NEW ERA INTERNATIONAL JOURNAL OF INTERDISCIPLINARY SOCIAL RESEARCHES
ISSN 2757-5608

Article Arrival Date
14.10.2021
Article Type
Research Article
Article Published Date
25.12.2021
Doi Number: http://dx.doi.org/10.51296/newera.133

AYKIRI DEĞER DURUMUNDA BAZI SAĞLAM REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRILMASI

COMPARISON OF SOME ROBUST REGRESSION METHODS IN CASE OF OUTLIER

Öznur İŞÇİ GÜNERİ*
Prof. Dr., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla/Türkiye
ORCID ID: https://orcid.org/0000-0003-3677-7121
Aynur İNCEKIRIK
Dr. Öğreti Üyesi, Manisa Celal Bayar Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü, Manisa/Türkiye
ORCID ID: https://orcid.org/0000-0002-5029-6036
Burcu DURMUŞ
Öğr. Gör, Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla/Türkiye
ORCID ID: https://orcid.org/0000-0002-0298-0802

Özet

Yapılan çalışmalarda veri setinde aykırı değerler ya da gözlemler bulunması istatistiksel analiz sonuçlarını ve modellemeyi önemli ölçüde etkileyebilir. Aykırı değerlere karşı duyarlı olan en küçük kareler yöntemi de varsayımlar sağlanmadığında yanıltıcı sonuçlar verebilmektedir. Bu durumunda çoklu doğrusal regresyona alternatif olarak sunulan sağlam regresyon yöntemleri kullanılmaktadır. Bu çalışmada örnek bir veri seti alınarak gözlem noktaları içerisinde aykırı değerler olması halinde regresyon tahmin edicilerinin veri setini ne kadar açıklayabildiğini araştırmak adına çalışma yapılmıştır. Bu amaçla aykırı gözlem olması durumunda sağlam regresyon yöntemlerinden kantil regresyon yöntemi, kantil regresyon yönteminin özel bir durumu olan en küçük mutlak sapmalar (LAD), sağlam tahminleyiciler arasında yaygın kullanılan M tahmin edicileri ile yüksek kırılma noktasına sahip S ve MM tahmin edicileri kullanılmıştır.

Anahtar Kelimeler: Kantil Regresyon, LAD Regresyon, M Regresyon, S Regresyon, MM Regresyon.

Abstract

The presence of outliers or observations in the data set in the studies can significantly affect the statistical analysis results and modeling. The least squares method which is sensitive to outliers can also give misleading results when the assumptions are not met. In this case robust regression methods which are presented as an alternative to multiple linear regression, are used. In this study a sample data set was taken and a study was conducted to investigate how much the regression estimators could explain the data set in case of outliers within the observation points. For this purpose, in case of outliers, quantile regression method from robust regression methods, smallest absolute deviations (LAD), which is a special case of quantile regression method, commonly used M estimators among robust estimators, with S and MM estimators with high breakpoints were used.

Keywords: Quantile Regression, LAD Regression, M Regression, S Regression, MM Regression.

1. GİRİŞ

Regresyon analizi iki ya da daha fazla değişken arasındaki modelin matematiksel fonksiyonunu inceleyerek parametrelerini tahmin etmeyi amaçlar. Regresyon analizi için en küçük kareler yöntemi (EKKY) yaygın olarak kullanılan en kolay yöntemlerden biridir. Ancak, EKKY tahmin edicilerinin etkin olabilmesi ve modelin tahmin amacıyla kullanılabilmesi için birtakım varsayımları sağlaması gereklidir. Bu varsayımlar; hataların normal dağılım göstermesi, ortalamasının sıfıra eşit stokastik bir değişken olması, eşit varyanslı olması, bağımsız değişkenlerin

tekrar eden örnek değerlerine göre sabit olması, modelin spesifikasyon hatası taşımaması ve bağımsız değişkenler arasında çoklu doğrusal bağlantı olmaması vb. sıralanabilir. Bu varsayımlar sağlanmazsa ana kütle parametrelerinin aralık tahminleri ve hipotez testleri geçerli olmayacaktır (Montgomery, Peck & Vining, 2013).

EKKY varsayımları sağlanırsa tahminciler en iyi doğrusal sapmasız tahminci (EDST)'ler olabilmektedir. EKKY'nin amacı hatalarının kareleri toplamının minimize edilmesidir. Fakat hataların dağılımı normal dağılım göstermediğinde ve aykırı gözlemler olduğunda EKK tahmin edicileri etkinlik özelliklerini kaybederler. Bu durumda sağlam (robust) regresyon yöntemleri önerilmektedir. En küçük mutlak sapmalar (LAD), en küçük medyan kareler (LMedS) ve kantil regresyon yöntemi bunlar arasında yer alır.

Kantil regresyon yöntemi tek değişkenli kantil kavramını verilen bir ya da daha fazla değişken için koşullu kantille genelleştirmektedir. Kantil regresyon yöntemi ya da dilim regresyon yöntemi, bağımlı değişkeni kantillere (%10, %20, …, %90) bölerek her bir kantil için ayrı ayrı tahmin ediciler sunar. Ayrıca bu yöntem aykırı değerlere karşı esnektir. Bu nedenle, yanlış fonksiyonel ilişkilerin oluşturduğu hatayı önler (Çınar, 2019). Araştırmacı aykırı değerlerin durumuna göre 0.05, 0.25, 0.50, 0.75, 0.95 gibi kantiller için farklı modeller kurabilir. Regresyon modelinin hatasını en düşük yapan kantil (dilim) değeri, tahmin modelinde kullanır. Farklı kantil değerlerine göre model kurulduğunda kantil değeri değiştiği için modelin fonksiyonel şekli değişebilmektedir.

Kantil regresyon, değişen varyansın (heteroscedasticity) belirlenmesine imkân vermektedir. Değişen varyans, modelde önemli bir değişkenin modelde bulunmadığı durumlarda ya da aykırı değerlerin bulunduğu durumlarda oluşabilir. Modelde değişen varyans olması halinde tahminciler düşük varyansa sahip olma varsayımını sağlayamazlar. Buna bağlı olarak, varyanslar büyür ve hatalar artar. Eğer hata teriminin dağılımı X'e bağlı değilse tüm kantil regresyonları birbirine paralel olacaktır. Diğer bir deyişle; hata terimlerinin dağılımında herhangi bir varsayım bozulması söz konusu değilse, kantiller her durumda ortanca değere (medyana) aynı uzaklıktadır. Yani regresyon doğrusuna paralel olmaktadır (Yu, Lu & Stander, 2003). Kantil regresyonda sabit varyans durumunda, açıklayıcı değişken katsayıları her bir kantil regresyonda birbiriyle aynı, fakat sabit terim farklı olacaktır. Bunun sonucunda EKK ile bulunan doğrusal regresyon modeli ile medyan regresyon modeli aynı olacaktır (Saçaklı, 2005).

Kantil regresyonda kantil değeri 0.50 olduğunda tahmin ediciler en küçük mutlak sapma (LAD) analizi ile elde edilir (Koenker & Basett, 1978). Kantil regresyon yönteminin en önemli avantajı dağılımının farklı noktalarındaki fonksiyonel ilişkileri ayrı ayrı ortaya koymasıdır. Ayrıca kantil regresyon yöntemi, dağılım ile ilgili güçlü varsayımlar sunması ve kısıt koymaması yönüyle yarı-parametrik bir yöntemdir.

EKKY yöntemi, hesaplama kolaylığı nedeniyle en yaygın olarak kullanılan regresyon yöntemidir. Fakat gözlem değerleri içerisinde aykırı değerler bulunması halinde veri setini temsil eden bir sonuç çıkarmaktan uzaktır. Bu nedenle çeşitli sağlam regresyon yöntemleri ile katsayılar tahmin edilebilir. Bu sağlam regresyon yöntemleri genel olarak 3 ana başlık altında toplanabilir. Bunlar: L tahmin ediciler, M tahmin ediciler ve R tahmin edicileridir. Bu tahmin edicilerin yanı sıra yüksek kırılma noktasına sahip olan LMS, LTS, S ve MM başlıklı sağlam regresyon tahmin edicileri vardır (Toy, 2014).

Bu çalışmada aykırı gözlem olması durumunda sağlam regresyon yöntemlerinden kantil regresyon yöntemi, kantil regresyon yönteminin özel bir durumu olan en küçük mutlak sapmalar

(LAD), sağlam tahminleyiciler arasında yaygın kullanılan M tahmin edicileri ve yüksek kırılma noktasına sahip S ve MM tahmin edicileri analiz amacıyla kullanılmıştır.

2. AYKIRI DEĞER

Temel olarak gözlemler genel dağılımın dışında yer alıyorsa aykırı gözlem (outlier) olarak ifade edilir. Aykırı gözlemler, verinin büyük bir kısmından belirgin bir şekilde sapmış olan değerlerdir. Bu nedenle sapan gözlemler olarak da adlandırılmaktadır. Bu değerler kayıt hatası, ölçüm hatası, üretim hatası, insan hatası gibi nedenlerden dolayı ortaya çıkabilir. Aykırı gözlemler hatalı model tahmini, hatalı parametre tahmini ya da hatalı analiz sonuçlarına neden olabilirler (Liu, Shah & Jiang, 2004). Bu nedenle aykırı değerlerin tespit edilmesi doğru model tahmini ve hatasız analiz için önem taşımaktadır.

Bir gözlemin aykırı değer olup olmadığı birden çok yöntemle incelenebilir. Aykırı değerlerin tespit edilmesi için grafiksel yöntemler (kutu grafiği, histogram, QQ plot, akış dizisi grafiği vb.), standartlaştırılmış artıklar, studentlaştırılmış artıklar, etki ölçüleri (Cook D değeri, DFBETAS, DFFITS) ya da istatistiksel testler (Dixon testi, Nalimov testi, Rosner testi, Weisberg t testi ve Walsh testi vb.) kullanılmaktadır.

Bu testler genel olarak aşağıdaki hipotezi test etmektedir:

$$\begin{cases} Ho: Veri\ seti\ aykırı\ değer\ içermemektedir \\ H1: Veri\ seti\ en\ az\ bir\ tane\ aykırı\ değer\ içermektedir \end{cases}$$

Aykırı değerlerin regresyon tahmin edicisi üzerindeki etkisinin ölçülmesi için kırılma noktası ve etki fonksiyonu gibi sağlamlık kriterleri geliştirilmiştir. Kırılma noktası (Breakdown Point- BP), tahmin edicinin tolerans gösterebileceği en yüksek aykırı değer miktarını gösterir ve bir tahmin edicinin sağlamlığı hakkında fikir verir. Yani, bir tahmin edicinin kırılma noktası %0' dan ne kadar büyükse, o tahmin edicinin sağlam olduğu söylenir. EKK tahmin edicisini sadece bir aykırı değer bile etkilenir. Dolayısıyla EKK tahmin edicisinin kırılma noktası 1/n dir. EKKY de n örneklemi büyüdükçe, 1/n değeri 0' a yaklaşacağından, kırılma noktası da 0' a yaklaşır (Rousseeuw & Leroy, 1987). Bu durumda da EKK tahmin edicisinin sağlam bir tahmin edici olmadığı söylenir. Etki fonksiyonu (Influence Function-IF) veya etki eğrisi (Influence Curve-IC) bir tahmin edicinin herhangi bir noktadaki az bir kirlenme miktarına, nasıl tepki vereceği konusunda kesin bir fikir verir (Staudte & Sheather, 1990).

Yaygın kullanılan aykırı değer tespit değerleri için formüller Tablo 1'de verilmiştir. Tabloda k bağımsız değişken sayısını ve n gözlem sayısını göstermektedir.

Tablo 1. Aykırı Değer Ölçüm Değerleri

Ölçüm Değerleri	Ölçüm	Değer
Kaldıraç nokta (etkili nokta)	leverage	> (2k+2) /n
Standartlaştırılmış artıklar	stdresid	> 2
Studentlaştırılmış artıklar	rstudent	> 2
Cook D değeri	Cook's D	> 4/n
Uyumlar arası uzaklık	DFITS	> 2*sqrt(k/n)
Uyumlar arası fark	DFBETA	> 2/sqrt(n)

Aykırı değerleri tespit ettikten sonra, analizin güvenilirliğini etkilememeleri için farklı yöntemler kullanılmaktadır. Aykırı değerler veri kümesinden çıkarılabilir, yeni değerler atayarak sınırlandırılır ya da dönüştürme işlemi yapılabilir. Örnek büyüklüğü yeterince büyük olduğunda veri setinin istatistiksel yöntemler ile analizi sırasında aykırı değerler analiz dışı bırakılabilir. Fakat

örnek büyüklüğü küçük olduğunda tek bir gözlemin bile analiz sonuçlarına katkısı önemli olabilir. Bu nedenle özellikle küçük örneklerde aykırı değerlerin tespiti ve giderilmesi önem arz eder.

Ayrıca aykırı değer tespiti mikro dizin verileri ve klinik biyokimya verileri gibi büyük veri setlerinin kalite kontrolü ve ilaç endüstrisi için büyük öneme sahiptir (Ovla & Taşdelen, 2012).

3. REGRESYON MODELLERİ

3.1 Doğrusal Regresyon Modeli (EKK)

Regresyon modelinde iki veya daha fazla bağımsız değişken ile bağımlı değişken arasındaki ilişki araştırılır. Y bağımlı değişken ve x_1, x_2, …, x_k bağımsız değişkenler olmak üzere Eşitlik-1 ile verilen modele çoklu doğrusal regresyon modeli denir.

$$Y_i = \sum_{i=1}^{k} \beta_i x_i + u \qquad 1)$$

Modelin regresyon katsayıları $\beta_0, \beta_1, \ldots, \beta_k$ tahminleri genellikle en küçük kareler yöntemi (EKKY) ile elde edilir. Bu modelde u hata terimini gösterir. Hata terimi u sıfır ortalamalı, normal dağılımlı ve eşit varyanslı u ~N $(0, \sigma_u^2)$ olmalıdır. Tahmin edilen model Eşitlik-2'deki gibidir:

$$\hat{Y}_i = \sum_{i=1}^{k} \hat{\beta}_i x_i + e_i \qquad 2)$$

En yaygın kullanılan ve bilinen yöntemlerden biri çoklu doğrusal regresyondur. Varsayımlar sağlandığında oldukça güçlü bir istatistiksel yöntemdir. Güvenilir tahminler elde edilmekte ve tahmin ediciler arzu edilen özelliklere sahip olabilmektedir. Bu yöntemde amaç ortalamadan sapmaların kareleri toplamının minimum yapılmasıdır (Eşitlik-3).

$$min \sum e_i^2 = (Y_i - \hat{Y}_i)^2 \qquad 3)$$

Ancak hata teriminin normal dağılmaması, çoklu doğrusal bağlantı sorunu, değişen varyans sorunu ve otokorelasyon gibi varsayımların sağlanmaması durumunda parametre tahminleri geçerli olmaz. Doğrusal regresyon modelinin önemli varsayımlarından biri de sabit varyansdır (homoscedasticity). Bu varsayıma göre hata terimlerinin varyansı, bağımsız değişkenlerdeki değişmelere bağlı değildir. Yani, bağımsız değişkendeki değişimlerden etkilenmemektedir. Sabit varyans Eşitlik-4 ile ifade edilmektedir:

$$var(u) = E[u - E(u)]^2 = E(u_i)^2 = \sigma_u^2 \qquad 4)$$

Burada hata terimi ile bağımsız değişken arasında ilişki yoktur. Sabit varyans varsayımının her zaman sağlanması mümkün değildir. Bu durumda değişen varyans (heteroscedasticity) durumu ile karşılaşılır. Değişen varyans durumunda hataların varyansları aynı kalmayıp bağımsız değişkenin değişiminden etkilenir. Değişen varyans Eşitlik-5 ile ifade edilir:

$$var(u) = E(u_i)^2 = \sigma_{ui}^2 \qquad 5)$$

Burada *i* indisi hata terimi varyanslarının farklı olduğunu göstermektedir. Yani varyanslar bağımsız değişkenle birlikte değişmektedir. Değişen varyans sonucunda EKK tahmincileri sapmasız ve tutarlı tahminciler olma özelliklerini korumakta ancak minimum varyanslı ve etkin olma özelliğini kaybetmektedir. Ayrıca tahmincilerin t ve F testleri anlamlarını yitirmektedirler. Bu durumda başka tahmin yöntemleri ile daha güvenilir tahminler elde etmek mümkündür. Değişen

varyansın tespiti için grafik yöntemleri ve Spearman Sıra Korelasyon Testi, Goldfeld-Quandt Testi, Glejser Testi, Park Testi, Barlett Testi ve Breusch-Pagan Testi gibi testler kullanılabilir.

Ayrıca veri kümesinde aykırı değer varsa, bu aykırı değerleri veriden çıkararak veya oldukları gibi dahil ederek EKKY kullanmak yanlış sonuçlar verebilir. Bu durumda aykırı değerlerin etkisini azaltacak regresyon yöntemlerinin kullanılması daha güvenilir sonuçlar elde etmeyi sağlayacaktır.

3.2 En Küçük Mutlak Sapmalar (Least Absolute Deviation; LAD)

En küçük mutlak sapmalar (LAD) yöntemi ilk kez 1757 tarihinde Roger Joseph Boscovich tarafından ileri sürülmüş daha sonra geliştirilmiş bir yöntemdir (Birkes & Dodge, 1993). LAD yöntemi, EKK varsayımları sağlanmadığı durumlarda alternatif sağlam regresyon yöntemlerden biridir. EKK yönteminde yalnızca gözlenen ve tahmin edilen uzaklıklar dikkate alındığında herhangi bir istatistiksel varsayım gerektirmez. Fakat aykırı değer varlığında veya hata teriminin varyansı sabit olmadığı zaman EKK kullanılması sonuçlar açısından tutarlı olmayabilir. Özellikle gerçek verilerin ele alındığı çalışmalarda aykırı gözlem varlığında, bu gözlemlerin verinin kalan kısmı ile özdeş dağılması beklenemez. Aykırı gözlemler verinin büyük bir kısmından uzakta bulunan değerlerdir. Aykırı değerlerin veri setinin büyük bir kısmının sahip olduğu dağılımdan farklı bir dağılıma sahip olduğu ya da farklı parametreler ile aynı dağılıma sahip olduğu varsayılır (Yorulmaz, 2009).

Bu yöntemde amaç diğer yöntemler de olduğu gibi veri kümesine en uyumlu ve mutlak sapmaların toplamını en küçük yapacak doğruyu belirlemektir. Bu işlem için net bir formül bulunmamaktadır. Tahmin yöntemi bir algoritma yardımıyla yapılmaktadır (Birkes & Dodge,1993).

EKK yönteminde parametrelerin tahmin edicilerini tahmin ederken hata karelerin toplamını minimize etmek esasken bu yöntemde hataların mutlak değerlerinin toplamını minimize etmek esastır (Ocak, 2019). Bu durum Eşitlik-6 ile ifade edilmektedir;

$$min \sum |e_i| \qquad 6)$$

Buradan, Eşitlik-7'deki fark bulunmak istenmektedir

$$\sum |e_i| = \sum |y_i - (\hat{\beta}_0 + \hat{\beta}_i X_i)| \qquad 7)$$

Burada amaç hataların mutlak değerleri toplamını minimize etmektir. Bu yöntem EKKY' ne göre daha kolaydır. Bunun nedeni EKKY hesaplanırken belli formüller ve algoritmalar kullanılırken, LAD hesaplanırken hiçbir formül bulunmamasından kaynaklanmaktadır. Bunun sonucunda yapılacak tek işlem mutlak sapmaların toplamını bulmak için bir algoritma oluşturulmasıdır. Bu algoritmanın temel amacı yine veri kümesine en uygun doğrunun seçilmesi üzerine kurulmalıdır. Bu nedenle sağlam regresyon yöntemleri içerisinde LAD en çok kullanılan tahmin yöntemlerinden biridir. Aynı zamanda bu yöntem kantil regresyon yönteminin özel bir durumudur. Kantil Regresyon analizinde, kantil değerinin τ=0.5 olması durumunda LAD regresyon elde edilmektedir.

LAD yöntemi, veriler %50'ye kadar aykırı değer içerse bile iyi tahminler veren güçlü bir regresyon yöntemidir (Rousseeuw & Leroy, 1987). Ancak bu yöntemde, artıkların medyan değeri en küçük yapılmaya çalışılırken geriye kalan (n–1) adet gözlem dikkate alınmaz. Bu nedenle örneklem büyüklüğü arttıkça regresyon katsayılarının kestiriminde LAD yöntemi EKK yöntemi kadar etkili olmamaya başlar (Ryan, 1997). Ayrıca LAD tahmincisi elde edilirken EKK tahmincisindeki gibi analitik çözüm olmadığından çeşitli iteratif yaklaşımlar kullanılır.

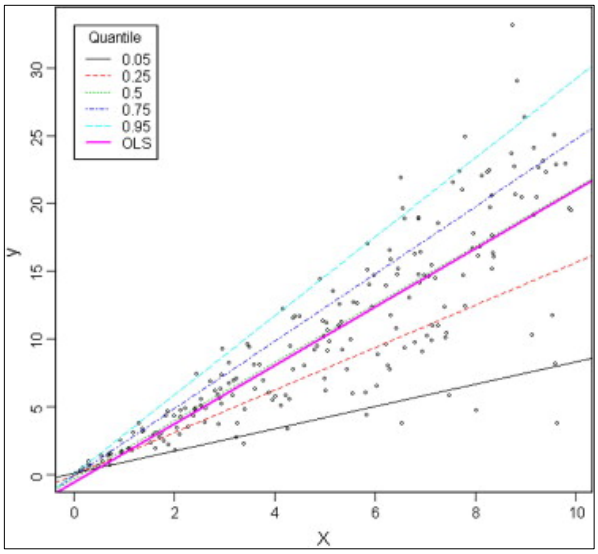
3.3 Kantil Regresyon

Koşullu kantil fonksiyonları kantil regresyonun başlangıç noktasıdır (Angrist & Pishke, 2008). Kantil regresyon modelleri koşullu ortalama ve koşullu kantil fonksiyonlarını tahmin etmek için kullanılır. Kantil Regresyon yöntemi ilk olarak Koenker ve Bassett (1978) tarafından iklim çalışmaları için önerilmiş bir regresyon modelidir. Doğrusal regresyon modeli EKKY ile tahmin edildiğinde varsayımlar sağlanmadığında tahminciler etkin değildir. Bu durumda kullanılacak alternatif regresyon modellerinden biri kantil regresyon yöntemidir.

Kantil regresyon yönteminde EKKY bulunan hataların dağılımının normal olması ve varyansın homojen olması varsayımı gerekli değildir. Bu nedenle bu yöntem EKKY'ne göre daha esnek bir yöntemdir. Kantil regresyon yöntemi farklı kantillere göre (örneğin 0.05, 0.10, …, 0.95) Y'nin koşullu dağılımının farklı bölgelerdeki ortak değişkenlerin etkilerini tahmin eder. Çoklu doğrusal regresyon doğrusu aşırı değerleri tespit etmede başarısız olurken, kantil regresyon doğruları farklı kantillere sahip olduğundan dolayı aşırı değerleri daha kolay tespit edebilir (Wang, 2007).

Kantil regresyon yönteminin, küresel iklim değişiklikleri, ekonomi, biyoloji, tıp, mühendislik, ücret eşitsizlikleri, gelir düzeyinin belirlenmesi, yaşam analizi gibi farklı alanlarda uygulamaları yapılmaktadır.

Uygulamalarda kantil değerleri çoğunlukla 0.25, 0.50 ve 0.75 olarak seçilir. Şekil 1'de farklı kantil değerlerine göre (0.05; 0.25; 0.50; 0.75; 0.95) regresyon grafiği görülmektedir.



Şekil 1. Farklı Kantil Değerlerine Göre Kantil Regresyon Grafiği

EKK ile tahmin edilen regresyon doğrusu dağılımın orta bölgesinden geçmektedir ve bu tahmin edicide uç değerler dikkate alınmamaktadır. Ancak kantil regresyonda 0.25'lik ve 0.75'lik dilimlerde bulunan tahmin değerleri aynı değildir. Böylece her bir bağımsız değişkenin ilgili değişkeni nasıl etkilediği konusunda daha eksiksiz bilgi edinilebilir (Babu & Hallam, 2017). Kantil Regresyon, özellikle koşullu kantillerin değişkenlik gösterdiği durumlarda kullanışlıdır. Doğrusal kantil regresyon modeli Eşitlik-8'deki gibi yazılır;

$$Q_{y_t}(\tau) = \sum_{i=1}^{k} \beta_{\tau,i} x_{ti} \qquad 8)$$

$\beta_{\tau,i}$ bilinmeyen parametreler ve τ kantil değerini gösterir. 0<τ<1 olmak üzere $Q_{y_t}(\tau)$ ise y_t'nin τ.'nin koşullu kantilini gösterir. Örneğin, τ=0.50 alınırsa $Q_{y_t}(0.50)$ dağılımın medyanını

ifade eder. τ=0.90 değeri bağımlı değişkenin en yüksek %90'lık kantil içerisinde yer aldığını, τ=0.10 değeri ise bağımlı değişkenin en düşük %10'luk kantil içerisinde yer aldığını gösterir. Regresyon katsayıları, asimetrik bir mutlak kayıp fonksiyonu kullanılarak tahmin edilir.

Katsayı tahminleri doğrusal regresyona benzer olarak görülmektedir. Fakat kantil regresyonda bağımlı değişkenin koşullu dağılımının farklı noktaları için tahmin yapılmaktadır. Bu modelde τ değerine göre kantil regresyon olmak üzere kantil regresyon tahmincileri doğrusal programlama şeklinde ifade edilip simpleks algoritması ile çözülebilir. F dağılım fonksiyonuna sahip bağımlı değişken Y için τ. regresyon kantili Eşitlik-9 ifadesinin minimize edilmesi ile elde edilir (Koenker & Bassett, 1978).

$$min_{\beta \epsilon R^K} \frac{1}{n}\left[\sum_{i\epsilon\{i:i\geq x_i\beta\}} \tau|y_i - x_i\beta| + \sum_{t\epsilon\{t:y_t<x_i\beta\}} (1-\tau)|y_i - x_i\beta|\right] \quad 9)$$

Kantil regresyonun bu şekilde gösterimi doğrusal programlama gösterimidir. Kantil regresyon parametre tahminleri, açıklayıcı değişkendeki bir birim değişme karşısındaki y'nin belli bir katilindeki değişmeyi gösterir (CSCU, 2007).

3.4. M Regresyon

EKK yöntemi, hata terimlerinin normal dağıldığı varsayımı ve olabilirlik fonksiyonunun (hata kareler toplamının) minimize edilmesiyle elde edilir. M tahminciler aynı fikirle hataların normal dağılmadığı durumlarda (çarpık, kirlenmiş, basık, uzun kuyruklu vb.) farklı bir fonksiyon kullanarak MLE tahmini yapar (Büyükkör & Şehirlioğlu, 2020). M tahmin edicisi ilk olarak Peter J. Huber (1964) tarafından sunulmuştur. M tahmincisi, y değişkeni aykırı değer içeriyorsa EKKY daha verimlidir. EKKY tahmin edicisinde amaç hata terimlerinin kareleri (e_i^2) toplamını minimum yapmak iken M tahmin edicisinde amaç hata terimlerinin fonksiyonu olan $\rho(e)$ fonksiyonunu minimum yapmaktır. Fox 2002'de $\rho(e)$ fonksiyonunun sahip olduğu bazı özellikleri ifade etmiştir. Bunlar:

- $\rho(e) \geq 0$
- $\rho(e) = 0$
- $\rho(e) = \rho(-e)$
- $|e_i| > |e_{i'}|$ *için* $\rho(e_i) > \rho(e_{i'})$ *dir. EKKY tahmini için* $\rho(e_i) = e_i^2$ *olur.*

Yukarıda sayılan özellikleri sağlayan bir $\rho(e_i)$ fonksiyonu varlığında M tahmin edicisinin amaç fonksiyonu Eşitlik-10'daki gibi tanımlanır.

$$min \sum_{i=1}^{n} \rho(e) = min \sum_{i=1}^{k} \rho(y_i - x_i\theta) \quad 10)$$

Eşitlik (10)'da ρ fonksiyonu sürekli ve türevi alınabilen bir fonksiyon özelliği taşımaktadır. Böylece eşitlik (10)'daki θ'ya göre ρ fonksiyonunun türevi alınacak olursa Eşitlik-11 elde edilir:

$$\sum_{i=1}^{k} \psi(e_i)x_i = 0 \quad 11)$$

Bu eşitlikte ρ fonksiyonunun türevini temsil eden fonksiyon ψ fonksiyonudur. Bu fonksiyonun gösterimi de $\psi(e_i) = \frac{\partial}{\partial\theta}\rho(e)$ şeklinde olmaktadır.

M tahmin ediciler için tanımlanan çeşitli fonksiyonlar söz konusudur. Genellikle dayanıklı (robust) regresyon analizinde hesaplama kolaylığı ve matematiksel olarak nispeten daha anlaşılır olması nedeniyle Huber ve Tukey M tahmincileri yaygın olarak kullanılmaktadır. Bunlar;

• *Huber'in tanımladığı ρ fonksiyonu,*

Huber (1964) tarafından geliştirilen M-tahminleme yöntemi için amaç fonksiyonu Eşitlik-12'de verilmektedir:

$$\rho(e) = \begin{cases} \frac{e^2}{2}, & |e| \le k \\ k|e| - \frac{k^2}{2}, & |e| > k \end{cases} \quad 12)$$

Bu fonksiyonda bulunan k sabiti dönüm noktası (turning constant) olarak adlandırılır. Fonksiyonun türevi alınacak olursa Huber'in ağırlık fonksiyonu Eşitlik-13'teki gibi olmaktadır (Huber, 1981). Burada k=1.345'dir.

$$\psi(e) = \begin{cases} e, & |e| \le k \\ ksign(e), & |e| > k \end{cases} \quad 13)$$

•Tukey'in tanımladığı ρ fonksiyonu,

Beaton ve Tukey (1974) tarafından geliştirilen Bisquare (Biweight) amaç fonksiyonu Eşitlik-14 ile ifade edilmektedir:

$$\rho(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[\left(\frac{e}{k}\right)^2\right]^3\right\}, & |e| \le k \\ \frac{k^2}{6}, & |e| > k \end{cases} \quad 14)$$

Huber M tahmincideki gibi *k,* dönüm noktası olarak kabul edilir ve değeri 4.685'tir. Bu fonksiyonun türevi alınacak olursa Tukey'in ağırlık fonksiyonu elde edilir (Eşitlik-15) (Fox, 2002).

$$\psi(e) = \begin{cases} \left[1 - \left[\left(\frac{e}{k}\right)^2\right]\right]^2, & |e| \le k \\ 0, & |e| > k \end{cases} \quad 15)$$

3.5. S Regresyon

S tahmin edicileri ilk olarak Rousseeuw ve Yohai (1984) tarafından ortaya atılmış tahmin edicilerdir. M tahmin edicilerinin eksik yönü, sadece medyanı kullanmasından dolayı dağılımı dikkate almamasıdır. Medyanın bu zayıflığını önlemek için S tahmin edicilerinde medyan yerine standart sapma kullanılır ve artıkların dağılımı minimize edilmeye çalışılır. S tahmin edicisi Eşitlik-16'daki gibi ifade edilir:

$$min_{\hat{\theta}} S[e_1(\theta), \dots, e_n(\theta)] \quad 16)$$

ve ölçek(scale) tahmincisi Eşitlik-17 ile elde edilir.

$$\hat{\theta} = S[e_1(\hat{\theta}), \dots, e_n(\hat{\theta})] \quad 17)$$

$S[e_1(\theta), \dots, e_n(\theta)]$ dağılımı Eşitlik-18'in çözümü olarak kabul edilir.

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = K \quad 18)$$

K genellikle $E_\phi[\rho]$ ifadesine eşit olacak biçimde alınır. Burada ϕ 'da standart normal dağılımı göstermektedir.

Eğer Eşitlik 18'in çözümü birden çok ise bu durumda çözümlerin supremumu alınır ve gösterimi Eşitlik-19'daki gibi olur (Rousseeuw & Leroy, 1987; Rousseeuw & Yohai,1984).:

$$S(e_1, \dots, e_n) = \left\{ s; \left(\frac{1}{n}\right) \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = K \right\} \quad 19)$$

Yukarıda eşitlikte kullanılan ρ fonksiyonu aşağıdaki özellikleri sağlaması gerekmektedir:

- ρ , $\rho(0) = 0$ *ve sürekli türevlenebilen bir fonksiyon olmalıdır.*
- ρ , *[0,c) aralığında kesinlikle artan, [c, ∞) aralığında da sabit olacak bir fonksiyon için bir c (c > 0) değeri mevcut olmalıdır.*
- $\frac{K}{\rho(c)} = \frac{1}{2}$ *eşitliğini sağlamalıdır (Rousseeuw & Leroy, 1987).*

Tukey'in iki ağırlıklı fonksiyonu yukarıdaki 3 özelliği gösteren bir ρ fonksiyonudur ve Eşitlik-20'deki gibi ifade edilmektedir.

$$p(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^2}{6c^2}, & |x| \le c \\ \frac{c^2}{6}, & |x| > c \end{cases} \quad 20)$$

S tahmin edicilerin kırılma noktası için $n \to \infty$ olabilecek en yüksek değer olan %50'ye ulaşır ve Eşitlik-21 ve 22 ile hesaplanır:

$$S(e_1, \dots, e_n) = \left\{ s; \left(\frac{1}{n}\right) \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = K \right\} \quad 21)$$

$$\epsilon^* = \frac{K}{\rho(c)} \quad 22)$$

S tahmin edicileri bazı özellikleri göstermesi gerektiğinden hesaplanması zor ve uzun süren bir yöntemdir.

3.6. MM Regresyon

Yüksek kırılma noktasına sahip olan MM tahmin edicileri Yohai (1987) tarafından sunuldu. MM tahmin edicileri üç aşamada tanımlanır:

- *İlk önce yüksek kırılma noktalı bir θ^* tahmini hesaplanır. Bu işlem için sağlam tahmin edicinin etkinlik göstermesine gerek yoktur.*
- *Daha sonra $e_i(\theta^*)$ artıklarından S_n gibi kırılma noktası %50 olan bir M ölçek tahmini, elde edilir.*
- *Son olarak, bir MM tahmin edicisi $\hat{\theta}$, Eşitlik-23'ün herhangi bir çözümü olarak ifade edilir.*

$$\sum_{i=1}^{k} \psi\left(\frac{e_i(\theta)}{s_n}\right) x_i = 0 \quad 23)$$

Buradaki ifade, Eşitlik-24 ifadesini sağlar.

$$S(\theta) = S(\theta^*) \quad 24)$$

$S(\theta)$ ise, Eşitlik-25 ile hesaplanır (Rousseeuw & Leroy, 1987).

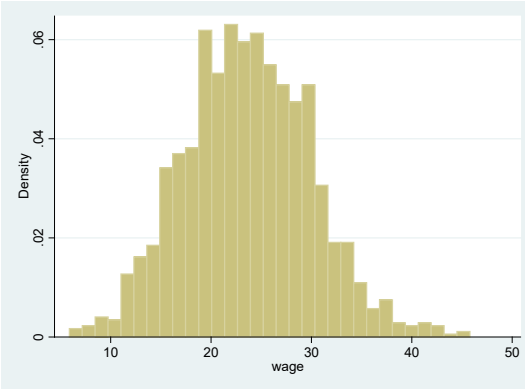
$$S(\theta) = \sum_{i=1}^{k} \psi(\frac{e_i(\theta)}{s_n})x_i \qquad 25)$$

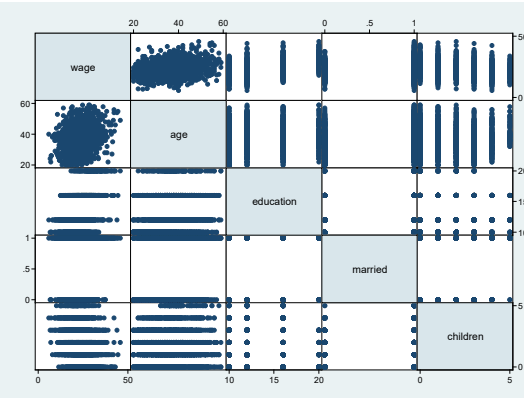
4. UYGULAMA

Aykırı gözlem durumunda EKKY yöntemi ve sağlam regresyon tekniklerinden Kantil regresyon LAD, M, S ve MM tahminlerini karşılaştırmak amacıyla gerçek bir veri seti ele alınmıştır. Veri setine http://www.stata-press.com/ adresinden ulaşılabilir. Bu çalışmada bağımlı değişken ücret (haftalık olarak alınmış, çalışmıyorsa kayıp veri) ve bağımsız değişkenler yaş, eğitim düzeyi, medeni durumu ve çocuk sayısıdır. 2000 kadın çalışanın bilgisinden oluşan veri setinde 657 gözlemin ücret bilgileri kayıptır. Analizler için SPSS 22 ve STATA 13 paket programı kullanılmıştır. Değişkenlere ilişkin tanımlayıcı istatistikler Tablo 2'de ve ücret değişkeninin histogramı ile değişkenlerin dağılımı Şekil 2'de verilmiştir.

Tablo 2. Tanımlayıcı İstatistikler

Değişkenler	n	Ortalama	Std. Sapma	Min	Maks.
Ücret	1343	23.69217	6.305374	5.88497	45.80979
Yaş	2000	36.208	828.656	20	59
Eğitim düzeyi	2000	13.084	3.045.912	10	20
Medeni durum	2000	.6705	.4701492	0	1
Çocuk sayısı	2000	1.6445	1.398963	0	5





Çalışan Kadınların Ücret Dağılımı (n=1343)

Değişkenlerin Dağımı

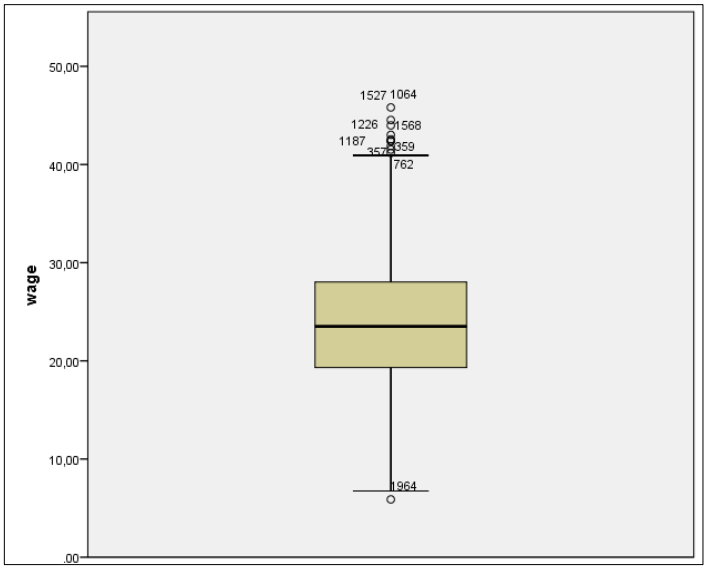
Şekil 2. Çalışan Kadınların Dağılımı

Tablo 3'te ücret değişkeni için aykırı gözlemler verilmektedir. Bu veri setinde 11 adet aykırı gözlem tespit edilmiştir.

Tablo 3. Aykırı Gözlemler

Gözlem	Ücret
1964	5.885
762	41.182
1286	41.525
359	41.868
357	42.376
1187	42.408
1729	42.565
1226	43.016
1568	43.979
1527	44.534
1064	45.810

Şekil 3'te Box plot kutu grafiğinde aykırı gözlemler görülmektedir.



Şekil 3. Aykırı Gözlemler için Box Plot Grafiği

Tablo 4 ile aykırı değer ölçümleri verilmiştir. Bu sonuçlara göre en az bir stdresid ve rstudent kalıntının 2'den büyük olduğu (793, 1901 ve 1568 gözlemler) görülmektedir. Ayrıca en az bir Cork'un mesafesi 4/1343 = 0.002'den daha büyüktür.

Tablo 4. Aykırı Değer Ölçümleri

Gözlem	stdresid	rstudent	dfbeta_1	cooksd	leverage
1964	-2.860	-2.867	-0.012	0.005	0.003
658	-2.809	-2.816	0.007	0.003	0.002
1019	-2.569	-2.574	-0.132	0.005	0.004
1625	-2.518	-2.523	-0.003	0.002	0.002
1865	-2.498	-2.503	-0.002	0.002	0.002
1527	2.893	2.901	0.157	0.008	0.005
357	2.905	2.913	0.146	0.010	0.006
793	3.002	3.011	0.036	0.005	0.003
1901	3.721	3.739	-0.226	0.015	0.005
1568	3.728	3.746	0.205	0.013	0.005

Aykırı değerlerin bulunduğu durumlarda değişen varyans problemi oluşabilir. Değişen varyans olması durumunda t ve F testleri anlamsızlaşır. Tahminciler, en düşük varyansa sahip olma özelliğini kaybeder. Böylece varyanslar büyür, hatalar artar. Çalışmada değişen varyans problemi, Breusch-Pagan testi ile test edilmiştir. Değişen varyans, hata teriminin varyansının tüm gözlem değerleri için eşit olmadığı durumlarda ortaya çıkmaktadır. Değişen varyans durumu ile önemli bir parametrenin model dışında kalması, aykırı değerlerin olması ya da hatalı model kurma gibi durumlarda karşılaşılır. Değişen varyans problemi değişken dönüşümü yapma, modele önemli bir parametre ekleme veya ağırlıklı EKK yöntemi kullanılarak giderilebilir (Bager & Odah, 2017). Değişen varyans için sıfır ve alternatif hipotezler şu şekildedir:

{Ho: Değişen varyans yoktur (sabit varyans)
{H1: Değişen varyans mevcuttur

Tablo 5. Breusch-Pagan Testi Çıktısı

Breusch-Pagan Test	Breusch-Pagan Test Değeri	p Değeri
Breusch-Pagan / Cook-Weisberg test for heteroscedasticity	chi2(4) = 16.23	Prob > chi2 = 0.0027

Modelde sabit varyans varsayımının geçerliliği Breusch-Pagan/Cook-Weisberg testi ile incelenmiştir. Hesaplanan değer 0.05'ten küçük olduğundan dolayı değişen varyans (heteroscedasticity) yoktur ve artıklar sabit varyansa sahiptir şeklinde oluşturulan sıfır hipotezi red edilmiştir (Tablo 5). Yani değişen varyans olduğu sonucuna ulaşılmıştır.

4.1. Doğrusal Regresyon Sonuçları

Çalışmada ilk adım olarak verilere doğrusal regresyon analizi uygulanmıştır. Katsayıların anlamlılığı için α=0.05 olarak kabul edilmiştir. Tablo 5 ile modele ilişkin varyans analizi tablosu verilmiştir. Belirlilik katsayısı R^2 çok düşük bulunmuştur.

Tablo 6. Veri Seti için Varyans Analiz Tablosu

Source	SS	df	MS		
				Number of obs =	1343
				F(4, 1338) =	128.55
Model	14812.5356	4	3703.1339	Prob > F =	0.0000
Residual	3854.23591	1338	28.8059485	R-squared =	0.2776
				Adj R-squared =	0.2755
Total	53354.8946	1342	39.7577456	Root MSE =	53.671

EKKY ile bulunan model katsayıları ve katsayı anlamlılık değerleri Tablo 7'de verilmiştir. Tablo 7'deki sonuçlara baktığımızda ücret, eğitim düzeyi ve çocuk değişkenlerinin katsayıları istatistiksel olarak anlamlı (p<0.05) iken medeni durum değişkeninin istatistiksel olarak anlamlı olmadığı (p>0.05) görülmektedir. Eğitim düzeyi ve yaş artışı ücret düzeyini arttırmaktadır.

Tablo 7. EKKY için Sonuçlar

Ücret	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
Yaş	.1514818	.0192717	7.86	0.000	.1136757	.1892879
Eğitim düzeyi	.8750694	.050243	17.42	0.000	.7765057	.973633
Medeni durum	-.5395024	.3574519	-1.51	0.131	-1.24073	.1617247
Çocuk sayısı	-.6862982	.1032256	-6.65	0.000	-.8887997	-.4837966
Sabit	7.934369	.9264515	8.56	0.000	6.116914	9.751825

Tahmin edilen denklem:

Ücret = 7.934 + 0.151 Yaş + 0.875 Eğitim düzeyi - 0.539 Medeni durum - 0.686 Çocuk sayısı

Elde edilen bu denklemin tahmin yapma amacıyla kullanılabilmesi için hataların normal dağılması ve aykırı değer içermemesi gerekmektedir. Ayrıca değişen varyans problemi, EKK yönteminde önemli bir sorundur. Çünkü hata terimi varyansı tüm gözlemler için eşit olmalıdır. Bu nedenle sağlam yöntemler ile analizlere devam edilmiştir.

4.2 En Küçük Mutlak Sapmalar (LAD) Regresyon Sonuçları

Kantil regresyonda .50 kantilde bulunan sonuçlar aynı zamanda medyan regresyon sonuçlarıdır. %50 kantil düzeyinde değer yaratmada medeni durum dışında tüm katsayıların istatistiksel olarak anlamlı olduğu görülmektedir ($p<0,05$). Tablo 8'de LAD sonuçları yer almaktadır.

Tablo 8. LAD için Sonuçlar

Median regression					Number of obs	= 1343
Raw sum of deviations	6.789.794 (about 23.511223)					
Min sum of deviations	5.749.108				Pseudo R2	= 0.1533
Ücret	**Coef.**	**Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Yaş	.1653391	.0238382	6.94	0.000	.1185748	.2121035
Eğitim düzeyi	.8940777	.0621482	14.39	0.000	.7721592	1.015996
Medeni durum	-.5971263	.4421509	-1.35	0.177	-1.464511	.2702581
Çocuk sayısı	-.6206084	.1276851	-4.86	0.000	-.8710931	-.3701237
sabit	6.906066	1.145976	6.03	0.000	4.657961	9.154172

Tahmin edilen denklem:

Ücret =6.609 + 0.165 Yaş + 0.894 Eğitim düzeyi - 0.597 Medeni durum - 0.620 Çocuk sayısı

4.3 Kantil Regresyon Sonuçları

Özellikle aykırı gözlemlerin önemli olduğu çalışmalarda kantil regresyonun kullanımı yararlı olmaktadır. Çünkü kantil regresyon yöntemi aykırı değerlere karşı esnektir ve bu yöntemin aykırı değerlerden etkilenecek varsayımları bulunmaz. Kantil regresyon yöntemi, değişen varyans durumu ile karşılaşıldığı durumlarda EKK yöntemine alternatif olarak geliştirilmiştir. Kantil regresyon kullanımını doğrulamak için bir değişen varyans testi yapmamız gerekiyor. Breusch-Pagan test istatistiğinin sıfırdan önemli ölçüde farklı olduğunu bulduk. Bu nedenle değişen varyansa sahibiz ve kantil regresyon kullanımında haklı olduğumuzu söylebiliriz. Tablo 9'da %25, %50 ve %75 değerleri için kantil regresyon analizi sonuçları verilmiştir.

Tablo 9. Kantil Regresyon için Sonuçlar

Bootstrap replications (100)						
1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5						
.....		50				
.....		100				
Simultaneous quantile regression		Number of obs	=	1.343		
bootstrap(100) SEs		.25 Pseudo R2	=	0.1515		
		.50 Pseudo R2	=	0.1533		
		.75 Pseudo R2	=	0.1562		
Ücret	**Coef.**	**Bootstrap Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
q25						
Yaş	.1222662	.0247427	4.94	0.000	.0737276	.1708049
Eğitim düzeyi	.9350447	.0510873	18.30	0.000	.8348249	1.035.265
Medeni durum	-1.100328	.3834612	-2.87	0.004	-1.852579	-.3480777

Çocuk sayısı	-.7710757	.1247789	-6.18	0.000	-1.015859	-.5262922
Sabit	5.113283	.9059688	5.64	0.000	3.336009	6.890557
q50						
Yaş	.1653391	.0245549	6.73	0.000	.1171689	.2135094
Eğitim düzeyi	.8940777	.0671558	13.31	0.000	.7623356	1.02582
Medeni durum	-.5971263	.4189391	-1.43	0.154	-1.418975	.2247228
Çocuk sayısı	-.6206084	.1418143	-4.38	0.000	-.8988109	-.3424059
Sabit	6.906066	1.261865	5.47	0.000	4.430617	9.381515
q75						
Yaş	.1497587	.024656	6.07	0.000	.1013902	.1981272
Eğitim düzeyi	.8252121	.0617805	13.36	0.000	.7040148	.9464093
Medeni durum	-.354559	.4628394	-0.77	0.444	-1.262.529	.5534108
Çocuk sayısı	-.6372464	.1050533	-6.07	0.000	-.8433335	-.4311592
Sabit	12.01653	1.098063	10.94	0.000	9.862419	14.17065

Kantil regresyonda iki tür anlamlı katsayı vardır: sıfırdan önemli ölçüde farklı olanlar ve EKK katsayılarından önemli ölçüde farklı olan kantil katsayıları (EKK güven aralığının dışında). Kantil regresyon sonuçları doğrusal regresyon sonuçları ile aynı şekilde yorumlanır. Fakat anlamlı olan değişkenler farklılık gösterebilir. Tablo 9'daki sonuçlara baktığımızda %25'de bütün değişkenler anlamlı iken %50 ve %75 de ücret, eğitim düzeyi ve çocuk sayısı değişkenlerinin anlamlı olduğu, medeni durum değişkenin anlamsız olduğu görülmektedir.

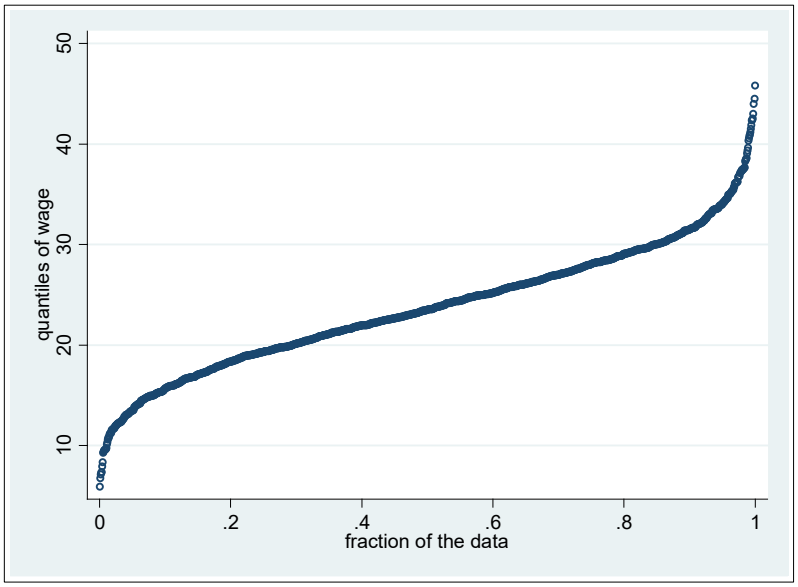
Tahmin edilen denklemler şu şekildedir:

$Ücret_{(Q25)}$ =5.113 + 0.122 Yaş + 0.935 Eğitim düzeyi - 1.100 Medeni durum -0.0771 Çocuk sayısı

$Ücret_{(Q50)}$ =6.609 + 0.165 Yaş + 0.894 Eğitim düzeyi - 0.597 Medeni durum -0.620 Çocuk sayısı

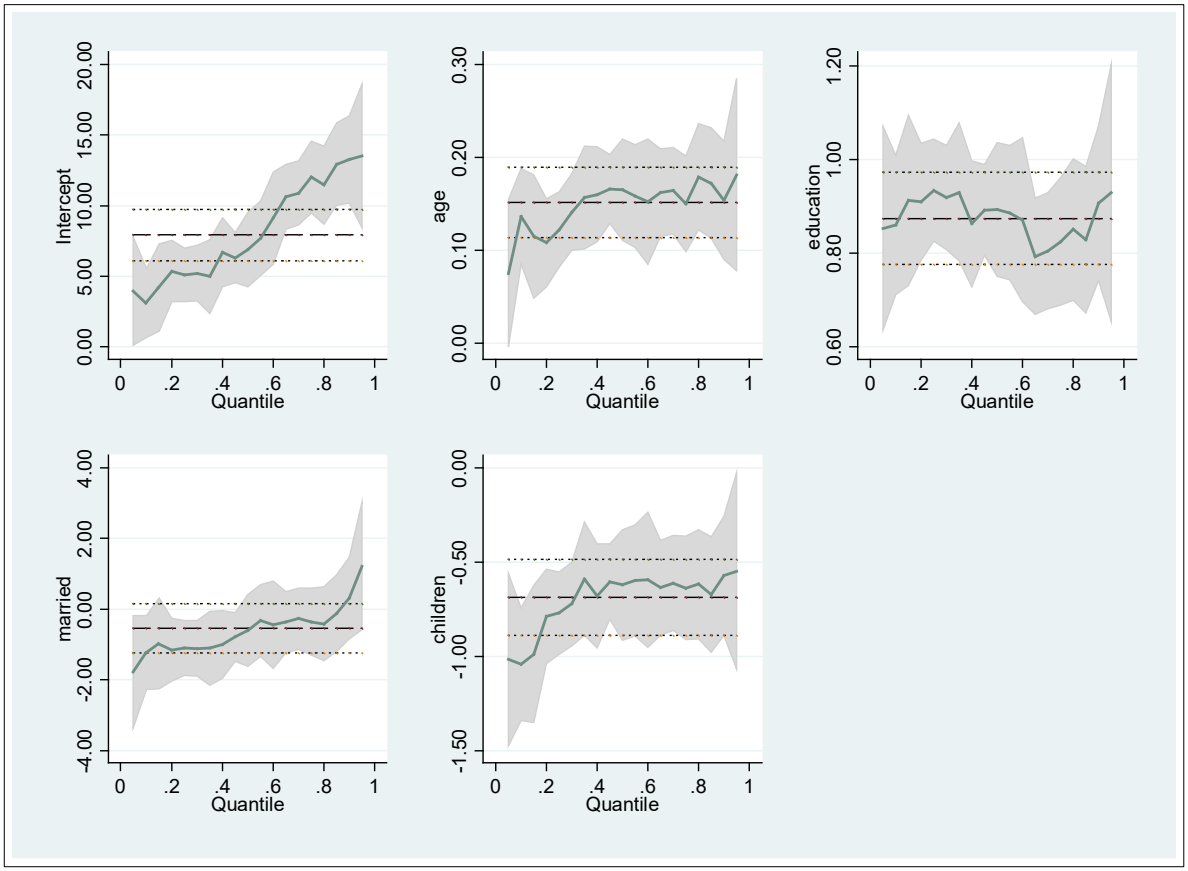
$Ücret_{(Q75)}$ =12.016 + 0.149 Yaş + 0.825 Eğitim düzeyi - 0.354 Medeni durum -0.637 Çocuk sayısı

EKKY aykırı değerlere karşı hassastır. Kantil regresyon, bağımsız değişkenlerin etkilerinin kantillere göre farklılık göstermesine izin verir. Örneğin, kantil (.25), koşullu dağılımın 25. yüzdelik dilimini (ilk çeyrek) tanımlayan parametreleri tahmin eder.



Şekil 4. Ücret Değişkenin Kantillere Göre Serpilme Diyagramı

Şekil 4'te kantillere göre serpilme diyagramı verilmiştir. Buna göre ücret çok düşük değerlerle başlıyor sonra yükseliyor. Şekil 5 ile her bir değişken için kantiller gösterilmektedir.



Şekil 5. Kantil Regresyon Katsayıları

Bağımlı değişkenin kantilleri yatay eksende, katsayı büyüklükleri dikey eksendedir. EKKY katsayıları kesikli çizgilerle (- - - -), katsayı çizgisi etrafında iki yatay çizgi (… … …) güven aralığını gösterir. EKKY katsayısı kantillere göre değişmez. Kantil regresyon katsayıları (kalın düz çizgiler), etraflarında güven aralıklarıyla kantiller arasında değişen çizgiler olarak çizilir. Kantil katsayısı EKKY güven aralığının dışındaysa, kantil ve EKK katsayıları arasında önemli farklar vardır. Ücret değişkeni yaş arttıkça yüksek kantillerde artmaktadır.

4.4. M Regresyon

M tahmin ediciler için Huber'in M tahmincisi, Tukey'in Biweight, Hampel'in M tahmincisi ve Andrews'un tahmincileri şu şekildedir (Tablo 10):

Tablo 10. Huber'in M Tahmincisi, Tukey'in Biweight Tahmincisi, Hampel'in M Tahmincisi ve Andrews'un Tahmincisi

M-Estimators	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
Ücret	23,5789	23,5071	23,5681	23,5042
a The weighting constant is 1,339.				
b The weighting constant is 4,685.				
c The weighting constants are 1,700, 3,400, and 8,500				
d The weighting constant is 1,340*pi.				

Huber ve Hampel'in k tahmin değerleri ile Tukey ve Andrew'un k tahmin değerleri birbirine yakın sonuçlar verdiği görülmektedir.

Tablo 11'de M regresyon için elde edilen sonuçlar verilmiştir. Buna göre medeni durum değişkeni anlamsız bulunmuştur.

Tablo 11. M Regresyon Sonuçları

M regression (95% efficiency)			Number of obs	=	1343	
			Wald chi2(4)	=	511.98	
			Prob > chi2	=	0.0000	
			Pseudo R2	=	0.2451	
			Huber k	=	1.3449975	
			Scale	=	5.3174692	
Ücret	**Coef.**	**Robust Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Yaş	.1515969	.0191875	7.90	0.000	.1139561	.1892378
Eğitim düzeyi	.8722044	.0521869	16.71	0.000	.7698274	.9745815
Medeni durum	-.6231222	.3514987	-1.77	0.076	-1.312.671	.0664264
Çocuk sayısı	-.679178	.1055953	-6.43	0.000	-.8863284	-.4720276
Sabit	7.975086	.9020128	9.84	0.000	6.205573	9.7446

Tahmin edilen denklem:

Ücret = 7.975 + 0.151 Yaş + 0.872 Eğitim düzeyi - 0.623 Medeni durum -0.679 Çocuk sayısı

4.5. S Regresyon

Tablo 12'de S regresyon için elde edilen sonuçlar verilmiştir. Buna göre medeni durum değişkeni yine anlamsız bulunmuştur. S tahminciler için etkinlik %28.7 olarak elde edilmiştir.

Tablo 12. S Regresyon Sonuçları

S regression (28.7% efficiency)			Number of obs	=	1343	
			Wald chi2(4)	=	160.87	
			Prob > chi2	=	0.0000	
			Pseudo R2	=	0.1204	
			Breakdown point	=	50	
			Biweight k	=	1.547645	
			Scale	=	5.3935471	
Ücret	**Coef.**	**Robust Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Yaş	.1683392	.0465501	3.62	0.000	.07702	.2596584
Eğitim düzeyi	.8963295	.1468705	6.10	0.000	.608208	1.184451
Medeni durum	-.8772324	.6833825	-1.28	0.199	-2.21785	.4633855
Çocuk sayısı	-.4549098	.2247165	-2.02	0.043	-.8957449	-.0140747
Sabit	6.636596	1.934008	3.43	0.001	2.842578	1.043061

Hausman test of S against LS: chi2(4) = 3.1377663 Prob > chi2 = 0.5350

Tahmin edilen denklem:

Ücret = 6.636 + 0.168 Yaş + 0.896 Eğitim düzeyi - 0.877 Medeni durum - 0.454 Çocuk sayısı

4.6. MM Regresyon

Tablo 13'te MM regresyon için elde edilen sonuçlar verilmiştir. Buna göre medeni durum değişkeni anlamsız bulunmuştur. MM regresyon için etkinlik %85 olarak bulunmuştur.

Tablo 13. MM Regresyon Sonuçları

MM regression (85% efficiency)	Number of obs	=	1343
	Wald chi2(4)	=	457.93
	Prob > chi2	=	0.0000
	Pseudo R2	=	0.2143

			Breakdown point	=	50	
			M-estimate: k	=	3.4436898	
			S-estimate: k	=	1.547645	
			Scale	=	5.3935471	
Ücret	**Coef.**	**Robust Std. Err.**	**t**	**P>t**	**[95% Conf.**	**Interval]**
Yaş	.1525892	.0204528	7.46	0.000	.1124662	.1927122
Eğitim düzeyi	.8734498	.0560005	15.60	0.000	.7635915	.9833081
Medeni durum	-.7139103	.367768	-1.94	0.052	-1.435375	.0075542
Çocuk sayısı	-.6589443	.1119495	-5.89	0.000	-.87856	-.4393286
Sabit	7.897596	.9611279	8.22	0.000	6.012115	9.783078

Hausman test of MM against S: chi2(4) = 3.0446457 Prob > chi2 = 0.5504

Tahmin edilen denklem:

Ücret = 7.897 + 0.152 Yaş + 0.873 Eğitim düzeyi - 0.713 Medeni durum - 0.658 Çocuk sayısı

5. SONUÇLAR

Regresyon analizi, istatistiksel çalışmalarda en yaygın kullanılan yöntemlerden biridir. Regresyon analizi, bir bağımlı değişkenin bir veya birden fazla bağımsız değişkenler arasındaki ilişkinin matematiksel bir fonksiyonu şeklinde yazılması olarak tanımlanabilir. Bu yöntem için EKKY yaygın olarak kullanılmaktadır. Fakat varsayımlar (hata teriminin dağılımın normal olması, otokorelasyon olmaması, eşit varyans olması vb.) sağlanmadığında elde edilecek sonuçlar güvenilir olmaz. Ayrıca gözlem değerlerinde aykırı gözlem olup olmaması da tahminlerin güvenirliliğini etkileyecektir. Bilindiği üzere EKK tahmin edicisi aykırı değerlere karşı hassasiyet gösterir ve güvenilir sonuçlar vermekten uzaktır. Sağlam regresyon tahmin edicilerinin aykırı değerlere karşı duyarlılığı azdır. Bu nedenle veri setinde aykırı değerler olsa bile sağlam regresyon tahmin edicileri etkili ve güvenilir sonuçlar vermektedir. Veri setinden aykırı değerler çıkarıldığında en küçük kareler tahmin edicisi de, sağlam regresyon tahmin edicileri gibi güvenilir sonuçlar vermektedir. Bu nedenle istatistiksel çalışmalarda değişkenler arasındaki ilişkiyi daha iyi açıklayabilmek için verinin yapısına göre farklı regresyon modelleri tercih edilmektedir.

Bu çalışmada sağlam regresyon yöntemlerinden, kantil regresyon analizi, kantil regresyonun özel bir durumu olan LAD yöntemi, M, S ve MM regresyon modelleri üzerinde çalışılmıştır. Hataların normal dağılmadığı ya da veri kümesinin aykırı değerlere sahip olması durumunda LAD yöntemi diğer klasik tahmin yöntemlerine göre üstünlük göstermektedir. LAD yöntemi kantil regresyonun özel bir durumudur. Kantil regresyonda %50 için bulunan değerler aynı zamanda LAD tahmin değerlerini vermektedir. Bu çalışmada, farklı oranlarda kantil regresyon modelleri, değerlendirilmiştir. Uygulama sonuçlarına baktığımızda; gözlem değerlerinde aykırı gözlem olması durumunda kantil regresyon ile doğrusal regresyona göre daha anlamlı katsayılar elde edildiği söylenebilir.

Gözlem değerleri arasında aykırı gözlemler olması durumunda bu değerlerin çıkarılması bilgi eksikliğine neden olabilir. Ayrıca gözlem değerleri yetersiz ve az sayıda olabilir. Bu gibi durumlarda sağlam yöntemler tercih edilmelidir. Bunun yanısıra çalışmada kullanılan verilerde değişen varyans durumu tespit edilmiştir ki zaten varsayım sağlanmamış olur. Bu nedenle sağlam regresyon yöntemlerinin kullanılması daha uygundur diyebiliriz. Bu çalışma veri setinden aykırı değerler çıkarılarak tekrar edilebilir.

KAYNAKLAR

Angrist, J. D. ve Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

Babu, S. C, Gajanan, S. N. ve Hallam, A. (2017). *Nutrition Economics: Principles and Policy Applications*. Academic Press, Boston.

Bager, A. S. M., Odah, M. H. ve Mohammed, B. K. (2017), Using Approach Quantile Regression to Determine the Factors Affecting Measuring Capacity in Iraq. *American Review of Mathematics and Statistics*, 5(1): 35-44.

Beaton, A. E. ve Tukey, J. W. (1974). The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data. *Technometrics*, 16(2), 147-185.

Birkes, D. ve Dodge, Y. (1993). *Alternative Methods of Regression*. NY: Wiley.

Büyükkör, Y. ve Şehirlioğlu, A. K. (2020). Dayanıklı (Robust) Regresyon: Karşılaştırmalı Simülasyon Çalışması. *Avrupa Bilim ve Teknoloji Dergisi*, 18, 188-195.

Çınar, U. K. (2019). En Küçük Kareler Regresyonuna Alternatif Bir Yöntem: Kantil Regresyon. *Avrasya Uluslararası Araştırmalar Dergisi*, 7(18), 57-71.

Fox, J. (2002). *Robust Regression. Appendix to An R and S-PLUS Companion to Applied Regression.* Sage Publications, Thousand Oaks, CA.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35, 73-101.

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

Koenker, R. W. ve *Bassett,* G. *(1978),* Regression Quantiles. *Econometrica*, 46(1), 33-50.

Liu, H., Shah, S. ve Jiang, W. (2004). On-line Outlier Detection and Data Cleaning. *Computers & Chemical Engineering*, 28(9), 1635-1647.

Montgomery, D. C., Peck, E. A. ve Vining, G. G. (2013). *Doğrusal Regresyon Analizine Giriş*. Ankara Nobel Akademik Yayıncılık.

Ocak, F. (2019). *En Küçük Kareler ve En Küçük Mutlak Sapmalar Yöntemlerinin Simülasyon Verileri ile Karşılaştırılması* (Yüksek Lisans Tezi). Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü.

Ovla, H. D. ve Taşdelen, B. (2012). Aykırı Değer Yöntemi. *Mersin Üniversitesi Sağlık Bilimleri Dergisi*, 5(3), 1-8.

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

Rousseeuw, P. J. ve Yohai, V. (1984). Robust regression by means of S-estimators, in Robust and Nonlinear Time Series Analysis. edited by J. Franke, W. Hardle, and R.D. Martin, *Lecture Notes in Statistics*. No.26, Spinger Verlag, New York, 256-272.

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

Saçaklı, İ. (2005). Kantil Regresyon ve Alternatif Regresyon Modelleri ile Karşılaştırması (YayımlanmamışYüksek Lisans Tezi), Marmara Üniversitesi, İstanbul.

Stata Press (2021). http://www.stata-press.com (Erişim Tarihi: 12.06.2021).

Staudte, R. G. ve Sheather, S. J. **(**1990). *Robust Estimation and Testing*. John Wiley and Sons, New York.

The Cornell Statistical Consulting Unit (CSCU) (2007). https://www.cscu.cornell.edu/, (Erişim Tarihi: 12.03.2021).

Toy, A. (2014). Sağlam Regresyon Tahmin Edicilerinin İncelenmesi ve Bir Uygulama (Yüksek Lisans Tezi). Fırat Üniversitesi, Fen Bilimleri Enstitüsü.

Wang, H. (2007). Quantile Regression: Overview and Applications to Risk Assessment. *North Caroline State University*, 1-26.

Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*.

Yorulmaz, Ö. (2009). Dayanıklı Regresyon Yöntemi ve Çeşitli Sosyal Veriler Üzerinde Aykırı Gözlemlerin Teşhisi. *Balikesir University Journal of Social Sciences Institute*,12(21).

Yu, K., Lu, Z. ve Stander, J. (2003). Quantile Regression: Applications and Current Research Areas. *Journal of Royal Statistical Socienty D (The Statistician)*. 52, 331-350.