

**T.C.
SİNOP ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANA BİLİM DALI**



**BASİT DOĞRUSAL REGRESYON MODELİNDE MODEL PARAMETRELERİNİN
DAYANIKLI TAHMİNİ**

ÖMER BAYRAKTAR

YÜKSEK LİSANS TEZİ

**DANIŞMAN
DOÇ. DR. DEMET HAN AYDIN**

SİNOP-2025

TEZ KABUL

ÖMER BAYRAKTAR tarafından hazırlanan “**BASİT DOĞRUSAL REGRESYON MODELİNDE MODEL PARAMETRELERİNİN DAYANIKLI TAHMİNİ**” başlıklı bu çalışma, **17.10.2025** tarihinde yapılan savunma sınavı sonucunda başarılı bulunarak, jürimiz tarafından oybirliği ile **YÜKSEK LİSANS tezi** olarak kabul edilmiştir.

Başkan	Prof. Dr. Fatma Zehra DOĞRU Giresun Üniversitesi / Fen-Edebiyat Fakültesi	İmza
Üye (Danışman)	Doç. Dr. Demet HAN AYDIN Sinop Üniversitesi /Fen-Edebiyat Fakültesi	İmza
Üye	Dr. Öğr. Üyesi Olcay ALPAY Sinop Üniversitesi /Fen-Edebiyat Fakültesi	İmza

Enstitü Müdürü

Prof. Dr. Onur ŞAHİN

.....

ETİK BEYANI

Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında; tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi, tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu, tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi, kullanılan verilerde herhangi bir değişiklik yapmadığımı, bu tezde sunduğum çalışmanın özgün olduğunu, Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Faaliyetlerinde Üretken Yapay Zekâ Kullanımına Dair Etik Rehber İlkelerine uyduğumu bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Ömer BAYRAKTAR

ÖZET

YÜKSEK LİSANS TEZİ

BASIT DOĞRUSAL REGRESYON MODELİNDE MODEL PARAMETRELERİNİN DAYANIKLI TAHMİNİ

ÖMER BAYRAKTAR

SİNOP ÜNİVERSİTESİ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

DANIŞMAN: DOÇ. DR. DEMET HAN AYDIN

Bu tez çalışmasında, basit doğrusal regresyon modelinde model parametrelerinin En Küçük Kareler (Least Squares-LS) tahmin edicilerinin performansı farklı senaryolar altında incelenmiştir. İlk senaryoda, hata terimlerinin dağılımı olarak normal dağılıma alternatif olabilecek farklı olasılık dağılımları ve çeşitli aykırı değer modelleri dikkate alınmıştır. İkinci senaryoda ise, hata terimlerinin sağa çarpık bir dağılım olan Gumbel dağılımını izlediği varsayılmıştır. Ayrıca, X-yönünde aykırı gözlemlerin bulunduğu durumlarda, LS tahmin edicilerinin performansları, literatürde yaygın olarak kullanılan En Küçük Mutlak Sapma (Least Absolute Deviation-LAD), Ağırlıklı En Küçük Mutlak Sapma (Weighted Least Absolute Deviation-WLAD) ve En Küçük Medyan Kareler (Least Median of Squares-LMS) gibi dayanıklı (robust) tahmin yöntemleri ile karşılaştırılmıştır.

Tahmin yöntemlerinin etkinliğini değerlendirmek amacıyla Monte-Carlo simülasyon tekniği kullanılmış; karşılaştırma ölçütleri olarak ise yan (Bias) ve hata kareler ortalaması (Mean Squared Error-MSE) kriterleri esas alınmıştır.

Simülasyon çalışmalarından elde edilen bulguları desteklemek amacıyla, literatürde yer alan gerçek veri seti ile analiz gerçekleştirilmiştir. Hem simülasyon hem de gerçek veri analizi sonuçları, hata terimlerinin normal dağılmadığı veya veri setinde aykırı gözlemlerin bulunduğu durumlarda, LS tahmin edicilerinin performansının önemli ölçüde azaldığını ortaya koymuştur. Buna karşın, dayanıklı tahmin yöntemlerinin bu tür veri bozulmalarına karşı daha kararlı ve güvenilir sonuçlar ürettiği belirlenmiştir.

Ayrıca, LMS tahmin edicisinin, model parametrelerinin tahmininde varsayım ihlalleri karşısında diğer yöntemlere kıyasla daha yüksek dayanıklılık gösterdiği sonucuna ulaşılmıştır. WLAD tahmin edicisi ise performans açısından LMS yönteminin ardından en başarılı tahmin yöntemi olarak tespit edilmiş olup, LMS'ye alternatif olarak tercih edilebilecek güvenilir bir seçenek durumundadır.

Sonuç olarak, veri setinin aykırı değerler içermesi veya hata terimlerinin normal dağılmaması durumlarında, model parametrelerinin tahmininde klasik LS tahmin yöntemi yerine LMS dayanıklı tahmin yönteminin kullanılması önerilmektedir.

ANAHTAR KELİMELER: Gumbel dağılımı; Aykırı değer; Dayanıklı tahmin.

Ekim, 2025; 55 Sayfa

ABSTRACT

MASTER'S THESIS

ROBUST ESTIMATIONS OF MODEL PARAMETERS IN SIMPLE LINEAR REGRESSION MODEL

ÖMER BAYRAKTAR

SINOP UNIVERSITY INSTITUTE OF GRADUATE PROGRAMS

DEPARTMENT OF STATISTICS

SUPERVISOR: ASSOC. PROF. DR. DEMET HAN AYDIN

In this thesis, the performance of the Least Squares (LS) estimators for model parameters in a simple linear regression model was examined under different scenarios. In the first scenario, various probability distributions that could serve as alternatives to the normal distribution, as well as several outlier models, were considered for the distribution of the error terms. In the second scenario, it was assumed that the error terms follow the right-skewed Gumbel distribution. Furthermore, in cases where outliers existed in the X-direction, the performance of LS estimators was compared with robust estimation methods commonly discussed in the literature, including the Least Absolute Deviation (LAD), Weighted Least Absolute Deviation (WLAD), and Least Median of Squares (LMS) estimators.

To evaluate the efficiency of the estimation methods, the Monte-Carlo simulation technique was employed, and the Bias and Mean Squared Error (MSE) criteria were used as performance measures.

To support the findings obtained from the simulation studies, a real dataset from the literature was also analyzed. The results of both the simulation and real data analyses revealed that the performance of LS estimators significantly deteriorates when the error terms deviate from normality or when the dataset contains outliers. In contrast, robust estimation methods were found to produce more stable and reliable results under such data contamination conditions.

Moreover, the findings indicated that the LMS estimator exhibits greater robustness against assumption violations compared to other estimation methods when estimating model parameters. The WLAD estimator demonstrated the second-best performance after LMS, making it a reliable alternative to the LMS method.

In conclusion, it is recommended to use the LMS robust estimation method instead of the ordinary LS estimation method for estimating model parameters in cases where the dataset contains outliers or the error terms do not have a normal distribution.

KEYWORDS: Gumbel distribution; Outlier; Robust estimation

October, 2025; 55 Page

TEŞEKKÜR

Tez çalışmam sırasında yaptığım mesleğin zorluklarını anlayan, karşılaştığım güçlükleri aşmamda bana yardımcı olan, değerli katkı ve önerileri ile beni yönlendiren, tez, makale ve seminer konusunda bana her konuda yardımcı olan danışmanım Sayın Doç. Dr. Demet HAN AYDIN'a, tezin değerlendirilmesi aşamalarında değerli yorumları ile bana farklı bakış açıları kazandıran jüri üyesi Giresun Üniversitesi öğretim üyesi Sayın Prof. Dr. Fatma Zehra DOĞRU'ya ve Sinop Üniversitesi öğretim üyesi Sayın Dr. Öğretim Üyesi Olcay ALPAY'a, bana her türlü fedakârlığı gösteren ve beni destekleyen sevgili eşim Yıldız BAYRAKTAR'a ve iki çocuğuma sonsuz teşekkür ederim.

Ömer BAYRAKTAR

İÇİNDEKİLER

Sayfa

TEZ KABUL	ii
ETİK BEYAN	iii
ÖZET	iv
ABSTRACT	v
TEŞEKKÜR	vi
İÇİNDEKİLER	vii
TABLolar DİZİNİ	ix
ŞEKİLLER DİZİNİ	x
SİMGELER VE KISALTMALAR DİZİNİ	xi
1. GİRİŞ	1
2. TEMEL KAVRAMLAR	4
2.1. Dayanıklı Tamir Edici	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.2. Aykırı Gözlem	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.3. Olasılık Yoğunluk Fonksiyonu	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.4. Dağılım Fonksiyonu	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.5. k . Moment	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.6. Normal Dağılım	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.7. Student T Dağılımı	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.8. Sürekli Uniform (Sürekli Tekdüze) Dağılım	HATA! YER İŞARETİ TANIMLANMAMIŞ.
2.9. Gumbel Dağılımı	HATA! YER İŞARETİ TANIMLANMAMIŞ.
3. MATERYAL VE YÖNTEM	13
3.1. Ölçülebilir Model	13
3.2. Basit Doğrusal Regresyon	13
3.3. En Küçük Kareler Yöntemi	15
3.4. En Küçük Mutlak Sapma Yöntemi	15
3.5. Ağırlıklı En Küçük Mutlak Sapma Yöntemi	16
3.6. En Küçük Medyan Kareler Yöntemi	16

4. BASİT DOĞRUSAL REGRESYON MODELİNDE NORMALLİK VARSAYIMINDAN SAPMALAR OLDUĞUNDA VEYA VERİ SETİ AYKIRI DEĞER İÇERDİĞİNDE PARAMETRELERİN TAHMİNİ	17
4.1. Parametre Tahmini	17
4.1.1. En küçük kareler tahmin edicileri	17
4.1.2. En çok olabilirlik tahmin edicileri	18
4.2. Simülasyon Çalışması I	20
4.3. Uygulama I	23
5. BASİT DOĞRUSAL REGRESYON MODELİNDE GUMBEL DAĞILIMI VARSAYIMINDA VE VERİ SETİ AYKIRI DEĞER İÇERDİĞİNDE PARAMETRELERİN TAHMİNİ	28
5.1. Parametre Tahmini	28
5.1.1. En küçük mutlak sapma tahmin edicileri.....	28
5.1.2. Ağırlıklı mutlak sapma tahmin edicileri	29
5.1.3. En küçük medyan kareler tahmin edicileri	29
5.2. Simülasyon Çalışması II	30
5.3. Uygulama II	34
6. BULGULAR.....	37
6.1. Simülasyon Çalışması I ve Uygulama I.....	37
6.2. Simülasyon Çalışması II ve Uygulama II.....	38
7. SONUÇ VE ÖNERİLER.....	40
KAYNAKLAR.....	42
ÖZGEÇMİŞ	44

TABLÖLAR DİZİNİ

Sayfa

Tablo 4.1 Basit doğrusal regresyon modelinde parametrelerin LS tahmin edicilerinin istatistiksel dayanıklılığı.....	22
Tablo 4.2 Standartlaştırılmamış artıklar Kolmogorov-Smirnov test istatistiği sonuçları.....	23
Tablo 4.3 Farklı aykırı değer senaryolarına göre elde edilen parametre tahminleri ve bu tahminlerin standart hataları.....	27
Tablo 5.1 Farklı örneklem büyüklükleri için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri ve verilerin %0 ve %5 oranında aykırı değer içermesi durumunda değişen örneklem büyüklüklerine karşılık gelen Bias ve MSE ve değerleri.....	32
Tablo 5.2 Farklı örneklem büyüklükleri için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri ve verilerin %10 ve %15 oranında aykırı değer içermesi durumunda değişen örneklem büyüklüklerine karşılık gelen Bias ve MSE değerleri.....	33
Tablo 5.3 Orijinal veriler ve farklı aykırı değer yüzdesi değerlerine sahip değiştirilmiş veriler için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri.....	36

ŞEKİLLER DİZİNİ

Sayfa

Şekil 2.1 Farklı konum parametresi (μ) değerlerine göre çizdirilen Normal dağılım oyf grafikleri.....	7
Şekil 2.2 Farklı ölçek parametresi (σ^2) değerlerine göre çizdirilen Normal dağılım oyf grafikleri.....	8
Şekil 2.3 Farklı serbestlik dereesi (ν) değerlerine göre çizdirilen Student-t dağılımı oyf grafikleri.....	10
Şekil 2.4 Farklı parametre değerlerine göre çizdirilen sürekli Uniform dağılımı oyf grafikleri.....	11
Şekil 2.5 Bazı seçilmiş σ değerleri için Gumbel dağılımının farklı yoğunluk fonksiyonları.....	12
Şekil 4.1 Orijinal veriye ilişkin regresyon doğrusu ve saçılım (scatter) grafiği.....	24
Şekil 4.2 X-yönünde bir tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği.....	25
Şekil 4.3 Y-yönünde bir tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği.....	25
Şekil 4.4 Hem X-yönünde hem de Y-yönünde birer tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği.....	26
Şekil 4.5 13. gözlemin hem X-yönünde hem de Y-yönünde birer tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği.....	26
Şekil 5.1 $T = 1$, $n = 100$ ve $\xi = 0,05$ için elde edilen simülasyon verilerinin saçılım grafiği.....	31
Şekil 5.2 Rüzgâr hızı verileri için Gumbel dağılımı QQ ve uyum grafikleri.....	34
Şekil 5.3 $\xi = 0,10$ olan veriler için dağılım grafiği.....	35
Şekil 5.4 (a) Veride aykırı değer bulunmadığında, (b) veride %5 oranında aykırı değer değer bulunduğunda, (c) veride %10 oranında aykırı değer bulunduğunda, (d) veride %15 oranında aykırı değer bulunduğunda β_0 ve β_1 parametrelerinin LS, LAD, WLAD ve LMS tahminleri ile elde edilen regresyon doğrusu ve saçılım grafiği.....	36

SİMGELER VE KISALTMALAR DİZİNİ

X	: Rasgele değişken
k	: Şekil parametresi
λ	: Ölçek parametresi
$erf(\cdot)$	: Gauss hata fonksiyonu
$F(x)$	: X rasgele değişkenin birikimli dağılım fonksiyonu
$F^{-1}(x)$	: Dağılım fonksiyonunun tersi
$f(x)$	: X rasgele değişkenin olasılık yoğunluk fonksiyonu
ϕ_1	: Çarpıklık katsayısı
ϕ_2	: Basıklık katsayısı
$E(X)$	: X rasgele değişkenin beklenen değeri
$\text{Var}(X)$	: X rasgele değişkenin varyansı
$\text{Med}(X)$	: X rasgele değişkenin medyan değeri
$\text{Mod}(X)$	: X rasgele değişkeninin modu
$M_X(t)$	: X rasgele değişkeninin moment çıkaran fonksiyonu
$E(X^k)$	: Dağılımın k . Momenti
$L(\cdot)$	: Olabilirlik fonksiyonu
$I(\cdot, \cdot)$	: Fisher bilgi matrisi
$B(\cdot, \cdot)$	: Beta fonksiyonu
γ	: EÜ'ler sabiti
$\Gamma(\cdot)$	: Gamma fonksiyonu
$I_{\nu/\gamma^2(x)+\nu}(\frac{\nu}{2}, \frac{1}{2})$	: Düzenli hale getirilmiş (regülenizde) tam olmayan Beta fonksiyonu
$w(x)$	: Ağırlık fonksiyonu
$f'(x)$	: $f(x)$ 'in x 'e göre birinci türevi
R^2	: Belirtme katsayısı
$\mathcal{R}$	: Reel sayılar
BLUE	: En İyi Lineer Yansız Tahmin Edici
Dg	: Dağılım fonksiyonu
ML	: En Çok Olabilirlik
LS	: En Küçük Kareler
MSE	: Hata Kareler Ortalaması
KS	: Kolmogorov-Smirnov Testi
Oyf	: Olasılık yoğunluk fonksiyonu

1. GİRİŞ

Regresyon analizi, değişkenler arasındaki ilişkiyi modelleyerek istatistiksel sonuç çıkarımları elde etmeyi sağlayan ve yaygın olarak kullanılan bir yöntemdir. Regresyon modelinde hata terimlerinin normal dağılımlı olduğu geleneksel bir varsayımdır. Ancak, pratik uygulamalarda normal olmayan dağılımlarla çok daha sık karşılaşılır (Zeckhauser ve Thompson, 1970; Tiku vd. 2001; Tiku ve Akkaya, 2004; Islam ve Tiku, 2005; Li, 2016; Pek vd. 2018; Acitas ve Senoglu, 2020).

En küçük kareler tahmin yöntemi, herhangi bir dağılım varsayımı olmaksızın, kolay uygulanabilirliği nedeni ile parametre tahmininde yaygın olarak kullanılmaktadır. Ancak en küçük kareler tahmin edicilerinin etkinliği, hata terimlerinin dağılımı normallikten uzaklaştıkça hızla düşer. Bu durumda, parametre tahminlerinin standart hataları da fazla tahmin edilir ve bu da nominal olandan daha büyük bir önem düzeyine sahip testlere ve oldukça geniş güven aralıklarının oluşmasına neden olur (Cox ve Hinkley, 1968; Bartolucci ve Scaccia, 2005). Ne yazık ki, uygulamalarda normallik varsayımının ihlal edilmesinin bu tahmin edicilerin etkinliği üzerindeki etkisi genellikle ihmal edilir.

Normal olmayan verilerle başa çıkmanın bir yolu, verilerin dağılımını normalleştirmeyi veya en azından yaklaşık olarak normalleştirmeyi amaçlayan Box-Cox dönüşümüdür (Box ve Cox, 1964). Ancak, normal olmayan verilerin tümü bu dönüşüme uygun olmayabilir. Ayrıca, dönüşümleri yorumlamak genellikle zor olduğundan, orijinal verilerle herhangi bir normalleştirici dönüşüme tabi tutmadan çalışmak tercih edilir (Bickel ve Doksum, 1981).

Birçok çalışmada, normallik varsayımının sağlanmadığı koşullarda regresyon analizinde model parametrelerinin tahmin edilmesi problemi incelenmiştir. Islam vd. (2001) doğrusal regresyon modelinde hataların asimetric dağılımlardan Weibull ve genelleştirilmiş lojistik dağılımlar olduğu varsayımı için parametrelerin tahmininde En Çok Olabilirlik (Maximum Likelihood-ML) yönteminin değiştirilmiş bir versiyonunu kullanmışlardır. Islam ve Tiku (2005) çoklu doğrusal regresyon modelinde, normal olmayan simetrik hata dağılımlarına sahip olunması durumunda parametrelerin tahminlerini, değiştirilmiş ML, En Küçük Kareler (Least Square-LS) ve M tahmin yöntemlerine göre elde etmişlerdir. Al-Adilee ve Mohamed (2014) hataların Gumbel dağıldığı varsayımı altında, doğrusal regresyon modelini incelemişler ve model parametrelerinin ML tahminlerini elde etmişlerdir. Acitas ve Senoglu (2020) hataların uzun kuyruklu simetrik dağıldığı varsayımında iki fazlı doğrusal regresyon modelinde

model parametrelerini tahmin etmek için deęişim noktası tahmin yöntemlerinin dayanıklı versiyonlarını geliřtirmişlerdir.

Dięer taraftan, pratikte aykırı gözlem(ler) içeren veri kümelerinin oldukça yaygın olduęu sıklıkla gözlemlenmektedir. ML ve LS gibi geleneksel tahmin tekniklerinin aykırı deęerler ve kirlenmiş veriler (contaminated data) varlığında dayanıklı olmadığı bilinmektedir. Başka bir deyişle, dayanıklı olmayan tahmin edicilerin performansları aykırı gözlem(ler)in varlığında önemli ölçüde deęişmektedir (Aydın, 2018b). Bu sorunu ele almak için, literatürde çok sayıda dayanıklı tahmin yöntemi geliştirilmiş ve önerilmiştir (Rousseeuw, 1984; Huber, 1992; Bai, 1995; Tiku ve Akkaya, 2004; Chan ve Peng, 2005; Olive, 2006). Bu yöntemler, aykırı deęerlerin etkisini azaltmayı ve veri aykırı gözlem(ler) içerdiğinde bile daha güvenilir ve doęru parametre tahminleri elde etmeyi amaçlamaktadırlar.

Bu çalışmada, X -yönünde aykırı gözlem (kaldıraç noktası) olması durumunda, Gumbel hata dağılımına sahip basit doğrusal regresyon model parametrelerinin tahmini incelenmiştir. Model parametrelerini tahmin etmek için, literatürde yaygın olarak kullanılan klasik tahmin yöntemlerinden en çok olabilirlik ve en küçük kareler aykırı gözlemlerden etkileneceğinden dolayı, alternatif olarak aykırı gözlemlerden etkilenmeyecek dayanıklı parametre tahmin etme yöntemlerinden En Küçük Medyan Kareler (Least Median of Square-LMS) önerilmiştir. Performans karşılaştırması amacıyla LS, En Küçük Mutlak Sapma (Least Absolute Deviation-LAD) ve Ağırlıklı En Küçük Mutlak Sapma (Weighted Least Absolute Deviation-WLAD) yöntemlerine göre basit doğrusal regresyon modelinin parametre tahminleri de elde edilmiştir. Rousseeuw (1984) tarafından önerilen LMS tahmin yöntemi, sıfır kırılma noktasına sahip LS ve LAD tahmin edicilerinin aksine, verideki %50 kirlenmenin etkisine dirençli olabilmesi nedeni ile yaygın olarak kullanılan kırılma noktası yüksek dayanıklı tahmin yöntemlerden biridir (Sapra, 2003). Çalışmada LS, LAD, WLAD ve LMS tahmin edicilerinin etkinlikleri aykırı gözlem yokluğunda ve varlığında karşılaştırılmıştır.

Çalışmada yer alan bölümler şu şekilde özetlenebilir. Bu çalışmada kullanılan temel kavramlar ve yöntemler 2. ve 3. Bölümlerde tanıtılmıştır. Basit doğrusal regresyon modelinde normallik varsayımından sapmalar olduğunda veya veri seti aykırı gözlem içerdiğinde parametrelerin LS ve ML tahminlerinin elde edilmesi 4. Bölümde anlatılmıştır. Ayrıca bu bölümde LS tahmin edicilerinin etkinlikleri simülasyon çalışması ve bir gerçek veri uygulaması ile incelenmiştir. Tezin 5. Bölümü'nde hata terimlerinin

Gumbel dağılımı varsayımında ve veri seti aykırı değer içerdiğinde parametrelerin tahminleri, simülasyon çalışması ve gerçek veri uygulaması yer almaktadır. Çalışma sonucunda elde edilen bulgular 6. Bölümdedir. Çalışmanın 7. Bölümü'nde ise sonuçlar ve öneriler yer almaktadır.



2. TEMEL KAVRAMLAR

Çalışmamızın bu bölümünde, bu tez çalışması için gerekli olan tanımlar ve temel bilgiler bu bölümde verilmiştir.

2.1. Dayanıklı Tahmin Edici

Dayanıklı (robust) tahmin edici, varsayılan model altında tam etkin ya da yaklaşık olarak tam etkin, varsayılan modelin makul alternatifleri altında da yüksek etkinliğe sahip tahmin edicidir. Tam etkin tahmin edici ise, yansız ve varyansı Rao-Cramer alt sınırına eşit olan tahmin edicidir. (Tiku ve Akkaya (2004).

Robust (sağlamlık, dayanıklılık) terimi, ilk olarak Box (1953) tarafından kullanılmıştır. Daha sonra Tukey (1960, 1962) Huber (1964, 1967) ve Hampel (1971, 1974) tarafından yapılan dayanıklı yöntemlerle ilgili çalışmalar literatüre kazandırılmıştır.

2.2. Aykırı Gözlem

Aykırı gözlemler, veri setinde yer alan diğer birimlerden istatistiksel olarak anlamlı düzeyde sapma gösteren gözlemler olarak tanımlanmaktadır. Bu tür gözlemler, veri üretim sürecinde meydana gelen ölçüm hataları, örnekleme kaynaklı sapmalar, modelleme sürecinde gerçekleştirilen hatalı varsayım veya yanlış model seçimi gibi nedenlerle ortaya çıkabilmektedir. Veri setinin dağılım özelliklerini bozarak yapısal bütünlüğü olumsuz etkileyen aykırı gözlemlerin uygun yöntemlerle tanımlanması ve sonrasında veri setinden çıkarılması, dönüştürülmesi ya da normalize edilmesi, gerçekleştirilecek istatistiksel analizlerin güvenilirlik ve geçerliliğinin sağlanması açısından kritik bir rol oynamaktadır (Rousseeuw ve Leroy, 1987). Regresyon analizinde aykırı gözlemler üç ana grupta değerlendirilmektedir:

Y-yönünde aykırı gözlem

Bu gözlemler, bağımsız değişkenler bakımından örneklem merkezine yakın konumda bulunmalarına rağmen, bağımlı değişkenin (y) değerleri açısından diğer gözlemlerden belirgin biçimde sapmaktadırlar. Söz konusu noktalar, yüksek derecede etkili gözlemler olarak değerlendirilebilseler de bağımsız değişkenler ekseninde uzaklık oluşturmadıkları için kaldıraç (leverage) etkisine sahip değildirler.

X-yönünde aykırı gözlem

Bazı durumlarda, bir veya birkaç gözlem, bağımsız değişkenler (x) ekseninde diğer gözlemlerden oldukça uzak bir konumda yer alabilir. Bu tür gözlemler, X -yönünde aykırı gözlemler olarak adlandırılmakta olup regresyon doğrusunun eğimini önemli ölçüde değiştirme potansiyeline sahiptirler. Bu sebeple literatürde, söz konusu noktalar kötü kaldıraç noktaları (bad leverage points) olarak tanımlanmaktadır.

Hem Y-yönünde hem de X-yönünde aykırı gözlem

Bu gruptaki gözlemler, veri setinin merkezinden uzaklaşmış olmakla birlikte, regresyon parametreleri üzerinde güçlü bir etki yaratmazlar. Başka bir ifadeyle, kaldıraç etkisi taşırsalar da modelin sonuçlarını bozucu nitelikte değildirler. Bu nedenle, bu tür gözlemler iyi kaldıraç noktaları (good leverage points) olarak adlandırılmaktadır (Rousseeuw ve Leroy, 1987).

2.3. Olasılık Yoğunluk Fonksiyonu

X sürekli bir rasgele değişken olmak üzere

(i) $f(x) \geq 0, \forall x \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$

koşullarını sağlayan $f(x)$ fonksiyonuna X rasgele değişkeninin olasılık yoğunluk fonksiyonu denir.

2.4. Dağılım Fonksiyonu

X sürekli bir rasgele değişken olmak üzere, herhangi bir gerçek x değeri için, X rasgele değişkeninin x 'e eşit ya da ondan küçük bir değer alma olasılığı dağılım fonksiyonu (df) olarak tanımlanır. Başka bir ifade ile, herhangi bir X rasgele değişkeninin df 'si,

$$F(x) = P(X \leq x) = \int_0^x f(t)dt$$

eşitliği ile elde edilmektedir.

2.5. k . Moment

Bir X rasgele değişkeninin oyf 'si $f(x)$ ve her $k \in \mathbb{N}$ için $E(X^k) < \infty$ olsun. $E(X^k)$ değerine X rasgele değişkeninin r . moment denir ve

$$E(X^k) = \int_{-\infty}^{\infty} x^k f(x)dx$$

formülü ile elde edilir. Bu eşitlik kullanılarak dağılımın beklenen değeri ve varyansı elde edilir. Ayrıca dağılımın çarpıklık (ϕ_1) ve basıklık (ϕ_2) katsayıları bu eşitlik yardımı ile

$$\phi_1 = \frac{E(X^3) - 3\mu E(X^2) + 2\mu^3}{\sigma^3}$$

$$\phi_2 = \frac{E(X^4) - 4\mu E(X^3) + 6\mu^2 E(X^2) - 3\mu^4}{\sigma^4}$$

olarak elde edilir. Burada $\mu = E(X)$ ve $\sigma = \sqrt{Var(X)}$ 'dir.

2.6. Normal Dağılım

Normal dağılıma sahip bir X rasgele değişkeninin olasılık yoğunluk fonksiyonu (oyf):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right); x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$$

Dağılım fonksiyonu (df) ise

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

dir. Burada, $\operatorname{erf}(\cdot)$ Gauss hata fonksiyonudur ve

$$\operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{x-\mu}{\sigma\sqrt{2}}} \exp(-t^2) dt$$

dir (Johnson vd., 1995a). Normal dağılımın bazı istatistiksel özellikleri aşağıdaki gibi verilmiştir.

Dağılımın k . Momenti,

$$E(X^k) = \begin{cases} \frac{\sqrt{2}}{\sigma^{k+1}\sqrt{\pi}} \int_{\mu}^{\infty} (x-\mu)^k \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx, & k = 2n \ (n \in \mathbb{Z}^+) \\ 0, & k = 2n + 1 \ (n \in \mathbb{Z}^+) \end{cases}$$

Yukarıdaki eşitlikte yer alan integralin sonucunda

$$E(X^k) = \begin{cases} \sum_{m=0}^k \binom{k}{m} \mu^m \sigma^{k-m} E(Z^{k-m}), & k = 2n \ (n \in \mathbb{Z}^+) \\ 0, & k = 2n + 1 \ (n \in \mathbb{Z}^+) \end{cases}$$

dir. Burada,

$$E(Z^{2m}) = \frac{(2m)!}{2^m m!}$$

dir (Winkelbauer, 2012). Moment çıkaran fonksiyonu,

$$M_x(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

dir. X rastgele değişkeninin beklenen değeri, varyansı, modu (tepe değeri), medyanı (ortanca), çarpıklık katsayısı ve basıklık katsayısı:

$$E(X) = \mu,$$

$$V(X) = \sigma^2,$$

$$Mod(X) = \mu,$$

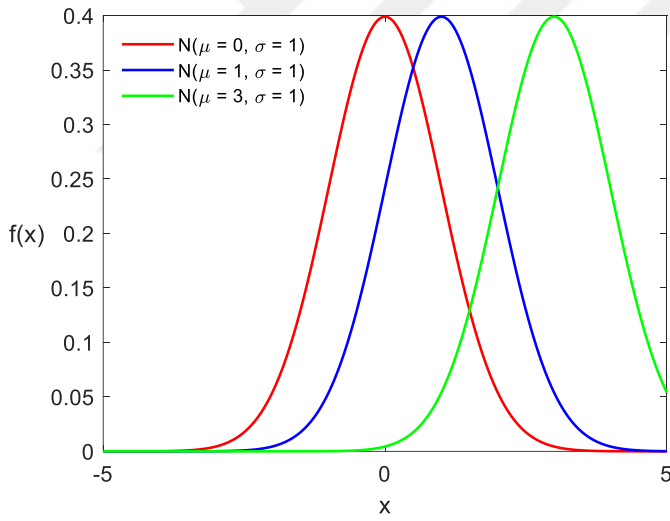
$$Med(X) = \mu,$$

$$\phi_1 = 0$$

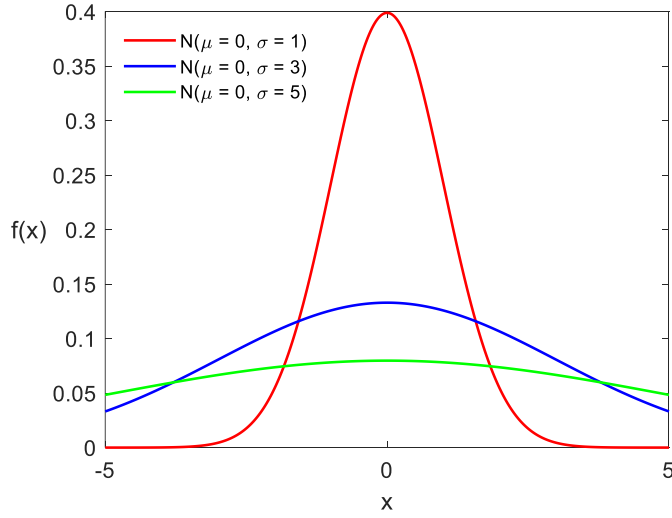
ve

$$\phi_2 = 3$$

tür. Farklı konum parametresi (μ) ve farklı ölçek parametresi (σ^2) değerlerine göre çizdirilen Normal dağılım oyf grafikleri Şekil 2.1 ve Şekil 2.2’de sırasıyla verilmiştir.



Şekil 2.1 Farklı konum parametresi (μ) değerlerine göre çizdirilen Normal dağılım oyf grafikleri



Şekil 2.2 Farklı ölçek parametresi (σ^2) değerlerine göre çizdirilen Normal dağılım oyf grafikleri

2.7. Student t Dağılımı

Student-t dağılımına sahip X rasgele değişkeninin oyf aşağıdaki gibidir:

$$f(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi\sigma^2}} \left(1 + \frac{(x-\mu)^2}{v\sigma^2}\right)^{-\frac{v+1}{2}}, \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0, v > 0.$$

Burada, v serbestlik derecesini, μ konum parametresini, σ ölçek parametresini, $\Gamma(\cdot)$ Gama fonksiyonunu ve göstermektedir ve

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) dx$$

dir. X rasgele değişkeninin df

$$F(x) = \int_{-\infty}^x \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sqrt{v\pi\sigma^2}} \left(1 + \frac{(x-\mu)^2}{v\sigma^2}\right)^{-\frac{v+1}{2}} dx$$

$$= \begin{cases} 1 - \frac{1}{2} I_{v/\gamma^2(x)+v} \left(\frac{v}{2}, \frac{1}{2}\right), & \gamma(x) = \frac{x-\mu}{\sigma} \leq 0 \\ \frac{1}{2} I_{v/\gamma^2(x)+v} \left(\frac{v}{2}, \frac{1}{2}\right), & \gamma(x) = \frac{x-\mu}{\sigma} > 0 \end{cases}$$

dir. Burada $I_{v/\gamma^2(x)+v} \left(\frac{v}{2}, \frac{1}{2}\right)$, düzenli hale getirilmiş (regularized) tam olmayan Beta fonksiyonudur ve

$$I_{v/\gamma^2(x)+v} \left(\frac{v}{2}, \frac{1}{2}\right) = \frac{B\left(v/\gamma^2(x)+v; \frac{v}{2}, \frac{1}{2}\right)}{B\left(\frac{v}{2}, \frac{1}{2}\right)}$$

dır (Johnson vd., 1995b). Ayrıca, $B\left(v/\gamma^2(x)+v; \frac{v}{2}, \frac{1}{2}\right)$ tam olmayan Beta fonksiyonu,

$$B\left(v/\gamma^2(x) + v; \frac{v}{2}, \frac{1}{2}\right) = \int_0^{v/\gamma^2(x)+v} t^{\frac{v}{2}-1} (1-t)^{\frac{1}{2}-1} dt$$

ve $B\left(\frac{v}{2}, \frac{1}{2}\right)$, Beta fonksiyonu

$$B\left(\frac{v}{2}, \frac{1}{2}\right) = \int_0^1 t^{\frac{v}{2}-1} (1-t)^{\frac{1}{2}-1} dt$$

dir. Student-t dağılımının bazı istatistiksel özellikleri:

Dağılımın k. momenti

$$E(X^k) = \begin{cases} 0, & k = 2n + 1 \ (n \in \mathbb{Z}^+), \ 0 < k < v \\ \frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{v-k}{2}\right)v^{\frac{k}{2}}}{\sqrt{\pi}\Gamma\left(\frac{v}{2}\right)}, & k = 2n \ (n \in \mathbb{Z}^+), \ 0 < k < v \\ \infty, & k = 2n \ (n \in \mathbb{Z}^+), \ 0 < v \leq k \end{cases}$$

eşitliğinden elde edilir. Student-t dağılımına sahip X rasgele değişkeninin sırasıyla beklenen değeri, varyansı, modu, medyanı, çarpıklık katsayısı ve basıklık katsayısı (Excess Kurtosis):

$$E(X) = \mu, \quad v > 1$$

$$Var(X) = \begin{cases} \text{tanımsız}, & v \leq 1 \\ \frac{v}{v-2} \sigma^2, & 1 < v < 2 \\ \infty, & v > 2 \end{cases}$$

$$Mod(X) = \mu,$$

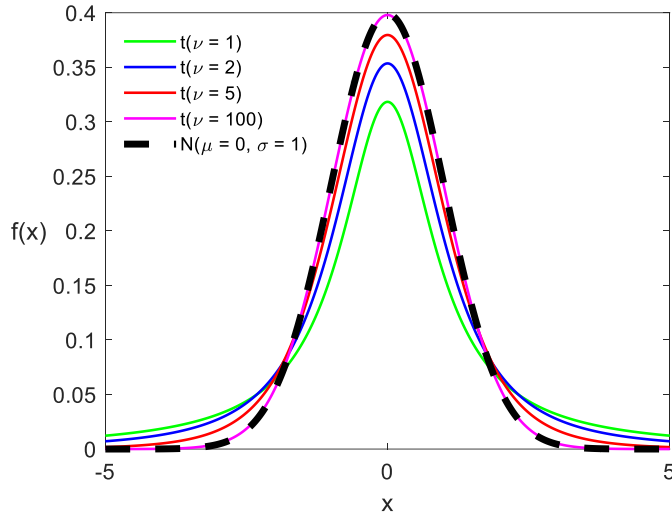
$$Med(X) = \mu,$$

$$\phi_1 = 0, \quad v > 2$$

ve

$$\phi_2 = \frac{6}{v-4}, \quad v > 4$$

dır. Student-t dağılımının şekli, Normal dağılım eğrisinin şekli ile benzerdir. Yani simetrik bir dağılımdır. Bu nedenle çarpıklık katsayısı sıfırdır. Şekil 2.3'te $\mu = 0$ ve $\sigma^2 = 1$ olan Student-t dağılımının farklı serbestlik derecelerine göre çizdirilen oyf grafikleri gösterilmiştir. Bu grafiğe göre, Student-t dağılımı Normal dağılıma göre daha kalın kuyruklu bir dağılımdır. Serbestlik derecesi arttıkça dağılımın Normal dağılıma yakınsadığı görülmüştür. Zaten bu bilinen bir sonuçtur yani, $v \rightarrow \infty$ için Student-t dağılımı Normal dağılıma yakınsar. Ayrıca, $v = 1$ için dağılım Cauchy dağılımı olur.



Şekil 2.3 Farklı serbestlik dereesi (ν) değerlerine göre çizdirilen Student-t dağılımı oyf grafikleri

2.8. Sürekli Uniform (Sürekli Tekdüze) Dağılım

Sürekli Uniform dağılımına sahip X rasgele değişkeninin oyf ve df aşağıdaki gibidir:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x < a, x > b \end{cases},$$

ve

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

Dağılımın k . momentini

$$E(X^k) = \frac{1}{k+1} \sum_{i=1}^k a^i b^{k-i}$$

dir. X rastgele değişkeninin moment çıkaran fonksiyonu,

$$M_x(t) = \frac{\exp(tb) - \exp(ta)}{t(b-a)}$$

biçimindedir. Sürekli Uniform dağılımın beklenen değeri, varyansı, modu, medyanı, çarpıklık katsayısı ve basıklık katsayısı (Excess Kurtosis):

$$E(X) = \frac{a+b}{2},$$

$$Var(X) = \frac{(b-a)^2}{12},$$

$Mod(X)$: $[a, b]$ kapalı aralığında herhangi bir değer,

$$Med(X) = \frac{(a+b)}{2},$$

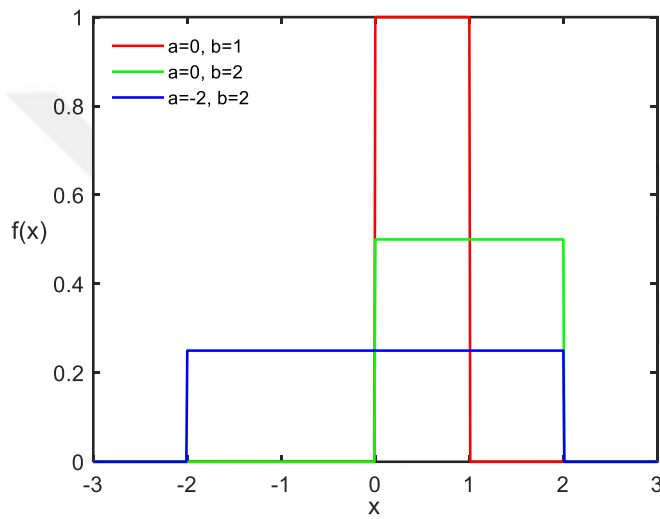
$$\phi_1 = 0$$

ve

$$\phi_2 = -\frac{6}{5}$$

dır. Farklı parametre değerlerine göre çizdirilen sürekli Uniform dağılımı oyf grafikleri

Şekil 2.4'te gösterilmiştir.



Şekil 2.4 Farklı parametre değerlerine göre çizdirilen sürekli Uniform dağılımı oyf grafikleri

2.9. Gumbel Dağılımı

Gumbel dağılıma sahip bir X rasgele değişkeninin olasılık yoğunluk fonksiyonu (oyf):

$$f(x) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma} - \exp\left(-\frac{x-\mu}{\sigma}\right)\right) \quad x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$$

df ise

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right)$$

dir (Aydın ve Şenoğlu, 2015; Aydın, 2018a; Aydın, 2018b). Burada μ konum parametresi ve σ ölçek parametresidir.

Dağılımın k . Momenti,

$$E(X^k) = \exp(\mu t) \Gamma(\sigma t + 1), \quad t > \frac{-1}{\sigma}$$

X Rastgele değişkeninin beklenen değeri, varyansı, modu (tepe değeri), medyanı (ortanca), çarpıklık katsayısı ve basıklık katsayısı:

$$E(X) = \mu + \sigma\gamma,$$

$$Var(X) = \frac{\pi^2}{6} \sigma^2,$$

$$Mod(X) = \mu$$

$$Med(X) = \mu - \sigma \ln(\ln(2))$$

$$\phi_1 = \frac{2\pi}{3}$$

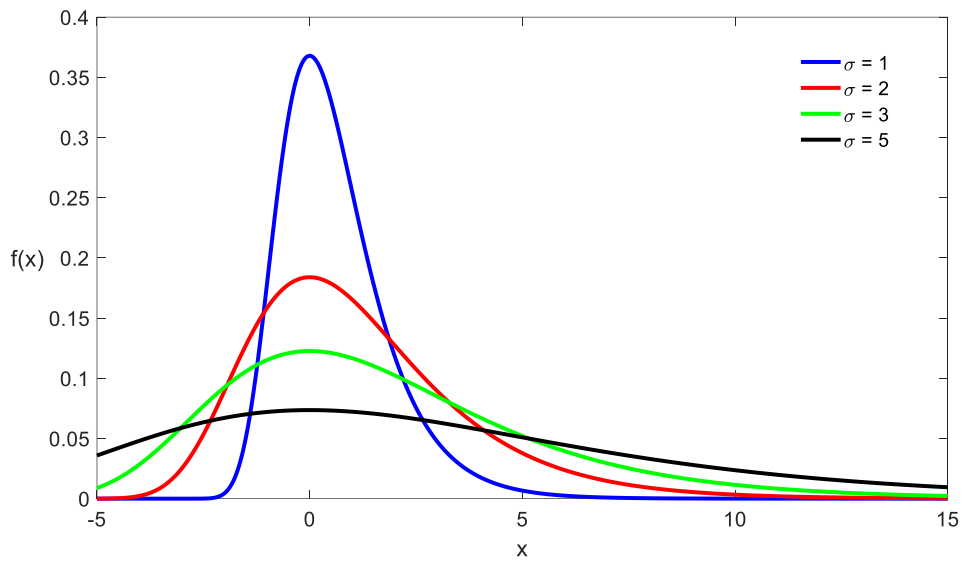
ve

$$\phi_2 = \frac{6}{5}$$

tir. Burada γ Euler sabitidir ve şu şekilde tanımlanır:

$$\gamma = - \int_0^{\infty} \ln(x) \exp(-x) dx$$

Ayrıca γ , yaklaşık olarak 0.5772 alınır. Şekil 2.5'de verilen $\mu = 0$ ve çeşitli ölçek parametresi σ değerleri için Gumbel dağılımının oyf grafikleri, Gumbel dağılımının tek modlu ve sağa çarpık olduğunu göstermektedir.



Şekil 2.5 $\mu = 0$ ve bazı seçilmiş σ değerleri için Gumbel dağılımının farklı olasılık yoğunluk fonksiyonları.

3. MATERYAL VE YÖNTEM

Bu bölümde bu tez çalışmasında kullanılan bazı yöntemler tanıtılacaktır.

3.1. Ölçülebilir Model

Bir $x \in \mathbb{R}^m$ değişkeni ile $y \in \mathbb{R}$ değişkeni arasında bilinmeyen bir $\theta \in \Theta \subseteq \mathbb{R}^p$ parametresine bağlı

$$y = f(x, \theta) \quad (3.1)$$

olarak gösterilen bir ilişki olduğu varsayalım. $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 'den oluşan bir dizi verildiğinde (3.1)'deki fonksiyonel ilişkiyi sağlayan $y_i = f(x_i, \theta^0)$ $i = 1, 2, \dots, n$ için $\theta^0 \in \Theta$ vardır. Ancak, (x_i, y_i) değişken değerleri fiziksel bir ölçüm sonucunda elde ediliyorlarsa, bahsedilen fonksiyonel ilişki kesin olarak belirlenemez. y değişkeninin ölçülen değeri ile $f(x, \theta)$ fonksiyonu ile elde edilen değeri arasındaki fark ε hatası olarak alınır. Bu nedenle bağımlı değişken için her zaman ε ölçüm hatası söz konusudur. Bu durumda x değişkeni ile y değişkeni arasındaki ölçülebilir model

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.2)$$

biçimindedir. ε_i 'ler bilinmeyen ve ölçülemeyen rasgele hata terimleri olarak da ifade edilirler. Hataların değerleri küçük olsalar bile herhangi bir $\theta \in \Theta$ $y_i = f(x_i, \theta) + \varepsilon_i$ eşitliği sağlanamayabilir. Bu nedenle amaç gerçek $f(x, \theta)$ fonksiyonu yerine yaklaşık $f(x, \theta^0)$ fonksiyonunu elde etmektir. Ancak, yaklaşık $f(x, \theta)$ fonksiyonu $f(x, \theta^0)$ 'yı elde ederken artık olarak adlandırılan $\varepsilon_i(\theta)$ ile gösterilen hatalar söz konusu olur (Ramponi, 2016).

3.2. Basit Doğrusal Regresyon

Doğrusal regresyon analizi, istatistiksel modelleme ve veri analizinde en sık kullanılan yöntemlerden biridir. Temel amacı, iki veya daha fazla değişken arasındaki doğrusal ilişkinin yönünü ve derecesini belirlemek, bu ilişkiyi matematiksel olarak ifade etmek ve tahminler yapmaktır. Bu bağlamda regresyon analizi, yalnızca değişkenler arasındaki ilişkinin varlığını saptamakla kalmaz, aynı zamanda bu ilişkinin büyüklüğü ve yönü hakkında da çıkarımlarda bulunulmasına olanak tanır. Dolayısıyla, araştırmalarda hem teorik modellemeler hem de pratik tahminler için güvenilir bir istatistiksel yöntem olarak kullanılmaktadır. Basit doğrusal regresyon analizinde, tek bir bağımsız değişkenin bir bağımlı değişken üzerindeki etkisi incelenmektedir. x bağımsız değişken (açıklayıcı

değişken) olarak y de bağımlı değişken (yanıt değişkeni) olarak adlandırılır. Basit doğrusal regresyon denklemi matematiksel olarak

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n \quad (3.3)$$

şeklinde ifade edilir. Bu denklemde, β_0 regresyon doğrusunun y eksenini kestiği noktayı, β_1 regresyon doğrusunun eğimini, bağımsız değişkendeki bir birimlik değişimin bağımlı değişken üzerindeki ortalama etkisini, ε_i rasgele hata terimini ve n gözlem sayısını göstermektedir. Basit doğrusal regresyon modelinde üç temel varsayım bulunmaktadır:

- (i) $E(\varepsilon_i) = 0$,
- (ii) $Cov(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$
- (iii) $Var(\varepsilon_i) = \sigma^2$.

Ayrıca, parametrik testlerin (örneğin t-testi, z-testi ve F-testi) kullanılabilmesi için hata terimlerinin Normal dağılıma sahip olduğu kabul edilmektedir. Bu varsayım, modelden elde edilen sonuçların güvenilirliğini ve özellikle anlamlılık testlerinin geçerliliğini doğrudan etkilemektedir. Ancak, söz konusu varsayımların ihlali durumunda çeşitli problemler ortaya çıkabilmektedir. Bunlar:

- (i) Hata terimlerinin beklenen değerinin sıfırdan farklı olması, modelin yanlı (biased) tahminler oluşturmaya neden olur.
- (ii) Hata terimlerinin ilişkili olması, standart hata tahminlerini etkileyerek test istatistiklerinin güvenilirliğini azaltır.
- (iii) Değişen varyanslılık (heteroscedasticity) durumu, tahmin edicilerin yansızlığını korumakla birlikte, etkinliklerini kaybetmesine ve yanlış anlamlılık sonuçlarına sebep olabilir.
- (iv) Hata terimlerinin Normal dağılmaması, özellikle küçük örneklerde modelden elde edilen sonuçların güvenilirliğini ve özellikle anlamlılık testlerinin geçerliliğini doğrudan etkilemektedir.

Regresyon analizinden elde edilen sonuçların güvenilirliğini sağlamak amacıyla bu varsayımların dikkatle test edilmesi ve sağlanmaması durumunda uygun dönüşümlerin ya da dayanıklı regresyon tekniklerinin uygulanması gerekmektedir.

3.3. En Küçük Kareler Yöntemi

En küçük kareler (Least square-LS) yönteminin amacı veriyi en iyi modelleyen $\hat{\theta} \in \Theta$ değerini bulmaktır. Burada en iyi ile kastedilen artık kareler toplamı $\phi_{LS}(\hat{\theta})$

$$\phi_{LS}(\hat{\theta}) = \sum_{i=1}^n \varepsilon_i^2(\hat{\theta}) = \sum_{i=1}^n (y_i - f(x_i, \hat{\theta}))^2$$

minimum olmasıdır. Burada $\phi_{LS}(\hat{\theta})$ küçük olması y_i değerlerinin $f(x_i, \hat{\theta})$ değerlerine yakın olduğu anlamına gelir. $\hat{\theta}, \theta^0 \in \Theta$ parametresinin LS tahmin edicisi olarak adlandırılır.

LS, yaygın olarak kullanılan istatistiksel bir tahmin yöntemidir.

3.4. En Küçük Mutlak Sapma Yöntemi

En küçük mutlak sapma (Least Absolute Deviation-LAD) yönteminde amaç Bölüm 3.1’de tanımı verilen artıkların mutlak değerlerinin toplamı olan $\phi_{LAD}(\hat{\theta})$ fonksiyonunun

$$\phi_{LAD}(\hat{\theta}) = \sum_{i=1}^n |\varepsilon_i(\hat{\theta})| = \sum_{i=1}^n |y_i - f(x_i, \hat{\theta})|$$

minimize edilmesidir. $\phi_{LAD}(\hat{\theta})$ fonksiyonunu minimum yapan $\hat{\theta}, \theta^0 \in \Theta$ parametresinin LAD tahmin edicisi olur.

Hataların normal dağılmaması ve/veya veri kümesinde aykırı değerlerin bulunması durumunda LAD yöntemi diğer klasik tahmin yöntemlerine göre daha dayanıklıdır (Temiz, 2006). LAD tahmin edicileri, özellikle kalın kuyruklu dağılımlar veya aykırı değerler söz konusu olduğunda, LS tahmin edicilerine kıyasla aykırı değerlere daha az duyarlıdır (Bai, 1995). Ancak, kaldıraç noktalarına duyarlıdırlar (Croux vd., 2003). LAD tahmin edicileri, hataların dağılımı hakkında herhangi bir varsayım gerektirmeden asimptotik olarak normal dağılım gösterir (Pollard, 1991; Dunsmuir ve Spencer, 1991; Davis ve Dunsmuir, 1997). Ayrıca, LAD tahmin edicileri X-yönündeki aykırı değerlere karşı dayanıklı değildir. Ancak, bazı durumlarda, LAD tahminleri LS tahminlerinden bile daha kötü performans gösterebilirler (Gao ve Fenk, 2018).

3.5. Ağırlıklı En Küçük Mutlak Sapma Yöntemi

Ağırlıklı en küçük mutlak sapma (Weight Absolute Deviation-WLAD) yönteminde, her bir gözlem için farklı bir ağırlık değeri atanır. Bu ağırlıklar, gözlemlerin modelin parametrelerinin tahmini üzerinde etkilidirler. WLAD yöntemi ile $\theta^0 \in \Theta$ parametresinin tahmini $\hat{\theta}$, $\phi_{WLAD}(\hat{\theta})$ fonksiyonunun

$$\phi_{WLAD}(\hat{\theta}) = \sum_{i=1}^n w_i |\varepsilon_i(\hat{\theta})| = \sum_{i=1}^n w_i |y_i - f(x_i, \hat{\theta})|$$

minimum yapılması ile elde edilir.

WLAD, özellikle aykırı değerler ve normal olmayan hata dağılımlarıyla başa çıkmada dayanıklı bir tahmin yöntemidir. Ayrıca, yoğunluk fonksiyonu ve türevinin düzgün sınırlı olması koşulu altında, asimptotik olarak normal bir dağılıma sahiptir. Bu özellik, WLAD yönteminin çok çeşitli koşullar altında güvenilir ve etkili kalmasını sağlayarak, pratik uygulamalarda istatistiksel analiz için değerli bir araç olmasını sağlar (Ling, 2005; Pan ve Wang, 2007).

3.6. En Küçük Medyan Kareler Yöntemi

En küçük medyan kareler (Least Median of Square-LMS) tahmin yöntemi Rousseeuw (1984) tarafından önerilmiştir. LMS tahmin edicileri özellikle aykırı değerlerin (outliers) etkisini minimize etmek için geliştirilmiş bir yöntemdir. Bir $\theta^0 \in \Theta$ parametresinin LMS tahmin edicileri

$$\phi_{LMS}(\hat{\theta}) = Med_i \varepsilon_i^2(\hat{\theta}) = Med_i \left(y_i - f(x_i, \hat{\theta}) \right)^2 \quad i = 1, 2, \dots, n$$

ile verilen hata karelerinin medyanı $\phi_{LMS}(\hat{\theta})$ fonksiyonunun minimum yapılması ile elde edilir.

LMS yöntemi, özellikle aykırı değerlere ve model varsayımlarından sapmalara karşı dayanıklı olduğundan daha güvenilir sonuçlar elde edilir. LMS, aykırı değer olması durumunda daha iyi performans gösterir. Normalde, aykırı değerler en küçük kareler yöntemiyle modelin parametrelerini bozabilirken, LMS bu tür ekstrem değerlere daha az duyarlıdır (Rousseeuw ve Leroy, 2003).

4. BASİT DOĞRUSAL REGRESYON MODELİNDE NORMALLİK VARSAYIMINDAN SAPMALAR OLDUĞUNDA VEYA VERİ SETİ AYKIRI DEĞER İÇERDİĞİNDE PARAMETRELERİN TAHMİNİ

Tezin bu bölümünde, Eş. (3.3)'de verilen basit doğrusal regresyon modelinde, hata terimlerinin dağılımının

$$f(\varepsilon_i|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right); \varepsilon_i \in \mathbb{R}, \sigma > 0$$

oyf ile verilen Normal dağılımına sahip olduğu varsayılmıştır. Model parametrelerinin tahmini bu varsayım altında yapılmıştır. Ayrıca $\sigma^2 = 1$ olarak bilindiği varsayılmıştır.

4.1. Parametre Tahmini

Bu bölümde, hata terimlerinin Normal dağıldığı varsayımı altında basit doğrusal regresyon modeli için, model parametrelerinin tahmin edicilerinin elde edilmesine yönelik en küçük kareler ve en çok olabilirlik yöntemi olmak üzere iki farklı yöntem ele alınmıştır. Ayrıca, Fisher Bilgi matrisi de elde edilmiştir.

4.1.1. En Küçük Kareler Tahmin Edicileri

En küçük kareler tahmin yöntemi Eş. (3.3)'de verilen basit doğrusal regresyon modeli için, model parametrelerinin LS tahmin edicileri,

$$\phi_{LS}(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \quad (4.1)$$

şeklinde verilen hata kareler toplamını minimum yapan β_0 ve β_1 değerleridir. Buna göre, Eş. (4.1)'de verilen $\phi_{LS}(\beta_0, \beta_1)$ fonksiyonunun ilgili parametrelere göre kısmi türev denklemleri

$$\frac{\partial \phi_{LS}(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0 \quad (4.2)$$

$$\frac{\partial \phi_{LS}(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) x_i = 0 \quad (4.3)$$

olarak elde edilir. Eş. (4.2) ve (4.3)'ün sırası ile β_0 ve β_1 'e göre çözümlenmesi sonucunda, model parametrelerinin LS tahmin edicileri,

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} \quad (4.4)$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.5)$$

olarak elde edilir. Burada, Eş. (4.1)'de verilen $\phi_{LS}(\tilde{\beta}_0, \tilde{\beta}_1)$ fonksiyonu kullanılarak σ^2 parametresinin LS tahmin edicisi, $\tilde{\sigma}^2$ elde edilemez. Bu nedenle $\tilde{\sigma}^2$ olarak

$$\tilde{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2 \quad (4.6)$$

ile verilen regresyon modeli hata kareler ortalaması alınmıştır. $\tilde{\beta}_0$, $\tilde{\beta}_1$ ve $\tilde{\sigma}^2$ tahmin edicileri, Eş. (3.3)'de verilen regresyon modelinin varsayımları sağlaması durumunda Gauss-Markov teoremine göre, diğer lineer yansız tahmin ediciler arasında en küçük varyansa sahip tahmin edicilerdir (BLUE-Minimum Variance Linear Unbiased Estimator) (Rao ve Toutenburg, 1999).

4.1.2. En Çok Olabilirlik Tahmin Edicileri

Eş. (3.3)'de verilen basit doğrusal regresyon modelinde, hata terimlerinin dağılımın $N(0, \sigma^2)$ olduğu varsayımı altında y_i 'lerin oyf

$$f(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right); y_i \in \mathbb{R}, \beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}, \sigma > 0$$

şeklinde dir. Burada, $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ 'dır. Olabilirlik fonksiyonu

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^n f(y_i|\boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right) \end{aligned}$$

şeklinde dir. Olabilirlik fonksiyonunun doğal logaritması (ln-olabilirlik fonksiyonunu) da

$$\ln L(\boldsymbol{\theta}) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (4.7)$$

dir. $\ln L(\boldsymbol{\theta})$ fonksiyonun birinci türevleri alınıp sifira eşitlendiğinde elde edilen kısmi türev denklemleri aşağıdaki gibidir:

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0, \quad (4.8)$$

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0, \quad (4.9)$$

$$\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 = 0. \quad (4.10)$$

Eş. (4.8)-(4.10) verilen denklemlerin ilgili parametrelere göre çözümlenmesi sonucunda β_0 , β_1 ve σ^2 parametrelerinin ML tahmin edicileri

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (4.11)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.12)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (4.13)$$

olarak elde edilir. Burada, $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahmin edicileri yansız tahmin edicileridir. Ancak, σ^2 parametresinin ML tahmin edicisi $\hat{\sigma}^2$ yanlı bir tahmin edicidir. Bu nedenle, (4.13)'da verilen $\hat{\sigma}^2$ 'da yan düzeltmesi yapıldığında,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (4.14)$$

olur. Bu çalışmada, σ^2 parametresinin ML tahmin edicisi olarak Eş. (4.14)'de yer alan tahmin edici kullanılmıştır. Ayrıca, hata terimlerinin Normal dağılıma sahip olması durumunda parametrelerin LS ve ML tahmin edicileri birbirine eşittir bu da beklenen teorik sonuçtur. Ayrıca; $\hat{\beta}_0$, $\hat{\beta}_1$ ve $\hat{\sigma}^2$ tahmin edicilerinin dağılımları sırasıyla

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right),$$

$$\hat{\sigma}^2 \sim \chi_{(n-2)}^2.$$

Hata terimleri Normal dağılıma sahip Eş. (3.3)'de verilen basit doğrusal regresyon modeli için Fisher bilgi matrisi

$$I(\boldsymbol{\theta}) = \begin{bmatrix} -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0^2}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \sigma^2}\right) \\ -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1^2}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \sigma^2}\right) \\ -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \sigma^2}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \sigma^2}\right) & -E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^2^2}\right) \end{bmatrix}$$

olarak tanımlanır. Burada $I(\boldsymbol{\theta})$, Fisher bilgi matrisinin öğeleri $\ln L(\boldsymbol{\theta})$ fonksiyonunun $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$ parametre vektörünün her bir bileşenine göre ikinci dereceden kısmi türevlerinin beklenen değerleri bulunup (-1) ile çarpılmasından elde edilir.

Eş. (4.7)'de verilen $\ln L(\boldsymbol{\theta})$ fonksiyonunun ikinci dereceden türevlerinin beklenen değerleri:

$$-E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0^2}\right) = -E\left(-\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}$$

$$-E\left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_1}\right) = -E\left(-\frac{\sum_{i=1}^n x_i}{\sigma^2}\right) = \frac{\sum_{i=1}^n x_i}{\sigma^2}$$

$$-E \left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_0 \partial \sigma^2} \right) = -E \left(\frac{-\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)}{\sigma^4} \right) = 0$$

$$-E \left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1^2} \right) = -E \left(-\frac{\sum_{i=1}^n x_i^2}{\sigma^2} \right) = \frac{\sum_{i=1}^n x_i^2}{\sigma^2}$$

$$-E \left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \beta_1 \partial \sigma^2} \right) = -E \left(-\frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i}{\sigma^4} \right) = 0$$

$$-E \left(\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial \sigma^2^2} \right) = -E \left(\frac{n}{2\sigma^4} - \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^6} \right) = \frac{n}{2\sigma^4}$$

olarak elde edilir. Bu durumda Fisher bilgi matrisi;

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \frac{n}{\sigma^2} & \frac{\sum_{i=1}^n x_i}{\sigma^2} & 0 \\ \frac{\sum_{i=1}^n x_i}{\sigma^2} & \frac{\sum_{i=1}^n x_i^2}{\sigma^2} & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

olur. Buradan Fisher bilgi matrisinin tersi alındığında

$$I^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} & \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & 0 \\ \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & 0 \\ 0 & 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

şeklinde elde edilir. Bu matrisin köşegen elemanları β_0 , β_1 ve σ^2 parametrelerinin tahmin edicilerinin varyansları için bir alt sınır vermek için kullanılır. Başka bir ifade ile, Fisher bilgi matrisinin tersinin köşegen elemanları β_0 , β_1 ve σ^2 parametrelerinin tahmin edicilerinin varyansları için Cramer-Rao alt sınırına eşittirler.

Bu tez çalışmasında $\sigma^2 = 1$ olarak bilindiği varsayıldığından, LS tahmin edicilerin etkinliklerinin karşılaştırıldığı simülasyon çalışması bölümünde $\hat{\sigma}^2$ ile ilgili sonuçlara yer verilmemiştir.

4.2. Simülasyon Çalışması I

Çalışmanın bu bölümünde, Eş. (3.3) ile verilen basit doğrusal regresyon modelindeki β_0 ve β_1 parametrelerinin LS tahmin edicilerinin, varsayılan modelden sapmalar olduğunda veya veri seti aykırı değer içerdiği durumlarda etkinlikleri Monte-Carlo simülasyon yöntemi ile incelenmiştir. Bu amaçla, bu simülasyon çalışmasında hata terimlerinin dağılımının standart Normal dağılım $N(0, 1)$, olduğu varsayılmıştır. Varsayılan modelin makul alternatifleri olarak alınan modeller ve aykırı değer modelleri:

Model I: Serbestlik derecesi $\nu = 3$ olan Student-t dağılımı, $t(\nu = 3, \sigma)$

Model II: Serbestlik derecesi $\nu = 5$ olan Student-t dağılımı, $t(\nu = 5, \sigma)$

Model III: Serbestlik derecesi $\nu = 10$ olan Student-t dağılımı, $t(\nu = 10, \sigma)$

Model IV: Dixon Aykırı değer modeli: $(n - r)N(0, 1) + rN(0, 4)$, $r = \llbracket 0,5 + n \times 0,1 \rrbracket$

Model V: Karma model: $0,90N(0, 1) + 0,10N(0, 4)$

Model VI: Bulaşık model: $0,90N(0, 1) + 0,10U(-3, 3)$

dir. Yani hata terimlerinin dağılımları gerçekte Model I-VI'deki gibidir. Ancak hata terimlerinin dağılımının $N(0, 1)$ olduğu düşünülüp model parametrelerinin ve σ^2 'nin LS tahminleri bu dağılıma göre elde edilmiştir. Ayrıca, LS tahmin edicilerinin etkinliklerini karşılaştırmak için yan (Bias), Hata Kareler Ortalaması (Mean Square Error-MSE) ve Göreli Etkinlik (Relative efficiency-RE) ölçülerine bakılmıştır. Bu ölçüler:

$$BIAS(\hat{\theta}) = \left| \frac{1}{T} \sum_{i=1}^T (\theta - \hat{\theta}_i) \right|,$$

$$MSE(\hat{\theta}) = \frac{1}{T} \sum_{i=1}^T (\theta - \hat{\theta}_i)^2,$$

$$RE(\hat{\theta}) = \frac{MSE(\tilde{\theta})}{MSE(\hat{\theta})} \times 100$$

biçimindedir. Burada, T simülasyon tekrar sayısını, $MSE(\hat{\theta})$ varsayılan modele göre θ parametresinin LS tahminin hata kareler ortalama değerini, $MSE(\tilde{\theta})$ modele göre θ parametresinin LS tahminin hata kareler ortalama değerini göstermektedir.

Bu simülasyon çalışması aşağıdaki adımlardan oluşmaktadır:

A.1. Parametre değerleri $\beta_0 = 0$ ve $\beta_1 = 1$ olarak alınmıştır.

A.2. Örneklem hacmi $n = 50, 100, 150$ ve 200 'dür.

A.3. Hata terimleri ε_i 'lerin dağılımının standart Normal dağılım $N(0, 1)$ olduğu durum için matlab “normrnd(0,1)” komutu kullanılarak hata terimleri ve bağımsız değişken değerleri için rasgele sayı üretilmiştir.

A.4. Alternatif modellerde yer alan Student-t (ν serbestlik derecesi) ve Uniform dağılımlarına göre rasgele sayı üretmek için matlab “trnd(ν)” ve “unifrnd(a, b)” komutları kullanılmıştır.

A.5. Üçüncü adımda elde edilen hata terimi ve bağımsız değişken değeri Eş. (3.3) denkleminde yerine yazılarak bağımlı değişken değerleri elde edilmiştir.

A.6. β_0 ve β_1 parametrelerinin LS tahminleri elde edilmiştir.

A.7. Bias, MSE ve RE ölçülerinin değerleri hesaplanmıştır.

Adım 3-5 arasındaki işlemler 5000 kez tekrarlanmıştır. Monte-Carlo simülasyon çalışmasının sonuçları Tablo 4.1’de sunulmuştur.

Tablo 4.1 Basit doğrusal regresyon modelinde parametrelerin LS tahmin edicilerinin istatistiksel dayanıklılığı

Model	Örneklem Hacmi (n)	β_0			β_1		
		Bias	MSE	RE	Bias	MSE	RE
$\varepsilon_i \sim N(0,1)$	50	0,0001	0,0205	100,00	0,0031	0,0213	100,00
	100	0,0013	0,0102	100,00	0,0003	0,0101	100,00
	150	0,0011	0,0067	100,00	0,0009	0,0070	100,00
	200	0,0026	0,0050	100,00	0,0000	0,0051	100,00
Model I	50	0,0021	0,0607	33,77	0,0021	0,0683	31,19
	100	0,0014	0,0300	34,00	0,0001	0,0306	33,01
	150	0,0021	0,0201	33,33	0,0038	0,0197	35,53
	200	0,0020	0,0153	32,68	0,0032	0,0150	34,00
Model II	50	0,0005	0,0344	59,59	0,0012	0,0357	59,66
	100	0,0011	0,0170	60,00	0,0009	0,0174	58,05
	150	0,0010	0,0112	59,82	0,0012	0,0116	60,34
	200	0,0007	0,0086	58,14	0,0014	0,0085	60,00
Model III	50	0,0008	0,0255	80,39	0,0010	0,0255	83,53
	100	0,0021	0,0124	82,26	0,0002	0,0130	77,69
	150	0,0018	0,0082	81,71	0,0004	0,0085	82,35
	200	0,0003	0,0062	80,65	0,0009	0,0061	83,61
Model IV	50	0,0002	0,0265	77,36	0,0002	0,0278	76,62
	100	0,0005	0,0135	75,56	0,0011	0,0132	76,52
	150	0,0005	0,0089	75,28	0,0004	0,0089	78,65
	200	0,0015	0,0065	76,92	0,0026	0,0067	76,12
Model V	50	0,0003	0,0266	77,07	0,0039	0,0279	76,34
	100	0,0003	0,0136	75,00	0,0014	0,0137	73,72
	150	0,0013	0,0085	78,82	0,0028	0,0089	78,65
	200	0,0025	0,0066	75,76	0,0006	0,0065	78,46
Model VI	50	0,0037	0,0241	85,06	0,0008	0,0261	81,61
	100	0,0010	0,0120	85,00	0,0022	0,0124	81,45
	150	0,0000	0,0083	80,72	0,0005	0,0081	86,42
	200	0,0009	0,0060	83,33	0,0015	0,0060	85,00

4.3. Uygulama I

Bu bölümde aykırı değerin regresyon parametrelerinin LS tahmini üzerindeki etkisini göstermek için literatürden alınan bir veri seti incelenmiştir.

Belirli bir sentetik lif üretiminde kullanılan hammadde, nem kontrolü yapılmaksızın bir yerde depolanır. Depo yerindeki bağıl nem ölçümleri ve hammaddenin bir örneğinin nem miktarı, aşağıdaki veri (yüzde olarak) şeklinde 15 gün içinde alınmıştır (Ross, 2009).

Bağıl Nem:	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Nem İçeriği:	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

İlk olarak veri setinin normal dağılıp dağılmadığı incelenmiştir. Tablo 4.2’de verilen Kolmogorov-Smirnov test istatistiği (KS_t) sonuçlarına göre $p > \alpha$ ($\alpha = 0.05$) olduğundan bu veri setinin Normal dağılım gösterdiği söylenebilir. Ayrıca, veriye ait saçılım grafiği ve regresyon doğrusu aynı grafik üzerinde gösterilmiştir ve Şekil 4.1’de verilmiştir. Şekil 4.1’e göre, gözlemlerin regresyon doğrusu boyunca rasgele bir biçimde saçılım gösterdiği gözlenmiştir.

Tablo 4.2 Standartlaştırılmamış Artıklar Kolmogorov-Smirnov test istatistiği sonuçları

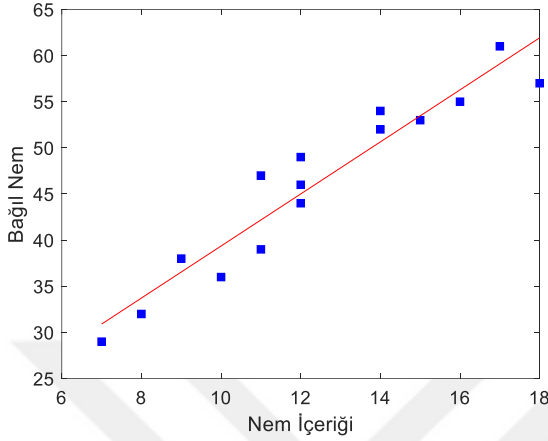
	KS_t	p-değeri
Artıklar	0,104	0,200

Yukarıda verilen veri setinde dört farklı durum için bazı aykırı değerler oluşturulmuştur. Aykırı değer elde etmek için,

$$U \left(X_{(n)} + 3\sqrt{Var(X)}, X_{(n)} + 5\sqrt{Var(X)} \right) \quad (4.15)$$

formülü kullanılmıştır. Burada, $X_{(n)} = \max(X_1, X_2, \dots, X_n)$ ’dir. Aynı zamanda aykırı değerin olmaması (Durum 0) ve olması (Durum 1-4) durumunda elde edilen parametre tahminleri ve bu tahminlerin standart hataları Tablo 4.3’te verilmiştir. Bu tabloda parantez içinde yer alan yüzdeler, bu çalışmada oluşturulan birtakım aykırı değer senaryoları karşısında parametre tahminlerinin ya da onların standart hatalarının ne ölçüde etkilendiğini göstermektedir. Ayrıca, içerisinde aykırı değer ya da değerlerin olduğu (Şekil 4.1-4.5) durumlar için veriye ait saçılım grafikleri (scatter plot) ve regresyon doğruları birlikte verilmiştir.

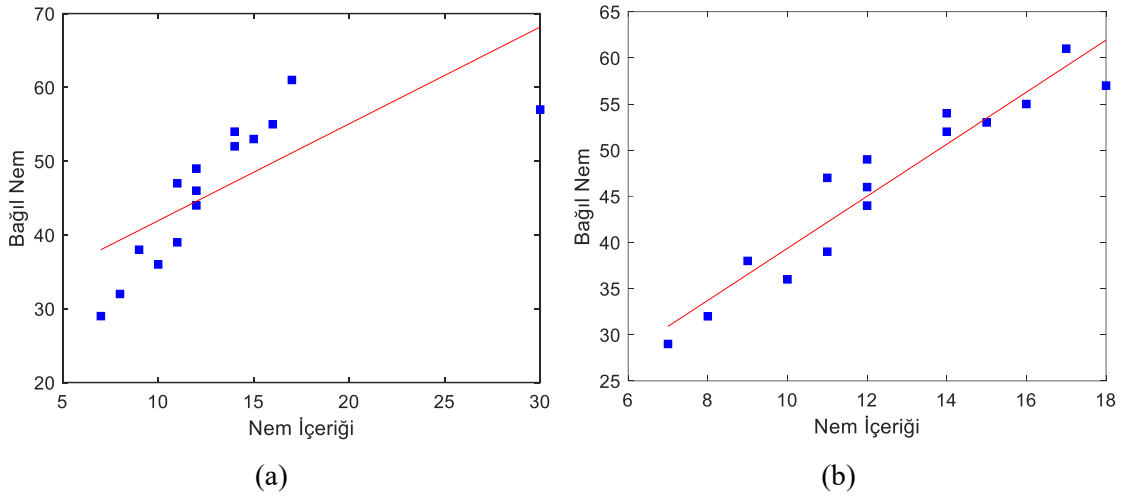
Çalışmanın bu kısmında orijinal veride yer alan bazı gözlem ya da gözlemler aykırı gözlem ya da gözlemler ile değiştirilerek modifiye bir veri seti elde edilmiştir. Aşağıda bu veri setinin nasıl elde edildiği açıklanmıştır. Dört farklı senaryoya göre elde edilen modifiye veri seti Durum 1-4’de açıklanmıştır.



Şekil 4.1 Orijinal veriye ilişkin regresyon doğrusu ve saçılım grafiği

Durum 1:

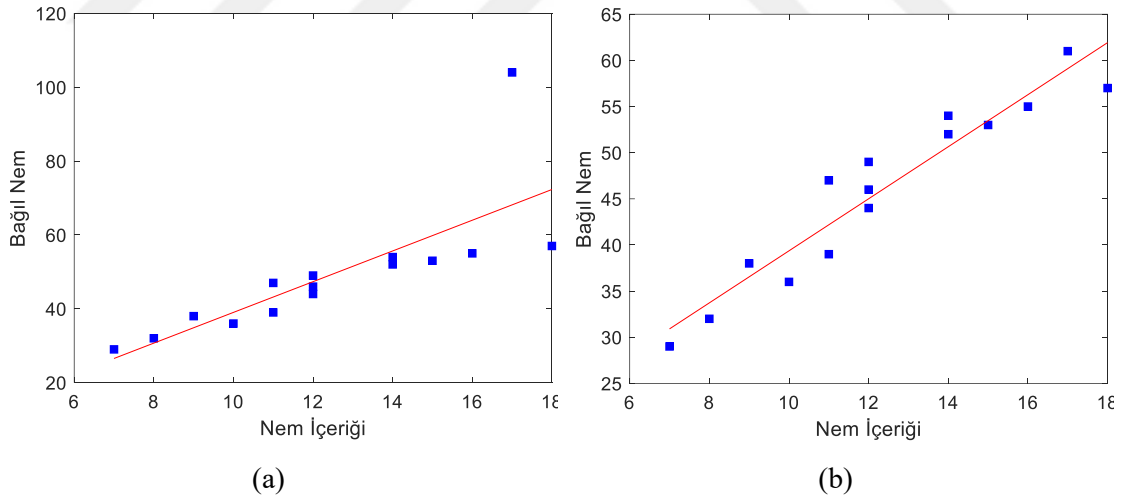
Burada, orijinal veride yer alan $x_{13} = 18$ gözlemi Eş. (4.15) kullanılarak $x_{13}^* = 30$ olarak değiştirilmiştir. X -yönünde bir tane aykırı gözlem olan veriye ilişkin saçılım grafiği ve regresyon doğrusu çizdirilmiştir ve Şekil 4.2’de gösterilmiştir. Şekil 4.2 incelendiğinde, aykırı değerın regresyon doğrusunu kendine doğru çektiği görülmektedir. Ayrıca bir tane aykırı gözlem olması durumunda bile parametrelerin tahminlerinin büyük ölçüde etkilendiği görülmüştür (bkz. Tablo 4.3).



Şekil 4.2 (a) X-yönünde bir tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği, (b) orijinal veriye ilişkin regresyon doğrusu ve saçılım grafiği

Durum 2:

$y_4 = 61$ iken $y_4^* = 104$ olarak değiştirildiğinde elde edilen saçılım grafiği ve regresyon doğrusunun yer aldığı Şekil 4.3'e göre; bu gözlem aykırı gözlemdir. Ancak etkili bir gözlem değildir.

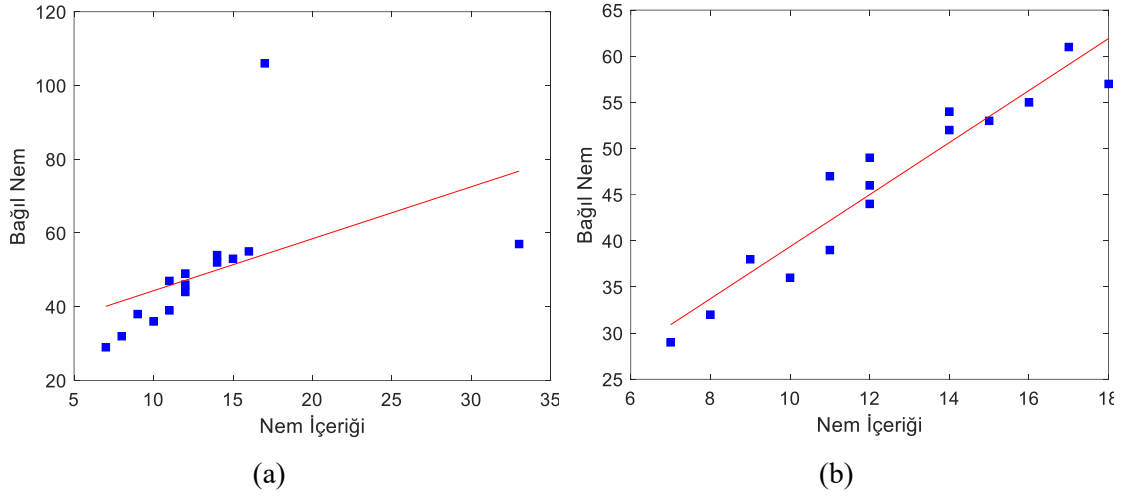


Şekil 4.3 (a) Y-yönünde bir tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği, (b) orijinal veriye ilişkin regresyon doğrusu ve saçılım grafiği

Durum 3:

$x_{13} = 18$ İken $x_{13}^* = 33$ olarak ve $y_4 = 61$ iken $y_4^* = 106$ olarak değiştirildiğinde bu duruma ilişkin grafik Şekil 4.4'te verilmiştir. Buna göre, özellikle X-yönünde elde edilen aykırı değer regresyon doğrusunun yönünü değiştirmekte etkili olduğu görülmektedir.

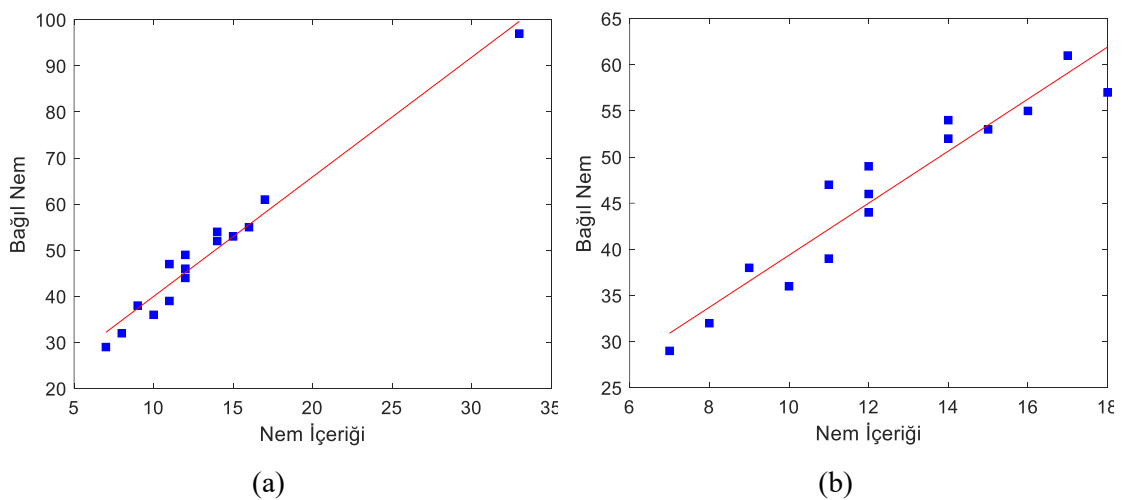
Doğruyu kendi yönüne doğru çekmektedir.



Şekil 4.4 (a) Hem X -yönünde hem de Y -yönünde birer tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği, (b) orijinal veriye ilişkin regresyon doğrusu ve saçılım grafiği

Durum 4:

$x_{13} = 18$ iken $x_{13}^* = 33$ olarak ve $y_{13} = 57$ iken $y_{13}^* = 97$ olarak değiştirildiğinde oluşan Şekil 4.5'e göre, hem X -yönünde hem de Y -yönünde edilen bu aykırı gözlem regresyon doğrusunun yönünü değiştirmemiştir ve doğrunun üzerinde yer almıştır.



Şekil 4.5 (a) 13. gözlemin hem X -yönünde hem de Y -yönünde bir tane aykırı değer olduğu duruma ilişkin regresyon doğrusu ve saçılım grafiği, (b) orijinal veriye ilişkin regresyon doğrusu ve saçılım grafiği

Tablo 4.3 Farklı aykırı değer senaryolarına göre elde edilen parametre tahminleri ve bu tahminlerin standart hataları

Durum	$\hat{\beta}_0$	$\hat{\beta}_1$	Shata($\hat{\beta}_0$)	Shata($\hat{\beta}_1$)
0	11,1680	2,8198	3,1196	0,2439
1	28,8060 (%157)	1,3127 (-%53)	4,6129 (%47)	0,3246 (%33)
2	-2,5827 (-%123)	4,1599 (%47)	12,1725 (%290)	0,9517 (%290)
3	30,2367 (%170)	1,4102 (-%49)	10,4403 (%234)	0,7127 (%192)
4	14,0457 (%25)	2,5936 (-%8)	1,8988 (-%39)	0,1296 (-%46)

5. BASİT DOĞRUSAL REGRESYON MODELİNDE GUMBEL DAĞILIMI VARSAYIMINDA VE VERİ SETİ AYKIRI DEĞER İÇERDİĞİNDE PARAMETRELERİN TAHMİNİ PARAMETRELERİN TAHMİNİ

Bu bölümde, regresyon analizinde hata terimlerinin klasik varsayımı olan Normal dağılıma uymadığı durum ele alınmaktadır. Hata terimlerinin simetrik olmayan dağılımlardan Gumbel dağılımına uyduğu varsayılmıştır. Bu durumda Eş. (3.3)'de verilen basit doğrusal regresyon modelinde, hata terimlerinin oyf

$$f(\varepsilon_i|\sigma) = \frac{1}{\sigma} \exp\left(-\frac{\varepsilon_i}{\sigma} - \exp\left(-\frac{\varepsilon_i}{\sigma}\right)\right); \varepsilon_i \in \mathbb{R}, \sigma > 0$$

olur. Model parametrelerinin tahmini bu varsayım altında yapılmıştır. Ayrıca $\sigma = 1$ olarak bilindiği varsayılmıştır.

5.1. Parametre Tahmini

Bu bölümde, regresyon analizinde hata terimlerinin klasik varsayımı olan Normal dağılıma uymadığı durumlarda model parametrelerinin tahmininde geleneksel yöntemler etkinliğini ve güvenilirliğini yitirmektedir. Bu nedenle, literatürde dayanıklı tahmin yöntemleri olarak adlandırılan alternatif yaklaşımlar geliştirilmiştir. Bu kapsamda, LAD WLAD ve LMS olmak üzere üç temel dayanıklı tahmin yöntemi kullanılmıştır. Söz konusu yöntemlerin parametre tahmin sürecinde nasıl uygulandıkları ele alınacak ve geleneksel yöntemlerle karşılaştırmalı değerlendirmelerine yer verilecektir.

5.1.1. En Küçük Mutlak Sapma Tahmin Edicileri

Bu yöntem ile model parametrelerinin LAD tahmin edicileri aşağıdaki gibi elde edilir.

$$\phi_{LAD}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|$$

$\phi_{LAD}(\hat{\beta}_0, \hat{\beta}_1)$ fonksiyonunu minimum yapan $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri, β_0 ve β_1 model parametrelerinin LAD tahmin edicileridir. Bu optimizasyon problemi kapalı formda çözümü olmadığından, sayısal optimizasyon yöntemleri kullanılarak çözüm elde edilmektedir. LAD tahmin yönteminin en önemli özelliği, aykırı gözlemlere karşı yüksek düzeyde dayanıklılık sağlamasıdır. Çünkü kareler toplamı yerine mutlak sapmaların minimize edilmesi, aykırı değerlerin tahminler üzerindeki etkisini önemli ölçüde azaltmaktadır.

5.1.2. Ağırlıklı Mutlak Sapma Tahmin Edicileri

WLAD yöntemi kullanılarak, β_0 ve β_1 parametrelerinin WLAD tahminleri

$$\phi_{WLAD}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n w_i |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|$$

ile verilen $\phi_{WLAD}(\hat{\beta}_0, \hat{\beta}_1)$ fonksiyonunu minimum yapan parametre değerleri olarak alınır. WLAD yönteminin uygulanmasındaki temel zorluklardan biri, w_i ile gösterilen ağırlık fonksiyonlarının seçimidir. Burada amaç, WLAD yönteminin hem doğruluğunu hem de aykırı gözlemlere karşı dayanıklılığını artıracak şekilde w_i ağırlıklarının belirlenmesidir. Bu çalışmada,

$$w_i = \frac{i(n-i+1)}{(n+2)(n+1)^2}$$

olarak alınmıştır. Özellikle, aykırı gözlemlerin yoğun olduğu veri setlerinde ağırlıklandırma sayesinde parametre tahminlerinin sapması azaltılmakta ve daha dayanıklı sonuçlar elde edilmektedir. Ayrıca, farklı ağırlık fonksiyonlarının seçimi, yöntemin esnekliğini artırarak değişik veri yapılarında daha uygun çözümler sunabilmektedir.

5.1.3. En Küçük Medyan Kareler Tahmin Edicileri

LMS ile model parametrelerinin tahmin edicileri aşağıdaki gibi elde edilir:

$$\phi_{LMS}(\hat{\beta}_0, \hat{\beta}_1) = \text{Med}_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

LMS yöntemi ile, $\phi_{LMS}(\hat{\beta}_0, \hat{\beta}_1)$ hata karelerinin medyanı bilinmeyen parametrelere göre minimize edilerek, regresyon parametrelerinin tahminleri elde edilir. Ancak, bu fonksiyonun kapalı formda analitik bir çözümü bulunmadığından, parametre tahminlerinin hesaplanmasında genellikle sayısal optimizasyon yöntemleri kullanılmaktadır.

LMS yöntemi ile elde edilen tahmin ediciler yüksek düzeyde dayanıklılığa sahiptir. Özellikle veri setinde aykırı gözlemlerin oranı %50'nin altında olduğu durumlarda, LMS tahmin edicileri güvenilir ve tutarlı sonuçlar verebilmektedir. Bu açıdan LMS yöntemi, aykırı değerlere karşı daha güçlü bir direnç göstermektedir. Ayrıca, hata dağılımının simetrik olmaması durumlarında dahi güvenilir tahminler sunabilmektedir. Bununla birlikte, LMS yönteminin aykırı gözlemlerin bulunmadığı veya az sayıda olduğu

durumlarda etkinliđi daha düşük olabilmektedir. Dolayısıyla, LMS yöntemi özellikle aykırı değerlerin fazla olduđu veri setlerinde tercih edilmesi daha uygundur.

5.2. Simülasyon Çalışması II

Bu bölümde hata terimlerinin Gumbel dağılımına sahip olduđu varsayımında basit bir doğrusal regresyon modelinin parametrelerini tahmin etmek için Bölüm 5.1’de tanıtılan tahmin yöntemlerinin ve LS yönteminin performansları bir Monte-Carlo simülasyon çalışması ile değerlendirilecektir. Tahmin edicilerin ekinliklerini değerlendirme kriteri olarak yan (Bias) ve hata kareler ortalaması (MSE) kullanılacaktır.

Simülasyon çalışması aşağıdaki adımlardan oluşmaktadır:

- A.1. Parametre değerleri $\beta_0 = 0$ ve $\beta_1 = 1$ olarak alınmıştır.
- A.2. Örneklem hacmi $n = 50, 100, 150$ ve 200 ’dür.
- A.3. Aykırı değer oranları $\xi = \%0, \%5, \%10$ ve $\%15$ olarak alınmıştır.
- A.4. Verideki aykırı değer sayısı $r = n \times \xi$ olarak alınmıştır.
- A.5. x_i bağımsız deđişken deđerleri için matlab

`normrnd(0,1)`

komutu kullanılarak rasgele sayı üretilmiştir.

- A.6. Gumbel dağılımına sahip hata terimleri için rasgele sayı üretmek için

$$\varepsilon_i = F^{-1}(u_i) = -\ln(-\ln u_i),$$

formülü ile verilen ters dağılım fonksiyonu tekniđi kullanılmıştır. Burada, $u_i \sim U(0,1)$, $i = 1, 2, \dots, n$.

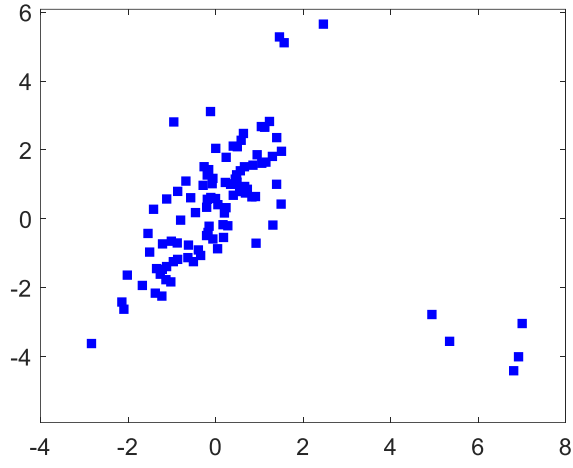
- A.6. y_i bağımlı deđişken deđerleri, A.5 ve A.6’da elde edilen bağımsız deđişken deđerleri ve hata terimi deđerleri Eş. (3.3)’de verilen denklemde yerine yazılarak elde edilmiştir.

- A.7. Rasgele seçilmiş r tane deđer, düzgün dağılım kullanılarak elde edilen X -yönündeki r tane aykırı deđerle deđiştirilmiştir (bkz. Şekil 5.1).

- A.8. β_0 ve β_1 parametrelerinin LS, LAD, WLAD ve LMS tahminleri elde edilmiştir.

- A.9. Bias, MSE ve RE ölçülerinin deđerleri hesaplanmıştır.

Burada Adım 5-8 arasındaki işlemler 5000 kez tekrarlanmıştır. Ayrıca, tüm hesaplamalar için MATLAB programlama dili kullanılmıştır. Farklı örneklem büyüklükleri ve X -yönündeki farklı aykırı deđer oranlarına göre elde edilen parametrelerin tahminleri, Bias, MSE ve RE deđerleri Tablo 5.1-5.2’de verilmiştir.



Şekil 5.1 $T=1$, $n = 100$ ve $\xi = 0,05$ için elde edilen simülasyon verilerinin saçılım grafiği.

Tablo 5.1 Farklı örneklem büyüklükleri için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri ve verilerin %0 ve %5 oranında aykırı değer içermesi durumunda değişen örneklem büyüklüklerine karşılık gelen Bias, MSE ve RE değerleri

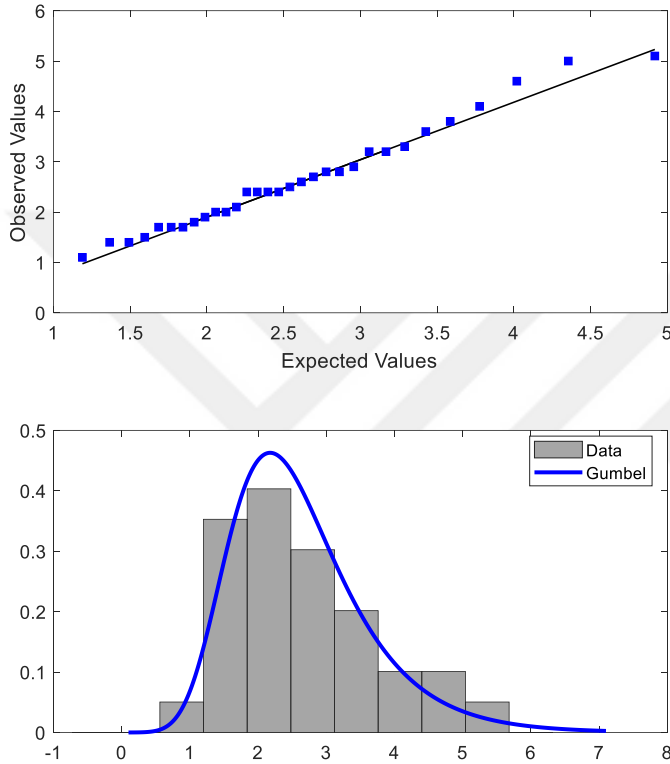
Aykırı Değer Oranı	Örneklem Hacmi	Yöntem	β_0			β_1		
			Bias	MSE	RE	Bias	MSE	RE
%0	n=50	LS	0,5739	0,3642	100,00	0,0012	0,0358	100,00
		LAD	0,3767	0,1864	100,00	0,0038	0,0471	100,00
		WLAD	0,3869	0,2008	100,00	0,3702	0,1823	100,00
		LMS	0,0097	0,0059	100,00	0,0164	0,0061	100,00
	n=100	LS	0,5708	0,3440	100,00	0,0026	0,0170	100,00
		LAD	0,3674	0,1556	100,00	0,0027	0,0212	100,00
		WLAD	0,3839	0,1714	100,00	0,3816	0,1673	100,00
		LMS	0,0037	0,0013	100,00	0,0207	0,0017	100,00
	n=150	LS	0,5716	0,3394	100,00	0,0003	0,0115	100,00
		LAD	0,3696	0,1505	100,00	0,0003	0,0145	100,00
		WLAD	0,3872	0,1660	100,00	0,3866	0,1635	100,00
		LMS	0,0023	0,0006	100,00	0,0177	0,0009	100,00
	n=100	LS	0,5736	0,3395	100,00	0,0021	0,0084	100,00
		LAD	0,3712	0,1484	100,00	0,0023	0,0108	100,00
		WLAD	0,3879	0,1625	100,00	0,3871	0,1604	100,00
		LMS	0,0018	0,0003	100,00	0,0153	0,0005	100,00
%5	n=50	LS	0,3984	0,2175	167,45	1,0140	1,0405	3,44
		LAD	0,3125	0,1550	120,26	0,6042	0,4616	10,20
		WLAD	0,2969	0,1419	141,51	0,4935	0,2985	61,07
		LMS	0,0172	0,0096	61,46	0,0180	0,0096	63,54
	n=100	LS	0,4143	0,2012	170,97	1,0282	1,0641	1,60
		LAD	0,3212	0,1307	119,05	0,5307	0,3261	6,50
		WLAD	0,3124	0,1239	138,34	0,4432	0,2201	76,01
		LMS	0,0057	0,0018	72,22	0,0158	0,0020	85,00
	n=150	LS	0,4138	0,1916	177,14	1,0821	1,1758	0,98
		LAD	0,3246	0,1245	120,88	0,6117	0,4136	3,51
		WLAD	0,3100	0,1126	147,42	0,4442	0,2132	76,69
		LMS	0,0038	0,0004	150,00	0,0176	0,0011	81,82
	n=200	LS	0,4180	0,1905	178,22	1,0843	1,1800	0,71
		LAD	0,3246	0,1196	124,08	0,5881	0,3769	2,87
		WLAD	0,3119	0,1099	147,86	0,4333	0,1992	80,52
		LMS	0,0026	0,0004	75,00	0,0171	0,0006	83,33

Tablo 5.2 Farklı örneklem büyüklükleri için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri ve verilerin %10 ve %15 oranında aykırı değer içermesi durumunda değişen örneklem büyüklüklerine karşılık gelen Bias, MSE ve RE değerleri

Aykırı Değer Oranı	Örneklem Hacmi	Yöntem	β_0			β_1		
			Bias	MSE	RE	Bias	MSE	RE
%10	n=50	LS	0,3669	0,2047	177,92	1,1807	1,4023	2,55
		LAD	0,2401	0,1469	126,89	1,1486	1,3990	3,37
		WLAD	0,2282	0,1115	180,09	0,6863	0,5455	33,42
		LMS	0,0162	0,0081	72,84	0,0156	0,0081	75,31
	n=100	LS	0,3726	0,1751	196,46	1,2325	1,5243	1,12
		LAD	0,2477	0,1148	135,54	1,2718	1,6544	1,28
		WLAD	0,2321	0,0834	205,52	0,6474	0,4569	36,62
		LMS	0,0086	0,0024	54,17	0,0204	0,0028	60,71
	n=150	LS	0,3790	0,1688	201,07	1,2576	1,5866	0,72
		LAD	0,2546	0,1037	145,13	1,3317	1,7907	0,81
		WLAD	0,2310	0,0726	228,65	0,6410	0,4369	37,42
		LMS	0,0051	0,0009	66,67	0,0197	0,0013	69,23
	n=200	LS	0,3813	0,1653	205,38	1,2728	1,6241	0,52
		LAD	0,2624	0,1005	147,66	1,3588	1,8570	0,58
		WLAD	0,2330	0,0693	234,49	0,6402	0,4291	37,38
		LMS	0,0041	0,0006	50,00	0,0178	0,0009	55,56
%15	n=50	LS	0,3342	0,1975	184,41	1,2982	1,6937	2,11
		LAD	0,1855	0,1509	123,53	1,4015	1,9834	2,37
		WLAD	0,0830	0,0809	248,21	1,1129	1,3012	14,01
		LMS	0,0229	0,0110	53,64	0,0168	0,0108	56,48
	n=100	LS	0,3540	0,1686	204,03	1,3237	1,7585	0,97
		LAD	0,2210	0,1150	135,30	1,4162	2,0175	1,05
		WLAD	0,1024	0,0527	325,24	1,0787	1,2130	13,79
		LMS	0,0089	0,0032	40,63	0,0199	0,0036	47,22
	n=150	LS	0,3614	0,1605	211,46	1,3454	1,8153	0,63
		LAD	0,2364	0,1020	147,55	1,4255	2,0415	0,71
		WLAD	0,0944	0,0387	428,94	1,1325	1,3168	12,42
		LMS	0,0073	0,0015	40,00	0,0178	0,0018	50,00
	n=200	LS	0,3661	0,1569	216,38	1,3541	1,8382	0,46
		LAD	0,2480	0,0961	154,42	1,4311	2,0559	0,53
		WLAD	0,1020	0,0330	492,42	1,1181	1,2831	12,50
		LMS	0,0050	0,0008	37,50	0,0171	0,0011	45,45

5.3. Uygulama II

Bu bölümde, çalışmada ele alınan tahmin yöntemlerinin uygulanabilirliğini göstermek için literatürden alınan bir veri seti analiz edilmiştir. Bu amaçla, Aydın (2018b) tarafından Gumbel dağılımı kullanılarak modellenen rüzgâr hızı verilerini kullanıyoruz (bkz. Şekil 5.2). Veri seti, Türkiye'nin Sinop kentinde 2015 yılının Ekim ayındaki günlük ortalama rüzgâr hızını içermektedir.



Şekil 5.2 Rüzgâr hızı verileri için Gumbel dağılımı QQ ve uyum grafikleri

Veriler için örnek işlem aşağıdaki adımlardan oluşur:

A.1. Günlük ortalama rüzgâr hızı verileri ε_i artıkları olarak alınmıştır.

A.2. x_i bağımsız değişken değerleri Standart Normal dağılımdan rasgele üretilmiştir.

Üretilen x_i değerleri:

2,9491,	-0,6300,	-0,0469,	2,6830,
-1,1467,	0,5530,	-1,0765,	1,0306,
0,3275,	0,6521,	-0,2789,	0,2452,
1,4725,	-2,2751,	-1,6333,	0,4155,
-0,6548,	-0,2963,	-1,4969,	-0,9048,
-0,4042,	-0,7258,	-0,8665,	-0,4218,
-0,9427,	1,3419,	-0,9884,	1,8179,
-0,3744,	-1,4517,	-0,6187,	

A.3. 1-2 Adımlardaki ε_i artıkları ve x_i değerleri aşağıdaki formülde yerine yazılarak y_i değerleri elde edilmiştir (Islam vd., 2001).

$$y_i = x_i + \varepsilon_{(i)}, \quad i = 1, 2, \dots, 31$$

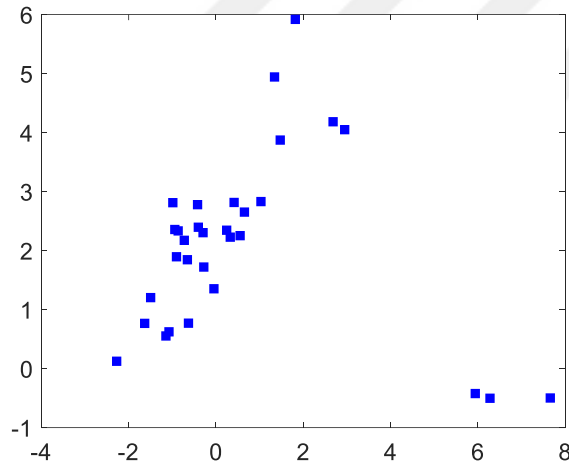
A.4. $\xi = \%0, \%5, \%10$ ve $\%15$ aykırı değer oranlarına göre verideki aykırı değer sayısı, $r = n \times \xi$ formülü ile belirlenmiştir.

A.5. Rasgele seçilmiş r tane veri, düzgün dağılım kullanılarak elde edilen X -yönündeki r tane aykırı değerle değiştirilmiştir (bkz. Şekil 5.3).

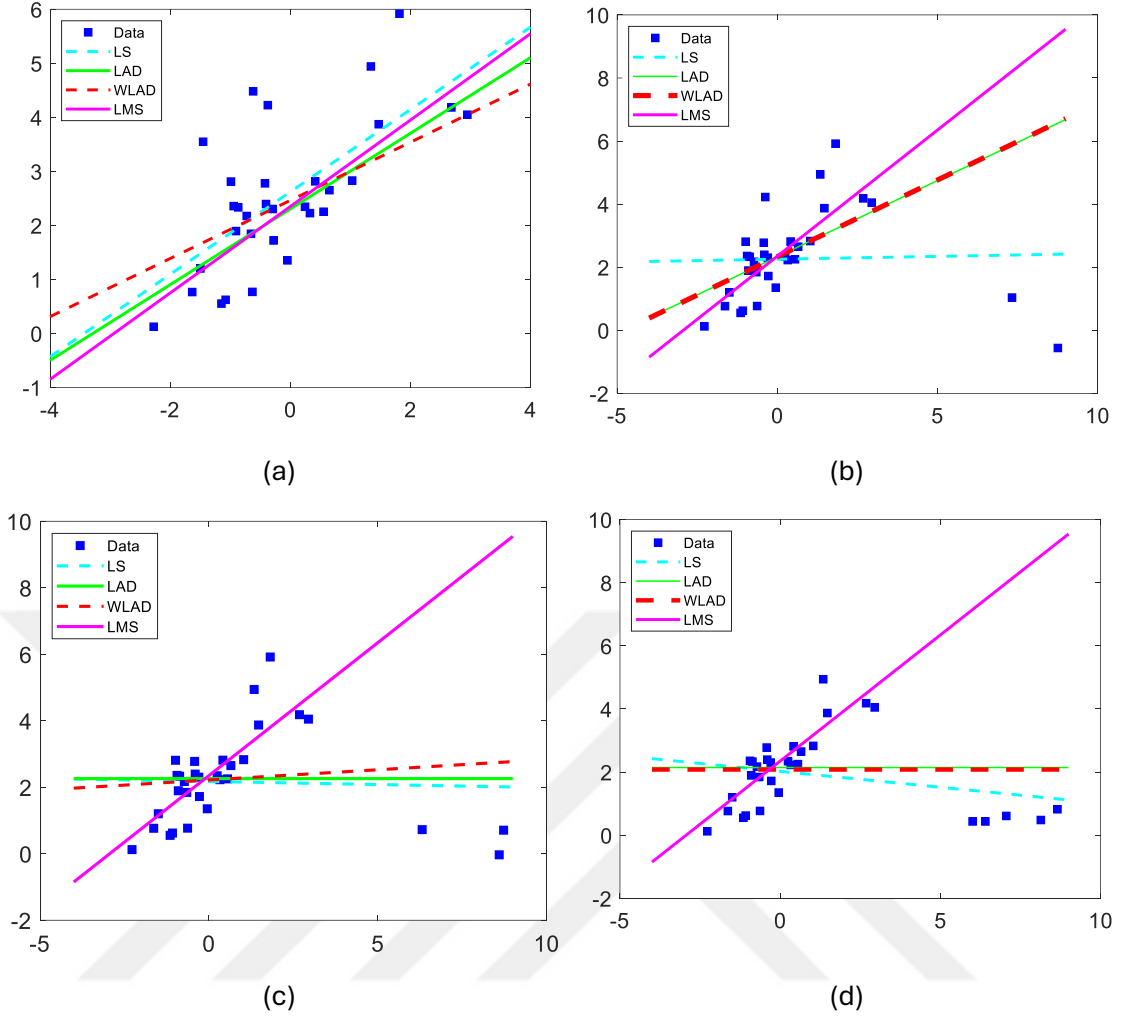
A.6. Farklı aykırı değer oranları için, model parametreleri β_0 ve β_1 'in LS, LAD, WLAD ve LMS tahminleri elde edilmiştir ve sonuçlar Tablo 5.3'te gösterilmiştir.

Ayrıca, veri setinde bulunan aykırı değer oranlarına göre elde edilen (bkz. Tablo 5.3) tahmin değerlerine göre çizilen regresyon doğrusu ve saçılım grafikleri Şekil 5.4'te sunulmuştur.

Tablo 5.3'te yer alan parantez içindeki değerler, veride aykırı değer bulunmadığında ve bulunduğu elde edilen tahminler arasındaki farkın mutlak değeridir.



Şekil 5.3 $\xi = 0,10$ olan veri için saçılım grafiği.



Şekil 5.4 (a) Veride aykırı değer bulunmadığında, (b) veride %5 oranında aykırı değer bulunduğunda, (c) veride %10 oranında aykırı değer bulunduğunda, (d) veride %15 oranında aykırı değer bulunduğunda β_0 ve β_1 parametrelerinin LS, LAD, WLAD ve LMS tahminleri ile elde edilen regresyon doğrusu ve saçılım grafiği

Tablo 5.3 Orijinal veriler ve farklı aykırı değer yüzdesi değerlerine sahip değiştirilmiş veriler için model parametrelerinin LS, LAD, WLAD ve LMS tahminleri

Aykırı Değer Oranı	LS		LAD		WLAD		LMS	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
%0	2,6197	0,7625	2,3039	0,7004	2,4630	0,5378	2,3494	0,7997
%5	2,2550 (0,3647)	0,0180 (0,7445)	2,3325 (0,0286)	0,4833 (0,2171)	2,3351 (0,1279)	0,4862 (0,0516)	2,3493 (0,0001)	0,7996 (0,0001)
%10	2,1744 (0,4453)	-0,0184 (0,7809)	2,2610 (0,0429)	-0,1022 (0,8026)	2,2190 (0,2440)	0,0615 (0,4763)	2,3493 (0,0001)	0,7996 (0,0001)
%15	2,0236 (0,5961)	-0,1007 (0,8632)	2,1453 (0,1586)	-0,2172 (0,9176)	2,0811 (0,3819)	-0,1283 (0,6661)	2,3493 (0,0001)	0,7996 (0,0001)

6. BULGULAR

Bu bölümde, tez kapsamında 4. ve 5. bölümlerde yapılan simülasyon çalışmaları ile gerçek veri analizlerinden elde edilen sonuçlar ayrıntılı olarak özetlenmiş ve değerlendirilmiştir.

6.1. Simülasyon Çalışması I ve Uygulama I

Bu bölümde, Eş. (3.3) ile verilen basit doğrusal regresyon modelinde yer alan β_0 ve β_1 parametrelerinin LS tahmin edicilerinin, model varsayımlarından sapmaların veya veri setinde aykırı değerlerin bulunması durumlarındaki performansları Monte-Carlo simülasyon yöntemiyle incelenmiştir. Simülasyon analizi sonucunda elde edilen bulgular Tablo 4.1’de sunulmuştur.

Elde edilen sonuçlar, $\tilde{\beta}_0$ ve $\tilde{\beta}_1$ parametrelerinin LS tahmin edicileri, incelenen tüm aykırı değer modelleri ve normal dağılıma alternatif olarak değerlendirilen tüm dağılımlar için yansızlık özelliğini koruduğunu göstermektedir. Diğer bir ifadeyle, LS tahmin edicilerinin Bias kriteri açısından normallikten sapmalara karşı belirli bir düzeyde dayanıklılık sergilediği söylenebilir.

MSE kriterine göre, hata terimlerinin gerçek dağılımı, Model I–III kapsamında farklı serbestlik derecelerine sahip Student-t dağılımları olduğunda, LS tahmin edicilerinin etkinliğinde belirgin bir düşüş gözlenmiştir. Bununla birlikte, serbestlik derecesinin artmasıyla birlikte LS tahmin edicilerinin etkinliğinde iyileşme olduğu tespit edilmiştir. Ayrıca, Model I durumunda β_0 ve β_1 parametrelerine ait LS tahmin edicilerinin etkinliğinin en düşük düzeyde olduğu belirlenmiştir.

Örnekleme büyüklüğünün artmasıyla birlikte LS tahmin edicilerine ait MSE değerlerinin azaldığı, dolayısıyla tahmin edicilerin etkinliklerinin arttığı görülmüştür.

Simülasyon sonuçlarına ait MSE değerleri, β_0 ve β_1 parametrelerinin LS tahmin edicilerinin performanslarının, Model IV–VI’da tanımlanan sapmalardan, Model I–III’te yer alan sapmalara göre daha az etkilendiğini göstermiştir.

Uygulama I kapsamında, basit doğrusal regresyon modeli için literatürde yer alan bir veri seti kullanılarak çeşitli aykırı gözlem senaryoları analiz edilmiştir. Bu kapsamda, Durum 1’de X-yönünde tek bir aykırı gözlem, Durum 2’de Y-yönünde tek bir aykırı gözlem, Durum 3’te hem X- hem de Y-yönünde birer aykırı gözlem (toplam iki aykırı değer) ve

Durum 4'te ise hem X hem de Y-yönlerinde aynı anda görülen bir aykırı gözlemin yer aldığı modeller ayrı ayrı incelenmiştir. Analiz sonuçları Tablo 4.3'te sunulmuştur.

Elde edilen bulgular, LS tahmin edicilerinin incelenen tüm aykırı gözlem senaryolarından etkilendiğini göstermektedir. Özellikle, tahminlerin standart hataları dikkate alındığında, LS tahmin edicilerinin Durum 2 ve Durum 3'teki aykırı değer modellerinden belirgin şekilde etkilendiği görülmüştür. Yanlılık açısından değerlendirildiğinde ise, β_0 parametresine ait LS tahmin edicisinin Durum 1-3'te tanımlanan aykırı değer modellerine karşı oldukça duyarlı olduğu gözlenmiştir. Buna karşılık hem yanlılık hem de standart hata birlikte değerlendirildiğinde, β_0 ve β_1 parametrelerine ait LS tahmin edicilerinin performanslarının Durum 4'teki aykırı gözlem modelinden nispeten daha az etkilendiği sonucuna ulaşılmıştır. Bu bulgu, literatürde yer alan benzer çalışmalarla da tutarlılık göstermektedir.

6.2. Simülasyon Çalışması II ve Uygulama II

Bu bölümde, hata terimlerinin Gumbel dağılımına uyduğu varsayımı altında, basit doğrusal regresyon model parametrelerinin tahmininde kullanılan dayanıklı tahmin yöntemleri LAD, WLAD ve LMS ile klasik LS yönteminin performansları, Monte-Carlo simülasyon analizi aracılığıyla değerlendirilmiştir. Simülasyon çalışması sonuçları Tablo 5.1-5.2'de sunulmuştur.

Monte-Carlo simülasyonu sonucunda elde edilen bulgular, veri seti aykırı değerler içermediğinde, β_0 parametresinin Bias ve MSE değerleri ile β_1 parametresinin MSE değerleri açısından LMS tahmin edicisinin, tüm örneklem büyüklüklerinde diğer tahmin ediciler arasında en iyi performansı gösterdiğini ortaya koymuştur. Ancak, β_1 parametresinin LAD ve LS tahmin edicileri, tüm örneklem büyüklüklerinde en düşük Bias değerlerine sahip olmaları nedeni ile diğer tahmin yöntemlerine kıyasla daha iyi performans göstermektedir. Bu sonuç, LMS tahmin edicisinin genel üstün performansına rağmen, geleneksel LS yönteminin Bias kriteri açısından özellikle β_1 parametresi için hâlâ güvenilir tahminler sağlayabildiğini göstermektedir. LAD ve WLAD dayanıklı tahmin ediciler olmalarına rağmen, β_1 parametresi için LS tahmin edicisine kıyasla MSE açısından daha düşük performans sergilemişlerdir (bkz. Tablo 5.1). Bu sonuç, LAD ve WLAD yöntemlerinin aykırı değerlerin parametre tahmin sürecindeki olumsuz etkilerini azaltmak amacıyla geliştirilmiş olsalar da bazı durumlarda belirli parametreler açısından

her zaman LS yönteminden daha iyi performans sergileyemeyebilecekleri anlamına gelmektedir.

Tüm aykırı değer senaryolarında, β_0 parametresine ait LS, LAD ve WLAD tahmin edicilerinin yanlı tahminler ürettiği gözlemlenmiştir. Bununla birlikte, söz konusu tahmin edicilerin Bias değerlerinin, aykırı değer içermeyen veri setlerinden elde edilen Bias değerlerine kıyasla daha düşük olduğu belirlenmiştir. LMS tahmin edicisi hem Bias hem de MSE değerleri bakımından diğer dayanıklı tahmin yöntemleri olan LAD ve WLAD'a göre daha yüksek bir performans sergilemiştir. Ancak, MSE kriteri açısından değerlendirildiğinde, LS tahmin edicisinin bazı durumlarda LAD tahmin edicisine kıyasla daha iyi performans göstermiştir. LMS tahmin edicisi, aykırı değerlere karşı duyarlılığının düşük olması nedeniyle farklı aykırı değer senaryolarında güçlü bir dayanıklılık sergilemekte olup, bu yönüyle en güvenilir yöntem olarak öne çıkmaktadır. WLAD tahmin edicisi ise dayanıklılık bakımından LMS'in ardından ikinci sırada yer almaktadır. Öte yandan, LS tahmin edicisi, aykırı gözlem içermeyen veri setlerinde, özellikle eğim parametresi β_1 için Bias açısından iyi bir performans gösterse de aykırı değerlere karşı dirençsizdir. Aykırı değerlerin oranı arttıkça, LS tahmin edicisinin Bias ve MSE değerleri, dayanıklı tahmin yöntemlerine kıyasla daha yüksek seviyelere ulaşmakta; bu durumda, performansı belirgin biçimde düşmektedir. Bununla birlikte, LAD ve WLAD tahmin edicileri, aykırı değerlerin olumsuz etkilerini azaltmak amacıyla geliştirilmiş dayanıklı tahmin yöntemleri olmasına rağmen, aykırı değer oranının yüksek olduğu durumlarda LMS tahmin edicisine kıyasla daha düşük performans sergilemektedirler.

Uygulama II'de, Bölüm 5'te ele alınan hata terimlerinin Gumbel dağılımına uyduğu varsayımı altında, X-yönünde aykırı değerlerin olduğu durumlarda model parametrelerinin tahmini için önerilen yöntemlerin uygulanabilirliğini değerlendirmek amacıyla literatürde yer alan bir veri seti analiz edilmiştir. Bu kapsamda, veri setine yapay olarak aykırı değerler eklenerek veri seti üzerinde gerekli düzenlemeler (modifikasyonlar) yapılmıştır. Yapılan analizler sonucunda elde edilen bulguların, simülasyon çalışmalarıyla tutarlılık gösterdiği belirlenmiştir. LMS yöntemi, hem hata terimlerinin normallik varsayımını sağlamadığı hem de veri setinde aykırı değerlerin bulunduğu koşullarda en yüksek performansı sergilemiştir (bkz. Şekil 5.4). Performans açısından WLAD yöntemi, LMS tahmin edicisinin ardından ikinci en başarılı yöntem olarak öne çıkmıştır.

7. SONUÇ VE ÖNERİLER

Bu çalışmada, basit doğrusal regresyon modelinde hata terimlerinin Gumbel dağılımına sahip olduğu varsayımı altında model parametrelerinin tahmin edilmesi problemi ele alınmıştır. Literatürde bu konu, Al-Adilee ve Mohamed (2014) tarafından incelenmiş ve söz konusu çalışmada model parametrelerinin ML tahmin edicileri elde edilmiştir. Bu çalışma, literatürdeki önceki çalışmadan farklı olarak, literatürde dayanıklı tahmin yöntemleri olarak kabul edilen LAD, WLAD ve LMS ile klasik LS yöntemi kullanılarak tahmin edilmiştir. Tahmin edicilerin dayanıklılık performansları, regresyon doğrusunun eğimini önemli ölçüde etkileyen X -yönündeki aykırı değerlerin, farklı yüzdeleri içeren veri senaryoları üzerinde yürütülen Monte-Carlo simülasyonu kullanılarak karşılaştırılmıştır. Elde edilen sonuçlar, tahmin yöntemlerinin performanslarını karşılaştırmak amacıyla Bias ve MSE ölçüleri temel alınarak analiz edilmiştir.

Simülasyon sonuçları, LMS tahmin edicisinin hem β_0 hem de β_1 parametrelerini tahmin etmede çoğu senaryo için en yüksek doğruluk ve güvenilirliği sağladığını göstermektedir. WLAD tahmin edicisi, performans açısından ikinci sırada yer almakta ve LMS'den sonra tercih edilebilir bir alternatif olarak öne çıkmaktadır.

Simülasyon ile elde edilen sonuçların, tutarlılığını göstermek için literatürden alınan gerçek veri analizi yapılmıştır. Veri setinde aykırı değer olmadığı için X -yönünde farklı yüzdelerde aykırı gözlemler eklenmiş ve model parametrelerinin tahminleri elde edilmiştir. Gerçek veri analiz sonuçları LMS tahmin edicisinin model parametrelerini tahmin etmede üstün performansının simülasyon sonuçları ile uyumlu olduğu görülmüştür. Simülasyon ve gerçek veri analiz sonuçları, yüksek bozulma noktalarına sahip veri setlerinde LMS tahmin edicisinin veri kirliliğine karşı daha dirençli olduğunu ve LS, LAD ile WLAD tahmin edicilerinden daha güvenilir sonuçlar verdiğini göstermiştir.

Sonuç olarak, LMS tahmin edicisinin sağladığı yüksek dayanıklılık, veri kirliliğinin önemli bir sorun oluşturduğu durumlarda, LAD ve WLAD yöntemlerine kıyasla onu öncelikli olarak tercih edilebilir bir dayanıklı tahmin yöntemi hâline getirmektedir.

Bu tez çalışması sonucunda elde edilen bulgulara göre, basit doğrusal regresyon modelinde hata terimleri simetrik olmayan bir dağılım gösterdiğinde ve veri seti X -yönünde yüksek oranda aykırı gözlem içerdiğinde LMS tahmin edicisinin kullanılması

önerilmektedir. Ayrıca, WLAD tahmin edicisi, uygun ağırlık fonksiyonları seçildiğinde orta düzeyde veri kirliliği içeren durumlarda etkili bir alternatif olarak değerlendirilebilir.

Gelecek çalışmalarda, LMS yönteminin çoklu regresyon modelleri ve farklı hata dağılımları altında performansının incelenmesi, yöntemlerin uygulanabilirliğini ve dayanıklılığını daha kapsamlı biçimde ortaya koyacaktır.



KAYNAKLAR

- Acitas, S., & Senoglu, B. (2020). Robust change point estimation in two-phase linear regression models: An application to metabolic pathway data. *Journal of Computational and Applied Mathematics*, 363, 337–349.
- Al-Adilee, A., & D Mohamed, K. (2014). Linear regression model related to Gumbel distribution. *Journal of Kerbala University*, 10(2), 70–78.
- Aydin, D., & Şenoğlu, B. (2015). Monte Carlo comparison of the parameter estimation methods for the two-parameter Gumbel distribution. *Journal of Modern Applied Statistical Methods*, 14(2), 123–140.
- Aydin, D. (2018a). Estimation of the lower and upper quantiles of Gumbel distribution: an application to wind speed data. *Applied Ecology and Environmental Research*, 16(1), 1–15.
- Aydin, D. (2018b). Alternative robust estimation methods for parameters of Gumbel distribution: an application to wind speed data with outliers. *Wind & structures*, 26(6), 383–395.
- Bai, J. (1995). Least absolute deviation estimation of a shift. *Econometric Theory*, 11(3), 403–436.
- Bartolucci, F., & Scaccia, L. (2005). The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, 48(4), 821–834.
- Bickel, P. J., & Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374), 296–311.
- Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4), 318–335.
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 26(2), 211–243.
- Cox, D. R., & Hinkley, D. V. (1968). A note on the efficiency of least-squares estimates. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 30(2), 284–289.
- Chan, N. H., & Peng, L. (2005). Weighted least absolute deviations estimation for an AR (1) process with ARCH (1) errors. *Biometrika*, 92(2), 477–484.
- Croux, C., Filzmoser, P., Pison, G., & Rousseeuw, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13, 23–36.
- Davis, R. A., & Dunsmuir, W. T. (1997). Least absolute deviation estimation for regression with ARMA errors. *Journal of Theoretical Probability*, 10, 481–497.
- Dunsmuir, W. T., & Spencer, N. M. (1991). Strong Consistency and Asymptotic Normality of L1 Estimates of The Autoregressive Moving-Average Model. *Journal of Time Series Analysis*, 12(2), 95–104.
- Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error.
- Gao, X., & Feng, Y. (2018). Penalized weighted least absolute deviation regression. *Statistics and its interface*, 11(1), 79–89.
- Hampel, F.R. (1971). A general definition of qualitative robustness. *The Annals of Mathematical Statistics*, 42(6): 1887–1896.
- Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346): 383–393.
- Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution* (pp. 492–518). New York, NY: Springer New York.
- Islam, M. Q., Tiku, M. L., & Yildirim, F. (2001). Nonnormal regression. I. Skew distributions. *Communications in Statistics-Theory and Methods*, 30(6), 993–1020.
- Islam, M. Q., & Tiku, M. L. (2005). Multiple linear regression model under nonnormality. *Communications in Statistics-Theory and Methods*, 33(10), 2443–2467.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions*, volume 2. John Wiley & Sons.
- Li, J. C. H. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavior Research Methods*, 48, 1560–1574.

- Ling, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(3), 381–393.
- Olive, D. (2006). Robust estimators for transformed location scale families. Southern Illinois University. Mailcode, 4408, 62901–4408.
- Pan, J., Wang, H., & Yao, Q. (2007). Weighted least absolute deviations estimation for ARMA models with infinite variance. *Econometric Theory*, 23(5), 852–879.
- Pek, J., Wong, O., & Wong, A. C. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology*, 9, 2104.
- Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2), 186–199.
- Ramponi, (2016) Lecture notes on system identification and data analysis. [https://federico-ramponi.unibs.it/teachingresources.html\(01.10.2025\)](https://federico-ramponi.unibs.it/teachingresources.html(01.10.2025)).
- Rao, C. R. & Toutenburg, H. (1999). *Linear models: least squares and alternatives*, Second Edition, Springer, New York.
- Ross, S. M. (2009). *Introduction to probability and statistics for engineers and scientists*. Forth Edition, Elsevier Science.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871–880.
- Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons, New York.
- Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. John Wiley & Sons.
- Sapra, S. K. (2003). High-breakdown point estimation of some regression models. *Applied Economics Letters*, 10(14), 875–878.
- Temiz, Sevil (2006). Hata paylarının normal dağılmaması durumunda tek değişkenli klasik regresyonda alternatif tahmin yöntemleri (yayın no:182650) [Yayınlanmış yüksek lisans tezi, Yıldız Teknik Üniversitesi]
- Tiku, M. L., Islam, M. Q., & Selçuk, A. S. (2001). Nonnormal regression. II. Symmetric distributions. *Communications in Statistics-Theory and Methods*, 30(6), 1021–1045.
- Tiku, M. L., & Akkaya, A. D. (2004). *Robust estimation and hypothesis testing*. New Age International.
- Tukey, J.W. 1960. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, I. Olkin, ed., Stanford University Press, Stanford.
- Tukey, J.W. 1962. The future of data analysis. *The Annals of Mathematical Statistics*, 33: 1-67.
- Winkelbauer, A. (2012). Moments and absolute moments of the normal distribution. arXiv preprint arXiv:1209.4340.
- Zeckhauser, R., & Thompson, M. (1970). Linear regression with non-normal error terms. *The Review of Economics and Statistics*, 280–286.

ÖZGEÇMİŞ

Adı Soyadı : Ömer BAYRAKTAR

Yabancı Dili : İngilizce

Eğitim Durumu

Lise : Samsun Mareşal Fevzi Çakmak lisesi, 2009

Lisans : İstanbul Mimar Sinan Güzel Sanatlar Üniversitesi, 2013

Yüksek Lisans :

Mesleki Deneyim

İş Yeri : Özer Teknik Isıtma Soğutma A.Ş, 2015-2018

İş Yeri : Sinop Emniyet Müdürlüğü-Rütbeli Personel, 2020-

Yayınlar:

Aydın, D., & Bayraktar, Ö. (2025). Robust Estimation of Parameters in Simple Linear Regression Model, presented at the 3rd International Conference: Constructive Mathematical Analysis, 02-05 July 2025, Konya.