyon modelinde, Peng vd. (2008) hata dağılımı konusunda herhangi bir varsayımda bulunmadan Theil-Sen tahminini çalışmışlardır (Yani, hata dağılımına ne simetri ne de sürekliliği varsayarak). Yine Peng vd. aynı çalışmalarında TSE'nin normal koşullar altında son derece tutarlı olduğunu ve asimptotik dağılıma sahip olduğunu ve hata dağılımının süreksiz olduğu durumlarda bile oldukça etkin olduğunu göstermişlerdir.

Doğal olarak, Dang vd. (2008) tarafından bu sonuçların ÇTSE'ye genişletilebilir olup olmadığı, eğer genişletilebilirse bunun hangi koşullar altında olacağı sorusu gündeme getirilmiştir. Bu sorunun cevabında özellikle karşımıza iki soru çıkmaktadır. Birincisi, hata dağılımının simetrik varsayımını çıkarılabilir mi? İkincisi ise, hata ε süreksiz olduğunda süper etkinlik geçerli olur mu? Bu iki sorunun cevabı da Wang vd. (2009) çalışmasında gösterildiği gibi olumlu olmuştur. Şimdi ÇTSE tahmin yöntemi için parametre tahmini algoritmasının işleyişini somut bir şekilde özetlemeye çalışalım.

Kurulan çoklu doğrusal regresyon modelinde bağımsız değişken sayısı iki $(p = 2)$ olarak alınsın ve örnek hacminin $n = 20$ olduğu varsayılsın. m keyfi sayısı $p + 1 \leq m \leq n$ aralığında seçilecek olan keyfi bir değerdir. Örnek vermek gerekirse, $m = 3$ olsun. Basitçe algoritmanın işleyişini Çizelge 3.3 'deki gibi özetlenebilir,

Çizelge 3.3. Çoklu doğrusal regresyonda ÇTSE tahmin yöntemi ile parametre tahmini

<table>
<tr><th>No</th><th>$\boldsymbol{Y}$</th><th>$\boldsymbol{X_1}$</th><th>$\boldsymbol{X_2}$</th><th colspan="3">$\boldsymbol{\hat{\beta}_i}$</th></tr>
<tr><td>1</td><td>Y_1</td><td>X_{11}</td><td>X_{12}</td><td rowspan="3">$\hat{\beta}_0$</td><td rowspan="3">$\hat{\beta}_1$</td><td rowspan="3">$\hat{\beta}_2$</td></tr>
<tr><td>2</td><td>Y_2</td><td>X_{21}</td><td>X_{22}</td></tr>
<tr><td>3</td><td>Y_3</td><td>X_{31}</td><td>X_{32}</td></tr>
<tr><td>1</td><td>Y_1</td><td>X_{11}</td><td>X_{12}</td><td rowspan="3">$\hat{\beta}_0$</td><td rowspan="3">$\hat{\beta}_1$</td><td rowspan="3">$\hat{\beta}_2$</td></tr>
<tr><td>2</td><td>Y_2</td><td>X_{21}</td><td>X_{22}</td></tr>
<tr><td>4</td><td>Y_4</td><td>X_{41}</td><td>X_{42}</td></tr>
<tr><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td><td>…</td></tr>
<tr><td>18</td><td>Y_{18}</td><td>X_{181}</td><td>X_{182}</td><td rowspan="3">$\hat{\beta}_0$</td><td rowspan="3">$\hat{\beta}_1$</td><td rowspan="3">$\hat{\beta}_2$</td></tr>
<tr><td>19</td><td>Y_{19}</td><td>X_{191}</td><td>X_{192}</td></tr>
<tr><td>20</td><td>Y_{20}</td><td>X_{201}</td><td>X_{202}</td></tr>
<tr><td>Toplam</td><td>1140 tane</td><td>kombinasyon</td><td>mevcut.</td><td colspan="3">$Çmed(\hat{\beta}_i)$</td></tr>
</table>

Çizelge 3.3 incelendiğinde ÇTSE tahmin yöntemi ile regresyon parametrelerinin tahmin edilmesi için toplam $\binom{n}{m} = \binom{20}{3} = 1140$ tane kombinasyon mevcuttur. Daha sonra her üçlü kombinasyon değerleri için EKK yöntemi ile regresyon parametreleri tahmin edilecektir. Hesaplanan her üçlü kombinasyon değeri için elimizde 1140 tane $\hat{\beta}_i$ tahmin değeri bulunacaktır. Tahmin edilen 1140 tane regresyon katsayılarının R paket programındaki "depth" paketi yardımıyla eş zamanlı olarak çok değişkenli medyan değeri hesaplanacaktır (URL-2). $\hat{\beta}_i$' ye ait olarak elde edilen bu çok değişkenli medyan değeri, ÇTSE tahmin yöntemine ait çoklu doğrusal regresyon tahmin değerleri olarak elde edilecektir.

4. ÇOK DEĞİŞKENLİ MEDYAN KAVRAMI

Bir veri setinin medyanı bir dağılımın merkezindeki doğal sağlam tahmin olarak parametrik olmayan problemlerde ortaya çıkar. Bir dağılımın merkezindeki parametrik olmayan sağlam tahmin olarak, farklı bozulma özellikleri ile gösterildiği üzere örnek ortalamasından daha farklı bir özelliğe sahiptir. Bu yüzden, bir veri setini kirleten ve sonsuzluğa giderken ortalamayı da sonsuzluğa gönderen tek bir nokta yeterlidir. Buna zıt olarak, medyanı aynısını yapmaya zorlamak için verinin en az %50 'si sonsuzluğa gönderilmelidir. Sibson (1984)' ün belirttiği gibi, ikinci dereceden çözülmelerin geometrisi, kare mesafelere dayalı yöntemlerin bir boyuttan daha yüksek boyutlara taşınırken genellemeye daha uygun hale getirir. Bu yüzden, çok değişkenli analize çok değişkenli normal dağılıma dayalı teori egemendir. Yine de, parametrik olmayan ve dağılımdan bağımsız olan yöntemler sınıfının daha yüksek boyutta veri analizinde bu zamana kadar olandan daha uygun bir yeri vardır.

Tek değişkenli medyanın iki veya daha fazla boyutdaki medyan ile benzerliği nedir sorusuna bir dizi farklı cevap aranmıştır. Uygun benzerliğin ne olduğuna dair kesin bir cevap olmasa da, kesinlikle sahip olması gereken bazı özellikler vardır. Örneğin, birinci boyutta, bir dağılımın simetri merkezi medyanıdır. Bu durumda, doğal olarak aynı şeyin daha yüksek boyutlarda da geçerli olması, daha yüksek boyutsal medyanın simetri merkezinin doğal bir tahmini olması beklenir. Yine de, burada bile biraz belirsizlik vardır: tek değişkenli bir dağılım simetrisinin merkezi farklı şekillerde genellenebilir. Small (1990) 'nın çalışmasında dört farklı simetri çeşidinin tahmini ve test edilmesinde çok boyutlu meydanların kullanımı ele alınmıştır: küresel, eliptik, merkezi ve açısal simetri noktaları. (Chaudhuri, 1992)

Çok değişkenli medyanlar bir çok değişkenli dağılımın simetri merkezini tahmin etmede kullanılan ortalama vektörünün sağlam rakipleridir. Çok değişkenli medyanların çeşitli tanımlamaları literatürde sunulmuş ve özellikleri (verimlilik, eşvaryanslılık, sağlamlık, sayısal uygunluk, doğruluk tahmini, vb.) geniş ölçüde incelenmiştir. Tek değişkenli medyan ve tek değişkenli işaret ve sıra kavramları, tek değişkenli gözlemlerin sıralamasına bağlıdır. Maalesef, çok değişkenli veri noktalarının doğal sıralaması yoktur. L_1 amaç fonksiyonlarını kullanan bir yaklaşım bu nedenle sıkça bu kavramlar çok değişkenli duruma genişletmekte kullanılır. Tez kapsamında, spatial medyan göz önünde bulundurulmuş ve literatürde bulunduğu haliyle istatistiksel özellikleri incelenmiştir (Becker vd., 2014). Çok değişkenli medyanın diğer incelemeleri için Small (1990), Chaudhuri ve Sengupta (1993), Niinimaa ve Oja (1999), Dhar ve Chaudhuri'ye (2011) çalışmalarına bakılabilir.

4.1. Tek Değişkenli Medyan

Becker, vd. (2014)'nin editörlüğünü yaptığı Robustness and Complex Data Structures adlı kitabın ilk bölümünde aktarıldığı gibi,
$\mathbf{x} = (x_1, \dots, x_n)$ dağılım fonksiyonu F olan tek değişkenli bir dağılımdan alınan rastgele bir örnek olsun. Medyan fonksiyonel olarak $T\,(F\,)$ ve buna karşılık gelen örnek istatistiği $T\,(\mathbf{x}) = T\,(F_n)$ bir kaç yolla tanımlanabilir. Buna karşılık tek değişkenli medyan için bazı mümkün tanımlar aşağıdaki gibidir.

1. Medyan fonksiyonel olarak şu şekilde tanımlanır.

$$T(F) = \inf\left\{x : F(x) \geq \frac{1}{2}\right\} \quad \mathbf{(4.1)}$$

2. Medyan $T\,(F\,)$ fonksiyonunun değerini en yüksek yapar. Şöyle ki;

$$t \ \rightarrow \text{enk}\{P(x_1 \leq \ t), P(x_1 \geq t)\} \ = \text{enk}\{F(t), 1 - \ F(t\, -)\} \quad \mathbf{(4.2)}$$

3. Medyan $T\,(F\,)$ fonksiyonunun değerinin olasılığını en yüksek yapar. Şöyle ki,

$$t \ \rightarrow \ P(\text{enk}\{x_1, x_2\} \leq \ t \ \leq \text{enb}\{x_1, x_2\}) = \ 2F\,(t)\big(1 - \ F\,(t\, -)\big) \quad \mathbf{(4.3)}$$

4. Medyan $T\,(F\,)$ fonksiyonunun beklenen değerini en küçük yapar. Şöyle ki;

$$E(|x_1 - t|) \ \text{ ya da } \ D(t) = E\{|x_1 - t| - |x_1|\} \quad \mathbf{(4.4)}$$

olur. Dikkat edilirse, $||x_1 - t| - \ |x_1|| \ \leq \ |t|$ olduğunda, $D(t)$' nin tanımında beklenti her zaman vardır.

5. Medyan, $T\,(F\,)$ fonksiyonunu

$$S(t) = \begin{cases} +1, & t > 0 \; ise, \\ \;\;0, & t = 0 \; ise, \\ -1, & t < 0 \; ise. \end{cases} \quad \mathbf{(4.5)}$$

tek değişkenli işaret fonksiyonu iken,

$$E[S(x_1 - t)] = 0 \qquad (\mathbf{4.6})$$

tahmin denklemini çözer.

Yukarıda tanımlanan kitle medyanı $T\ (F)$'nin farklı tanımları, bir F dağılımı için sınırlı ve sürekli yoğunluğu μ' de $f\ (\mu)$ olan aynı özel μ değerine sahiptir. Öyleyse amaç fonksiyonu $D(t)$ için eşitlik (4.7) doğrudur.

$$D(t) = D(\mu) + \frac{\delta}{2}(t-\mu)^2 + o((t-\mu)^2) \qquad \delta = 2f(\mu) \text{ ile} \qquad (\mathbf{4.7})$$

Örnek medyan $\hat{\mu}$ işaret fonksiyonu $S(t)$'yi temel alan tek değişkenli işaret testi ile ilişkilidir. Tek değişkenli işaret fonksiyonundan başlayarak, tek değişkenli (merkez) sıra fonksiyonu şu şekilde tanımlanır.

$$\hat{R}(t) = \frac{1}{n}\sum_{i=1}^{n} S(t - x_i) \qquad (\mathbf{4.8})$$

$\hat{R}(t) \in [-1, 1]$' a ve örnek medyan için tahmin denkleminin $\hat{R}(\hat{\mu})$= 0 olduğuna dikkat edilmelidir. Sıfır hipotezi $H_0 : \mu = 0$' ı test etmek için işaret test istatistiği $\hat{R}(0)$'dır. Test istatistiği açık şekilde ve asimptotik olarak dağılımdan bağımsız, gerçek medyan µ ise,

$$n\frac{R(\mu) + 1}{2} \sim Bin\left(n, \frac{1}{2}\right), \quad \text{ve} \quad \sqrt{n}\hat{R}(\mu) \to {}_dN(0,1) \quad (\mathbf{4.9})$$

$\delta = 2f\ (\mu)$ olduğunda şu da gösterilebilir ki

$$\hat{\mu} = \mu + \delta^{-1}\hat{R}(\mu) + o_P\left(n^{-\frac{1}{2}}\right) \qquad (\mathbf{4.10})$$

ve sonuç olarak,

$$\sqrt{n}(\hat{\mu} - \mu) \to {}_dN_p(0, \delta^{-2}) \qquad (\mathbf{4.11})$$

bulunur.

4.1.1 Hesaplama

Örnek dağılım fonksiyonu F_n' ye uygulandığında, yukarıdaki farklı tanımlar farklı ve çok tek olmayan çözümlere sahiptir. Öyle ise kitle medyan $T\ (F\ ) = \mu$' nün tahmini örnek medyan $\hat{\mu}$ genellikle şu şekilde tanımlanır. Önce, $x_{(1)}, \dots, x_{(n)}$ sıralı gözlemler olsun. (Çok değişkenli durumda veri noktalarının doğal sıralanışının olmadığına dikkat edilmedilir). Öyle ise örnek medyanı; n çift sayı ise

$$\hat{\mu} = \frac{x_{[n/2]} + x_{[(n+2)/2]}}{2} \qquad (\mathbf{4.12})$$

ve n tek sayı ise $\hat{\mu} = x_{\left(\frac{n+1}{2}\right)}$' ye eşittir.

4.1.2 Sağlamlık

Medyanın asimptotik bozulma noktası $1/2$ ile sınırlı ve etki fonksiyonu $IF(x;\ T, F) = \delta^{-1}S(x - T\ (F\ ))$ ile oldukça sağlam bir tahmin olduğu bilinir.

4.1.3 Asimptotik etkinlik

Eğer F dağılımı sınırlı bir ikinci moment σ^2 'ye sahipse, o zaman kitle ortalaması $\mu = E(x_i)$' yi tahmin eden örnek ortalaması $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, sınırlayıcı bir normal dağılıma sahiptir $\sqrt{n}(\bar{x} - \mu) \to N(0, \sigma^2)$ ve simetrik F için, örnek medyanı ve örnek ortalaması arasındaki asimptotik göreli etkinlik (ARE) sınırlayıcı varyansların oranı olarak tanımlanır.

$$ARE = 4f^2(\mu)\sigma^2 \qquad (\mathbf{4.13})$$

Eğer F normal dağılımlı $N(\mu, \sigma^2)$ ise, bu ARE = 0.64 kadar küçüktür. Fakat yavaş hareket eden kuyruklu dağılımlar için, medyanın asimptotik etkinliği daha iyidir; serbestlik derecesi 3 olan bir t-dağılımı, Cauchy dağılımı ve örneğin 1.62 ve 2 olan bir Laplace dağılımı için ARE değerleri daha etkindir.

4.1.4. Tahminin varyansının tahmini

Verilen bir $\delta = 2f\ (\mu)$' yü tahmin etmek zordur. Tartışma için, Hettmansperger ve McKean (2010) bakılabilir. Fakat dikkat çekicidir ki işaret testini ters çevirerek açık şekilde μ için dağılımdan bağımsız güven aralıkları elde etmek mümkündür. Bu sürekli bir F dağılımı için şu şekildedir,

$$P\left(x_{(i)} < \mu < x_{(n+1-i)}\right) = P\left(i \leq \frac{nR(\mu) + 1}{2} \leq n - i\right) = \sum_{j=i}^{n-i} \binom{n}{j} 2^{-n} \qquad (\mathbf{4.14})$$

4.1.5. Eşvaryans

Bir lokasyon dönüşümü için, fonksiyonelin lineer dönüşümler altında eş değişkenli olması umulur, yani,

$$T(F_{ax+b}) = aT(F_x) + b, \quad \forall \text{ a ve b için} \qquad (\mathbf{4.15})$$

Bu, medyanda sınırlı ve sürekli türevi olan dağılımlar ailesindeki medyan dönüşümü için doğrudur. Medyanın aslında çok daha büyük dönüşüm setleri altında eş değişkenli olduğuna dikkat edin. Eğer $g(x)$ açıkça monoton bir fonksiyon ise, o zaman $T\ (F_{g(x)})\ =\ g(T\ (Fx\ ))$ yazılabilir.

4.2. Spatial Medyan

Spatial olarak adlandırılan **T(X)** medyanı $\sum_i \|\mathbf{x}_i - \mathbf{t}\|$ kıstas fonksiyonunu ya da $\|t\| = (t_1^2 \ + \cdots + t_p^2)^{1/2}$ Öklid normunu belirtecek şekilde

$$D_n(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^{n}\{\|\mathbf{x}_i - \mathbf{t}\| - \|\mathbf{x}_i\|\} \qquad (\mathbf{4.16})$$

ile ifade edilir. Buna karşılık gelen fonksiyonel, spatial medyan T(F) ise

$$D(t) = E_F\{\|\mathbf{x} - \mathbf{t}\| - \|\mathbf{x}\|\} \qquad (\mathbf{4.17})$$

ifadesini en küçük yapar. Asimptotik sonuçlar için

1. $D(\mathbf{t})$'yi en küçük yapan μ spatial medyanı tektir.
2. F_x dağılımı sınırlı ve μ de sürekli yoğunluğu vardır. Öyle olsa da

$$D(\mathbf{t}) = D(\mu) + \frac{1}{2}(\mathbf{t} - \mu)'\Delta(\mathbf{t} - \mu) + o(\|\mathbf{t} - \mu\|^2) \quad (\mathbf{4.18})$$

Burada,

$$\Delta = E\left(\frac{1}{\|\mathbf{x} - \mu\|}\left[\mathbf{I}_p - \frac{(\mathbf{x} - \mu)(\mathbf{x} - \mu)'}{\|\mathbf{x} - \mu\|^2}\right]\right) \qquad (\mathbf{4.19})$$

şeklindedir.

Çok değişkenli spatial işaret ve merkezsel sıra fonksiyonları şu şekilde verilebilir,

$$\mathbf{S}(\mathbf{t}) = \begin{cases} \frac{\mathbf{t}}{\|\mathbf{t}\|}, & \mathbf{t} \neq 0 \text{ ise} \\ \mathbf{0}, & \mathbf{t} = 0 \text{ ise} \end{cases} \qquad (\mathbf{4.20})$$

ve

$$\widehat{\mathbf{R}}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{S}(\mathbf{t} - \mathbf{x}_i) \qquad (\mathbf{4.21})$$

$\mathbf{S}(\mathbf{t})$ spatial işaretinin sadece **t, t** ≠ 0 yönünde bir vektör olduğuna dikkat edilmelidir. $\widehat{\mathbf{R}}(\mathbf{t})$ merkezi sırası p-top B^p biriminde uzanır.

$H_0\colon \mu = \mathbf{0}$ hipotezini test etmek için spatial işaret test istatistiği $\mathbf{R}(\mathbf{0})$ ve sınırlayıcı sıfır dağılımı

$$\sqrt{n}\widehat{\mathbf{R}}(\boldsymbol{\mu}) \rightarrow {}_{d}N_p(\mathbf{0}, \Omega) \qquad (\mathbf{4.22})$$

ile verilir. Burada,

$$\Omega = E\left(\frac{(\mathbf{x}-\mu)(\mathbf{x}-\mu)'}{\|\mathbf{x}-\mu\|^2}\right) \qquad (\mathbf{4.23})$$

şeklindedir. Ayrıca

$$\hat{\mu} = \mu + \Delta^{-1}\widehat{\mathbf{R}}(\mu) + \mathbf{o}_P\left(n^{-\frac{1}{2}}\right) \qquad (\mathbf{4.24})$$

olup buradan

$$\sqrt{n}(\hat{\mu}-\mu) \rightarrow {}_{d}N_p(0, \Delta^{-1}\Omega\Delta^{-1}) \qquad (\mathbf{4.25})$$

elde edilir. Tahminin özellikleri için Oja (2010), Möttönen vd., (2010) çalışmaları incelenebilir.

4.2.1 Tahminin hesaplanması

Eğer veri en az iki-boyutlu uzaya düşmüşse spatial medyan tektir. Spatial medyanın hesaplanması için Weisfeld denilen algoritmanın işlem akış adımı vardır.

$$\mu \leftarrow \mu + \left[\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\mu\|^{-1}\right]^{-1}\mathbf{R}(\boldsymbol{\mu}) \qquad (\mathbf{4.26})$$

Algoritma bazen başarısız olabilir ama Vardi ve Zhang (2000) tarafından düzenlenmiş bir algoritma hızlı ve monoton bir şekilde yakınsar. Tahmin edilmiş kovaryans matrisi ile tahmin R programının MNM paketi kullanılarak elde edilebilir. Örnek çalışma için Nordhausen ve Oja (2011)'e bakılabilir.

4.2.2 Tahminin sağlamlığı

Spatial medyan asimptotik bozulma noktası $1/2$ ile oldukça sağlamdır. $S(\mathbf{t})$ spatial işaret fonksiyonu olduğunda, etki fonksiyonu $IF(x;\, T, F) = \Delta^{-1}\boldsymbol{S}(\boldsymbol{x} - \boldsymbol{T}(F))$ sınırlıdır.

4.2.3 Tahminin asimptotik etkinliği

Eğer Σ gibi bir kovaryans matrisi varsa ve aynı $\boldsymbol{\mu}$ kitle değerini tahmin ederlerse o zaman medyan ve ortalama vektörü arasındaki asimptotik göreli etkinlik (ARE) şu şekildedir:

$$\text{ARE} = \left(\frac{|\Sigma|}{|\Delta^{-1}\Omega\Delta^{-1}|}\right)^{\frac{1}{p}} \qquad \mathbf{(4.27)}$$

x' in küresel bir p −değişkeni dağılımı durumunda $p > 1$ iken, bu ARE

$$\text{ARE}_p = \left(\frac{p-1}{p}\right)^2 E(\|\mathbf{x}\|^2)E^2(\|\mathbf{x}\|^{-1}) \qquad \mathbf{(4.28)}$$

şeklinde yazılır. p −değişkenli küresel normal dağılım durumda ise,

$\text{ARE}_2 = 0.785,$ $\text{ARE}_3 = 0.849,$ $\text{ARE}_6 = 0.920$ ve $\text{ARE}_{10} = 0.951$

olur ve etkinlik $p \to \infty$ iken 1'e gider. Yavaş hareket eden yanlı dağılımlar için, spatial medyan örnek ortalama vektöründen daha iyidir.

4.2.4 Kovaryans matrisi tahmininin tahmin edilmesi

Bu durumda,

$$\frac{1}{n}\Delta^{-1}\Omega\Delta^{-1} \qquad \mathbf{(4.29)}$$

Buna karşılık gelen kovaryans matrisi için,

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\|\mathbf{x}_i - \hat{\mu}\|}\left[\mathbf{I}_p - \frac{(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})'}{\|\mathbf{x}_i - \hat{\mu}\|^2}\right]\right) \qquad \mathbf{(4.30)}$$

ve

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\frac{(\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})'}{\|\mathbf{x}_i - \hat{\mu}\|^2} \qquad \mathbf{(4.31)}$$

'i kullanarak kolayca bir tahminde bulunulur. Spatial medyanın kovaryans matris tahmini R paket programı MNM paketinde yapılmıştır.

4.2.5. Tahminin afin eşvaryansı

Spatial medyan

$$\mathbf{T}(F_{\mathbf{Ax+b}}) = \mathbf{AT}(F_{\mathbf{x}}) + \mathbf{b} \qquad \mathbf{(4.32)}$$

sadece A ortogonal matrisleri için doğru olduğundan dolayı afin eş değişkenli değildir.

4.2.6. Dönüşüm-tekrar dönüşüm (TR) tahmini

Afin, bir eşdeğişkenli dönüşümü tekrar dönüşüm (TR) spatial medyanı şu şekilde bulunur. $\boldsymbol{S}(F)$ bir saçılım fonksiyoneli olsun ve öyle bir $p \times p$ −matris değerli $\boldsymbol{G}(F) = \boldsymbol{S}^{-1/2}(F)$ fonksiyoneli bulunur ki

$$\mathbf{G}(F)\mathbf{S}(F)\mathbf{G}(F)' = \mathrm{I}_p \qquad (\mathbf{4.33})$$

olacaktır. $\boldsymbol{G}(F)$' nin sabit koordinat fonksiyoneli olması gerekmediğine dikkat edilmelidir. Sonra dönüşüm tekrar dönüşüm (TR) medyanı ise $p \times p$ matris değerli fonksiyonel $\boldsymbol{G}(F)$ sabit koordinat sistemi (ICS) fonksiyonu olarak adlandırılır. O zaman dönüşüm-tekrar dönüşüm (TR) medyan fonksiyonu

$$\mathbf{T}_{TR}(F\mathrm{x}) = \mathbf{G}(F\boldsymbol{x})^{-1}\mathbf{T}\left(F_{\mathbf{G}(F_{\mathbf{x}})\mathbf{x}}\right) \qquad (\mathbf{4.34})$$

Bu konu için Chakraborty vd. (1998), Ilmonen vd. (2012) çalışmalarına bakılabilir.

Spatial medyan ile Tyler'ın saçılım matrisini birleştiren TR medyanı Hettmansperger ve Randles (2002) tarafından önerilmiştir ve Hettmansperger–Randles medyanı olarak adlandırılmaktadır. Bu medyan R paket programında MNM paketinde hesaplanabilmektedir.

Spatial medyanın tüm bu özellikleri dikkate alındığında diğer çok değişkenli medyan yöntemlerine göre üstünlüğünü göstermektedir. Bu bağlamda spatial medyanın varlığı ve eşsizliği hakkında bazı bilgiler kısaca verilecek olursa,

Z, $\mathbb{R}^d$'de Q dağılımlı rastgele bir değişken olsun. Eğer aşağıdakilerden birisi doğruysa Z, tek spatial bir medyana sahiptir.

a) Q bir doğru üzerinde yoğunlaşmamıştır (Milasevic ve Ducharme, 1987, Ann. Statist.).
b) Hiç birisi $d \geq 2$ için nokta kitlesi olmayan iki tek boyutlu marjinal dağılım vardır.
c) En az iki kesin sürekli tek boyutlu marjinal dağılım vardır.
d) Q medyanı etrafında açısal simetriktir.
e) Q medyanı etrafında merkezi simetriktir (URL-1).

Ayrıca spatial medyanın güçlü tutarlılık ve asimptotik özellikleri literatüre yerleşmiştir (bknz. Bose (1998), Chaudhuri (1992), Niemiro (1992)).

Daha farklı çok değişkenli medyan yöntemleride vardır. Ancak tez kapsamında spatial medyan yöntemi kulanıldığı için, diğer medyan yöntemlerine değinilmemiştir. İleriki çalışmalarda spatial medyan yerine diğer medyan yöntemleri de kullanılarak çeşitli araştırmalar yapılacaktır.

5. ÇAKI YÖNTEMİ

Çakı**,** yanlılık ve standart hatanın tahmini için bilgisayar tabanlı yöntemlerin ilkidir. Çakı yeniden örnekleme yöntemi, istatistiksel yanlılıkları ortadan kaldırmak amacıyla 1956'da Maurice Quenouille tarafından kullanılmıştır. Daha sonra hipotez testleri ve güven aralıkları oluşturmak için 1958' de John Tukey tarafından genişletilmiş ve "Jackknife" (Çakı) adı uyarlanmıştır (Tukey, 1958). Tukey ve Quenouille'nin her ikiside, yanı azaltmayı ve sağlam aralık tahminini ele almışlardır (Miller, 1974). Doğrusal regresyon ayarında daha da genişletilmesi Miller (1974), Gray ve Schucany (1972), Hinkley (1977) ve Reeds (1978) tarafından sağlanmıştır. Roberts ve Martin (2010) ise jackknife-after-bootstrap yöntemi ile önemli gözlemlerin belirlenmesi çalışmalarında bulunmuşlardır.

Örnekleme yöntemi, genel anlamda kitleden örneklem seçme sürecidir. Bu nedenle bilimsel bir araştırmanın vazgeçilmez aşamalarından biridir. Örneklemin en önemli özelliği, kitleye ilişkin geçerli ve tutarlı tahminlere ulaşmak için örnekleme hatasının en küçük olmasını sağlayabilmektedir. Fakat parametrik tahmin etme yöntemleri, örneklem sayısı az olduğunda güvenilir sonuçlar vermemekte, aynı zamanda parametrik yöntemlerin varsayımlarını da bozabilmektedir.

Son yıllarda verilerin parametrik yöntemlerle değerlendirilmesi uygun olmadığı durumlarda yeniden örnekleme yöntemleri kullanılmaya başlanmıştır. Bunlardan birisi de çakı yöntemidir. Çakı yöntemi, kitle parametrelerinin tahmin edilmesinde dar güven aralıkları elde edilmesi ile ilgili olarak örnekleme hatasının en aza indirilmesi amacına yönelik olarak geliştirilmiştir. Ayrıca, çakı yöntemi, parametre tahmini gerektiren birçok alanda veri setindeki değişkenler arasındaki ilişkiyi ortaya koymayı amaçlayan istatistiksel bir süreç olarak değerlendirilmektedir (Swingler, 1995). Çakı, çok çeşitli işlerde kullanacağımız bir el aracıdır. Bu yöntemde, çakı gibi, bir çok işi gerçekleşmede kullanılabilecek bir yöntemdir. Bu açıdan isim benzerliği çok uygundur. Ayrıca çakı kullanarak yapılabilen işlerden biri de yontma işidir. Çakı yönteminde de yontmaya benzer bir işlem yapılmaktadır.

İstatistiksel tahmin çalışmalarında farklı yöntemler ile çalışıldığında, elde edilen sonuçların benzer olmaları beklenmektedir. Farklı yöntemler ile aynı veriye analiz yapılması çoğu zaman uzun süre alabileceği ve her yöntemin farklı veri yapısında uygulama alanı olmasından dolayı tercih edilmeme durumları olmaktadır.

Ayrıca tahmin sonuçlarının benzer olması ve kitle için genelleyebilmenin yapılabilmesi için, yeni örnek grubu ile çalışmayı tekrarlamak ile mümkün olabilmektedir. Bu durum ise uzun zaman almakta, fazladan finans veya personel gerektirdiğinden araştırmacılar tarafından pek tercih edilmemektedir. Buna karşılık yeni bir örneklem oluşturmadan da mevcut bir çalışmayı tekrarlayabilmek mümkün olmaktadır. Aynı çalışmayı yeni bir örneklem grubu ile tekrarlamadan, kitle için genelleştirebilecek güvenli tahminler elde etme yollarını araştıran bazı yöntemler bulunmaktadır. Bunlardan birisi de çakı yöntemidir Kayri ve Büyüköztürk, 2009; Bekiroğlu vd., 2013. Zaman ve Alakuş, 2015).

Yöntemin temel mantığı, veri setinden her bir gözlem değerini bir kez dışarıda bırakarak geride kalan gözlemlerden örneklem istatistikleri hesaplamaya dayanmaktadır. Bu şekilde n tane gözlemden sadece n tane farklı örneklem oluşturulabilir.

$X = (x_1, x_2, \dots, x_n)$ örneklemine sahip olalım ve $\hat{\theta} = s(X)$ tahmin edici olsun. ÇakıYönteminde göre $i.$ gözlem dışarıda bırakıldığında elde edilen örneklem;

$$x_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \ ; \ i = 1,2, \dots, n \qquad \mathbf{(5.1)}$$

şeklindedir.Bu ifade $i = 1,2, \dots, n$ olmak üzere çakı örnekleri olarak adlandırılır. Yani, $i.$ çakı örneği orijinal veri kümesi olarak adlandırılan kümeden $i.$ kümenin kaldırılmasıyla elde edilir. Örneğin orijinal veri kümesinin ortalaması $\bar{x}$ ve $i.$ gözlem işlem dışı bırakıldıktan sonra hesaplanan ortalama $\bar{x}_{(i)}$ biliniyor ise $i.$ gözlem değerini hesaplamak mümkündür (Walsh, 2000). Bu işlem aşağıdaki şekilde ifade edilebilir.

$$x_{(i)} = n\bar{x} - (n-1)\bar{x}_{(i)} \qquad \mathbf{(5.2)}$$

$\hat{\theta}_{(i)} = s(x_{(i)})$ $\hat{\theta}$' nın $i.$ çakı tekrarı olarak düşünüldüğünde yanlılığın çakı tahmini aşağıdaki şekilde tanımlanır,

$$\widehat{Yan}_{çakı} = (n-1)\big(\hat{\theta}_{(.)} - \hat{\theta}\big) \qquad \mathbf{(5.3)}$$

Burada $\hat{\theta}_{(.)}$, θ'nın yanlılığı azaltılmış çakı tahmini olup $\hat{\theta}_{(.)} = \frac{\sum_{i=1}^{n} \hat{\theta}_{(i)}}{n}$ eşitliği ile hesaplanır (Friedl ve Stampfer, 2001).

Standart hatanın çakı tahmini ise;

$$\widehat{se}_{çakı} = \frac{\left\{\frac{\left[\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2\right]}{n-1}\right\}^{1/2}}{\sqrt{n}} \qquad (5.4)$$

eşitliği kullanılarak elde edilir. Çakı tahmininde sözde değerlerin (pseudo value) hesaplanması ise,

$$Sözdedeğerler_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} \qquad (5.5)$$

eşitliği ile hesaplanır (Zaman ve Alakuş, 2015; Abdi ve Williams, 2010; URL-3).

Buraya kadar çakı tahmini ile ilgili açıklamalar sadece tek örnek durumu dikkate alınarak yapılmıştır. Örneğin, örnek sayısı iki olduğunda standart hatanın çakı tahmininin nasıl hesaplanacağı sorusunun cevabı Arseven (1969) 'dan gelmiştir (Arseven, 1969). Şöyle ki;

$x^1 = (x_1^1, x_2^1, \ldots, x_{n1}^1)$ ve $x^2 = (x_1^2, x_2^2, \ldots, x_{n2}^2)$ ile gösterilen iki bağımsız örneğin bulunduğu ve bu iki bağımsız örneğin anakütle parametrelerinin ortak bir fonksiyonu ile ilgilendiğimizi varsayalım. Bu ortak fonksiyonu $\theta = g(\theta^1, \theta^2)$ ile ifade edelim. θ' ya ait tahmin değerini ise $\hat{\theta} = g(\hat{\theta}_1, \hat{\theta}_2)$ ile gösterirsek, her iki örnekteki tüm değerlere dayalı olarak elde edilen θ'nın tahmini olsun ve $\hat{\theta}_{(i)}^1 = g\left(\hat{\theta}_{(i)}^1, \hat{\theta}^2\right)$; $i = 1, 2, \ldots, n_1$ ve $\hat{\theta}_{(j)}^2 = g\left(\hat{\theta}^1, \hat{\theta}_{(j)}^2\right)$; $i = 1, 2, \ldots, n_2$ ile ifade edilsin. İki örneğe ait $\tilde{\theta}_{(i)}^1 = n_1\hat{\theta} - (n_1 - 1)\hat{\theta}_{(i)}^1$ ve $\tilde{\theta}_{(j)}^2 = n_2\hat{\theta} - (n_2 - 1)\hat{\theta}_{(j)}^2$ sözde değerleini kullanarak, θ'nın iki örneğe dayalı yanlılığı azaltılmış çakı tahmini aşağıdaki eşitlikle ifade edilir:

$$\tilde{\theta} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}\tilde{\theta}_{(i)}^1 + \sum_{j=1}^{n_2}\hat{\theta}_{(j)}^2\right] \qquad (5.6)$$

Öte yandan, her iki jackknife varyansının toplamı ikili örnek için çakı varyansı olarak kullanılır. Söz konusu varyansın karekökü iki örnek için çakı standart sapmasını

$$\widehat{var}_{çakı}(\tilde{\theta}) = \frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\left[\tilde{\theta}_{(i)}^1 - \frac{1}{n_1}\sum_{k=1}^{n_1}\tilde{\theta}_{(k)}^1\right]^2 + \frac{1}{n_2(n_2-1)}\sum_{i=1}^{n_1}\left[\tilde{\theta}_{(j)}^2 - \frac{1}{n_2}\sum_{l=1}^{n_2}\tilde{\theta}_{(l)}^2\right]^2 \qquad (5.7)$$

verilir (Friedl ve Stampfer, 2001).

5.1. Deleted-d Çakı

Deleted-d çakı yönteminin az önce ifade edilen delete-one çakı yönteminden farkı, her defasında bir gözlemi dışarıda bırakmak yerine işlemlerin her defasında **d** tane gözlemi dışarıda bırakarak gerçekleştirilmesidir (Wu, 1990) Detaylı olarak çakı yönteminin asimptotik özelliklerini görmek için Wu (1990), çalışması incelenebilir. $\frac{\sqrt{n}}{d} \to 0$ ve $n - d \to \infty$ olduğu durumda, deleted-d çakı medyanı için oldukça uygun ve tutarlı sonuçlar verir (Efron ve Tibshranni, 1993). Daha açık bir ifade ile uygulamayı yapan kişi $d = \sqrt{n}$'den daha fazla gözlemi dışarıda bırakmak zorunda kaldığında daha temsili bir standart hatanın çakı tahmini n'den daha az gözlem yapmak zorundadır.

s_* , $(1,2, \ldots, n)$ tane eleman içinden d elemanlı bir alt küme olsun ve $\hat{\theta}(s_*) = \hat{\theta}(x_{s_*})$, x orijinal veri kümesinden $\{x_i, i \epsilon s\}$ dışarıda bırakılmak suretiyle geriye kalan $n - d$ tane veri noktasına dayanarak bulunan istatistiği göstersin. Toplamda $N = \binom{n}{d}$ tane farklı alt küme mevcuttur ve $s_j, j = 1, 2, \ldots, N$ ile gösterilsin. Standart hatanın deleted-d çakı tahmini,

$$\widehat{se}_{çakı-d}(\hat{\theta}) = \left[\frac{n-d}{dN}\sum_{j=1}^{N}\left(\hat{\theta}_{(s_j)} - \frac{1}{N}\sum_{k=1}^{N}\hat{\theta}_{(s_k)}\right)\left(\hat{\theta}_{(s_j)} - \frac{1}{N}\sum_{k=1}^{N}\hat{\theta}_{(s_k)}\right)\right]^{0.5} \quad \mathbf{(5.8)}$$

şeklinde ifade edilir. Bu ifadeye göre $d = 1$ için deleted-d çakı, deleted-bir çakı'ya eşittir. Deleted-d çakı yönteminin en önemli avantajı, $\hat{\theta}$' nın örnek dağılımının uygun bir tahminin yorumlanmasında yardımcı olmasıdır. $\Pr\left(\sqrt{n}\left(\hat{\theta} - \theta\right) \leq y\right)$, aynı zamanda kümülatif çakı histogramı olarak adlandırılan histogram yardımıyla da tahmin edilebilir;

$$Pr_*\left\{\sqrt{\frac{n(n-d)}{d}}\left(\hat{\theta}_{s_*} - \hat{\theta}\right) \leq y\right\} = \frac{1}{N}\sum_{j=1}^{N} I\left\{\sqrt{\frac{n(n-d)}{d}}\left(\hat{\theta}_{s_*} - \hat{\theta}\right) \leq y\right\} \quad \mathbf{(5.9)}$$

Burada Pr_* , $x_{(s)}$'i ifade eder ve x'den iadesiz olarak alınan $(n - d)$ büyüklüğündeki basit tesadüfi örnek olarak düşünülür. I göstergesi ise 1 değerine sahip bir çarpandır. $\hat{\theta}$'nın deleted-d çakı standart hata tahmini, çakı histogramındaki varyansa karşılık gelir ve

$$\widehat{se}_*\left(\sqrt{\frac{n-d}{n}}\left(\hat{\theta}_{(s)}-\hat{\theta}\right)\right)=\left[\frac{n-d}{dN}\sum_{j=1}^{N}\left(\hat{\theta}_{(s_j)}-\hat{\theta}-\frac{1}{N}\sum_{k=1}^{N}(\hat{\theta}-\hat{\theta}_{(s_k)})\right)^2\right]^{0.5} \quad \mathbf{(5.10)}$$

şeklinde ifade edilir (Yay, 2003).

Çakı yönteminin yetersiz kaldığı durumları kısaca açıklayalım. Yeniden örnekleme yöntemi olan çakı, bootstrap yöntemine, hem basit hem de oldukça iyi bir yaklaşım sunmaktadır. Fakat çakı, $\hat{\theta}$ istatistiği düzgün olmadığı taktirde başarısız sonuçlar vermektedir. Buradaki düzgün kelimesinin anlamı, veri kümesindeki küçük değişikliklerin istatistiğin hesabında önemsiz değişmelere yol açmasıdır. Düzgün olmayan bir istatistiğe en iyi örnek medyandır. Düzgünlüğün olmaması medyan için hesaplanacak standart hatanın çakı tahmininde uygun olmayan durumlara yol açacaktır (Yay, 2003).

Ayrıca bağımsız değişkenlerin kitleye genellenebilme durumları çakı yöntemi ile incelenebilir. Bununla ilgili literatürde çeşitli çalışmalar mevcuttur (bknz. Kayri ve Büyüköztürk, 2009; Bekiroğlu vd., 2013. Zaman ve Alakuş, 2015).

Zhang vd. (2016a), çakı yöntemi ile küme örneklemesinde ortalama tahmin yöntemi önermişlerdir. Zhang vd. (2016b) ise yüksek boyutlu regresyon katsayı tahminleri için çakı yönteminde olabilirlik testi ile ilgili incelemelerde bulunmuşlardır.

Lineer regresyonda regresyon katsayılarının örnekleme dağılımını tahmin etmek için çakı yöntemlerinin kullanılması Efron (1979) tarafından ilk kez önerildi ve Freedman (1981) ve Wu (1986) tarafından geliştirilmiştir. Bu bağlamda çoklu doğrusal regresyon analizinde çakı yönteminin nasıl kullanılacağı ile ilgili bölüm 5.1.1'de örnek verilmiştir.

5.1.1 Çoklu doğrusal regresyon analizinde çakı yöntemi için bir uygulama

Bu alt başlıkta çakı yönteminin işleyişi hakkında bir örnek verilecektir.

Çalışmada, Ondokuz Mayısı Üniversitesi Araştırma Görevlilerinin tükenmişlik düzeylerini Maslach Tükenmişlik ölçeği kullanarak çakı yöntemine göre incelenmiştir. Bu amaçla Ondokuz Mayıs Üniversitesindeki görev yapan Araştırma Görevlilerine bir anket çalışması uygulanmıştır. Örneklem sayısının az tutulması özellikle istenmiştir. Çünkü çakı yöntemi sayıca az olan örneklemlerden hesaplanan istatistiklerden elde edilen parametre değerlerinin kitleye genellenebilme durumlarını test eden istatistiksel yöntemlerden biridir. Araştırma Görevlilerinin Maslach

Tükenmişlik Ölçeğinde bulunan toplam 22 sorudan 9 madde duygusal tükenme ile, 5 madde duyarsızlaşma ile ve geriye kalan 8 madde de kişisel başarıda tükenme durumuyla ilgilidir. Bu maddeler 0: Hiçbir Zaman, … , Her Zaman şeklinde 7 farklı şekilde derecelendirilmiştir (Sucuoğlu ve Kuloğlu, 1996). Bu sorulara verilen cevapların toplanmasıyla elde edilen tükenmişlik puanı kurulacak olan regresyon modelindeki bağımlı değişken değeri olarak ifade edilir. Bu bağımlı değişken değerini açıklamaya çalışan bağımsız değişkenler ise; Yaş (x_1), Haftalık girilen ders saati (x_2), Aylık ortalama kredi kartı ekstreniz (x_3), yayınlanan toplam makale ve bildiri sayınız (x_4), Cinsiyet (x_5), Medeni durum (x_6), Çocuk sayısı (x_7), Bölüm (x_8) olmak üzere, bu değişkenlere ait değerler yapılan anket çalışması sonucunda elde edilmiştir. Burada Cinsiyet, Medeni durum, Çocuk sayısı, Bölüm değişkenleri kukla değişken olarak alınmıştır. Şöyle ki Cinsiyet: 1 (bay), 0 (bayan); Medeni durum: 1 (evli), 0 (bekar); Çocuk sayısı: 1 (var), 0 (yok); Bölüm: 1 (istatistik), 0 (diğer) şeklinde alınmıştır. Dolayısıyla kurulmak istenen regresyon modeli;

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon_i \quad \textbf{(5.11)}$$

şeklinde ifade edilir.

Anket çalışması sonucunda elde edilen veri seti Çizelge 5.1'de verilmiştir.

Çizelge 5.1. Veri seti ($p = 8$)

Sıra No	Y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	71	26	3	317	2	1	0	0	0
2	31	32	5	600	4	1	0	0	0
3	27	27	6	500	2	0	0	0	1
4	20	27	9	1500	1	1	0	0	1
5	27	28	10	2000	13	1	1	0	1
6	43	26	10	750	8	1	0	0	1
7	45	26	6	1200	6	1	0	0	1
8	45	29	5	2500	10	1	1	1	1
9	52	27	7	2000	15	1	1	1	1
10	47	36	7	750	2	0	1	1	1
11	28	33	14	350	6	1	1	1	0

Burada kullanılan değişkenler aşağıdaki gibi ifade edilir.

Y =Tükenmişlik

x_1 = Yaş

x_2 = Girilen ders saati

x_3 = Aylık ortalama kredi kartı ekstreniz (TL)

x_4 = Yayınlanan makale sayısı

x_5 = Cinsiyet

x_6 = Medeni durumu

x_7 = Çocuk sayısı

x_8 = Bölüm

Bu veri setine ait çoklu doğrusal regresyon analizleri R paket programı kullanılarak çözümlenmiş ve model için anlamlı olan bağımsız değişkenlere geriye doğru eleme yöntemi ile karar verilmiştir. Sonuçlar çizelge 5.2'de özetlendiği gibi elde edilmiştir.

Çizelge 5.2. Çoklu doğrusal regresyon sonuçları

Değişken	Standardize Olmayan Katsayılar		Standardize Katsayılar		
	B	Std. Hata	Beta	t-testi	Önem Olasılığı
Sabit	43,632	7,141		6,110	,002
x_2	-7,134	,907	-1,489	-7,865	,001
x_3	-,044	,008	-2,259	-5,794	,002
x_5	56,505	9,866	1,550	5,727	,002
x_6	42,111	6,933	1,491	6,074	,002
x_8	46,099	9,587	1,460	4,809	,005
R^2: 0.927					
$F = 12.779$ olup; p= 0.007'dir					

Çoklu doğrusal regresyon modeli için önemli olan bağımsız değişkenlere ait sonuçlar çizelge 5.2' de verilmiştir. Çizelge 5.2 incelendiğinde model için X_2, X_3, X_5, X_6 ve X_8 bağımsız değişkenleri istatistiksel olarak önemli olduğu gözlenmiştir. $(p < 0.05)$. Bu değişkenler yardımıyla kurulan çoklu doğrusal regresyon modeli de istatistiksel olarak anlamlı bulunmuştur $(p = 0.007 < 0.05)$. Model için $R^2 = 0.927$ olarak hesaplanmıştır. Yani modelde anlamlı olarak bulunan bağımsız değişkenlerin (X_2, X_3, X_5, X_6 ve X_8) modeli açıklama oranı 0.927 ile gayet iyi olarak elde edilmiştir.

Çizelge 5.3'te, çakı yöntemi ile sırayla birer gözlem dışarıda bırakılarak uygulanan regresyon analizi sonucu elde edilen standardize beta katsayıları gözlenmektedir.

Çizelge 5.3. Çakı yöntemi uygulanarak elde edilen beta katsayıları

Çıkarılan Gözlem	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_8$	R^2
Hiç Çıkartılmamış	-1.489	-2.259	1.550	1.491	1.460	0.927
1. GözlemiÇıkart	-1.499	-2.525	1.831	1.790	1.665	0.949
2. Gözlemi Çıkart	-1.415	-1.924	1.377	1.249	0.954	0.974
3. Gözlemi Çıkart	-1.591	-2.395	1.330	1.604	1.583	0.927
4. Gözlemi Çıkart	-1.780	2.752	1.841	1.803	1.731	0.927
5. Gözlemi Çıkart	-1.489	-2.139	1.585	1.455	1.477	0.925
6. Gözlemi Çıkart	-1.430	-2.158	1.483	1.416	1.369	0.928
7. Gözlemi Çıkart	-1.568	-2.471	1.707	1.487	1.636	0.951
8. Gözlemi Çıkart	-1.433	-1.893	1.568	1.410	1.482	0.930
9. Gözlemi Çıkart	-1.530	-2.156	1.571	1.429	1.467	0.923
10. Gözlemi Çıkart	-1.564	-2.389	1.161	1.528	1.539	0.931
11. Gözlemi Çıkart	-1.155	-2.192	1.599	1.421	1.359	0.925

Çizelge 5.3 ‘teki standardize edilmiş beta tahmin değerleri kullanılarak X_2, X_3, X_5, X_6 ve X_8 bağımsız değişkenlere ait sözde değerler eşitlik (5.5) kullanılarak, çizelge 5.4 ‘ deki şekilde elde edilmiştir.

Çizelge 5.4. Sözde değerler

Çıkarılan Gözlem	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_8$	R^2
Hiç Çıkartılmamış	-1.489	-2.259	1.550	1.491	1.460	0.927
1. GözlemiÇıkart	-1.389	0.401	-1.260	-1.499	-0.59	0.707
2. Gözlemi Çıkart	-2.229	-5.609	3.280	3.911	6.520	0.457
3. Gözlemi Çıkart	-0.469	-0.899	3.750	0.361	0.230	0.927
4. Gözlemi Çıkart	1.421	2.671	-1.360	-1.629	-1.25	0.927
5. Gözlemi Çıkart	-1.689	-3.459	1.20	1.851	1.290	0.947
6. Gözlemi Çıkart	-2.079	-3.269	2.220	2.241	2.370	0.917
7. Gözlemi Çıkart	-0.699	-0.139	-0.020	1.531	-0.30	0.687
8. Gözlemi Çıkart	-2.049	-5.919	1.370	2.301	1.240	0.897
9. Gözlemi Çıkart	-1.079	-3.289	1.340	2.111	1.39	0.967
10. Gözlemi Çıkart	-0.739	-0.959	5.440	1.121	0.670	0.887
11. Gözlemi Çıkart	-4.829	-2.929	1.060	2.191	2.470	0.947
Çakı Tek.ile Elde Edilen Ort. Sözde değerler	-1.439	-2.127	1.547	1.317	1.276	0.842
Sözde Standart Hata Ort.	0.459	0.785	0.623	0.503	0.632	0.048
Hesaplanan t değeri	-3.135^*	-2.710^*	2.483^*	2.618^*	2.019	17.54^*
*Tahmin edilen parametre değerlerinin sağlamlığını ve değişmezliğini gösterir.						

Güven aralığının sıfırı kapsamayan durumlarında, tahmin edilen regresyon parametrelerinin sağlamlığı ve değişmezliği görülmektedir. Bu ifadeye göre x_2, x_3, x_5, x_6 bağımsız değişkenlerin parametre tahmin değerlerinin orijinal değerlerle kararlılık gösterdiği saptanmıştır. Bunun anlamı çakı yöntemi ile elde edilen beta katsayılarının sadece bu örnekleme ait olmadığı, aynı zamanda kitleye de genelleştirebilecek tahminler olduğudur.

Çizelge 5.5. Çakı tekniği ile hesaplanmış parametre tahmin değerlerinin %95 güven aralıkları

	x_2	x_3	x_5	x_6	x_8	R^2
Orijinal Katsayı	-1.489^*	-2.259^*	1.550^*	1.491^*	1.460^*	0.927^*
Çakı Tek.ile Elde Edilen Ort. Sözde değerler	-1.439	-2.127	1.547	1.317	1.276	0.842
Sözde Standart Hata Ort.	0.459	0.785	0.623	0.503	0.632	0.048
Alt Sınır % 95 GA	-2.462	-3.876	0.159	0.196	-0.132	0.735
Üst Sınır % 95 GA	-0.416	-0.378	2.935	2.437	2.684	0.948
*Orijinal tahmin değerlerinin, çakı tekniği ile elde edilmiş olan % 95 güven aralıkları içinde kaldığını göstermektedir.						

Çizelge 5.5 incelendiğinde, çakı yöntemi ile elde edilen haftada ortalama girilen ders saati tahmin değerinin $\hat{\beta}_{2çakı} = -1.439$, hesaplanan ilk orijinal değeri olan $\hat{\beta}_{2orj} = -1.489$ değerini doğruladığı ve çakı parametre değerinin % 95 güven aralığı içerisinde yer aldığı görülmüştür. Bunun anlamı modelde bağımsız değişken olarak bulunan haftada ortalama girilen ders saati değişkeninin sadece bu örnekte geçerli olmadığı, aynı zamanda kitleye genellenebilme özelliğine sahip olduğudur. Diğer açıklayıcı değişkenlerden aylık ortalama kredi kartı ekstreniz, cinsiyet, medeni durum değişkenleri ile ilgili olarak da çakı tahmin değerleri, hesaplanan ilk orijinal $\hat{\beta}$ değerlerini doğruladığı çizelge 5.5‘ den ve aynı şekilde aylık ortalama kredi kartı ekstreniz, cinsiyet, medeni durum bağımsız değişkenlerine ait çakı parametre değerlerinin de % 95 güven aralığı içerisinde yer aldığı çizelge 5.5‘ den görülmekte olup, bu değişkenlerin de kitleye genellenebilme özelliğine sahip oldukları söylenebilir.

Çizelge 5. 5’e göre modelde, tükenmişlik bağımlı değişkeni üzerinde etkisi incelenen bölüm bağımsız değişkeninin önemli olmadığı görülmektedir. Bu durumda bölüm değişkeninin bir etkisinden bahsedilecek olursa bu etkinin örneğe özel olduğu şeklinde düşünülebilir.

Son olarak çakı yöntemi ile hesaplanan $R^2 = 0.842$ ve standart hata değeri de 0.048 olarak elde edilmiştir. Çizelge 5.5’e göre R^2’nin de genelleştirilebileceği hesaplanan güven aralığı yardımıyla anlaşılmaktadır. Yani modelin belirtme katsayısı örneğe özel olmayıp, kitleye genellenebilme özelliğine sahiptir. Yani farklı zamanda yapılacak benzer çalışmalarda da bağımsız değişkenlerle bağımlı değişkeni açıklama yüzdesi bu çalışmadan elde edilen değer ile aynı ya da bu değere yakın olabileceği söylenebilir. Benzer şekilde elde edilen x_2, x_3, x_5, x_6 bağımsız değişkenlerinde kitleye genellenebilme özelliğinden dolayı farklı çalışmalarda da istatistiksel olarak anlamlı bir değişken olarak bulunabileceği söylenebilir.

Sonuç olarak, örneklem genişliği küçük olan veri seti için çoklu doğrusal regresyon yönteminin kullanıldığı kitlede, gerçek değere yakın, güçlü ve sağlam tahminlerin elde edilmesi çakı yöntemi ile incelenmiştir. Uygulama için Maslach tükenmişlik ölçeği kullanılmıştır. Bu tükenmişlik ölçeği yardımıyla çoklu doğrusal regresyon modeli kurularak, model için önemli olan bağımsız değişkenlerin çakı yöntemi ile parametre tahminlerinin değişmezlik ve sağlamlık özelliklerinin yanı sıra bu tahminlere ait güven aralıkları da oluşturulmuş olup, orijinal beta tahmin değerlerinin bu güven aralıklarının içinde olduğu sonucuna varılmıştır. Bu amaçla çakı yönteminin uygulanması amacına hizmet etmesi açısından örneklem genişliği 11 alınarak, özellikle sayıca az tutulmuştur. Bu sebeple, yeni bir örneklemle çalışmayı tekrar etmek yerine çakı tekniği yardımı ile elde edilen sonuçların, değişmezliğinin, sağlamlığının ve araştırma sonuçlarına olan güvenin güçlü olduğu sonucu çıkarılabilir. Ayrıca bağımlı değişken üzerinde etkisi bulunan 5 bağımsız değişken için çakı yöntemi yardımıyla standardize beta katsayıları, belirtme katsayısı ve bu değerlere ait % 95 güven aralıkları hesaplanmıştır. Bu bağımsız değişkenlerden haftada girilen ders saati, aylık ortalama kredi kartı borcu, cinsiyet ve medeni durum bağımsız değişkenleri çakı parametre tahminleyicisinin güven aralığı içinde yer aldığı gözlenmiş olup, bu değişkenlerin kitleye genellenebileceği sonucuna ulaşılmıştır. Ancak bölüm bağımsız değişkeninin sadece bu örneğe ait özellikte olduğu sonucuna varılır (Zaman ve Alakuş, 2015).

5.2. Önerilen Çakı-Theil Tahmin Yöntemi (ÇÇTSE) ile Parametre Tahmini

Çoklu doğrusal regresyon analizinde çok değişkenli medyan yaklaşımı kullanılarak model parametrelerinin tahmini yeni bir problem olmamasına karşın çoklu doğrusal regresyon analizinde model parametrelerinin önerilen Çakı-Theil tahmin yöntemiyle elde edilmesi literatürde hiç çalışılmamıştır. Önerilen Çakı-Theil tahmin yöntemi literatürde mevcut olan Theil-Sen tahminine çakı yönteminin kombinasyonu ile elde edilen bir yöntem olup, güven aralığını daraltıp, aşırı uç değerlerin etkisini azaltan bir tekniktir. Önerilen Çakı-Theil tahmin yönteminin analizinde örneklem verisi yerine herbir gözlem değerleri sırasıyla örneklemden atılarak elde edilmiş olan alt örneklemler kullanır. Önerilen Çakı-Theil tahmin yöntemi için yapılacak işlemler algoritması aşağıdaki gibi olacaktır;

1.Adım: Öncelikle veri setinde bulunan n tane gözlemi sırasıyla her biri veriden atılarak $(n-1)$ büyüklüğünde n tane farklı alt örneklem oluşturulur.

2.Adım: m keyfi sayısı, $p+1 \leq m < n$ olacak şekilde belirlenir. Burada p bağımsız değişken sayısını ve n örnek hacmini gösterir.

3.Adım: β_i, $(i = 1, 2, \ldots, p)$ parametre tahmin değerlerini hesaplamak için belirlenen m keyfi değerine göre tüm olası $\binom{n-1}{m}$'li kombinasyonuna, EKK yöntemi uygulanarak regresyon katsayıları hesaplanır. Elde edilen bu regresyon katsayıları L_{ij} ile gösterilecek olursa, regresyon parametre tahminleri L_{ij} değerlerinin çok değişkenli medyanı alınarak bulunur. Burada kullanılan çok değişkenli medyan spatial medyandır. Yani $\beta_{i\,(-j)} = Çmed(L_{ij})$ $(i = 1, 2, \ldots, p)$; $(j = 1, 2, \ldots, n)$ şeklindedir. Böylece, her bir alt örneklem için bu işlem tekrarlanacağından elimizde n tane $\hat{\beta}_{i(-j)}$ değeri hesaplanacaktır.

4.Adım: $\hat{\beta}_{0(-j)}$ tahmininin hesaplanabilmesi için farklı alternatifler mevcuttur. Bunlar;

$i)$ Hata terimi yaklaşık olarak sıfır etrafında simetrik olan bir dağılıma sahipse, tüm olası $\left(y_i - \hat{\beta}_{i(-j)} x_i\right)$ değerleri hesaplanarak medyanı alındığında $\hat{\beta}_{0(-j)}$ tahmini hesaplanır. Yani, $\hat{\beta}_{0(-j)} = med\left(y_i - \hat{\beta}_{i(-j)} x_i\right)$ şeklinde bulunur.

$ii)$ Aynı şekilde tüm olası $\left(y_i - \hat{\beta}_{i(-j)} x_i\right)$ değerleri hesaplanarak ortalaması alınarakta $\hat{\beta}_{0(-j)}$ tahmini hesaplanır. Yani, $\hat{\beta}_{0(-j)} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_{i(-j)} x_i\right)}{n-1}$ 'dir.

$iii)$ Buna alternatif olarak daha az kısıtlayıcı bir durumla, çok değişkenli medyan ile eş zamanlı olarak (β_0, β_i) değerleri tahmin edilebilir. Yani, $\left(\hat{\beta}_0, \hat{\beta}_i\right) = Çmed\left(\hat{\beta}_{0(-j)}, \beta_{i\,(-j)}\right)$ şeklinde hesaplanabilir.

5.Adım: Her alt örneklem için hesaplanan bu katsayıların ortalamaları önerilen Çakı-Theil tahminleri olur. Şöyle ki, $\hat{\beta}_0^{ÇT} = \frac{\sum_{j=1}^{n} \hat{\beta}_{0(-j)}}{n}$ ve $\hat{\beta}_i^{ÇT} = \frac{\sum_{j=1}^{n} \hat{\beta}_{i(-j)}}{n}$ olacaktır.

ÇTSE ve önerilen Çakı Theil yöntemleri ile tahmin yaparken dikkat edilmesi gereken bazı hususlar vardır. Örneğin ele alınan bağımsız değişkenler kategorikse m keyfi sayısına göre seçilen alt örneklemler sıfır değerli olabilir. Bu durumda EKK yöntemi ile parametre tahmin değerleri (L_{ij}) $\pm\infty$ gibi değerler alabilir. Bu durumda yapılacak analizlerde belirsizlik söz konusu olur. m keyfi sayısına bağlı hesaplanan

L_{ij} değerleri analiz dışı bırakılır ise veri kaybı olacağından analize olan güven azalacaktır. Basit doğrusal regresyon modellerinde bu konudaki problemin çözümü hakkında Erilli ve Alakuş (2014 ve 2016)' da öneriler sunmuşlardır. Bu önerilerin Çakı-Theil tahminine uyarlanmasında izlenilmesi gereken bazı yollar şu şekilde önerilebilir;

1. Sonsuz çıkan (L_{ij}) değerindeki regresyon katsayılarının yerine, m keyfi sayısına bağlı olarak hesaplanan tüm olası L_{ij} değerleri içindeki en büyük ya da en küçük değeri yazılarak (yani $+\infty$ değeri yerine max (L_{ij}), $-\infty$ yerine min (L_{ij}) yazılarak) $\hat{\beta}_i$ hesaplanır.

2. Sonsuz çıkan (L_{ij}) değerindeki regresyon katsayılarının yerine sonsuz çıkan değerler veriden çıkartıldıktan sonra L_{ij} değerleri içindeki regresyon katsayılarının yerine verinin medyanı yazılabilir.

3. Sonsuz çıkan (L_{ij}) değerindeki regresyon katsayılarının yerine kırpılmış ortalama değerleri yazılabilir.

Bu yöntemler kullanılarak $\hat{\beta}_i$ tahminleri hesaplanıp yukarıda belirtilen algoritma dikkate alınarak çok değişkenli medyan yardımıyla önerilen Çakı-Theil tahmin yöntemi için β_i, $(i = 1, 2, \dots, p)$ değerleri hesaplanabilir. Hangi yöntemin en iyi sonuç vereceği hakkında kesin bişey söylenememektedir. Araştırmacının eldeki veri yapısına göre en iyi sonucu elde edebilmesi için olası tüm yöntemlere ait tahmin sonuçlarını bulup karşılaştırması tavsiye edilir.

Tez kapsamında yapılan simülasyon ve gerçek veri çalışmalarında önerilen Çakı-Theil tahmin yöntemine ait regresyon katsayılarının tahmininde $(\beta_0; \beta_i, (i = 1, 2, \dots, p$ eş zamanlı olarak sonuçlar elde edilip, yorumlanmıştır.

Önerilen Çakı-Theil tahmin yöntemini aşağıdaki çizelge ile kısaca özetleyecek olursak,

Çizelge 5.6'da ele alınan 20 gözlem için her defasında bir gözlem silinerek veriye ait üçlü kombinasyonlar ele alınıp regresyon parametreleri EKK yöntemi ile tahmin edilir. Her $\binom{19}{3}$ kombinasyonundan sonra tahmin edilen regresyon katsayılarına ait çok değişkenli medyan hesaplanıp, sonunda çok değişkenli medyanların ortalaması alınarak önerilen Çakı-Theil tahmin yöntemine ait regresyon katsayıları hesaplanır.

Çizelge 5.6. Parametrelerin önerilen Çakı –Theil tahmin (ÇÇTSE) yöntemiyle elde edilmesi

	No	$\boldsymbol{Y}$	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$		
1. Gözlem Silindi	2	Y_2	X_{21}	X_{22}			
	3	Y_3	X_{31}	X_{32}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	4	Y_4	X_{41}	X_{42}			
	2	Y_2	X_{21}	X_{22}			
	3	Y_3	X_{31}	X_{32}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	5	Y_5	X_{51}	X_{52}			
	…	…	…	…	…	…	…
	18	Y_{18}	X_{181}	X_{182}			
	19	Y_{19}	X_{191}	X_{192}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	20	Y_{20}	X_{201}	X_{202}			
…	…	…	…	…	…	…	…
20. Gözlem Silindi	1	Y_1	X_{11}	X_{12}			
	2	Y_2	X_{21}	X_{22}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	3	Y_3	X_{31}	X_{32}			
	1	Y_1	X_{11}	X_{12}			
	2	Y_2	X_{21}	X_{22}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	4	Y_4	X_{41}	X_{42}			
	…	…	…	…	…	…	…
	17	Y_{17}	X_{171}	X_{172}			
	18	Y_{18}	X_{181}	X_{182}	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
	19	Y_{19}	X_{191}	X_{192}			
	Toplam $20 * \binom{19}{3} = 19380$ tane kombinasyon mevcut.				$Çmed(L_{i,j})$		

6. ÖNERİLEN YÖNTEMİN ETKİNLİĞİ

Uygulama kapsamında, simülasyon çalışması ve gerçek veriler yardımıyla tez kapsamında incelenen tahmin yöntemlerinin birbirlerine göre durumları değerlendirilip, önerilen yöntemin üstün olduğu durumlar araştırılmıştır. Simülasyon çalışmasında veriler EKK varsayımlarını sağladığı durumda, veride çeşitli oranlarda kirlilik mevcut olduğunda, hata terimi dağılımı normal, 3 serbestlik dereceli t tağılımı ve 1 serbestlik dereceli t (Cauchy) dağılımından örnekler üretildiği durumda ve ayrıca hata terimi dağılımı kesikli düzgün dağılımdan örnekler üretilip tahmin yöntemlerine ait regresyon katsayıları ve deneysel hata kareler ortalamaları hesaplanarak sonuçlar incelenmiştir. Önerilen Çakı-Theil tahmin yöntemi ve incelenen diğer tahmin yöntemlerinin karşılaştırılması için kullanılan gerçek veri, coleman verisi ve eğitim giderleri verisi olarak araştırılmıştır. Ayrıca Beetle marka aracın Türkiye piyasasındaki değerini etkileyen faktörlerin belirlenebilmesi için bir veri seti göz önüne alınarak, incelenen yöntemler değerlendirilmiştir.

6.1. Simülasyon Çalışması

Bu bölümde; EKK, ÇTSE ve ÇÇTSE ‘nin davranışının sağlamlığını incelemek için , bir simülasyon çalışması yapılmıştır.

Çizelge 6.1. Sağlamlık Sonuçları

	Gerçek parametreler (2.5, 3, 1.5)		
	EKK	ÇTSE	ÇÇTSE
$n = 20$	(2.480 3.001 1.521)	(2.488 3.043 1.523)	(2.469 3.001 1.509)
$n = 30$	(2.487 3.004 1.530)	(2.466 3.023 1.494)	(2.488 3.017 1.501)
$n = 40$	(1.894 2.242 1.084)	(2.792 2.544 0.417)	(2.226 2.202 0.875)
n=20; kirlilik oranı=%1	(2.264 1.862 -0.436)	(2.420 2.864 1.550)	(2.462 2.961 1.482)
n=20; kirlilik oranı=%5	(1.532 1.795 0.432)	(4.086 0.896 0.677)	(1.811 3.224 2.814)
n=20; kirlilikoranı=%10	(0.855 0.949 -0.219)	(2.180 3.144 1.370)	(2.563 3.001 1.221)
n=20; kirlilikoranı=%20	(-0.732 -2.403 -3.249)	(1.210 2.363 -0.849)	(0.826 1.183 1.178)
n=20; kirlilikoranı=%30	(-0.935 -3.601 -6.696)	(0.912 0.480 -2.213)	(1.633 1.667 -1.413)
n=20; kirlilikoranı=%40	(-0.348 -0.736 -0.860)	(1.644 -0.225 0.369)	(-0.187 -0.188 -0.394)

Bu simülasyon çalışması için $Y_i = 2.5 + 3X_{1i} + 1.5X_{2i} + \varepsilon_i$ çoklu doğrusal regresyon modelinden örnekler üretilmiştir. Burada $X_{1i} \sim N(0,1)$, $X_{2i} \sim U(0,1)$ ve hatalar ise farklı amaçlar için farklı dağılımlardan alınmıştır.

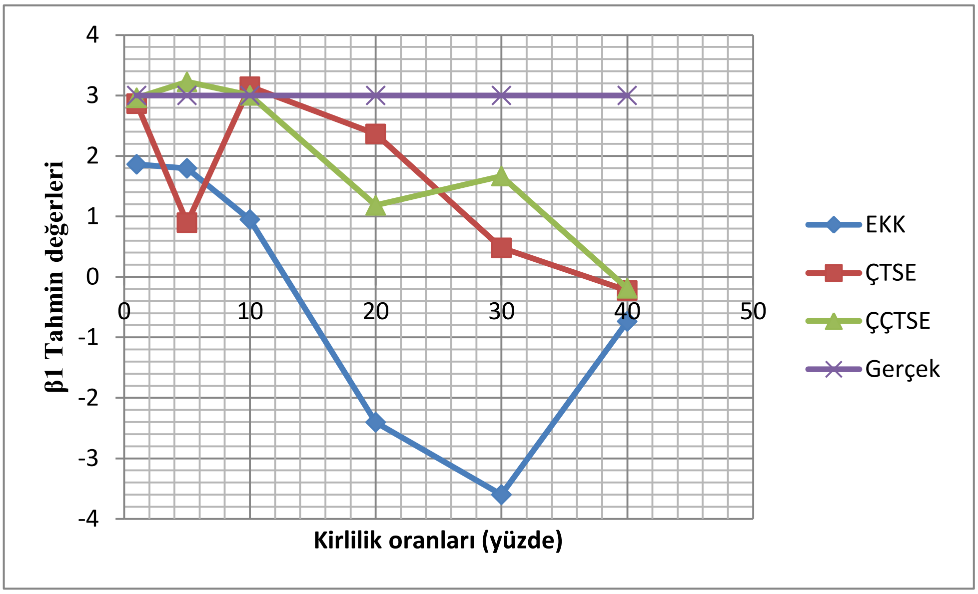
Sağlamlık sonuçları üzerine simülasyon: Hata dağılımı $\varepsilon_i \sim N(0, 0.5)$ olan çoklu doğusal regresyon modelinden örneklem hacmi $n = 20, 30, 40$ olan örnekler üretilmiştir. EKK, ÇTSE ve ÇÇTSE' ye ait tahmin edilen beta katsayıları Çizelge 6.1. 'deki gibi hesaplanmıştır. Çizelge 6.1 incelendiğinde, ilk üç satırda $n = 20, 30, 40$ için ele alınan tahmin yöntemleri için $Y_i = 2.5 + 3X_{1i} + 1.5X_{2i} + \varepsilon_i$ çoklu doğrusal regresyon modelinden örnekler üretilerek söz konusu tahmin yöntemlerine ait regresyon katsayıları tahmin edilmiştir. İkinci kısımda ise örnek hacmi $n = 20$ alınarak ve modele çeşitli oranlarda kirli veri katılarak tahmin edilen regresyon katsayılarına ait sonuçlar gözlenmiştir. Veriyi kirletmek için kullanılan çoklu doğrusal regresyon modeli $Y_i = -10 - 20X_{1i} - 25X_{2i} + \varepsilon_i$ 'dir. Burada hata dağılımı $\varepsilon_i \sim N(0, 0.5)$ olarak alınmıştır. $n = 20$ birimlik örnek hacmini $\%1, \% 5, \% 10, \% 20, \% 30$ ve $\% 40$'ı, $Y_i = -10 - 20X_{1i} - 25X_{2i} + \varepsilon_i$ modeli ile kirletilmiştir.

Önce veride herhangi bir kirlenme söz konusu değilken elde edilen tahmin sonuçlarını inceleyelim. Kirletilmemiş gözlemlerde, EKK, ÇTSE ve ÇÇTSE yöntemleri kullanılarak hesaplanan katsayı tahmin değerleri gerçek betalar olan (2.5; 3; 1.5) değerlerine oldukça yakın değerler almıştır. Yani incelenen tahmin yöntemleri veride aykırı değer yokken ve çoklu doğrusal regresyon model varsayımları sağlanıyorken oldukça iyi sonuçlar vermiştir. Ancak var olan veriye çeşitli oranlarda aykırı değerler eklendiğinde EKK yöntemi tamamen bozularak, kullanışsız hale geldiği gözlenmiştir . ÇTSE tahmini $\% 20$ aykırı değere kadar gerçek regresyon katsayılarına yakın değerler aldığı gözlenmiştir. Yani bu simülasyon çalışması için ÇTSE, veride yaklaşık % 20 aykırı değer oranından sonra sağlamlığını kaybettiği gözlenmiştir. Yeniden örnekleme yöntemi olan çakı yöntemi ile Çoklu Theil-Sen tahmininin hibritlenmesiyle önerilen ÇÇTSE ile gösterilen Çakı-Theil tahmini ise hemen hemen % 30 aykırı değer oranına kadar sağlamlığını koruduğu ve bu simülasyon çalışması için % 30 aykırılıktan sonra etkinliğini kaybettiği görülmüştür.

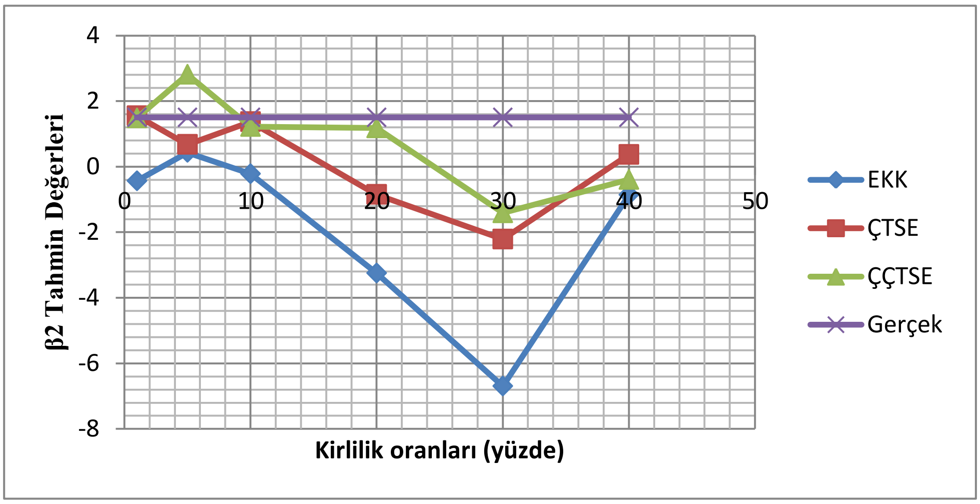
Çizelge 6.1 incelendiğinde ayrıca şu yorum yapılabilir, veride oranı ne olursa olsun kirlilik varken klasik tahmin olan EKK yöntemi gerçek regresyon katsayılarından uzakta değerler almıştır. Kirlenme oranı arttıkça ÇÇTSE ile tahmin edilen regresyon katsayıları, ÇTSE' ye göre hesaplanan regresyon katsayılarına göre daha iyi sonuçlar verdiği gözlenmiştir. Yani literatürde yer alan ÇTSE tahminine yeniden örnekleme yöntemi olan çakı yönteminin hibritlenmesiyle elde edilen regresyon katsayıları bu similasyon çalışması sonucunda literatüre katkı niteliği taşımaktadır.

İlaveten bir başka yeniden örnekleme yöntemi olan bootstrap yöntemi dikkate alınarak sonuçlar yeniden tartışılabilir.

Çizelge 6.1 dikkate alınarak örnek hacmi 20 iken çeşitli kirlilik oranlarında incelenen üç yöntem için izleme sinyalleri aşağıdaki şekilde verilebilir.



Şekil 6.1. İncelenen yöntemlere ait çeşitli kirlilik oranlarına göre β_1 tahmin değerine ait izleme sinyali.



Şekil 6.2. İncelenen yöntemlere ait çeşitli kirlilik oranlarına göre β_2 tahmin değerine ait izleme sinyali.

Şekil 6.1 incelendiğinde, çeşitli kirlilik oranlarında incelenen yöntemlerin gerçek beta değerlerine göre durumları görülmektedir. Genel olarak değerlendirildğinde önerilen Çakı-Theil (ÇÇTSE) tahmin yöntemi gerçek beta katsayılarına daha yakın tahminler yaptığı açıktır.

Hesaplama ve Simülasyon 2.

Bu bölümde, ÇTSE ve önerilen ÇÇTSE ‘nin davranışlarını incelemek için, simülasyon çalışması yapılmıştır. Burada $p = 2$ ve $m = p + 1 = 3$ olarak alınmıştır. Örneklemler $Y_i = 2.5 + 3X_{1i} + 1.5X_{2i} + \varepsilon_i$ çoklu doğrusal regresyon modelinden üretilmiştir. Burada $X_{1i} \sim N(0,1)$, $X_{2i} \sim N(0,1)$ ve $\varepsilon_i \sim N(0,1)$ dağılımdan çekilmiştir. Bu model, $Y_i = -10 - 20X_{1i} + 25X_{2i} + \varepsilon_i$, $X_{1i} \sim N(0,1)$, $X_{2i} \sim N(0,1)$ çoklu doğrusal regresyon modelinden aykırı (X_i, Y_i) gözlem değerleri ile kirletilmiştir. Örnek hacmi $n = 10, 20, 25, \ldots, 50$ ‘e kadar ve kirlilik oranları da % 5, %10 ve % 20 olarak belirlenmiş ve 100 simülasyon sonucunda çizelge 6. 2’deki değerler elde edilmiştir.

Çizelge 6.2. Bazı aykırılık oranları durumunda EKK, ÇTSE, ÇÇTSE ait parametre tahminleri için deneysel hata kareler ortalaması

Örnek hacmi	Aykırı değer %								
	5%			10%			20%		
	EKK	ÇTSE	ÇÇTSE	EKK	ÇTSE	ÇÇTSE	EKK	ÇTSE	ÇÇTSE
10	28.80	0.4422	0.5483	32.0605	0.5326	0.5677	66.5179	4.2187	5.2909
20	4.8272	0.1373	0.1381	15.7596	0.2152	0.1484	26.1113	0.7689	0.7158
25	8.145	3.85	1.754	12.0568	0.1455	0.1468	31.2967	1.2035	0.8902
30	4.8467	0.0697	0.0698	10.3119	4.5238	5.8034	27.5454	0.3196	0.3271
35	4.2601	0.0499	0.0499	12.5515	9.9076	0.3733	25.7719	9.1562	9.4418
40	3.2691	0.0509	0.0509	8.734	4.5144	1.8785	26.3822	6.2388	2.8074
45	4.3719	0.0477	0.0478	7.9876	5.4876	4.6048	22.2376	8.6498	7.3089
50	3.8026	0.3682	0.1505	7.1731	0.0631	0.0632	24.4306	10.3123	8.7384
Ort	7.7903	0.6270	0.3512	13.3294	3.1737	1.6983	31.2867	5.1085	4.4401

Çizelge 6.2’ deki değerler; 100 simülasyon sonucunda tahmin edilen regresyon katsayılarına ait deneysel hata kareler ortalamalarını göstermektedir. Regresyon katsayılarına ait deneysel hata kareler ortalaması denklem (6.1)’deki formül yardımıyla hesaplanmıştır.

$$DHKO = \frac{\left\| \sum_{i=1}^{k} \hat{\beta}_i - \beta \right\|}{k} \qquad (\mathbf{6.1})$$

Burada $k = 100$, simülasyon sayısını göstermektedir. $\beta = (2.5, 3, 1.5)$ kirletilmemiş çoklu doğrusal regresyon modeline ait olan gerçek regresyon katsayı değerleridir. $\hat{\beta}_i$ ise incelenen yönteme ait i. örneklem için tahmindir.

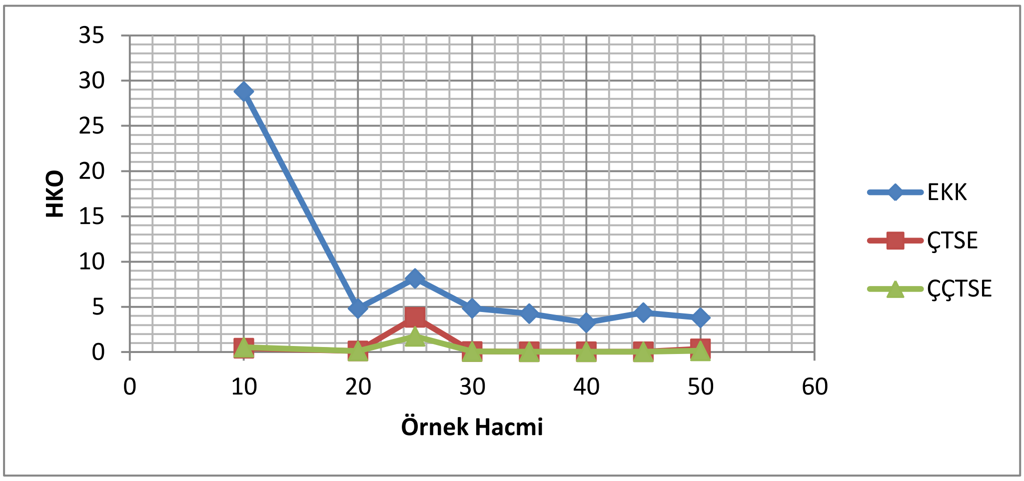
Çizelge 6.2 incelendiğinde hem aykırı değer oranına göre hem de örnek hacminin tüm durumları dikkate alındığında EKK yöntemi incelenen diğer yöntemlere göre kötü sonuçlar vermiştir. Çoklu Theil-Sen tahmini ile önerilen Çakı-Theil tahmin yöntemleri karşılaştırıldığında ise birbirlerine çok yakın sonuçlar elde edilmekle birlikte, kirlilik oranı ve örnek hacmine bağlı olarak üstünlükleri tartışılabilir. Yani aralarında $DHKO_{ÇÇTSE}$, $DHKO_{ÇTSE} < DHKO_{EKK}$ ilişkisi olduğu söylenebilir.

Örnek hacmi sabit alınıp, aykırı değer oranları arttırıldığında incelenen yöntemlere ait deneysel hata kareler ortalamaları büyümektedir. Örneğin örnek hacmi 20 alındığında ve aykırılık oranları sırasıyla $\%\,5, \%10$ ve $\%\,20$ alındığında, EKK yöntemine ait deneysel hata kareler ortalaması sırasıyla aykırılık, $\%5$ iken, 4.8272; $\%\,10$ iken, 15.7596 ve $\%\,20$ iken, 26.1113 olarak hesaplanmıştır. Benzer şekilde ÇTSE tahmin yöntemine ait deneysel hata kareler ortalaması, sırasıyla aykırılık, $\%5$ iken, 0.1373; $\%\,10$ iken, 0.2152 ve $\%\,20$ iken, 0.7689 olarak hesaplanmıştır. Son olarak önerilen ÇÇTSE tahmin yöntemine ait deneysel hata kareler ortalaması ise sırasıyla aykırılık, $\%5$ iken, 0.1381; $\%\,10$ iken, 0.1484 ve $\%\,20$ iken, 0.7158 olarak hesaplanmıştır. Yani incelenen yöntemler kendi içinde değerlendirildiğinde örnek hacmi sabitken aykırılık yüzdesi artırıldığında tahmin yöntemlerine ait deneysel hata kareler ortalaması büyümektedir.

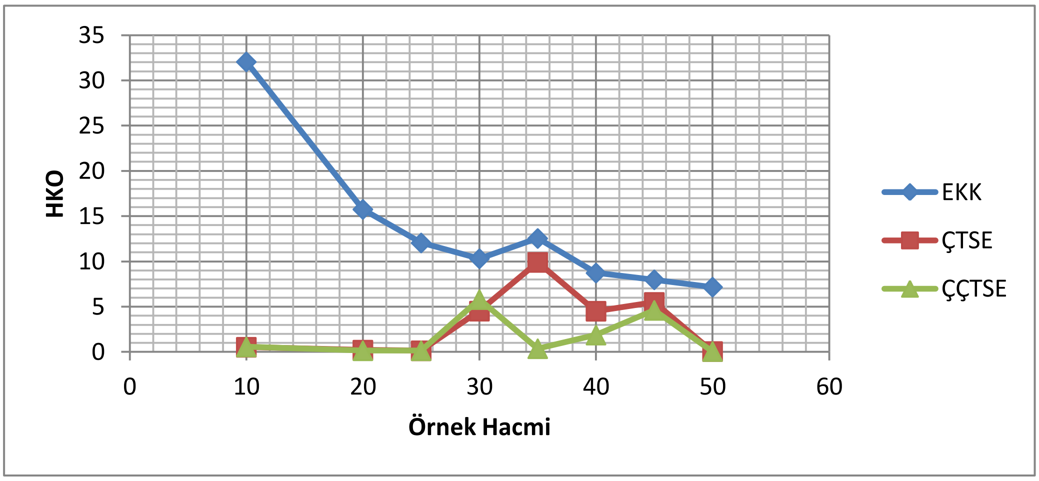
İncelenen yöntemlere ait tahmin edilen regresyon katsayı değerleri aykırı değer yüzdesi sabit alınıp, tüm örnek hacimlerine göre ortalama değerleri hesaplandığında, yani $n = 10, 20, 25, \ldots, 50$ hacminde, $\%\,5$ kirlilik oranında tahminlere ait ortalama DHKO değerleri aykırı değer oranı; $\%5$ iken EKK için 7.7903; ÇTSE için 0.6270 ve ÇÇTS için 0.3512 ve $\%10$ iken EKK için 13.3294; ÇTSE için 3.1737 ve ÇÇTSE için 1.6983 ve $\%20$ iken EKK için 31.2867; MTSE için 5.1085 ve ÇÇTSE için 4.4401 olarak hesaplanmıştır. Dikkat edilecek olursa aykırı değer oranları sabitken incelenen tahmin yöntemlerine ait ortalama DHKO'ları arasında $DHKO_{ÇÇTSE} < DHKO_{ÇTSE} < DHKO_{EKK}$ ilişkisi mevcuttur. Bu kapsamda kirlilik oranı sabit ve örnek hacmi $n = 10, 20, 25, \ldots, 50$ olduğunda önerilen ÇÇTSE tahmin yönteminin daha etkin sonuçlar verdiği görülmektedir.

Aynı şekilde örnek hacmi sabit tutulup aykırı değer oranı artırıldığında da yine ÇÇTSE tahmin yönteminin EKK ve ÇTSE tahmin yöntemlerine nazaran gerçek regresyon katsayılarının tahmininde daha doğru sonuçlar verdiği söylenebilir. Yani söz konusu bu durumlar için önerilen ÇÇTSE tahmin yöntemi gerçek parametreler tahmininde daha etkili olduğu görülmektedir.

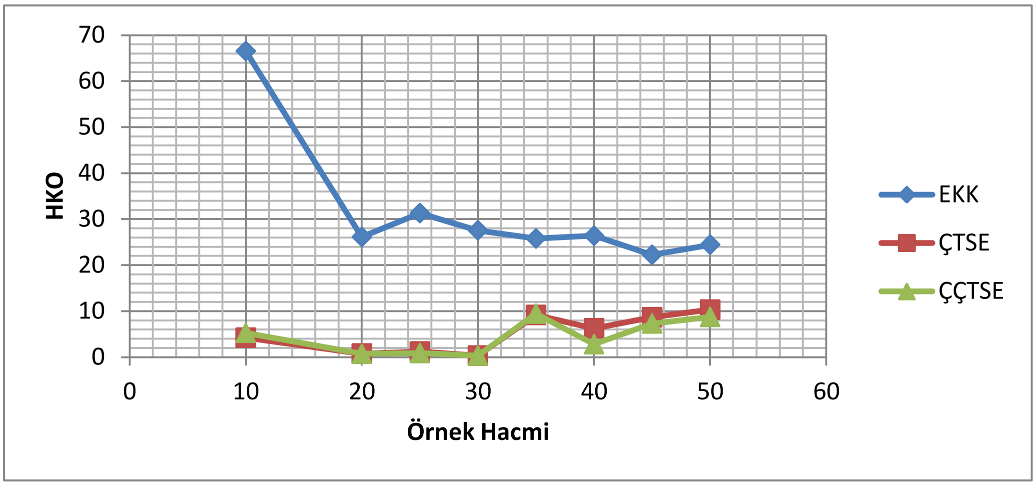
Çizelge 6.2 dikkate alınarak çeşitli örnek hacimleri ve kirlilik oranlarında incelenen üç yöntem için izleme sinyalleri aşağıdaki şekilde verilebilir.



Şekil 6.3. % 5 aykırı oranına göre incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri



Şekil 6.4. % 10 aykırı oranına göre incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri.



Şekil 6.5. % 20 aykırı oranına göre incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri

Şekil 6.3, Şekil 6.4 ve Şekil 6.5 incelendiğinde, çeşitli örnek hacimleri ve kirlilik oranlarında incelenen yöntemlere ait deneysel hata kareler ortalamaları görülmektedir. Şekil 6.3, Şekil 6.4 ve Şekil 6.5 genel olarak değerlendirilecek olursa, incelenen tahmin yöntemleri arasında önerilen Çakı-Theil (ÇÇTSE) tahmin yöntemi en küçük deneysel hata kareler ortalaması değerine sahip olduğu görülmektedir.

Şimdi ise hata dağılımı farklı dağılımlardan geldiğinde incelenen tahmin yöntemlerine ait deneysel hata kareler ortalamaları Çizelge 6.3'deki gibi hesaplanmıştır.

Çizelge 6.3. Etkili karşılaştırma

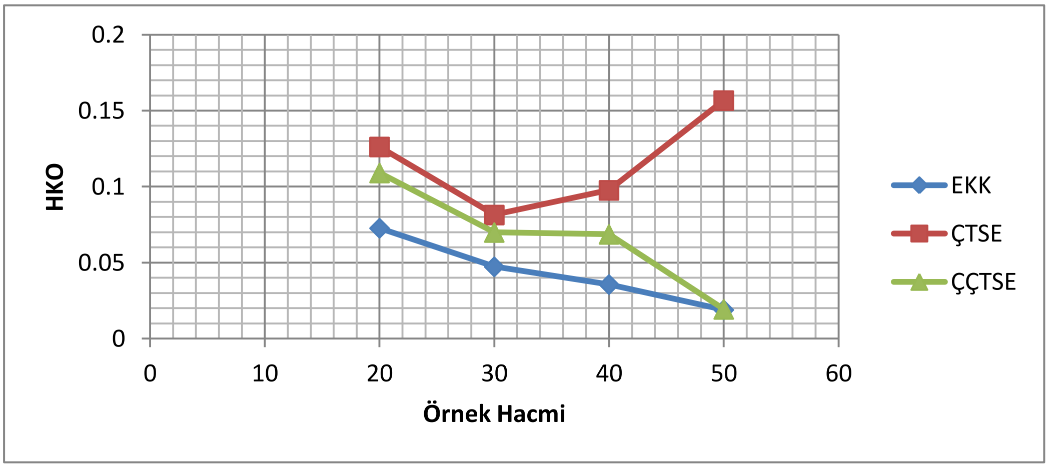
Örnek hacmi	Normal			T3			Cauchy		
	EKK	ÇTSE	ÇÇTSE	EKK	ÇTSE	ÇÇTSE	EKK	ÇTSE	ÇÇTSE
20	0.0727848	0.1262657	0.1091303	0.9418105	0.5330842	0.5343455	802.218717	2.513192	2.364063
30	0.04741218	0.08149654	0.0699978	0.523894	0.3724878	0.3657349	169.801073	0.8567677	0.8024617
40	0.03576168	0.09772447	0.06876328	0.4512333	0.5205301	0.351836	163.714456	4.880104	3.658022
50	0.01906063	0.15678564	0.0193932	0.3197961	0.2671266	0.2543545	2.67E+07	3.51E-01	3.28E-01
Ort	0.04375482	0.11556808	0.0668211	0.55183475	0.42330717	0.376567	8901877.84	2.02936017	1.59605607

Sağlam tahminlerle ilgili daima etkinlik kaybı olabilir. Bu kapsamda etkinliğin araştırılması için bir simülasyon yapılmıştır. Bu simülasyon çalışması için hata dağılımları; $N(0,1)$, serbestlik derecesi 3 olan t dağılımı ve serbestlik derecesi 1 olan t dağılımdan (Cauchy), $n = 20, 30, 40, 50$ örnek hacimli 100 örnek üretilmiştir. EKK, ÇTSE ve ÇÇTSE yöntemlerine ait $\hat{\beta}_i$ değerleri hesaplandı. Daha sonra denklem (6.2) eşitliği kullanılarak tahmin yöntemlerine ait deneysel hata kareler ortalamaları hesaplanmıştır.

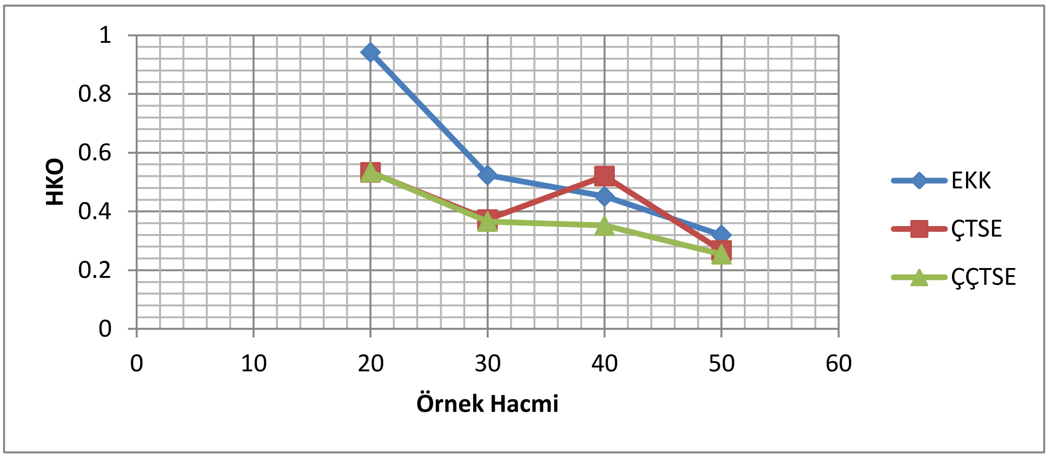
$$DHKO = \frac{\left\| \sum_{i=1}^{z} \hat{\beta}_i - \beta \right\|}{z} \qquad (6.2)$$

Burada, $z = 100$ simülasyon sayısını, $\beta = (2.5, 3, 1.5)$ gerçek regresyon katsayı değerlerini ve $\hat{\beta}_i$ ise incelenen yönteme ait $i.$ örneklem için tahmini göstermektedir. Bu koşullar altında Çizelge 6.3 incelendiğinde hatalar normal dağılımdan geldiğinde incelenen yöntemlere ait deneysel hata kareler ortalamaları tüm örnek hacimleri için yakın sonuçlar aldığı gözlenmiştir. Ancak hatalar ağır kuyruklu dağılımlardan geldiğinde, hatta hata dağılımı özellikle Cauchy dağılımı olduğunda tahmin yöntemlerine ait deneysel hata kareler ortalamaları arasındaki ilişki $DHKO_{ÇÇTSE} < DHKO_{ÇTSE} < DHKO_{EKK}$ şeklindedir. Ayrıca $n = 20, 30, 40, 50$ örnek hacmine ait deneysel hata kareler ortalama değerlerinin ortalaması da dikkate alındığında önerilen ÇÇTSE yönteminin üstünlüğü açıkça görülmektedir.

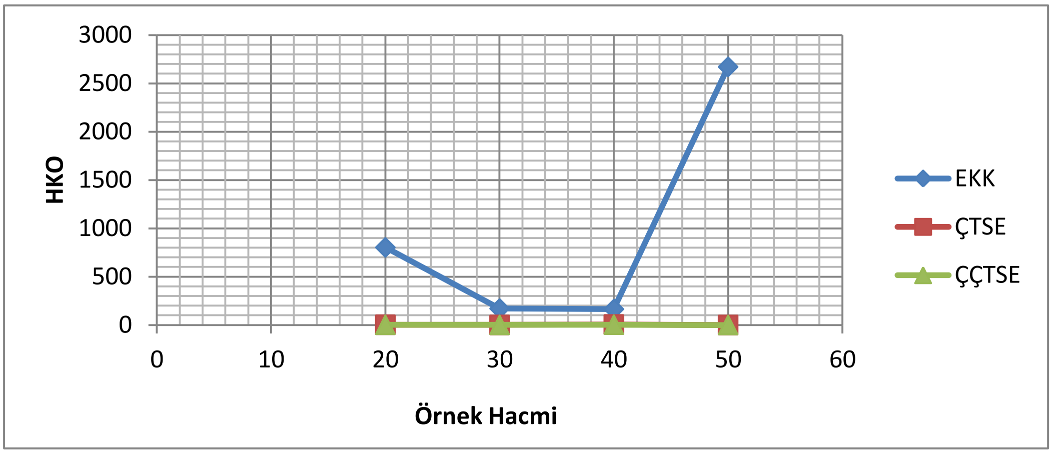
Çizelge 6.3 dikkate alınarak çeşitli örnek hacmi ve hata dağılımları dikkate alınarak incelenen üç yöntem için izleme sinyalleri aşağıdaki şekilde verilebilir.



Şekil 6.6. Hata terimi Normal dağılımdan geldiğinde incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri



Şekil 6.7. Hata terimi T3 dağılımdan geldiğinde incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri



Şekil 6.8. Hata terimi Cauchy dağılımdan geldiğinde incelenen yöntemlerin hata kareler ortalamalarına ait izleme sinyalleri

Şekil 6.6, Şekil 6.7 ve Şekil 6.8'de, çeşitli örnek hacimleri ve hata terimi dağılımına göre incelenen yöntemlere ait deneysel hata kareler ortalamaları görülmektedir. Şekil 6.6, Şekil 6.7 ve Şekil 6.8 genel olarak değerlendirildiğinde, veride kirlilik yokken hata terimi dağılımı ağır kuyruklu dağılımlardan (Cauchy) geldiğinde önerilen Çakı-Theil (ÇÇTSE) tahmin yöntemi en küçük deneysel hata kareler ortalama değerine sahip olduğu görülmektedir.

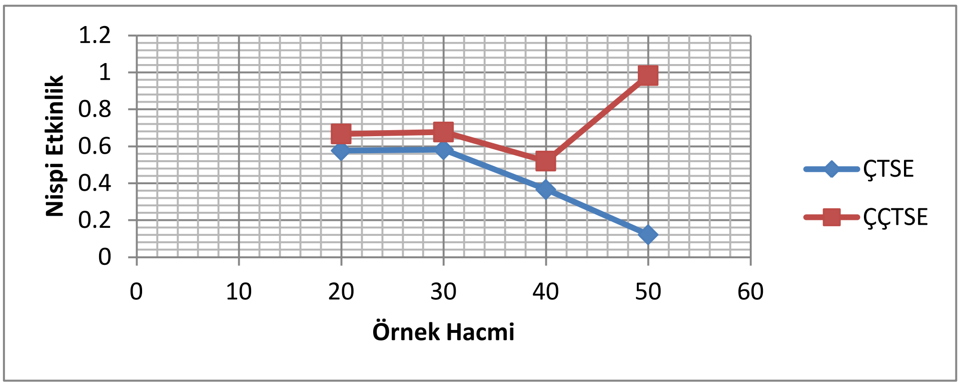
Dolayısıyla ağır kuyruklu dağılımlarda önerilen ÇÇTSE tahmininin önemli bir fark yarattığı açıktır. Bu sonuçlara göre önerilen ÇÇTSE yönteminin sağlamlık ve etkinlik yönünden avantajlı olduğu söylenebilir. Ayrıca etkili karşılaştırma için incelenen yöntemlerin göreli etkinliği çizelge 6.4'de verildiği gibidir.

Çizelge 6.4. Nispi etkinlik

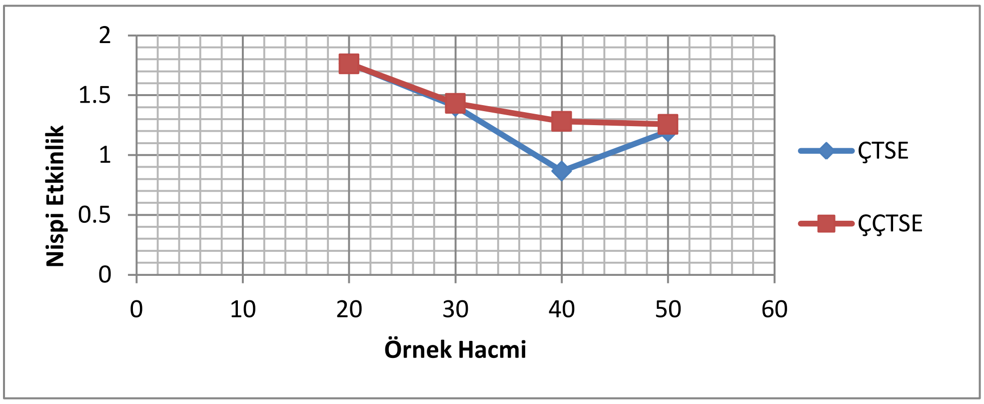
		Normal			T3			T1 (Cauchy)		
		EKK	**ÇTSE**	**ÇÇTSE**	**EKK**	**ÇTSE**	**ÇÇTSE**	**EKK**	**ÇTSE**	**ÇÇTSE**
n=20	**DHKO**	0.072785	0.126266	0.10913	0.941811	0.533084	0.534346	802.2187	2.513192	2.364063
	GE	1	0.576442	0.666953	1	1.76672	1.76255	1	319.2031	339.339
n=30	**DHKO**	0.047412	0.081497	0.069998	0.523894	0.372488	0.365735	169.8011	0.856768	0.802462
	GE	1	0.581769	0.677338	1	1.406473	1.432442	1	198.188	211.6002
n=40	**DHKO**	0.035762	0.097724	0.068763	0.451233	0.52053	0.351836	163.7145	4.880104	3.658022
	GE	1	0.365944	0.520069	1	0.866873	1.28251	1	33.54733	44.75491
n=50	**DHKO**	0.019061	0.156786	0.019393	0.319796	0.267127	0.254355	2.67E+07	3.51E-01	3.28E-01
	GE	1	0.121571	0.982851	1	1.197171	1.257285	1	76068376	81402439
Ort.			0.411432	0.711803		1.309309	1.433697		19017232	20350759

Çizelge 6.4'deki nispi etkinlik değerleri incelenen yöntemlere ait deneysel hata kareler ortalamalarının bulunmasından sonra elde edilmiştir. $\hat{\beta}$'nın nispi etkinliği, $\hat{\beta}$ için hesaplanan EKK değerinin incelenen diğer yöntemlere ait deneysel hata kareler ortalamasına bölünmesiyle elde edilmiştir. Çizelge 6.4 incelendiğinde, Gaussian model altında, ÇTSE 'nin nispi etkinliğinin ortalaması 0.41 civarında, ÇÇTSE'nin nispi etkinlik ortalaması yaklaşık olarak 0.71 olarak hesaplanmıştır. Ancak hata terimi ağır kuyruklu dağılımlardan geldiğinde (özellikle Cauchy) önerilen ÇÇTSE, EKK ve ÇTSE yöntemlerine nazaran başarılıdır. Cauchy dağılımı altında ÇÇTSE'nin nispi etkinliği çok büyüktür. Ayrıca dağılımlara ait ortalama nispi etkinlik değerleri 1'den büyük olduğunda, incelenen ÇTSE ve önerilen ÇÇTSE yöntemlerinin EKK yöntemine göre üstünlüğü mevcuttur. Bakıldığında hatalar özellikle ağır kuyruklu dağılımlardan geldiğinde ÇTSE ve ÇÇTSE, EKK'ya göre oldukça üstündür. Ayrıca nispi etkinlik değerleri göz önüne alındığında hatalar T3 ve Cauchy dağılımından geldiğinde, önerilen ÇÇTSE'nin EKK' ya göre nispi etkinlik değerleri, ÇTSE'nin EKK' ya göre nispi etkinlik değerlerinden daha büyük olduğu ve dolayısıyla önerilen ÇÇTSE tahmin yönteminin ÇTSE tahmin yöntemine nazaran daha etkili olduğu sonucuna varılır.

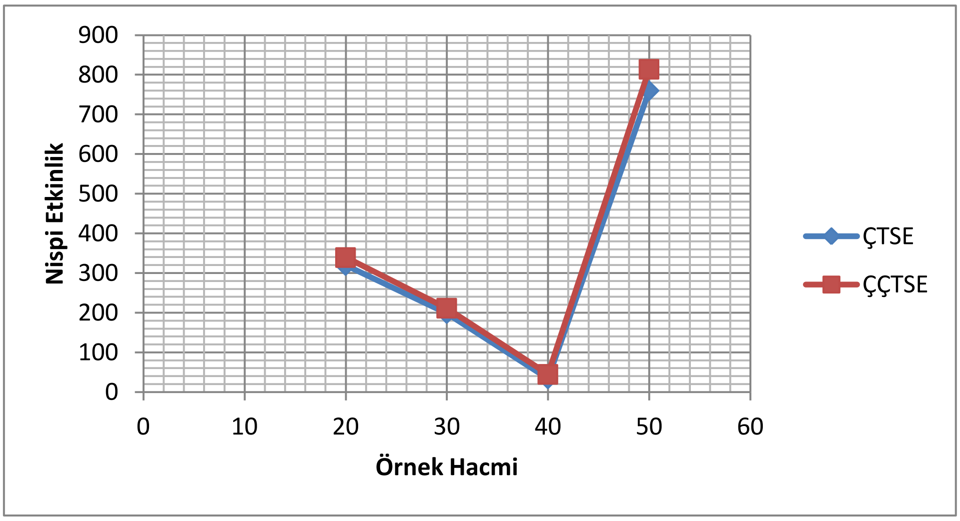
Çizelge 6.4 dikkate alınarak EKK yöntemine göre, örnek hacmi $20, 30, 40$ ve 50 iken çeşitli hata terimi dağılımları için incelenen yöntemlerin nispi etkinlik değerlerine ait izleme sinyalleri aşağıdaki şekilde verilebilir.



Şekil 6.9. Hata terimi Normal dağılımdan geldiğinde incelenen yöntemlerin EKK'ya göre nispi etkinlik değerlerine ait izleme sinyalleri



Şekil 6.10. Hata terimi T3 dağılımından geldiğinde incelenen yöntemlerin EKK'ya göre nispi etkinlik değerlerine ait izleme sinyalleri



Şekil 6.11. Hata terimi Cauchy dağılımdan geldiğinde incelenen yöntemlerin EKK'ya göre nispi etkinlik değerlerine ait izleme sinyalleri

Şekil 6.9, Şekil 6.10 ve Şekil 6.11' de, çeşitli örnek hacimleri ve hata terimi dağılımlarına göre incelenen yöntemlere ait nispi etkinlik değerleri görülmektedir. Şekil 6.9, Şekil 6.10 ve Şekil 6.11 genel olarak değerlendirildiğinde, önerilen Çakı-

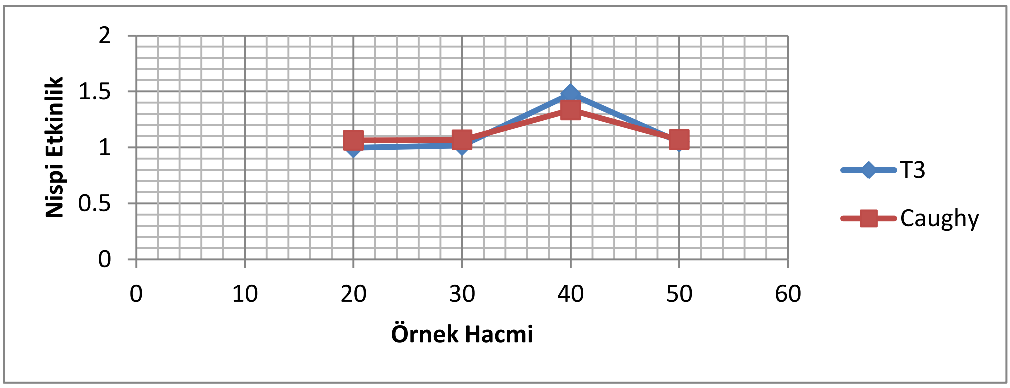
Theil tahmin yöntemi hata terimi dağılımı T3 ve Cauchy dağılımından geldiğinde nispi etkinlik değeri 1'den büyük bulunmuştur. Dolayısıyla bu durum önerilen Çakı-Theil (ÇÇTSE) tahmin yönteminin daha küçük deneysel hata kareler ortalamasına sahip olduğunu göstermektedir.

Çizelge 6.5. ÇTSE'nin ÇÇTSE'ye göre nispi etkinliği

		T3		Cauchy	
		ÇTSE	**ÇÇTSE**	**ÇTSE**	**ÇÇTSE**
n=20	**DHKO**	0.533084	0.534346	2.513192	2.364063
	GE	1	0.99764	1	1.063082
n=30	**DHKO**	0.372488	0.365735	0.856768	0.802462
	GE	1	1.018464	1	1.067674
n=40	**DHKO**	0.52053	0.351836	4.880104	3.658022
	GE	1	1.479468	1	1.334083
n=50	**DHKO**	0.267127	0.254355	3.51E-01	3.28E-01
	GE	1	1.050214	1	1.070122
Ort.			1.136446		1.13374

Çizelge 6.5 'deki nispi etkinlik değerleri, $\hat{\beta}$ için hesaplanan ÇTSE'ye ait deneysel hata kareler ortalamasının önerilen ÇÇTSE'ye ait deneysel hata kareler ortalamasına bölünmesiyle elde edilmiştir. Hatalar ağır kuyruklu dağılımlardan geldiğinde önerilen ÇÇTSE yöntemi, ÇTSE yöntemine nazaran daha üstündür. Bu bağlamda çizelge 6.4 ve çizelge 6.5 dikkate alındığında, önerilen ÇÇTSE çok tutarlıdır, sağlamlık ve etkinlik arasında iyi bir dengeye sahip olduğu söylenebilir. Yani incelenen diğer yöntemlere göre hem sağlamdır, hem de etkindir.

Çizelge 6.5 dikkate alınarak hata terimi T3 ve Cauchy dağılımından geldiğinde önerilen Çakı-Theil (ÇÇTSE) yönteminin ÇTSE yöntemine göre nispi etkinliğine ait izleme sinyali aşağıdaki şekilde verilebilir.



Şekil 6.12. Farklı hata dağılımları için ÇTSE'ye göre ÇÇTSE'nin nispi etkinlik değerlerine ait izleme sinyali

Şekil 6.12' de, önerilen Çakı-Theil (ÇÇTSE) yönteminin, ÇTSE yöntemine göre nispi etkinlik grafiği görülmektedir. Bu grafik incelendiğinde, hata terimi dağılımı T3 ve Cauchy dağılımından geldiğinde nispi etkinlik değeri 1'den büyük çıktığı gözlenmektedir. Dolayısıyla bu durum önerilen Çakı-Theil (ÇÇTSE) tahmin yönteminin daha küçük deneysel hata kareler ortalamasına sahip olduğunu göstermektedir.

Şimdi ÇÇTSE tahmin yönteminin süper etkinliğini, hata terimi kesikli düzgün dağılımdan geldiğinde inceleyelim. $Y_i = 2.5 + 3X_{1i} + 1.5X_{2i} + \varepsilon_i$ regresyon modelindeki hata terimi kesikli düzgün dağılımdan geldiğinde ÇÇTSE tahmin yöntemiyle $\tilde{\beta}$ hesaplanmıştır. Simülasyon sayısı 100 olarak alınmıştır. $\tilde{\beta} = \beta$ olabilmesi için ne kadar büyüklükte örnek hacminin gerekli olduğu araştırılmış ve çizelge 6.6'da gösterilmiştir.

Çizelge 6.6. ÇÇTSE yöntemine ait süper etkinlik

n	10			20			30			50		
	β_{0se}	β_{1se}	β_{2se}	β_{0se}	β_{1se}	β_{2se}	β_{0se}	β_{1se}	β_{2se}	β_{0se}	β_{1se}	β_{2se}
$\varepsilon_i \sim unif(-1,+1$	0.68	0.90	0.47	0.78	0.99	0.63	0.92	1	0.65	0.97	1	0.79

Çizelge 6.6 incelendiğinde, hata terimi $\{-1,+1\}$ kesikli düzgün dağılımdan geldiğinde ve örnek hacmi artığında, önerilen ÇÇTSE tahmin yöntemi yardımıyla tahmin edilen regresyon katsayılarına ait süper etkinlik değerleri belirgin şekilde artmaktadır. Örnek hacmi $n = 10, n = 20, n = 30$ ve $n = 50$ değerlerine ait katsayı değerlerinin ortalaması alındığında, β_0 tahmin değerine ait ortalama süper etkinlik değeri 0.84; β_1 tahmin değerine ait ortalama süper etkinlik değeri 0.97 ve β_2 tahmin değerine ait ortalama süper etkinlik değeri ise 0.64 olarak hesaplanmıştır. Dolayısıyla hata terimi kesikli dağılımlardan geldiğinde ve örnek hacmi arttığında önerilen ÇÇTSE yöntemine ait katsayı tahmin değerlerinin etkinliği artmaktadır. Örneğin özellikle örnek hacmi $n = 30$ ve $n = 50$ için β_1 tahmin değeri süper etkin bulunmuştur.

Sonuç olarak, önerilen ÇÇTSE yönteminde, hata dağılımının simetrik varsayımının sağlanması gerekli değildir ve hata ε süreksiz olduğunda bile süper etkin durumunda olduğu söylenebilir.

Hata terimi dağılımı farklı kesikli dağılımlar alınarak ayrıca incelenebilir.

Simülasyon çalışmasının hemen akabinde incelenen tahmin yöntemlerinin birbirlerine göre durumlarını gerçek veriler yardımıyla araştıralım.

6.1.Gerçek Veri Uygulaması

6.1.1. Coleman veri seti

Veri seti, Coleman ve arkadaşları (1966) tarafından Mid-Anlantic ve New England eyaletlerinden bir kitleden çekilen 20 okul hakkında bilgi içermektedir. Mosteller ve Tukey (1977) birisi bağımlı olarak değerlendirilecek olan altı farklı değişken üzerindeki ölçümleri içeren bu örneklemi analiz etmektedir.

x_1 = Öğrenci başına düşen personel maaşı

x_2 = Memur babaların yüzdesi

x_3 = Sosyo-ekonomik statüden sapma durumları: Aile büyüklüğü, ailenin zarar görmüşlüğü, baba eğitimi, anne eğitimi ve ev eşyaları.

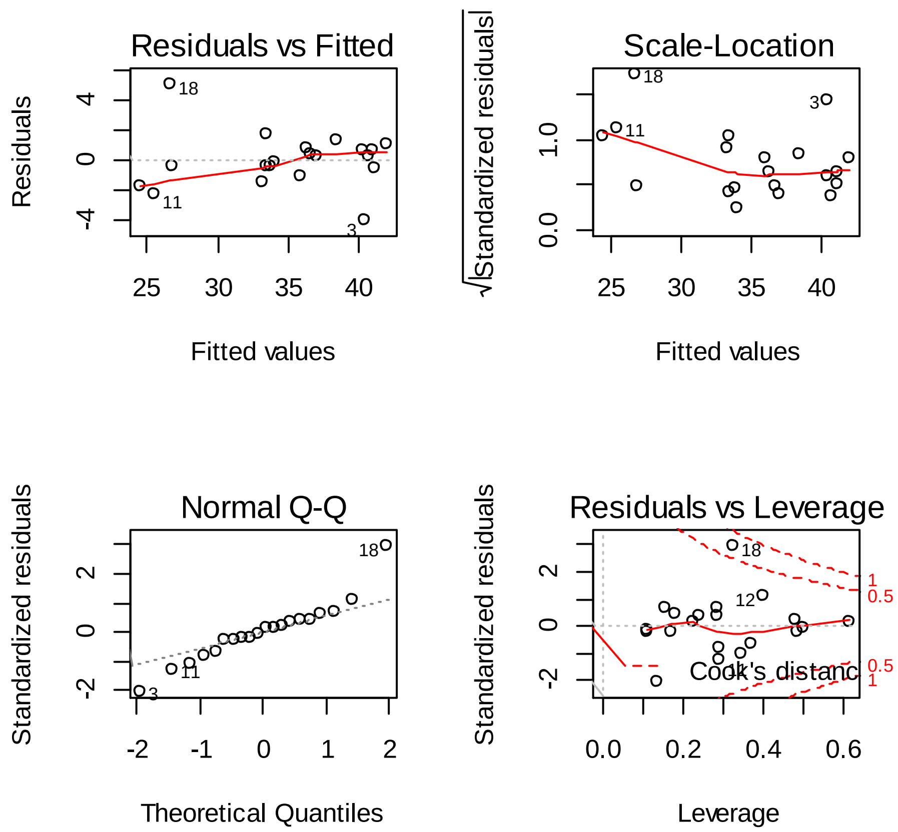
x_4 = Öğretmenlerin sözlü test puanı ortalaması

x_5 = Anne eğitim seviyesi ortalaması (bir birim iki okul yılına eşit)

y = Sözlü test ortalama puanı (tüm altıncı sınıflar).

Çizelge 6.7. Mid-Atlantic ve New England eyaletlerinden 20 okulun bilgilerini içeren Coleman veri seti

Dizin	x_1	x_2	x_3	x_4	x_5	y
1	3,83	28,87	7,20	26,60	6,19	37,01
2	2,89	20,10	-11,71	24,40	5,17	26,51
3	2,86	69,05	12,32	25,70	7,04	36,51
4	2,92	65,40	14,28	25,70	7,10	40,70
5	3,06	29,59	6,31	25,40	6,15	37,10
6	2,07	44,82	3,16	21,60	6,41	33,90
7	2,52	77,37	12,70	24,90	6,86	41,80
8	2,45	24,67	-0,17	25,01	5,78	33,40
9	3,13	65,01	9,85	26,60	6,51	41,01
10	2,44	9,99	-0,05	28,01	5,57	37,20
11	2,09	12,20	-12,86	23,51	5,62	23,30
12	2,52	22,55	0,92	26,60	5,34	35,20
13	2,22	14,30	4,77	28,01	5,80	34,90
14	2,67	31,79	-0,96	23,51	6,19	33,10
15	2,71	11,60	-16,04	23,60	5,62	22,70
16	3,14	68,47	10,62	24,51	6,94	39,70
17	3,54	42,64	2,66	25,80	6,33	31,80
18	2,52	46,70	-10,99	25,20	6,01	31,70
19	2,68	86,27	15,03	25,01	7,51	43,10
20	2,37	76,73	12,77	24,51	6,96	41,01



Şekil 6.13. Coleman veri seti yardımıyla elde edilen artık terimine ait grafikler

Şekil 6.13 incelendiğinde, sol üsteki grafik artıkların, tahmini $\hat{Y}$ değerlerine karşı olan grafiğini göstermektedir. Bu grafiğin sıfırın etrafındaki hataları temsil eden yatay doğru etrafında rastgele dağıtılmış olması beklenir. Yani noktaların dağılımında açık bir trend olmaması gerekir. İncelendiğinde 3,11 ve 18 okullarının doğrusal modelden en uzak okullar olduğu sonucu elde edilir. Sol alttaki grafik, hataların normal dağılışlı olup olmadığını gösteren standart Q-Q grafiğidir. İncelendiğinde hataların normal dağılışa sahip olmadığı gözlenmektedir. Sağ üstteki grafik standardize artıkların karekökü ile $\hat{Y}$ tahmin değerlerinin grafiğini göstermektedir. Yine bu noktalarında açık bir trende sahip olmaması gerekir. Son olarak sağ alttaki grafik, regresyon sonuçlarını değerlendirmede önemli bir ölçüm olan kaldıraç kuvvetini gösterir. Ayrıca regresyonda her bir gözlemin bir diğer önemli ölçümü olan Cook's uzaklığı da gösterilmiştir. 4 grafikte birlikte incelendiğinde 3,11 ve 18 okullarının regresyon doğrusundan sapmış olduğu gözlenmektedir.

Çakı-Theil yöntemi, literatürde mevcut olan Theil-Sen tahminine çakı yönteminin kombinasyonu ile elde edilen yeni bir tahmin yöntemi olup, daha dar güven aralığı ve aşırı uç değerlerin etkesini azaltan bir teknik olduğu düşünülerek bu gerçek yaşam verisi üzerinde araştırma yapılmıştır.

Bu veri seti dikkate alınarak EKK, ÇTSE ve önerilen ÇÇTSE tahmin yöntemlerine ait regresyon modeli ve güven aralıkları hesaplanacaktır. Bu veri seti incelenirken çapraz geçerlilik yöntemi kullanılmıştır. Yani önce veri seti, eğitim ve test verisi olmak üzere iki gruba ayrılmıştır. Verinin $\%\,80$ ‘ lik ilk kısmı eğitim verisi, geri kalan $\%\,20$’ lik kısmı ise test verisi olarak alınmıştır. Eğitim verisi kullanılarak incelenen üç yöntem içinde regeresyon modeli kurulmuştur. Tahmin edilen bu regresyon modeli üzerinden test verisi kullanılarak bağımlı değişkene ait tahmin değerleri elde edilip, yöntemlere ait hata kareler ortalamaları hesaplanmıştır. Çapraz geçerlik yöntemi literatürde sıklıkla kullanılmaktadır. Çünkü verinin tamamı kullanılarak model kurulup hata kareler ortalaması hesaplanınca model aşina olduğu veriyi iyi tanıdığı için yanıltıcı sonuçlar elde edilebilir. Amaç veriye dışardan yeni bir veri gelince de modelin etkinliğini artırmaktır. Bu sebeple veri setine çapraz geçerlik yöntemi uygulanarak tahmin yöntemlerine ait sonuçlar değerlendirilmiştir.

Tüm bu bilgiler dikkate alınarak çizelge 6.6’daki veri setinin analizi yapılmıştır. Çizelge 6.7’de EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait regresyon katsayıları için $\%\,95$ güven aralıkları elde edilmiştir.

Çizelge 6.8. Coleman verisi: EKK ve ÇÇTSE modellerine ait güven aralıkları

	EKK		ÇÇTSE	
Katsayı	Alt sınır	Üst sınır	Alt sınır	Üst sınır
Sabit terim	16.743	59.624	35.175	37.619
x_1	-3.013	1.242	-0.675	-0.377
x_2	-0.004	0.159	0.058	0.086
x_3	0.540	0.835	0.654	0.684
x_4	0.323	1.619	0.881	0.975
x_5	-8.221	-1.629	-4.723	-4.342

İncelenen tahmin yöntemlerine ait güven aralıkları yardımıyla çoklu doğrusal regresyon modeli için regresyon katsayılarının istatistiksel olarak önemli olup olmadığı kontrol edilebilir. Tahmin edilen regresyon modeline ait regeresyon katsayılarının testi için sıfır hipotezi $H_0\colon \beta_j = 0,\ \ j = 1,2,\ldots,5$ ve alternatif hipotez $H_1\colon \beta_j \neq 0,\ \ j = 1,2,\ldots,5$ şeklinde kurulur. Güven aralığı sıfırı kapsamıyorsa tahmin edilen regresyon katsayısı model için istatistiksel olarak önemlidir şeklinde yorum yapılır.

Çizelge 6.7 incelendiğinde, sol taraf klasik tahmin yöntemi olan EKK yöntemine ait güven aralıklarını göstermektedir. % 95 güven düzeyinde EKK tahmin yöntemi için Coleman verisi incelendiğinde x_3, x_4 ve x_5 bağımsız değişkenlerine ait regresyon katsayılarının güven aralığı sıfırı kapsamadığından dolayı sıfır hipotezi reddedilir. Yani x_3, x_4 ve x_5 bağımsız değişkenlerine ait regresyon katsayıları kurulan regresyon modeli için istatistiksel olarak önemli olduğu bulunmuştur. Çizelge 6.7 'in sağ tarafı önerilen Çakı-Theil tahmin yöntemine ait elde edilen regresyon katsayılarının % 95 güven düzeyindeki güven aralıklarını göstermektedir. İncelendiğinde önerilen çakı-theil tahmin yöntemi ile elde edilen regresyon katsayıları tüm değişkenler için sıfırı kapsamadığından dolayı sıfır hipotezi reddedilir. Dolayısıyla x_1, x_2, x_3, x_4 ve x_5 şeklinde ifade edilen bağımsız değişkenlere ait regresyon katsayıları kurulan regresyon modeli için istatistiksel olarak önemli bulunmuştur.

Şimdi ise incelenen tahmin yöntemlerine ait hata kareler ortalamalarını çizelge 6.8.'de özetleyelim.

Çizelge 6.9. Coleman verisi: EKK, ÇTSE ve Önerilen ÇÇTSE tahmin yöntemlerine ait hata kareler ortalaması

Metot	HKO
EKK	16.423
ÇTSE	15.209
ÇÇTSE	14.724

Çizelge 6.8 incelendiğinde önerilen Çakı-Theil tahmin yöntemine ait hata kareler ortalaması EKK ve ÇTSE tahmin yöntemleri ile elde edilen hata kareler ortalamasından daha küçük bulunmuştur.

Dolayısıyla çizelge 6.7 ve çizelge 6.8 birlikte değerlendirildiğinde veride aykırı değer olması durumunda önerilen Çakı-Theil Tahmin yöntemi, hem tahmin edilen regresyon modelindeki regresyon katsayılarının hepsini model için önemli buldu, hem de hata kareler ortalaması incelenen diğer tahmin yöntemlerine göre daha küçük olarak hesaplanmıştır.

6.1.2. Eğitim giderleri veri seti

Veri seti değişen varyansın ilginç bir uygulamasını verir. Bu veri Amerika'daki 50 tane eyaletin eğitim harcamalarından oluşur. Veri çizelge 6.9'de verildiği gibidir.

Chatterjee ve Price (1977) birisi yanıt olarak değerlendirilecek olan dört farklı değişken üzerindeki ölçümleri içeren bu örneklemi analiz etmiştir.

x_1 =1970’de kentte oturan her 1000 kişiye düşen ev sayısı

x_2 =1973’de kişi başına düşen gelir

$x_3 =$ 1974’de 18 yaşının altında her 1000 kişiye karşılık gelen ev sayısı.

$y =$ Bir eyelatteki halk eğitimi için kişi başına düşen gider (1975 için tasarlanmıştır).

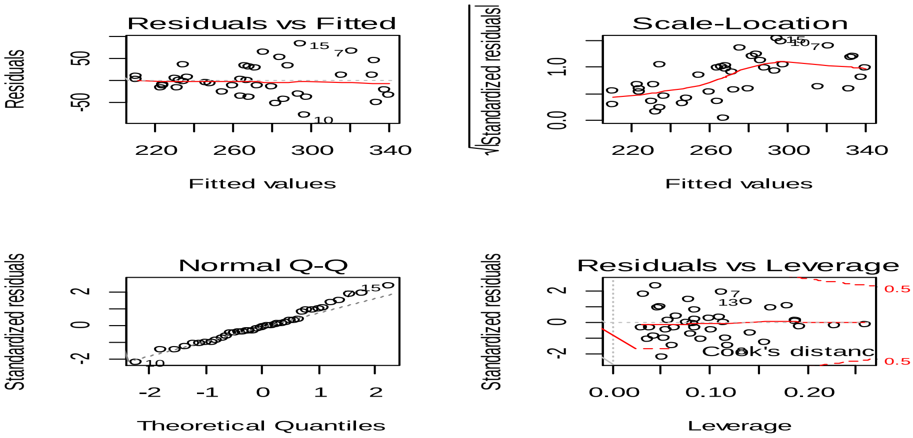
Burada amaç y bağımlı değişkenini x_1, x_2 ve x_3 bağımsız değişkenler yardımıyla açıklamaktır.

Çizelge 6.10. Eğitim giderleri verisi

No	Yer	Bölge	x_1	x_2	x_3	y
1	ME	KB	508	3944	325	235
2	NH	KB	564	4578	323	231
3	VT	KB	322	4011	328	270
4	MA	KB	846	5233	305	261
5	RI	KB	871	4780	303	300
6	CT	KB	774	5889	307	317
7	NY	KB	856	5663	301	387
8	NJ	KB	889	5759	310	285
9	PA	KB	715	4894	300	300
10	OH	KM	753	5012	324	221
11	IN	KM	649	4908	329	264
12	IL	KM	830	5753	320	308
13	MI	KM	738	5439	337	379
14	WI	KM	659	4634	328	342
15	MN	KM	664	4921	330	378
16	IA	KM	572	4869	318	232
17	MO	KM	701	4672	309	231
18	ND	KM	443	4782	333	246
19	SD	KM	446	4296	330	230
20	NB	KM	615	4827	318	268
21	KS	KM	661	5057	304	337
22	DE	G	722	5540	328	344
23	MD	G	766	5331	323	330
24	VA	G	631	4715	317	261
25	WV	G	390	3828	310	214
26	NC	G	450	4120	321	245
27	SC	G	476	3817	342	233
28	GA	G	603	4243	339	250
29	FL	G	805	4647	287	243
30	DY	G	523	3967	325	216

31	TN	G	588	3946	315	212
32	AL	G	584	3724	332	208
33	MS	G	445	3448	358	215
34	AR	G	500	3680	320	221
35	LA	G	661	3825	355	244
36	OK	G	680	4189	306	234
37	TX	G	797	4336	335	269
38	MT	B	534	4418	335	302
39	ID	B	541	4323	344	268
40	WY	B	605	4813	331	323
41	CO	B	785	5046	324	304
42	NM	B	698	3764	366	317
43	AZ	B	796	4504	340	332
44	UT	B	804	4005	378	315
45	NV	B	809	5560	330	291
46	WA	B	726	4989	313	312
47	OR	B	671	4697	305	316
48	CA	B	909	5438	307	332
49	AK	B	831	5309	333	311
50	HI	B	484	5613	386	546

Çizelge 6.9 ‘daki veri seti coğrafi bölgelerle 4 gruba ayrılmıştır. Kuzey-Doğu eyaletleri (1-9 ile gösterilir), kuzey-merkez eyaletleri (10-21ile gösterilir), güney eyaletleri (22-37 ile gösterilir)ve batı eyaletleri (38-50 ile gösterilir) şeklindedir (Chatterjee ve Price, 1977).



Şekil 6.14. Eğitim giderleri veri seti kullanılarak elde edilen artık terimine ait grafikler

Yukarıdaki hata terimine ait grafikler birlikte değerlendirildiğinde veri setinde 7. ve 15. gözlemlerin sapan değer olduğu görülmektedir.

Yine çapraz geçerlilik yöntemiyle eğitim harcamaları verisinin ilk % 80’lik kısmı eğitim verisi olarak alınmış ve bu kısma ait çoklu doğrusal regresyon modeli tahmin edilmiştir. Test verisi olarak geri kalan % 20 'lik kısımdaki bağımsız değişkenler, eğitim verisi yardımıyla tahmin edilen regresyon modelinde yerine yazılmış ve tahmini $\hat{Y}_i$ değerleri elde edilmiştir. Daha sonra test verisinde gerçekte var olan bağımlı değişkene göre durumları incelenmiştir. Belirtilen bu durumlara göre çizelge 6.9'daki veri setinin analizi yapılmıştır. Çizelge 6.9'da EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait regresyon katsayıları için % 95 güven aralıkları hesaplanmış ve aşağıdaki şekilde ifade edilmiştir.

Çizelge 6.11. Eğitim giderleri verisi: EKK ve ÇÇTSE tahmin yöntemlerine ait güven aralıkları

	EKK		ÇÇTSE	
Katsayı	Alt sınır	Üst sınır	Alt sınır	Üst sınır
Sabit terim	-589.951	77.250	-346.571	-329.127
X1	-0.113	0.134	-0.024	0.008
X2	0.034	0.088	0.066	0.073
X3	-0.142	1.625	0.879	0.954

Çizelge 6.11'de, sol taraf EKK tahmin yöntemi ve sağ taraf önerilen Çakı-Theil tahmin yöntemleri için % 95 güven düzeyinde regresyon katsayılarına ait elde edilen güven aralıklarını göstermektedir. İncelenen tahmin yöntemlerine ait hesaplanan güven aralıkları yardımıyla çoklu doğrusal regresyon modeli için regresyon katsayılarının istatistiksel olarak önemli olup-olmadığı araştırılabilir. Tahmin edilen regresyon modeline ait regresyon katsayılarının testi için sıfır hipotezi $H_0: \beta_j = 0, \ j = 1,2,3$, alternatif hipotez $H_1: \beta_j \neq 0, \ j = 1,2,3$ şeklinde kurulur.

Çizelge 6.11 dikkate alındığında, % 95 güven düzeyinde, EKK tahmin yöntemi için eğitim giderleri verisi incelendiğinde sadece x_2 bağımsız değişkenine ait regresyon katsayısının güven aralığı sıfırı kapsamadığından dolayı sıfır hipotezi reddedilir. Dolayısıyla x_2 bağımsız değişkenine ait regresyon katsayısı tahmin edilen regresyon modeli için % 95 güven düzeyinde istatistiksel olarak önemli bulunmuştur. Çizelge 6.11 'in sağ tarafı, önerilen Çakı-Theil tahmin yöntemine ait hesaplanan regresyon katsayılarının % 95 güven düzeyindeki güven aralıklarını göstermektedir. İncelendiğinde önerilen Çakı-Theil tahmin yöntemi ile bulunan regresyon katsayıları x_2 ve x_3 bağımsız değişkenlerine ait regresyon katsayılarının güven aralığının sıfırı kapsamamasından dolayı sıfır hipotezi reddedilir. Dolayısıyla x_2 ve x_3 ile ifade edilen bağımsız değişkenlere ait regresyon katsayıları tahmin edilen regresyon modeli için % 95 güven düzeyinde istatistiksel olarak önemli bulunmuştur.

Eğitim giderleri verisi dikkate alınarak çalışma kapsamında incelenen EKK, ÇTSE ve önerilen ÇÇTSE tahmin yöntemlerine ait hata kareler ortalamaları çizelge 6.12 ‘de özetlenecektir.

Çizelge 6.12. Eğitim giderleri verisi: EKK, ÇTSE ve önerilen ÇÇTSE tahmin yöntemlerine ait hata kareler ortalaması

Metot	HKO
EKK	4293.951
ÇTSE	3645.489
ÇÇTSE	3564.373

Çizelge 6.12 incelendiğinde, önerilen Çakı-Theil tahmin yöntemine ait hata kareler ortalaması, EKK ve Çoklu Theil–Sen tahmin yöntemleri ile hesaplanan hata kareler ortalamalarından daha küçük olarak bulunmuştur.

Bu bağlamda çizelge 6.11 ve çizelge 6.12 birlikte değerlendirildiğinde, veride aykırı değer mevcutken önerilen Çakı-Theil tahmin yöntemi, hem tahmin edilen regresyon modelindeki regresyon katsayılarından x_3 değişkeninine ait regresyon katsayısını model için önemli bulmuş, hem de hata kareler ortalaması incelenen diğer tahmin yöntemlerine göre daha küçük olarak hesaplamıştır.

6.1.3. Vosvos otomobil markasının Türkiye piyasasındaki yeri

Çalışmada kullanılan veriler Sahibinden.com internet sitesinde bulunan 52 ilandan derlenmiştir (URL-4). Kullanılan değişkenler incelendiğinde kategorik değişkenlerin söz konusu olduğu ve kukla değişken kullanılması gerektiği görülmektedir. Kukla değişken kullanılırken kategorik değişkenin düzeylerinden bir tanesi referans olarak model dışında tutulur ve düzey sayısının bir eksiği kadar yeni değişken tanımlanır (Gujarati ve Porter, 2012).

Birisi bağımlı olarak değerlendirilecek olan sekiz farklı değişken üzerindeki ölçümleri içeren aşağıda belirtilen örneklem analiz edilerek sonuçlar değerlendirilecektir.

Bağımlı değişken, Türkiye piyasasında vosvos otomobil markasının birim fiyatı olarak alınmıştır. Bu bağımlı değişkeni açıklamaya çalışan bağımsız değişkenler ise; Yakıt (x_1), Öncam (x_2), Far (x_3), Torpido (x_4), Döşeme (x_5), Yaş (x_6) ve Motor Hacmi (x_7) şeklindedir. Burada Yakıt, Öncam, Far, Torpido, Döşeme ve Motor Hacmi değişkenleri kukla değişken kullanılarak yeni bir değişken olarak tanımlanmışlardır. Şöyle ki kullanılan bağımsız değişkenler ve düzeyleri çizelge 6.13’de verilmiştir.

Çizelge 6.13. Kullanılan bağımsız değişken ve düzeyleri

DeğişkenAdı	**Kod**	**Düzeyler**
Yakıt	Fuel	Benzin*, LPG
Öncam	Gl	Bombe*, Düz
Far	Lamp	Normal*, Badem
Torpido	Torp	Plastik*, Metal
Döşeme	Uph	Kumaş*, Deri
Yaş		Nicel Değişken
MH	MV	1300* ,1600

x_1 =Yakıt

x_2 =Öncam

$x_3 =$ Far

$x_4 =$ Torpido

$x_5 =$ Döşeme

$x_6 =$ Yaş

$x_7 =$ Motor Hacmi

$y =$ Fiyat

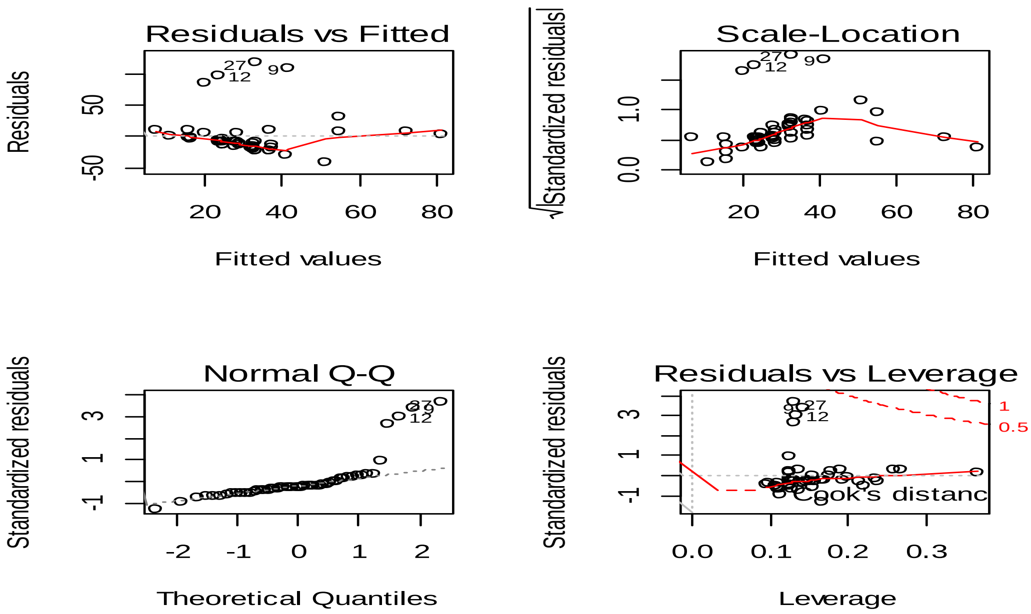
şeklinde ifade edilir. Bu değişkenlere ait sahibinden.com'dan alınan 52 gözleme ait veriler çizelge 6.14'deki gibi verilmiştir.

Çizelge 6.14. Türkiye'de 52 Vosvos otomobil markası bilgilerini içeren veri kümesi

Sıra No	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	20.75	1	0	0	0	1	42	1
2	20.5	1	0	0	0	0	43	1
3	81	0	1	1	1	1	56	0
4	85	0	1	1	1	1	58	0
5	17.75	1	0	0	0	0	42	1
6	17.25	1	0	0	0	0	42	1
7	85	0	1	1	1	1	52	0
8	62	0	1	1	1	1	52	0
9	150	1	1	0	0	1	45	1
10	14.5	0	1	0	1	1	44	1
11	12.5	1	0	1	1	1	44	0
12	120	0	1	0	1	0	44	1
13	10.5	0	1	0	0	1	45	1
14	10	1	1	1	1	1	51	0
15	25	0	0	0	1	1	42	0
16	19.99	1	0	0	0	1	43	1
17	16	0	1	0	1	1	44	1
18	46.5	0	0	1	0	1	43	1
19	13.25	1	1	0	1	0	45	0

20	13	1	0	0	0	1	42	1
21	24.75	1	1	1	0	1	43	1
22	24.5	0	0	0	1	1	43	0
23	18.25	1	1	0	0	1	42	1
24	16	0	0	1	1	0	42	0
25	23	0	1	0	1	1	48	0
26	16.5	0	1	0	1	1	46	0
27	150	1	0	1	0	1	42	1
28	14.2	0	1	0	1	0	46	1
29	20.5	1	0	0	0	1	43	0
30	16.5	1	1	1	1	0	44	1
31	16	1	1	0	0	1	44	0
32	15	0	1	1	1	1	45	1
33	14.5	0	1	0	0	1	44	1
34	12.5	1	0	1	1	1	44	0
35	12	0	1	1	0	0	44	1
36	10.5	0	0	0	0	1	45	0
37	17	1	1	1	1	0	44	1
38	14.75	0	0	0	0	1	44	0
39	14.7	1	1	1	1	0	44	1
40	105	0	1	0	1	1	44	0
41	14.75	0	1	1	1	1	44	1
42	14.7	1	1	0	1	0	44	1
43	10.5	0	1	0	1	0	44	0
44	9.5	1	1	1	1	1	44	1
45	18	0	0	0	0	1	43	0
46	16	0	1	0	1	1	42	1
47	10.95	1	0	1	0	0	43	0
48	20	0	0	0	1	1	42	1
49	19	0	1	1	0	1	42	1
50	15	1	0	0	1	1	42	0
51	14	0	1	0	0	0	43	1
52	34.5	1	0	1	1	1	42	1

Bu değişkenler kullanılarak elde edilen tahmin modelinden yararlanılarak hesaplanan artık terime ait grafikler aşağıdaki gibidir.



Şekil 6.15. Vosvos otomobil veri kümesi yardımıyla elde edilen artık terimine ait grafikler

Şekil 6.15 incelendiğinde 9., 12. ve 27. gözlemlerin sapan değer olduğu görülmektedir.

Çalışma kapsamında incelenen tahmin yöntemlerine ait özet bilgiler aşağıdaki şekilde sunulacak olursa,

Çizelge 6.15'de EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait regresyon katsayıları için % 95 güven aralıkları aşağıdaki şekilde bulunmuştur.

Çizelge 6.15. Vosvos otomobil markası veri seti: EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait güven aralıkları

	EKK		ÇÇTSE	
Katsayı	Alt sınır	Üst sınır	Alt sınır	Üst sınır
Sabit terim	-352.47377	8.098526	-177.7433590	-167.790042
Yakıt	-21.02222	21.856875	-0.6127248	1.169465
Öncam	-30.2792	21.353672	-5.4951212	-3.345970
Far	-21.76697	22.161663	-0.9421815	1.036236
Torpido	-26.99735	18.306549	-4.9900690	-3.541635
Döşeme	-13.64817	31.226825	6.9830234	8.729435
Yaş	0.27505	8.455732	4.2833561	4.505045
MH	-12.65042	37.367661	11.2000360	12.957397

Çizelge 6.15'in sol tarafı EKK tahmin yöntemini ve sağ tarafı önerilen Çakı-Theil tahmin yöntemi için % 95 güven düzeyinde regresyon katsayılarına ait hesaplanan güven aralıklarını göstermektedir. İncelenen tahmin yöntemlerine ait

oluşturulan güven aralıkları yardımıyla çoklu doğrusal regresyon modeline ait regresyon katsayılarının istatistiksel olarak önemli olup-olmadığı araştırılabilir.

Tahmin edilen regresyon modeline ait regeresyon katsayılarının testi için sıfır hipotezi $H_0: \beta_j = 0,\ j = 1,2,\dots,7$ ve alternatif hipotez $H_1: \beta_j \neq 0,\ j = 1,2,\dots,7$ şeklinde kurulur.

Çizelge 6.15 incelendiğinde, % 95 güven düzeyinde EKK tahmin yöntemi için vosvos otomobil marka verisi incelendiğinde sadece yaş bağımsız değişkenine ait regresyon katsayısının güven aralığı sıfırı kapsamadığından dolayı sıfır hipotezi reddedilir. Dolayısıyla yaş bağımsız değişkenine ait regresyon katsayısı % 95 güven düzeyinde istatistiksel olarak önemli bulunmuştur. Çizelge 6.15 ‘in sağ tarafı önerilen Çakı-Theil tahmin yöntemine ait hesaplanan regresyon katsayılarının % 95 güven düzeyindeki güven aralıklarını göstermektedir. Bu güven aralığı incelendiğinde önerilen Çakı-Theil tahmin yöntemi ile öncam, torpido, döşeme, yaş ve motor hacmi bağımsız değişkenlerine ait regresyon katsayıları için güven aralığı sıfırı kapsamadığından dolayı sıfır hipotezi reddedilir. Bu sebeple x_2 ve x_4, x_5, x_6 ve x_7 ile ifade edilen bağımsız değişkenlere ait regresyon katsayıları % 95 güven düzeyinde istatistiksel olarak önemli bulunmuştur.

Şimdi Vosvos otomobil marka verisi dikkate alınarak çalışma kapsamında incelenen EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait ortalama mutlak sapma değerleri çizelge 6.16 ‘da verilmiştir.

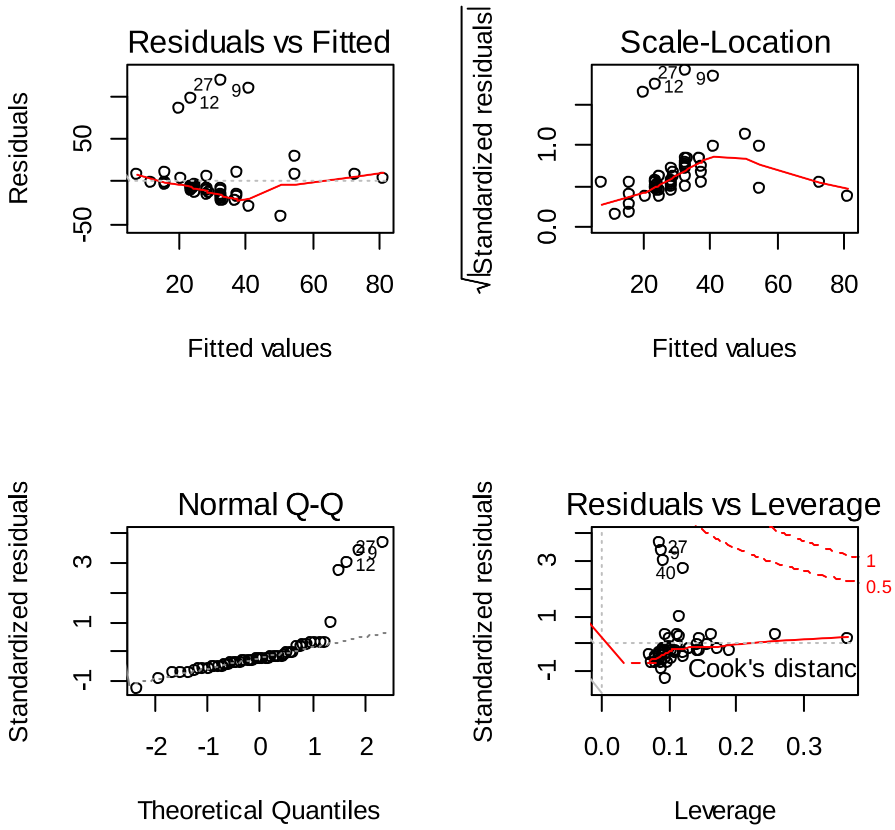
Çizelge 6.16. Vosvos otomobil marka verisi: EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait ortalama mutlak sapma

Metot	OMS
EKK	19.049
Çakı-Theil	18.961

Çizelge 6.16 incelendiğinde, önerilen Çakı-Theil tahmin yöntemine ait ortalama mutlak sapma, EKK tahmin yöntemi ile hesaplanan ortalama mutlak sapma değerinden daha küçük olarak hesaplanmıştır.

Bu bağlamda çizelge 6.15 ve çizelge 6.16 birlikte değerlendirildiğinde veride sapan değer mevcutken önerilen Çakı-Theil tahmin yöntemi, hem tahmin edilen regresyon modelindeki regresyon katsayılarından öncam, torpido, döşeme, yaş ve motor hacmi bağımsız değişkenlerine ait regresyon katsayılarını tahmin edilen regresyon modeli için önemli bulmuş, hem de önerilen Çakı-Theil tahmin yöntemiyle elde edilen ortalama mutlak sapma değeri incelenen diğer yöntemlere göre daha küçük olarak hesaplanmıştır.

Burada, önerilen Çakı-Theil yöntemi ile tahmin edilen model için önemsiz bulunan yakıt (x_1) ve far (x_3) bağımsız değişkenleri modelden çıkarılıp, veri seti tekrar analiz edilirse;



Şekil 6.16. Tahmin modeli kullanılarak elde edilen artık terimine ait grafikler

Şekil 6.16. incelendiğinde 9., 12. ve 27. gözlemlerin sapan değer olduğu görülmektedir.

Çizelge 6.17 incelendiğinde, EKK yöntemi ile elde edilen sonuçlarda yine sadece yaş bağımsız değişkeni model için önemli bulunmuştur. Öneilen Çakı-Theil yöntemi ile elde edilen sonuçlarda tüm bağımsız değişkenler model için önemli bulunmuştur.

Çizelge 6.17. Vosvos otomobil markası veri seti: EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait güven aralıkları

	EKK		ÇAKI-THEIL	
Katsayı	Alt sınır	Üst sınır	Alt sınır	Üst sınır
Sabit terim	-336.0734710	-8.354417	-182.902625	-170.191963
Öncam	-29.1751302	19.978484	-6.633599	-4.159943
Torpido	-25.9040883	17.210234	-4.534255	-2.924486
Döşeme	-12.8930356	30.306236	6.669938	8.889776
Yaş	0.6425917	8.105330	4.321954	4.620484
MH	-11.3807199	36.297863	10.959257	13.184807

Ayrıca bu veri seti için EKK ve önerilen Çakı-Theil tahmin yöntemlerine ait ortalama mutlak sapma değerleri çizelge 6.18 ‘de verilmiştir.

Çizelge 6.18. Vosvos otomobil marka verisi: EKK ve önerilen ÇÇTSE tahmin yöntemlerine ait ortalama mutlak sapma

Metot	OMS
EKK	19.024
Çakı-Theil	18.365

Çizelge 6.18 incelendiğinde, önerilen Çakı-Theil tahmin yöntemine ait ortalama mutlak sapma değeri, EKK tahmin yöntemi ile hesaplanan ortalama mutlak sapma değerinden daha küçük olarak hesaplanmıştır.

7. SONUÇ VE ÖNERİLER

Çoklu doğrusal regresyonda Theil-Sen yönteminin tanıtılıp, simülasyon çalışması ve gerçek veri uygulamalarının yapılması suretiyle EKK yöntemi, Çoklu Theil-Sen yöntemi, çakı yötemi ile TSE yönteminin birleşimi olarak önerilen Çakı-Theil yöntemlerinin karşılaştırmalı olarak değerlendirilmesinin ve önerilen Çakı- Theil tahmin yönteminin üstünlüğünün ispatlanmasının amaçlandığı bu tezin birinci bölümünde; parametrik ve parametrik olmayan regresyon yöntemlerine kısaca bir giriş yapılmıştır. İkinci bölümde; doğrusal regresyon modeli ve varsayımları, klasik tahmincilerden EKK yöntemi ve aykırı değerlerin teşhisi ve dışarda tutulması gibi konular üzerinde durulmuştur. Üçüncü bölümde, basit doğrusal regresyonda çeşitli theil tahmin yöntemleri tanıtılmış, asimptotik özellikleri araştırılmış ve bir gerçek yaşam verisi üzerinde bu tahminlerin avantajları yorumlanmıştır. Daha sonra çoklu doğrusal regresyonda Theil-Sen tahmin yöntemi ve bu yönteme ait regresyon katsayılarının tahmini için gerekli koşullar, Theil-Sen tahmininin tutarlılığı, asimptotik dağılımı, derinlik fonksiyonları ve derinlik medyanları, spatial derinlik ve spatial medyan, spatial medyanın varlığı ve eşsizliği, süper etkinliği, asimptotik normalliği ve sağlamlığı konuları sunulmuştur. Dördüncü bölümde; yeniden örnekleme yöntemlerinden biri olan çakı yöntemi incelenmiş, çakı yönteminin çoklu doğrusal regresyondaki işleyişi bir uygulama ile gösterilmiştir. Daha sonra Çoklu Theil-Sen tahmin yöntemi ve çakı yöntemi hibritlenerek önerilen Çakı-Theil tahmin yöntemine ait regresyon katsayılarının tahmini, katsayılara ait güven aralıkları ele alınmış ve uygulaması 6. bölümde yapılmıştır. Beşinci bölümde, yukarıda sözü edilen EKK, Çoklu Theil-Sen ve önerilen Çakı-Theil tahmin yöntemine ait regresyon tahmin edicilerinin çeşitli uygulamaları yapılmıştır.

Öncelikle simülasyon çalışması yapılmıştır. Simülasyon çalışmasının ilk safhasında yöntemlerin sağlamlığı araştırılmıştır. Daha sonra veride aykırı değerlerin varlığında incelenen yöntemlere ait deneysel hata kareler ortalamaları hesaplanarak çeşitli yorumlar yapılmıştır. Simülasyon çalışmasının son kısmında ise hata terimine ait dağılım normal dağılım ve ağır kuyruklu dağılımlardan geldiğinden tahmin edicilerin birbirlerine göre durumları karşılaştırılmıştır. Simülasyon çalışmasının ardından gerçek yaşam verileri dikkate alınarak farklı iki orijinal veri ile incelenen tahmin yöntemlerine ait regresyon katsayılarının model için önemlilikleri, % 95 güven aralıkları ve hata kareler ortalamaları hesaplanarak çeşitli yorumlar yapılmıştır.

Ayrıca Vosvos marka aracın Türkiye piyasasındaki değerini etkileyen çeşitli faktörler dikkate alınarak oluşturulan orijinal veri seti ile çalışma kapsamında araştırılan tahmin yöntemlerinin regresyon modeli tahmin edilip, regresyon katsayılarına ait % 95 güven aralıkları oluşturulmuş ve ortalama mutlak sapma değeri hesaplanarak, çeşitli yorumlamalar yapılmıştır.

Çoklu Theil-Sen, EKK ve önerilen Çakı-Theil tahmin yöntemlerinden elde edilen sonuçların karşılaştırmalı olarak incelenmesinden sonra aşağıdaki sonuçlara ulaşılmıştır.

Simülasyon çalışmasında tez kapsamında incelenen tahmin yötemlerinin, sağlamlık yönünden karşılaştırılması sonucunda veride herhangi bir aykırı değer olmadığında ve hata terimlerine ait dağılım normal dağılımdan geldiğinde EKK, ÇTSE ve ÇÇTSE tahmin yöntemlerinin hepsi, gerçek regresyon katsayılarına yakın değerler almıştır.

Simülasyon çalışması sonucunda üretilen veri setinde hatalar yine normal dağılımdan geldiğinde ancak veriye çeşitli oranlarda kirlilik katıldığında EKK yöntemi hemen gerçek regresyon katsayılarından uzakta tahmin yapmış ve kullanılamaz hale gelmiştir. Ancak Çoklu-Theil tahmini ve önerilen Çakı-Theil tahmin yöntemleri belli bir yüzdeye kadar bozulmadan dayanabilmiştir. Bu oran Çoklu Theil-Sen tahmini için % 20 ve önerilen Çakı-Theil tahmini için % 30 civarında olduğu görülmüştür.

Simülasyon çalışmasının ikinci kısmında hatalar normal dağılımdan geldiğinde ancak veride kirlenme mevcut olduğunda tahmin edilen regresyon katsayılarına ait deneysel hata kareler ortalamaları hesaplanmıştır. Sonuçta önerilen Çakı-Theil tahmin yönteminin incelenen diğer yöntemlere göre daha doğru sonuçlar elde ettiği gözlenmiştir.

Simülasyon çalışmasının üçüncü kısmında üretilen veriye herhangi bir kirlilik katılmayıp, ancak hata terimleri dağılımı; normal dağılım, üç serbestlik dereceli t dağılımı ve bir serbestlik dereceli t (Cauchy) dağılımı göz önünde bulundurularak üretilip incelenen tahmin edicilerinin durumları araştırılmıştır. Araştırma sonucunda hata terimi ortalaması 0 ve varyansı 1 olan normal dağılımdan geldiğinde incelenen tahmin yöntemleriyle tahmin edilen regresyon katsayılarına ait deneysel hata kareler ortalamaları birbirlerine çok yakın bulunmuştur. Yani hatalar normal dağılımdan geldiğinde ve veride herhangi bir sapan değer olmadığında incelenen tahmin yöntemleri birbirlerine çok yakın bulunmuştur. Ancak hata terimi dağılımı özellikle ağır kuyruklu bir dağılım olan Cauchy dağılımından geldiğinde önerilen Çakı-Theil yöntemi ve Çoklu Theil-Sen tahmin yöntemi ile hesaplanan deneysel hata kareler ortalamaları, EKK yöntemi ile hesaplanan deneysel hata kareler ortalaması sonucuna göre oldukça küçük sonuçlar verdiği gözlenmiştir. Önerilen Çakı-Theil tahmin yöntemi, ÇTSE yöntemine göre de daha iyi sonuçlar verdiği gözlenmiştir. Ayrıca hata dağılımı ağır kuyruklu dağılımdan geldiğinde incelenen tahmin yöntemleri, $n = 20{,}30{,}40{,}50$ olarak alındığında hesaplanan deneysel hata kareler ortalamasının ortalaması alındığında, tahmin yöntemleri arasında $DHKO_{ÇÇTSE} < DHKO_{ÇTSE} < DHKO_{EKK}$ ilişkisi olduğu görülmüştür.

Simülasyon çalışmasının son kısmında ise hata terimi kesikli dağılımlardam geldiğinde ve örnek hacmi ne kadar büyüklükte alınırsa önerilen ÇÇTSE yöntemiyle hesaplanan regresyon katsayılarının süper etkin olduğu incelenmiştir. Bu bağlamda örnek olması sebebiyle hata terimi dağılımı kesikli düzgün dağılımdan alınmıştır. Sonuçta, örnek hacmi artırıldığında önerilen yöntem yardımıyla hesaplanan süper etkinlik değerlerinde belirgin bir artış olduğu gözlenmiştir. Örneğin β_1 tahmin değeri için örnek hacmi $n = 30$ alındığında tam etkin olarak hesaplanmıştır.

Dang vd. (2008) çoklu doğrusal regresyon modelinde, hata dağılımının süreksiz olduğu durumlarda bile süper etkin olduğunu ve hata dağılımının simetrik varsayımının çıkarılabileceğini göstermişlerdi. Bu sonuçların önerilen ÇÇTSE'ye genişletilebilir olup olmadığı incelenmiş ve hata dağılımının simetrik varsayımının sağlanmasının gerekli olmadığı ve hata ε süreksiz olduğunda bile süper etkin durumda olduğu gözlenmiştir.

Bu bağlamda simülasyon çalışması sonucunda, önerilen Çakı-Theil tahmin yönteminin gerçek regresyon katsayılarının tahmininde daha etkili olduğu gözlenmiştir.

Simülasyon çalışmasının akabinde üç tane gerçek veri uygulaması yapılmıştır. Verilerden ilki literatürde Coleman verisi ve diğeri ise eğitim giderleri verisi olarak bilinmektedir. Bu verilerin ikisinde de aykırı değer mevcut olup, artıklara ait normallik varsayımı sağlanmamaktadır.

Coleman veri seti dikkate alındığından EKK yöntemi ile $\hat{\beta}_3$, $\hat{\beta}_4$ ve $\hat{\beta}_5$ tahmini regresyon katsayıları model için önemli bulunmuştur. Önerilen Çakı-Theil tahmin yönteminde ise tahmin katsayılarının hepsi çoklu doğrusal regresyon modeli için önemli olduğu görülmüştür ve incelenen tahmin yöntemlerine ait hata kareler ortalamaları arasındaki ilişki $HKO_{ÇÇTSE} < HKO_{ÇTSE} < HKO_{EKK}$ şeklinde bulunmuştur.

Eğitim giderleri veri seti dikkate alındığında ise EKK yönteminde sadece $\hat{\beta}_2$ tahmin değeri önemli bulunmuştur. Önerilen Çakı-Theil tahmin yönteminde ise $\hat{\beta}_2$ ve $\hat{\beta}_3$ tahmin değerleri regresyon modeli için önemli bulunmuş olup, incelenen tahmin yöntemlerine ait hata kareler ortalamaları arasındaki ilişki $HKO_{ÇÇTSE} < HKO_{ÇTSE} < HKO_{EKK}$ şeklindedir.

Türkiye piyasasındaki Vosvos otomobil marka verisi kullanılarak gerçekleştirilmiş olan uygulama kapsamında, önerilen Çakı-Theil tahmin yöntemine ait ortalama mutlak sapma değeri karşılaştırma kriteri olarak dikkate alınmış ve sonuçta EKK tahmin yöntemine göre daha küçük bulunmuştur.

Sonuç olarak incelenen regresyon tahmin yöntemleri sağlamlık, deneysel hata kareler ortalaması, ortalama mutlak sapma, hata dağılımlarının durumlarına göre ve etkili karşılaştırmalar yönünden incelenmiştir. Elde edilen sonuçlara göre EKK yönteminin tüm durumlar için çok dayanıklı sonuçlar vermediği gözlemlenmiştir. Önerilen Çakı-Theil tahmin yöntemi özellikle gerçek regresyon katsayılarının hesaplanmasında, kirlilik oranı yüksek olduğu durumlarda diğer iki yönteme göre daha sağlam ve dayanıklı kalmıştır. Önerilen Çakı-Theil tahmin yöntemi hata terimi ağır kuyruklu dağılımlardan geldiğinde özellikle 1 serbestlik dereceli t dağılımından geldiğinde, diğer diğer yöntemlere göre oldukça etkili olduğu gözlenmiştir. Ayrıca önerilen ÇÇTSE'nin süreksiz hata dağılımı için süper etkin olduğu görülmüştür.

Yapılan tüm bu yorumlar burada yapılan uygulamalar için geçerlidir. Farklı simülasyon çalışmalarıyla daha farklı sonuçlar elde edilip yorumlanabilir.

Burada belirlenmiş keyfi m sayısı dikkate alınarak, $\binom{n}{m}$ kombinasyonu kadar veriye, EKK yöntemi uygulanarak hesaplanan regresyon parametreleri tahmin edilip, tahmin edilen bu katsayılar üzerinden çok değişkenli medyan hesabı yapılarak, ÇTSE ve ÇÇTSE tahmin yöntemlerine ait regresyon katsayıları hesaplanmıştır.

Ayrıca burada, m keyfi olarak seçilen bir sayıdır ve $p+1 \leq m \leq n$ arasında değişmektedir. Simülasyon çalışmasında $m = p+1$ alınmıştır. Ancak m' nin değişik versiyonları da kullanılabilir. Örneğin; m, $n/2$ olarak alınabilir. Eğer $m = n/2$ olarak alınırsa en fazla kombinasyona ulaşılır. Ancak, eğer m değeri küçük seçilirse, veriden seçilen $\binom{n}{m}$'li kombinasyonunda verideki aykırı değerlerin sisteme dahil edilmesi olasılığı, m'nin büyük seçilmesi durumuna kıyasla daha azdır. Yani $p + 1 \leq m \leq n$ arasından keyfi seçilen m sayısının küçük seçilmesiyle analize dahil edilecek olan aykırı değer sayısı daha az olur.

Çoklu Theil-Sen tahmini ve önerilen Çakı-Theil tahmin yöntemi ile tahmin yaparken, verinin belli bir yüzdesi alınarak farklı kombinasyonlar seçiliyor. Dolayısıyla bu kombinasyonlara bağlı olarak farklı sonuçlar elde edilebilir. Bu bağlamda, belli varsayımlar altında öyle bir alt örneklem seçilmeli ki amaca uygun sonuçlar elde edilebilsin. Sonraki çalışmalarda bu konu üzerine yoğunlaşılacaktır.

Bir başka yeniden örnekleme yöntemi olan bootstrap yöntemi de Theil-Sen tahminine hibritlenerek sonuçlar değerlendirilebilir. Ayrıca diğer önerilen sağlam regresyon modellerine çakı yöntemi ve bootstrap yöntemi ikame edilerek uygulama üzerinde elde edilen sonuçlar tartışılabilir.

Sonraki çalışmalarda önerilen Çakı-Theil yöntemi ile literatürdeki sağlam tahmin yöntemlerine ait algoritmalar arasındaki farklılıklar ve benzerlikler göz önüne alınarak detaylıca incelenebilir ve birbirlerine göre üstünlükleri tartışılabilir.

KAYNAKLAR

Abdi, H. and Williams, L. J. 2010. Jackknife. In Neil Salkind (Ed.). Encyclopedia ofResearchDesign.ThousandOaks,CA:Sage.https://www.utdallas.edu/~herve/abdi-Jackknife2010-pretty.pdf (Erişim Tarihi: 12.11.2014)

Akritas, M.G., Murphy S.A. and LaValley, M.P. 1995. The Theil–Sen estimator with doubly censored data and applications to astronomy. *Journal of the American Statistical Association*, 90, 170–177.

Alakuş, K. 2015. Ondokuz Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü Parametrik Olmayan İstatistiksel Yöntemler Yayınlanmamış Doktora Dersi Notları, Samsun

Alpar, R. C. 2013. *Uygulamalı Çok değişkenli İstatistiksel Yöntemler*. Detay Yayıncılık. ISBN. 978-605-5437-42-8. Ankara.

Arvesen, J. N. 1969. Jackknife U-Statistics. *Annals of Mathematical Statistics*. S.2076.

Becker, C, Fried, R. and Kuhnt S. (Editors). 2014. *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather.* Springer Science ve Business Media.

Bekiroğlu, N., Konyalıoğlu, R. ve Karahan, N. 2013. Çoklu Doğrusal Regresyon Sonuçlarının Jackknife Tekniği ile Tekrarlanabilirliğinin Değerlendirilmesi. *Marmara Medical Journal*. 26:63-7.

Bose, A. 1998. Bahadur representation of Mm estimates. *Annals of statistics*, 771-777.

Bowerman, B. L., O'Connell, R.T., Murphree, E.S. and Orris, J. B. 2013. *İşletme İstatistiğinin Temelleri*. (Çev. Neyran Orhunbilge, Mustafa Can ve Şebnem Er). Nobel yayınları, Ankara.

Candan, M. (1995). Doğrusal Regresyon Çözümlemesinde Sağlam Kestiriciler. H. Ü. Fen Bilimleri Enstitüsü. Ankara.

Chakraborty, B., Chaudhuri, P. and Oja, H. 1998. Operating transformation retarnsformation on spatial median and angle test. *Statistica Sinica*, *8*, 767–784.

Chatterje, S. and Price, B. 1977. *Regression Analysis by Example*, John Wiley and Sons, New York.

Chaudhuri, P. 1992. Multivariate Location Estimation Using Extension of R-Estimates Through U-Statistics Type Approach. *Ann. Statist*. 20, 897-916.

Chaudhuri, P. and Sengupta, D. 1993. Sign tests in multidimension: Inference based on the geometry of data cloud. *Journal of the American Statistical Association*, *88*, 1363–1370.

Dang, X., Peng, H., Wang, X. and Zhang, H. 2008. Theil-Sen Estimators in a Multiple Linear Regression Model. Submitted paper. http://home.olemiss.edu/~xdang/papers/

Daniel, C. and Wood, F.S. 1971. Fitting Equations to Data, John Wiley and Sons, New-York.

Daniel, W.W. 2000. *Applied Nonparametric Statistics*. Georgia State University, Boston, 2nd edition, p: 18-20, 426-443.

Dhar, S. S. ve Chauduri, P. 2011. On the statistical efficiency of robust estimators of multivariate location. *Statistical Methodology*, *8*, 113–128.

Dietz, E. J. 1989. Teaching Regression in a Nonparametric Statistics Course. *The American Statistician*. 43, 35-40.

Efron, B. 1979. Bootstrap Method; Another Look at Jackknife. *Annals of Statistics*, Vol. 7, pp. 1-26.

Efron, B. and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York, s.202-210

Erilli, N. A. and Alakus, K. 2014. Non-Parametrıc Regression Estimation for Data With Equal Values. *European Scientific Journal*. February. Edition Vol. 10, No.4 ISSN:1857-7881 (Prınt) e-ISSN 1857-7434.

Erilli, N. A. and Alakus, K. 2016. Parameter Estimation In Theil-Sen Regression Analysis With Jackknife Method. *Eurasian Academy of Sciences Eurasian Econometrics, Statistics ve Emprical Economics Journal*. Vol. 5, pp.28-41.

Eubank, R. L. 1988. Spline Smoothing and Nonparametric regression. *Marcel Dekker Inc*., New York, USA.

Fernandes, R. and Leblanc, S. 2005. Parametric (modified least squares) and nonparametric (Theil-Sen) lineer regression for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*. 95, 303-316.

Freedman, D.A. 1981. Bootstrapping Regression Models, *Annals of Statistics*. Vol.1, No. 6, pp. 1218-1228.

Friedl,H.andStampfer,E.2001.JakknifeResampling.http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=6C19CCCF139906FE02B2AD4DB8AD2FCA?doi=10.1.1.57.6702verep=rep1vetype=pdf (Erişim Tarihi: 22.03.2016)

Granato, G. E. 2006. *Kendall-Theil Robust Line (KTRLine-version 1.0*) *–A Visual Basic Program for Calculating and Graphing Robust Nonparametric Estimates of Linear-Regression Coefficients Between Two Continous Variables*. Chapter 7, Section A, Statistical Anlaysis, Book 4, Hydrologic Analysis and Interpretation. U. S. Geological Survey Techniques and Methods 4-A7.

Gray, H. L. and Schucany, W. R. 1972. The Generalized Jackknife Statistic. Mercel Dekker Inc.., New-York.

Gujarati, D. N ve Porter, D. C. 2012. (Çev. Şenesen, Ü., ve Günlük-Şenesen, G.) Temel ekonometri. Literatür Yayıncılık.

Hettmansperger, T. P., and McKean, J. W. 2010. *Robust nonparametric statistical methods*. CRC Press.

Hettmansperger, T. P. and Randles, R. 2002. A practical affine equivariant multivariate median. *Biometrika, 89*, 851–860.

Hinkley, D. V. 1977. Jackknife confidence limits using Student t approximations. *Biometrika*, *64*(1), 21-28.

Hollander, M. and Wolfe, D.A 1973. *Nonparametric Statistical Methods (1st edn)* (NY: JohnWiley ve Sons), S. 205–206.

Hollander, M. and Wolfe, D.A. 1999. *Nonparametric Statistical Methods*. 2nd Ed. NY: JohnWiley ve Sons, 421–423.

Hussain, S.S and Sprent, P 1983. Non-parametric Regression. *J.Roy Statist. Soc.*, *Series A*, 146, 182-191.

Ilmonen, P. Oja, H. and Serfling, R. 2012. On invariant coordinate system (ICS) functionals. *International Statistical Review*, *80*, 93–110.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An introduction to statistical learning* (Vol. 6). New York: springer.

Jones, M.P. 1997. A class of semi-parametric regressions for the accelerated failure time model. *Biometrika*, 84, 73–84.

Kalaycı, Ş. 2016. *SPSS uygulamalı çok değişkenli istatistik teknikleri*. Asil Yayın Dağıtım (7. Baskı).

Kayri, M. ve Büyüköztürk, Ş. 2009. Nicel Bilimsel Bulgulara Ait Genellebilirliğin Jackknief Yöntemi ile İncelenmesi: Uygulamalı Bir Çalışma. *Kuram ve Uygulamada Eğitim Bilimleri/ Educational Science: Theory Practice* 9(4), 1751-1780.

Kıroğlu, G. 2001. *Uygulamalı Parametrik Olmayan İstatistiksel Yöntemler*. Mimar Sinan Üniversitesi Fen Edebiyat Fakültesi., İstanbul.

Koshevoy, G. and Mosler, K. 1997. Zonoid trimming for multivariate distributions. *Ann. of statist*. 95,1998-2017.

Lavagnini, I., Badocco, D., Pastore, P. and Magno, F. 2011. Theil-Sen nonpara-metric regression technique on univariate calibration, inverse regression and detection limits *Talanta*, Volume 87, Pages 180-188.

Lehmann, E.L. 2006. *Nonparametrics: Statistical methods based on ranks.* New-York: Springer, pp. 463.

Lia, F., Penga, H. and Tanb, F. 2009. The Asymptotics of the Mahalanobis Depth-Based Theil-Sen Estimators in Multiple Linear Models.

Liu, R. 1990. On a notion of data depth based on random simplices. *Ann. Stat*. 18, 405-414.

Maritz, J. S. 1979. On Theil's Method in Distribution-Free Regression. Australian *Journal of Statistics*, 21, 30-35.

Milasevic, P. and Ducharme, G. R. 1987. Uniqueness of the spatial median. *Ann. Statist*. 15, 1332-1333.

Miller, R. G. 1974. The jackknife-a review. *Biometrika*, *61*(1), 1-15.

Mount, D. M. and Netanyahu, N.S. 2001. Efficient randomized algorithms for robust estimation of circular arcs and aligned ellipses. *Computational Geometry – Theory and Applications*, 19, 1–33.

Mosteller, F. and Tukey, J. W. 1977. *Data Analysis and Regression*. Addison-Wesley, Reading, MA.

Möttönen, J., Nordhausen, K. and Oja, H. 2010. Asymptotic theory of the spatial median. In *IMS collections: Vol. 7. Festschrift in honor of professor Jana Jureckova* (pp. 182–193).

Niemiro, W. 1992. Asymptotics for M-estimators defined by convex minimization. *The Annals of Statistics*, 1514-1533.

Niinimaa, A. and Oja, H. 1999. *Multivariate median*. In S. Kotz, N. L. Johnson, and C. P. Read (Editors.), *Encyclopedia of statistical sciences* (Vol. 3). New York: Wiley.

Nordhausen, K. and Oja, H. 2011. Multivariate L1 methods: the package MNM. *Journal of Statistical Software*, *43*, 1–28.

Oja, H. 1983. Descriptives tatisticsf or multivariate distributions. *Statist.Prob. Letters*.1 , 327-332.

Oja, H. and Niinimaa, A. (1984). *On robust estimation of regression coefficients*. Oulun Yliopisto. Department of Mathematics.

Oja, H. 2010. *Multivariate nonparametric methods with R. An approach based on spatial signs and ranks*. New York: Springer.

Peng, H., Wang, S. and Wang, X. 2008. Consistency and asymptotic distribution of the Theil–Sen estimator. *Journal of Statistical Planning and Inference*, 138(6), 1836-1850.

Roberts, S., and Martin, M. A. 2010. Bootstrap-after-bootstrap model averaging for reducing model uncertainty in model selection for air pollution mortality studies. *Environmental health perspectives*, *118*(1), 131.

Quenouille, M. H. 1956. Notes on bias in estimation. *Biometrika*, *43*(3/4), 353-360.

Randles, R. H., and Wolfe, D. A. 1979. *Introduction to the Theory of Nonparametric Statistics*. Wiley ve Son, New York.

Rao, K. S. M. and Gore, A.P. 1982. Nonparametric Tests for Intercept in Linear Regression Problems. *Australian J.of Statistics*, 24(1), 42-50

Reeds, J. A. 1978. Jackknifing maximum likelihood estimates. *The Annals of Statistics*, 727-739.

Rousseeuw, P. J., and Leroy, A. M. 2005. *Robust regression and outlier detection* (Vol. 589). John wiley ve sons.

Rousseeuw, P. J. and Aelst, S. V. 1999. Positive-Breakdown Robust Methods in Computer Vision. *Computing Science and Statistics*, Vol. 31.

Scholz, F. W. 1978. Weighted Median Regression Estimates. *The Annals of Statistics*, 6(3): 603-609.

Sen, P. K. 1968. Estimates of the regression coefficient based onKendall's tau. *Journal of the American Statistical Association*, 63, 1379–1389.

Serfling, R 1984. Generalized L-, M-, and R-statistics. *Ann. Statist*. 1, 76-86.

Shen, G. 2009. Asymptotics of a Theil-type estimate in multiple linear regression. *Statistics ve Probability Letters*, 79(8), 1053-1064.

Sibson, R. 1984. Multivariate analysis. *J. R. Statist. Soc. A* 147, 198-207.

Sievers, G. L. 1978. Weighted Rank Statistics For Simple Linear Regression. *J.Amer.Statist.Ass*., 73, 628-631.

Small, C. 1990. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique* 58, 263-277.

Sprent, P. 1993. *Applied Nonparametric Statistical Methods* (2nd edn) (NY: CRC Press).

Srivastava, A. and Shalabh, K. 1997. Improved Estimation of The Slope Parameter in A Lineer Ultratructural Model When Measurement Are Not Necessarily Normal. *Journal Of Econometrics*. 78. pp:853-864.

Sucuoğlu, B. ve Kuloğlu, N., (1996). Özürlü Çocuklarla Çalışan Öğretmenlerde Tükenmişliğin Değerlendirilmesi. *Türk Psikoloji Dergisi: Türk Psikologlar Derneği Yayınları*, Cilt:10, No:36.

Swingler, D. N. 1995. On the application of the jackknife to the estimation of the parameters of short multisinusoidal signals using the data-matrix formulation. *Iee Transactions on Signal Processing*, 43(9), 2135-2143.

Terzi, Y. (2015). Ondokuz Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü Regresyon Çözümlemesi Yayınlanmamış Ders Notları, Samsun.

Theil, H. 1950. *A rank-invariant method of linear and polynomial regression analysis*. I. Proc. Kon. Ned. Akad. v. Wetensch. A53, 386-392.

Toka, O., Çetin, M., ve Altunay, S. A. 2011. Basit Doğrusal Regresyonda Sağlam ve Theil Kestiricilerinin Karşılaştırılması. *Tüik, İstatistik Araştırma Dergisi*. Volume: 08. Number: 03. Category: 02. Page: 45-53. ISSN No: 1303-6319.

Topal, M., Yıldız, N., ve Bilgin, Ö. C. (2004). Farklı Dağılış Gösteren Verilerde Parametrik ve Nonparametrik Regresyon Metotlarının İncelenmesi. 4. Ulusal Zootekni bilim Kongresi, s. 556-559.

Tukey, J. W. 1958. Bias and confidence in not-quite large sample. *Ann Math Stat*, *29*, 614.

Tukey, J. W. 1975. *Mathematics and picturing data*. In Proceedings of the 1974 International Congress of Mathematicians (R. James, ed.) 523-531.

Vardi, Y. and Zhang, C. H. 2000. The multivariate *L*1 median and associated data depth. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 1423–1426.

Walsh, B. 2000. Resampling Methods: Randomization Tests, Jackknife and Bootstrap Estimators. Lecture Notes for EEB 596z. http://nitro.biosci.arizona.edu/courses/EEB581-2006/handouts/random.pdf.

Wang, X., Dang, X., Peng, H. and Zhang, H. (2009). The Theil-Sen estimators in multiple linear regression models. Manuscript available at: http:// home. olemiss. edu/~ xdang/ papers/ MTSE. pdf.

Wang, X. 2005. Asymptotics of the Theil-Sen estimator in a simple linear regression model with a random covariate. *Journal of Nonparametric Statistics*. 17:1, 107-120.

Wang, X. and Yu, Q. 2005. Unbiasedness of the Theil–Sen estimator. *Journal of Nonparametric Statistics*. 17:6, 685-695.

Wilcox, R. R. 1998. Simulations on the Theil–Sen regression estimator with right-censored data. *Statistics and Probability Letters*, 39, 43–47.

Wilcox, R. R. 2013. A Heteroscedastic Method for Comparing Regression Lines at Specified Design Points When Using a Robust Regression Estimator. *Journal of Data Science* 11,281-291.

Wu, C. F. J. 1986. Jackknife, Bootstrap and other Resampling Methods in Regression Analysis, *Annals of Statistics*, Vol. 14, No. 4, pp. 1261-129.

Wu, C. F. J. 1990. On the Asymptotic Properties of the Jackknife Histogram. *The Annal of Statistics*. Vol: 18, No: 3. 1438-1452.

Yay, M. (2003). Bootstrap ve Jackknife Yöntemlerinin Otomotiv Sanayi Üzerine Uygulanması. Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı. 130423.

Yıldız, N. ve Topal, M. 2001. Nonparametrik Regresyon Metodlarının İncelenmesi. Atatürk Üniversitesi, *Ziraat Fakültesi Dergisi*, 32(4), 429-435.

Zaman, T. and Alakuş, K. (2015). Some Robust Estimation Methods And Their Applications. *Alphanumeric Journal*. Vol. 3, Issue. 2, pp. 73-82.

Zhang, J. 2002. Some extensions of Tukey depth functions. *J. of Multi. Analysis* 82, 134-165.

Zhang, Z., Liu, T. and Zhang, B. 2016a. Jackknife Emperical Likelihood Inferences for The Population Mean with Ranked Set Samples. *Statistics and Probability Letters* 108, 16-22.

Zhang, Y., Zhang, S., Li, Q. and Zhang, Q. 2016b. Jackknife Empirical Likelihood Test for High Dimensional Regression Coefficients. *Computational Statistics and Data Analysis* 94, 302-316.

Zhou, W. and Serfling, R. 2006. Multivariate spatial U-quantiles: A Bahadur–Kiefer representation, a Theil–Sen estimator for multiple regression, and a robust dispersion estimator. *Journal of Statistical Planning and Inference*, *138*(6), 1660-1678.

Zou, Y. and Serfling, R. 2000. General notions of statistical depth functions. *Ann. Statist.* 28, 461-482.

Zaman, T. and Alakuş, K. 2015. Analysis of the invariance and generalizability of multiple regresssion model results obtained from maslach burnout scale through jackknife method. *Open Journal of Statistics*, 2015, 5, 645-651.

URL-1.http://math.iupui.edu/~hpeng/Talks/TSETalk.pdf (Erişim Tarihi: 10.11.2015)

URL-2.https://cran.r-project.org/web/packages/depth/depth.pdf (Erişim Tarihi: 06.07.2016)

URL-3.https://web.as.uky.edu/statistics/users/pbreheny/621/F10/notes/9-9.pdf (Erişim Tarihi: 10.09.2015).

URL-4. https://www.sahibinden.com/ (Erişim Tarihi: 12.04.2016).

ÖZGEÇMİŞ

Adı ve Soyadı	: Tolga Zaman
Doğum Yeri	: Tonya/ Trabzon
Doğum Tarihi	: 23/09/1989

Eğitim Durumu

Lise	: Trabzon Yunus Emre Lisesi (2006)
Lisans	: Ondokuz Mayıs Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü (2007-2011)
Yüksek Lisans	: Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı (Eylül 2011 – Ağustos 2013)

Çalıştığı Kurum/Kurumlar ve Yıl

Ondokuz Mayıs Üniversitesi İstatistik Bölümü, Araştırma görevlisi (2012- 2014)

Çankırı Karatekin Üniversitesi İstatistik Bölümü Araştırma Görevlisi (2014-2015)

Ondokuz Mayıs Üniversitesi İstatistik Bölümü, Araştırma görevlisi (2015-…)