DOĞRUSAL REGRESYONDA BOZULMA NOKTASINA SAHİP TAHMİNLER

BREAKDOWN POINT ESTIMATIONS IN LINEAR REGRESSION

Joaquim Jorge Da Costa Khálau

PROF. DR. SÜLEYMAN GÜNAY

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2017

JOAQUIM JORGE DA COSTA KHALAU' ın hazırladığı "**Doğrusal Regresyonda Bozulma Noktasına Sahip Tahminler**" adlı bu çalışma aşağıdaki jüri tarafından **İSTATİSTİK ANABİLİM DALI**' nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Hamza GAMGAM
Başkan .....

Prof. Dr. Süleyman GÜNAY
Danışman .....

Prof. Dr. Turhan MENTEŞ
Üye .....

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak hazırlanmıştır.

Prof. Dr. Salih Bülten ALTEN
Fen Bilimleri Enstitüsü Müdürü

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezim herhangi bir bölümünü bu üniversitede veya başka bir üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim

05/01/2017

Joaquim Jorge Da Costa Khalau

ÖZET

DOĞRUSAL REGRESYONDA BOZULMA NOKTASINA SAHİP TAHMİNLER

Joaquim Jorge Da Costa Khalau

Yüksek Lisans, İstatistik Bölümü

Tez Danışmanı: Prof. Dr. Süleyman Günay

Ocak 2017, 40 Sayfa

Bu çalışmada doğrusal regresyonda kullanılan bazı sağlam yöntemler incelendi ve bu yöntemlerden, Huber, Hampel, Andrew ve Tukey´ın M-tahminleri karşılaştırıldı.

Çalışmanın birinci bölümünde konuya giriş yapıldı. Regresyon modelleri, kullanılan yöntemler, sağlam ve yüksek bozulma noktasına sahip yöntemler kısaca tanıtıldı.

Tahminler üzerinde etkili olan gözlemlerin ortaya çıkartılmasında kullanılan yöntemler ve diğer tanısal ölçümlere de ikinci bölümde ayrıntılı biçimde yer verildi. Önemli sağlamlık unsurlarından olan bozulma noktası tanımlandı. Bu bölümde ayrıca doğrusal regresyon modeli ve En Küçük Kareler yöntemi ayrıntılı olarak incelendi.

Üçüncü bölümde ise yüksek bozulma noktasına sahip yöntemler verildi ve bu yöntemlerin özellikleri tartışıldı.

Çalışmada uygulama kısmı dördüncü bölümde, tartışma ve sonuçlar ise beşinci bölümde verildi.

Anahtar Kelimeler: Doğrusal regresyon, bozulma noktası, sağlam tahmin ediciler

ABSTRACT

BREAKDOWN POINT ESTIMATIONS IN LINEAR REGRESSION

Joaquim Jorge Da Costa Khalau

Master, Department of Statistics

Supervisor: Prof. Dr. Süleyman Günay

January 2017, 40 pages

In this study Robust methods used in Linear Regression are briefly examined. For this study, Huber, Hampel, Andrew and Tukey´s M estimation methods are compared each other.

For the first part of study, there is a brief introduction. Robust methods and high breakdown point methods used in Linear Regression has briefly introduced.

The residual obtained from used methods and the other diagnostic measurements which are used to detect the observations effecting the estimations are discussed in the second chapter and also the important robustness measurements breakdown point and influence function are presented. In this chapter also considered the linear regression model and the Least Square method in detail.

For the following chapter is presented in detail the breakdown point methods used in linear Regression analyses.

Finally an application of those methods, discussion and results are given.

Keywords: Linear Regression, Breakdown Point, Robust Estimators

TEŞEKKÜR

Çalışmanın gerçekleşmesinde, bana yardımcı olan, hem değerli katkı ve eleştirilerileri ile, emeğini ve zamanını esirmeden yol gösterdi hem de çalışma için gerekli kaynakları sunan hocam ve danışmanım Sayın Prof. Dr. Süleyman Günay'a;

Tüm bölüm hocalarıma ve arkadaşlarıma;

Bana maddi destek sağlayan Yurtdışı Türkler ve Akraba Topluluklar Başkanlığına, babama ve tüm aileme;

Türkiye'de okuyan tüm Mozambikli öğrencilere ve kardeşim İnrareque Khalau'a;

İÇTENLİKLE TEŞEKKÜR EDİYORUM.

Vefat eden anneme Allah rahmet eylesin.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

ÇİZELGELER DİZİNİ

Sayfa

KISALTMALAR

EKK	En Küçük Kareler
EMK	En Küçük Medyan Kareler
KT	Kareler Toplamı
EKKK	En Küçük Kesilmiş Kareler
AKT	Artık Kareler Toplamı

BİRİNCİ BÖLÜM

1.GİRİŞ

Regresyon analizinde parametre tahmininde çok sık kullanılan yöntemlerden biri En Küçük Kareler (EKK)´dir. EKK kullanılarak, parametreler için yansız ve en küçük varyanslı tahminler elde edilmektedir. Bu yönteme alternatif olarak aykırı değerlerden çok fazla etkilenmeyen sağlam regresyon yöntemleri önerilmektedir.

Bu çalışmada, regresyon modellerinin analizinde kullanılan EKK, sağlam En Küçük Medyan kareler (EMK), M ve MM tahmin yöntemleri incelenmiş ve bu yöntemlerden, Huber, Hampel, Andrew ve Tukey´ın M-tahminleri karşılaştırılmıştır.

En küçük karelere seçenek olarak, sağlam bir regresyon tahminine doğru ilk adım Edgeworth`den gelmiştir (Rosseeuw 1984). Bu seçenek, en küçük mutlak değerler ($L_1$ kriteri) olarak adlandırılır. Daha sonra artıkların daha farklı bir amaç fonksiyonunun kullanılması fikrine dayalı M tahminler geliştirilmiştir. Bu tahminler, normal hatalara sahip bir modelde $L_1$`den daha etkindir (Candan 1995). Yohai (1988) tarafından, σ^2`nin M ve S tahminleri önerilmiştir. Ancak bu tahminler normal hatalara sahip bir regresyon modeli için küçük etkinliğe sahiptir. Aykırı değerlere karşı dayanıklı olmasına karsın uç değerler sözkonusu olduğunda bu tahminler için aynı iddialarda bulunmak mümkün olmamaktadır. Uç değerlere karşı daha dayanıklı tahminler elde etmek amacıyla M, L, R ve GM yöntemleri önerilmektedir. Bu yöntemlerden hiçbirinde %30`dan çok bozulma noktası elde edilememiştir. Yüksek etkinlik, yüksek bozulma noktası elde etmek için, medyan en küçük kareler tahminini Rouseeuw (1984) önermiştir. Bu durumda artıkların kareleri toplamı yerine artıkların karelerinin ortancası minimum yapılmaktadır (Candan 1995).

Yohai (1987), Rousseeuw ve Leroy (1987), en küçük ortanca için progress algoritması, MM yöntemi için de ayrı bir algoritma önermiştir. MM tahmin edicileri, hataların normal dağıldığı durumda yüksek etkinliğe sahiptir ve bozulma noktaları % 50`dir.

İKİNCİ BÖLÜM

2. GENEL BİLGİLER

Bu bölümde doğrusal regresyon, doğrusal regresyon modelleri, doğrusal regresyonda en küçük kareler yöntemi, doğrusal regresyonda artıklar etkili gözlemler ve aykırı değer, tanısal yöntemler, doğrusal regresyonda bazı önemli sağlamlık unsurları ve doğrusal regresyonda sağlam yöntemler incelenmiştir.

2.1 Doğrusal Regresyon

Ross (1987) tarafından Regresyon Analizinde kullanılan değişkenler arasında fonksiyonel bir ilişkinin var olduğu ve bu ilişkinin de regresyon modeli olarak bilindiği belirtilmektedir.

Regresyon analizinde uygulanan işlemler ve elde edilen sonuçlar, doğrusal olarak sınıflanan bu modeller üzerindeki varsayımlara bağlıdır.

Çok değişkenli regresyon modelleri gibi doğrusal model örnekleri, özel bir veri kümesine en iyi uyumu sağlamaktadır. Bu modeller, fiziksel, kimyasal ya da biyolojik modeller için çok ender kullanılmaktadır. Çok sayıda veri kümesine uyum sağlayamamalarına karşın basit ve anlaşılır olmaları en büyük avantajdır.

2.2 Doğrusal Regresyon modelleri

Doğrusal regresyon Y olarak isimlendirilen sayısal bir bağımlı değişkenle X olarak belirtilen bir veya daha çok bağımsız değişken arasındaki ilişkiyi modelleme yaklaşımıdır.

Regresyon modelindeki bağımsız değişken sayısı bir ise model basit doğrusal regresyon olarak tanımlanır. Modeldeki bağımsız değişken sayısı birden çok ise bu model çoklu doğrusal regresyon olarak isimlendirilir.

Regresyon değişkenleri arasındaki ilişkinin, bağımsız değişkenlere ve parametrelerine göre doğrusal olan bir yapıyla iyi bir şekilde temsil edildiği varsayıldığında, uygun model aşağıdaki gibi verilebilir:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i , \quad i = 1, 2, \ldots, n \qquad (2.1)$$

Yukardaki model, hem bağımsız değişkenlere hem de β parametre vektörüne göre doğrusal olan modeller sınıfındandır. Ayrıca, değişkenlerine göre doğrusal olmayıp parametrelerine göre doğrusal olan modeller de vardır.

$$Y_i = \beta_0 + \beta_1 X_{1i}^2 + \beta_2 X_{2i}^3 + \varepsilon_i , \quad i = 1,2, \ldots, n \qquad (2.2)$$

Eşitlik (2.1) ile verilen regresyon modelinin matris gösterimi

$$\mathbf{Y = X \beta + \varepsilon} \qquad (2.3)$$

şeklinde verilebilir.

Burada:

- **Y**: nx1 boyutlu cevap vektörü,
- **X:** nx(k + 1) boyutlu girdi matrisi,
- **β**: (k + 1) x 1 boyutlu katsayılar vektörü,
- **ε**: n x 1 boyutlu hata vektörüdür.

Hata terimi ε_i `nin sıfır ortalama ve σ^2 varyansıyla normal dağılımlı ve bağımsız oldukları varsayılır. β´nın tahmini için en çok kullanılan yöntem En Küçük Kareler yöntemidir.

2.3 Doğrusal Regresyonda En Küçük Kareler Yöntemi

Bir modelin parametrelerinin tahmininde en çok kullanılan yöntem En Küçük Kareler (EKK)´dir. Bu yöntemin çok yaygın biçimde kullanılmasının nedeni, hesaplama ve anlaşılma kolaylığıdır. Bu yöntemle elde edilen sonuçların yanlış yorumlanmasının önlenmesi amacıyla yöntemin çok iyi anlaşılması gerekmektedir.

Eşitlilik (2.3)´den EKK tahmini $\hat{\beta}$,

$$\mathbf{e = Y - X\hat{\beta}}$$

olmak üzere,

$$AKT_{\beta} = \mathbf{e}'\mathbf{e} \qquad (2.4)$$

AKT fonksiyonunu minimum yapan β`nın değeri olarak belirlenir. Bu eşitlikte (Y - X β)', nx1 boyutlu (Y - Xβ) vektörünün transpozu ve AKT de artıkların kareler toplamını göstermektedir.

Eşitlik (2.4)`den $\hat{\beta}$,

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

biçiminde elde edilir.

2.4 Doğrusal Regresyonda Artıklar, Etkili Gözlemler ve Aykırı Değer

Alt Bölüm (2.2)´de değinildiği gibi hataların (ε_i) normal dağılıma sahip olduğu varsayımı altında, EKK tahmin edicileri en iyi tahmin edicilerdir. Ancak, aykırı değerlerin varlığı, bu varsayımı bozduğundan yanlı ve büyük varyansılı tahminlerin elde edilmesine neden olabilir.

Eşitlik (2.3)´de ε hata terimi yerine **e** artık terimi aşağıdaki gibi tanımlanır

$$\mathbf{e} = \mathbf{Y} - \widehat{\boldsymbol{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \qquad (2.5)$$

Veri kümesi verilen Eşitlik (2.3) modelinin uygun olup olmadığına karar vermek için ε_i´lere ilişkin varsayımlar incelenmelidir. ε_i´ler gözlenemediğinden bu doğrudan Eşitlik (2.5) kullanılarak yapılır.

Rousseeuw ve Leroy (1987) tarafından, tek bir aykırı değer olduğunda, tanı yöntemleri bir gözlemin çıkarılarak etkisine bakma yoluyla iyi sonuçlar elde edilebildiği

gösterilmiştir. Ancak, birbiri ile ilişki çok aykırı değerleri teşhis etmek zordur ve gelişmiş programları gerektirir.

2.4.1 En Küçük Kareler Artıkları ve Şapka Matrisi

Şapka matrisi, veri kümesindeki X_i bağımsız değişkenler tek bir aykırı değer içerdiğinde daha kullanışlıdır. Bununla birlikte, veri kümesi birkaç aykırı değer içerdiğinde, bu değer h_{ii}`de görülmeyebilir. Burada h_{ii} matris gösterimi ile aşağıdaki gibi belirtilebilir.

$$\mathbf{H = X(X'X)^{-1} X'} \tag{2.6}$$

Verilen Eşitlik (2.6), **H**, Şapka matrisi olarak adlandırılır. Bu şapka matrisi nxn boyutlu simetrik izdüşum matrisidir. Köşegen öğeleri $h_i = h_{ii}$ ile gösterildiğinde, $0 \leq h_i \leq 1$ ve İz(**H**) = p özelliklerini sağlar.

Şapka matrisi **H**, bağımsız değişkenlerin singüler olmayan doğrusal dönüşümlerinden de etkilenmemektedir: A, pxp boyutlu ve tam ranklı matris olmak üzere nxp boyutlu X matrisi,

$$\widetilde{X} = \mathbf{XA}$$

biçiminde dönüştürüldüğünde

$$\widetilde{H} = \widetilde{X}\,(\widetilde{X}'\,\widetilde{X})^{-1}\,\widetilde{X} = \mathbf{X\,(XX)^{-1}\,X = H}$$

elde edilir (Rousseeuw ve Leory, 1987).

Şapka matrisinin köşegen öğeleri, i. gözlemin tahmini üzerindeki etkisini ölçtüğünden çok önemlidir. h_{ii}'nin sıfır olması i. gözlemin tahmini üzerinde etkili olmadığını gösterir. Regresyon analizinde aykırı değerleri saptamak için tek başına h_{ii} yeterli değildir; çünkü şapka matrisi, y_i'deki aykırı değerleri göz önüne almaz. Gözlem etkilerini ortaya çıkaran

başka istatistiksel ölçütler vardır; i. gözlemin $\hat{Y}_i$ tahmin değeri üzerinde yarattığı etkiyi incelemek için DFFITS ölçüsü kullanılır:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{ii}}{S_{i(i)}\sqrt{h_{ii}}} \qquad i = 1, \ldots, n \qquad (2.7)$$

Burada $\hat{Y}_i$ değeri n sayıda gözleme dayalı olarak bulunan tahmin değerini, $\hat{Y}_{i(i)}$ ve $S_{i(i)}$ değerleri de i. gözlem veriden atıldıktan sonra geriye kalan n-1 sayda gözleme deyalı olarak bulunan tahmin değeri ve artıklar için standart sapmayı gösterir (Gamgam ve Altunkaynak, 2015).

Şapka matrisi, $\hat{Y}$ ´nın kovaryans matrisi üzerinde de etkilidir. $e = [e_1, e_2, \ldots, e_n]'$, nx1 boyutlu EKK artıklar vektörü olmak üzere,

$$Cov(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H} \qquad (2.8)$$

$$Cov(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H}) \qquad (2.9)$$

dır.

Eşitlik (2.9) den h_{ii} yaklaşık 1 olduğunda, i. artığın varyansı sıfıra yaklaşır. i. gözlemin EKK tahminleri üzerinde etkili olup olmadığı incelenmelidir (Rousseeuw ve Leroy, 1987).

2.4.2 Mahalanobis Uzaklığı

Mahalanobis uzaklık ile sabit terimli bir regresyon modelinin uç değerleri bulunabilir.

Bu uzaklık Ertaş (2011) tarafından, x_i lerin oluşturduğu çok değişkenli bir veri kümesinde bir gözlemin veri kümesinin merkezine olan uzaklık şeklinde belirtilmiştir. $x_i^{'}$, gözlem vektörü,

$$x'_i = 1x_{i1}x_{i2}\ldots x_{ik} = 1z_i$$

biçiminde tanımlanırsa, sırasıyla z_i nin ortalama vektörü ve kovaryans matrisi aşağıdaki gibidir:

$$\bar{z}=\frac{1}{n}\sum_{i=1}^{n} z_i \qquad (2.10)$$

ve

$$C=\frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})'(z_i-\bar{z}) \qquad (2.11)$$

z_i`nin $\bar{z}$ `dan uzaklığının hesaplanmasında kullanılan bir ölçüm, Mahalanobis uzaklığının karesidir:

$$M_i^2=(z_i-\bar{z})C^{-1}(z_i-\bar{z})'$$

Rousseeuw ve Leroy (1987) tarafından, M_i^2, α anlamlılık düzeyinde p-1 serbestlik dereceli ki-kare değeri ile karşılaştırılır ve bu değeri aşan gözlemler uç değer olarak tayin edilir.

2.5 Tanısal Yöntemlere Alternatif Olarak Sağlam Yöntemler

Aykırı değerlerin ortaya çıkarılması ve tahminler üzerindeki etkilerinin giderilmesi amacıyla kullanılan yöntemlere alternatif olarak sağlam yöntemler önerilmektedir. Klasik EKK yöntemiyle, normal dağılmasa bile normal görünen artıklar elde edilmeye çalışılır. Ancak, sağlam yöntemler kullanılarak önce regresyon modeli verilerin büyük bir kısmına uydurulur sonra artıkların büyük değerleri aykırı değerler olarak belirlenir. Böylece, sağlam yöntemlere aykırı ve üç değerlerin etkisine karşı sağlam tahminler, bu gözlemler veri kümesinden çıkartılmadan elde edilebilir (Candan 1995).

Myers (1986) tarafından, sağlam yöntemler kullanılarak elde edilen artıkların aykırı değerleri tanımlamada kullanılması çok yararlı olsa da, tanısal yöntemler yerine hiçbir zaman geçemediği iddia edilmektedir. Bu durumda tanısal yöntemlerle, sağlam uyumdan elde edilmeyen birçok bilgiye ulaşılabilmektedir. Ancak, tanısal yöntemlerle

ortaya çıkarılan aykırı değerlerin atılmasıyla elde edilen klasik tahmin ediciler, sağlam tahmin ediciler kadar yüksek etkinliğe sahip olmayabilir.

2.6 Doğrusal Regresyonda Bazı Önemli Sağlamlık Unsurları

Doğrusal regresyonda kullanılan sağlam yöntemler, hem veri kümesinden hem de modellemeden kaynaklanan sorunları çözmek amacıyla kullanılabilmektedir. Bu yöntemler, veri kümesine iyi bir yaklaşımın sağlandığı durumda çok iyi sonuç verebilmektedir.

Sağlam bir yöntem kullanılarak, elde edilen tahmin edicinin ne kadar sağlam olduğuna karar vermede, bozulma noktası ve etkinlik fonksiyonunun önemi büyüktür.

Doğrusal regresyon modellerinin sağlam tahminlerinin elde edilmesinde **M, L, R** tahmin edicilerinin kullanılması önemli bir adımdır.

Huber ve Ronchetti (2009), aşağıdaki örnek regresyon modellerinin tahmininde aykırı değerler sorununa yer vermiştir. Kullanılan veri kümesinde aykırı değerler üç değişik regresyon modeli üzerinde incelenmiştir ve EKK yöntemiyle elde edilen artıklar ve grafikleri verilmiştir:

Çizelge 2.1 n = 6 gözlemli veri kümesi ve EKK uyumları ve artıkları

			Uyum 1		Uyum 2		Uyum 3	
Gözlem	x	y	$\hat{y}$	$y-\hat{y}$	$\hat{y}$	$y-\hat{y}$	$\hat{y}$	$y-\hat{y}$
1	-4	2.48	0.39	2.09	2.04	0.44	2.23	0.25
2	-3	0.73	0.32	0.42	1.06	-0.33	0.99	-0.26
3	-2	-0.04	0.23	-0.27	0.08	-0.12	-0.09	-0.13
4	-1	-144	0.15	-1.59	-0.9	-0.54	-1	-0.44
5	0	-1.32	0.07	-1.39	-1.87	0.55	-1.74	0.42
6	10	0	-0.75	0.75	-11.64	11.54	0.01	-0.01
			$\hat{\sigma}=1.55$		$\hat{\sigma}=0.55$		$\hat{\sigma}=0.41$	
			$e/\hat{\sigma}=1.35$		$e_{max}/\hat{\sigma}=1.00$		$e_{max}/\hat{\sigma}=1.08$	

Aşağıda verilen grafiklerden,

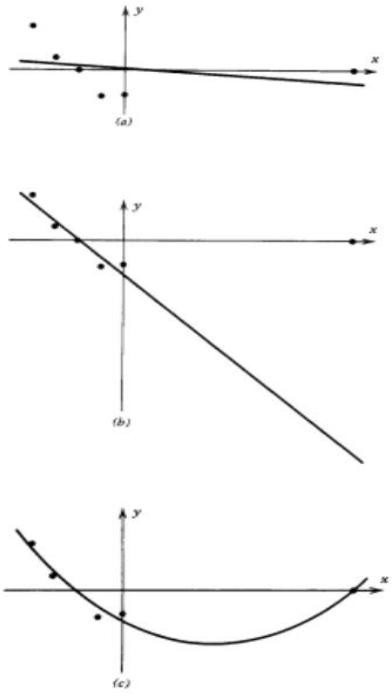
(a) grafik doğrusal bir modele uyduğu kabul edilen verilerin EKK uyumuna aittir;

(b) grafik doğrusal bir modele uyduğu kabul edilen verilerin sağlam uyumuna aittir;

(c) grafik ise ikinci dereceden polinomial bir modele uyuduğu kabul edilen verilerin EKK uyumuna aittir.

Burada, iki değişkenli bir veri kümesinde, en iyi uyumu sağlayan sağlam model belirlendikten sonra aykırı değerlerin ortaya çıkarılmasının kolay olduğu görülebilir.

Verilen artıklar incelenerek, verilere en iyi uyum sağlayan modelin 3. model olduğu söylenebilir ve 6. gözlem dışında 5 gözlemin $y = -(2 + x) + \varepsilon$ ($\varepsilon \sim N(0, 0.36)$) modeline göre elde edildiği bilindiğinde, gerçek modele en iyi yaklaşım gösteren modelin 2. model olduğuna karar verilir.



Şekil 2.1 Regresyon modellerinin veri kümesine uyum grafikleri

2.6.1 Bozulma Noktası

Bozulma noktasının bir boyutlu durumda konum parametrelerin tahmin için tanımlandığı bilinmektedir. Candan (1995) tarafından, eğer bir tahmin aykırı değerlerin küçük bir oranı tarafından yalnızca sınırlı miktarda değiştirilirse dirençli olduğu belirtilmektedir. Fakat bu oran büyürse tahmin bozulur. Bozulma noktası için elde edilebilecek en yüksek değer, %50'dir. Bu nedenle %50'yi aşan bozulma noktası ile normal gözlemlerle aykırı değerler arasında ayırım yapılamamaktadır.

Bozulma noktası, nicel sağlamlığın bir göstergesidir ve bir tahmin edicinin etkilenmediği aykırı değerlerin sınırlı bir miktarını verir. Küçük bir olasılıkla büyük hataların küçük bir bölümüdür ve model dağılımından olan uzaklıktır. Bozulma noktaları, uzak aykırı değerlerin ne kadarının reddedilebileceğini söyler. Eğer bozulma sıfırdan büyükse bozulma noktası dirençlidir.

n gözlemden oluşan **Z** örneklemi,

$$\mathbf{Z} = \{(x_{11},\ldots, x_{1p}, y_1),\ldots,(x_{n1},\ldots,x_{np}, y_n)\}$$

ve **Z** örnekleminin regresyon tahmin edicisi T olsun. T tahmin edicisinin **Z** örneklemine uygulanması sonucu regresyon katsayıları vektörü $T(\mathbf{Z}) = \hat{\beta}$ elde edilir. m tane orjinal veri noktası **Z** örnekleminden çıkarılıp, yerine m tane keyfi bozulmuş değerler konularak **Z´** örneklemi elde edilsin. Burada bozulmuş **Z´** örneklemleri için yan miktarı:

$$Yan(m;T,\mathbf{Z}) = Sup_{z'}\|T(\mathbf{Z}')-T(\mathbf{Z})\|$$

şeklinde belirtilebilir. Bu yan değeri sonsuza yaklaştığında, m aykırı değerin T üzerinde büyük etkisi olduğunu gösterir ve tahmin edicinin bozulduğu söylenir. Böyle bir **Z** örnekleminde T tahmin edicisinin bozulma noktası aşağıdaki biçimde tanımlanır:

$$\varepsilon_n^*(T,\mathbf{Z}) = Min_z\{\frac{m}{n};Yan(m;T,\mathbf{Z})= \infty\}$$

Bir başka ifade ile, T tahmin edicisinin T(**Z**)´den uzak değerler almasına neden olan en küçük bozulum miktarını verir.

Rousseeuw ve Leroy (1987) tarafından, tahmin edicilerin sağlamlığının önemli bir öğesi olan bozulma noktası, bu tahmin edicileri hesaplamada kullanılan yöntemlerle birlikte anılması gerekliği belirtilmektedir.

EKK ve sağlam olarak önerilen L_1 yöntemi için,

$$\varepsilon_n^*(T,\mathbf{Z}) = 1/n$$

dir.

Burada n sonsuza giderken, bu oranın sıfıra yakınsadığı açıktır.

2.6.2 Etkinlik Fonksiyonu

Bir F dağılımlı kitleden çekilen $Y = \{y_1,\ldots,yn\}$ örnekleminden elde edilen T tahmin edicisi için etkinlik fonksiyonu,

$$IF(y;T,F)= \lim_{\varepsilon\to 0} \frac{T\left[(1-\varepsilon)F+ \varepsilon\delta_y\right]-T(F)}{\varepsilon} \qquad (2.12)$$

biçiminde tanımlanır.

Veri kümesinde, ölçme hatasının bulunmadığı, yaklaşık normalliğin sağlandığı ve etkili gözlemin olmadığı bir durumda bu veri kümesine sağlam regresyon uygulandığında, regresyon modelinin hemen hemen EKK ile aynı olması istenir, çünkü EKK böyle bir veri için en uygun tahmin yöntemidir.

Ortalama tahmin edicisi için etkinlik fonksiyonu yukarıdaki fonksiyondan

$$IF(y; \bar{y},F)=\lim_{\varepsilon\to 0} \frac{(1-\varepsilon)\mu + \varepsilon y - \mu}{\varepsilon} = y - \mu$$

olarak elde edilir.

IF(y) fonksiyonu $y \to \infty$ iken tanımlanamayan biçimde artmaktadır.

Etkinlik fonksiyonu bazı önemli sağlamlık ölçümlerinin elde edilmesinde de kullanılmaktadır. Huber (1981) tarafından, tek bir gözlemin tahmin edicisi veya test istatistiği üzerindeki göreli etkisini değerlendirilmesinin sağlandığı belirtilmektedir. Eğer fonksiyon sınırlı değilse bir aykırı değer sorun çıkarabilir.

Tüm tahmin ediciler bozulma noktasına sahiptir, ama tümü bir etki fonksiyonuna sahip değildir. Ayrıca etki fonksiyonuna sahip tahmin ediciler genelde asimptotik normal dağılır.

$$\sqrt{n}(T_n - T(F)) \sim N(0, V(T,F))$$

Sağlam tahmin edicilerin belirli bir bölgedeki sağlamlıklarıyla ilgili bilgi edinebilmek amacıyla etkinlik fonksiyonunun sonlu örneklemlere uyarlanmasıyla elde edilen duyarlılık eğrilerinden yararlanılır. Bu duyarlık eğrileri (2.12) eşitliğinde, F yerine F_{n-1} ve ε yerine 1/n yazılarak elde edilir:

$$SC_{n-1}(y) = \frac{T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_y\right) - T(F_{n-1})}{1/n}$$

$$= n[T_n(y_1,\ldots,y_n, y) - T_{n-1}(y_1,\ldots,y_n)] \qquad (2.13)$$

2.7 Doğrusal Regresyonda Sağlam Yöntemler

Regresyon için sağlam yöntemler, varsayılan modelden küçük sapmalara karşı duyarsız parametre tahminlerini verir.

Doğrusal regresyon modellerin sağlam tahminlerinin elde edilmesinde **M, L, R** tahmin edicilerin kullanılması önemli bir adımdır.

Bir örneklem durumda ise, p = 1 ve tüm i´ler için $x_1 = 1$ konularak elde edilen genel regresyon modelin özel bir durumu gibi düşünülmektedir. Bu durumda y_i örnekleminde edilen model denklemi,

$$y_i = \beta + \varepsilon_i \ , \qquad i = 1, 2, \ldots, n$$

dir.

Bir y_i örneklemine, bir v sabiti eklendiğinde T_n´de aynı miktarda artarsa, yani,

$$T_n (y_1 + v,\ldots,y_n + v) = T_n (y_1,\ldots,y_n) + v$$

ise bir T_n tahmin edicisi konum eşdeğişkendir. y_i örnekleminin tüm gözlemlerinin bir c sabitiyle çarpılması durumu da sözkonusu olabilir.

Regresyonda uzun hata dağılımları, bilinmeyen parametrelerin EKK tahminlerini aykırı değerlere aşırı duyarlı kılar. Bu nedenle, bir boyutlu durumda, ortalamaya seçenek olarak önerilen sağlam tahmin yöntemleri, çok değişkenli regresyon modellerinin tahmininde kullanılmak üzere genelleştirilmiştir. Bu bölümde, çok parametreli regresyon modellerinin sağlam tahminlerinin elde edilmesinde kullanılan bazı yöntemler incelenmiştir.

2.7.1 M yöntemi ile sağlam tahmin ediciler

Huber (1985) tarafından, M-tahmin edicileri olarak bilinen bir tahmin edici sınıfı önerilmiştir. Bu tahmin yöntemi, artıkların kareleri yerine, artıkların başka bir fonksiyonunu minimum yapma fikrine dayanmaktadır. M-tahmin edicisinin amaç fonksiyonu, aşağıda biçimde verilir:

$$Min_{\beta_j} \ \rho \ (y_i - \sum_{j=1}^{p} x_{ji}\beta_j)$$

ρ artıkların simetrik bir fonksiyonudur ve sıfır noktasında en küçük değerini alır.

Yukardaki amaç fonksiyonun β regresyon parametresine göre türevi alındığında

$$\sum_{i=1}^{n} \Psi(y_i - \sum_{j=1}^{p} x_{ji}\beta_j)x_{ji} = 0 \qquad (2.14)$$

denklem sistemi elde edilir.

M tahminlerinin Eşitlik (2.14) ile elde edilmesinin bir dezavantajı vardır: tahminin ölçek değişmez olması için bir σ ölçeği ile β´nın,

$$\sum_{j=1}^{p} \Psi(\frac{e_i}{\sigma})x_{ji} = 0$$

eşitliğinden tahmini önerilmiştir. Ancak, hangi ölçek tahmininin kullanılması gerektiği açık biçimde belirtilememektedir.

M tahmin ediciler, sağlam tahminde önemli bir adımdır. Yapılan çalışmaların çoğunda p ve Ψ fonksiyonlarını oluşturma üzerine yoğunlaşmıştır.

En küçük mutlak değer tahmininin EKK tahminine göre üstünlüğü aykırı değerlere karşı duyarlı olmamasıdır. Ancak, aykırı değer olmadığında EKK tahmini daha doğru olabilir. Bir boyutlu durumda kullanılan medyan mutlak sapmaya benzer bir tahmin çok boyutlu durumda medyan mutlak artıklar,

$$\hat{\sigma} = med\{|e_i|\}$$

biçiminde tanımlanır.

Huber (1985) tarafından, β ve σ parametrelerin M tahminlerinin birlikte hesaplanmasını sağlayan **H** algoritması önerilmiştir. Regresyon artıklarının değiştirilmesine dayanan bu algoritmanın adımları aşağıda verilmektedir:

a) e_i′nin değeri,

$$e_i^m = y_i - f_i(\theta^{(m)})$$

b) σ´nın yeni değeri,

$$\left(\sigma^{(m+1)}\right)^2 = \frac{1}{a} \sum_{i=1}^{n} x \left(\frac{e_i^m}{\sigma^m}\right) \left(\sigma^{(m)}\right)^2$$

c) e_i artıkları aşağıdaki gibi değiştirilir:

$$Z_i = \Psi \left(\frac{e_i^{(m)}}{\sigma^{(m+1)}}\right) \sigma^{(m+1)}$$

d) Kısmi türevleri hesaplamak için:

$$x_{ij} = \frac{\partial f_i(\theta^{(m)})}{\partial f_j}$$

e) X`Xτ = X`z`den $\hat{\tau}$ elde edilir.

f) $\theta^{(m+1)} = \theta^{(m)} + q\hat{\tau}$ bulunur.

g) $\bar{X} = (X'X)^{-1}$ matrisinin köşegen elemanları $\bar{x}_{jj}$ olmak üzere $\left|\bar{\tau}_i\right| < \varepsilon\sqrt{\bar{x}_{jj}}\sigma^{(m+1)}$

ise iterasyon durdurulur ve $\sigma^{(m+1)}$, θ´nın tahmini olarak bulunur. Aksi halde m = m + 1 alınarak süreç tekrarlanır. Bu yöntemler için başlangıç tahminlerin ($\theta^{(0)}$, $\sigma^{(0)}$) seçilmesi sorun olabilir.

2.7.2 R yöntemi ile sağlam tahmin ediciler

R tahminlerinin M tahminlerine göre önemli bir avantajı, bu tahminlerin ölçek eşdeğişken olmasıdır. Bu tahminler rank testlerinden türetilmişdir. Tek örneklem yalnızca konum problemleri için bulunur.

EKK yaklaşımı kullanılarak R tahminleri aşağıdaki şekilde elde edilir.

$$Min_{\beta_j} \sum_{i=1}^{n} a_n(R_i)e_i$$

Burada R_i, (e_1, **...** , e_n)`de $e_i$`nin rankıdır ve $a_n(.)$, $\sum_{i=1}^{n} a_n(i) = 0$ sağlayan monoton skorlar fonksiyonudur.

Yukardaki fonksiyonun β_i`lere göre kısmi türevi alınarak

$$\sum_{i=1}^{n} a_n(R_i)X_{ji} \cong 0$$

denklem sistemi elde edilir. Bu denklem sistemi

$$Min_{\beta_j} \sum_{j=1}^{p} \left| \sum_{i=1}^{n} a_n(R_i)x_{ji} \right|$$

biçiminde tekrar bir optimization problemine dönüştürülebilir (Rousseeuw and Leroy 1987).

2.7.3 L yöntemi ile sağlam tahmin ediciler

Bir boyutlu durumdan çok boyutlu duruma uyarlanabilen diğer bir tahmin edici sınıfı da L`dir. Bickel (1973), doğrusal bir model için tek iterasyon süren bir L tahmin edici sınıf önermiştir (Rousseeuv ve Leroy, 1987).

Veri kümesinin dağılım fonksiyonunun simetrik olması durumunda, L tahminleri M ve R tahminlerine eşit olduğunu Jaekel göstermiştir (Rousseuv ve Leroy, 1987). Ancak,

dağılımla ilgili simetrik özelliği yoksa bir L tahmin edicisini M ve R tahmin edicileriyle eşleştirmek mümkün değildir.

Örneklem yüzde değerlerine dayalı bir L tahmin edici Koenker ve Basset önermiştir (Rousseeuw ve Leroy 1987). a, bir regresyon yüzde değeri (0 < a < 1) olmak üzere, $\hat{\beta}$ tahmini.

$$Min_{\beta} \sum_{i=1}^{n} p_a(e_i)$$

sağlayan β değerdir. Burada,

$$p_a(e_i) = \begin{cases} ae_i \ , & e_i \geq 0 \\ (a-1)e_i \ , & e_i \leq 0 \end{cases}$$

dir.

$x_1 \leq x_2 \leq \ldots \leq x_n$ sıralı örneklem olsun. L tahmin edici sıralı istatistiklerin doğrusal birleşimleridir ve

$$T_n(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} a_i x_i$$

biçiminde ifade edilir. α_i ağırlıkları,

$$a_i = \frac{\int_{i-1/n}^{i/n} h d\lambda}{\int_0^1 h d\lambda}$$

biçiminde verilir. Burada h

[0,1] → R ve $\int_0^1 h d\lambda \neq 0$ özelliklerini sağlayan bir fonksiyonudur. Bu tahmin edici asimptotik olarak normallik özelliği gösterir.

- L karşılık gelen fonksiyonu:

$$T(G) = \frac{\int xh\big(G(x)\big)dG(x)}{\int hF(y)df(y)}$$

- L etkinlik fonksiyonu:

$$IF(x;T;F)=\frac{\int_{[0,x]} h(Fy))d\lambda(y)-\int[\int_{(0,t)} h\big(F(y)\big)d\lambda]dF(t)}{\int h\big(F(y)\big)dF(y)}$$

  biçiminde verilir.

Hampel (1971) tarafından L tahmin için IF, dağılıma bağlıdır ve küçük değişikliklerde IF uygun değildir.

ÜÇÜNCÜ BÖLÜM

3. DOĞRUSAL REGRESYONDA YÜKSEK BOZULMA NOKTASINA SAHİP YÖNTEMLER

Bu bölümde, bir boyutlu regresyon analizinde yüksek bozulma noktasına sahip yöntemler incelenmiştir.

Doğrusal regresyon modellerinin parametrelerinin yüksek bozulma noktalı, çok sağlam parametre tahminleri elde edilmesi amacıyla, bir boyutlu durumda 0.50 bozulma noktasına sahip medyana dayalı yöntemlere bazı seçenek önerilmiştir. En küçük medyan kareler tahmin edicisi, en küçük kesilmiş ortalamalar yöntemi ve S-tahmin edicileri bu sınıfta yer alır. Ryan (1997) tarafından, tüm yüksek bozulma noktasına sahip yöntemler tam-uygun doğruyu vermesine rağmen, EKK yönteminin kullanılmasının gerektiği veri kümesinde göreli olarak zayıf bir performans gösterdiği belirtilmektedir.

3.1 Bir Boyutlu Durumda Yüksek Bozulma Noktasına Sahip Yöntemler

Alpu ve arkadaşları (2010) tarafından, yüksek bozulma noktalı regresyon tahmin edicilerinin çok sayıda aykırı değerin varlığında güvenilir tahminler elde etmek için geliştirildiği ifade edilmektedir. Bu tahmin ediciler 0.50 bozulma noktasına sahiptir ve dirençli (resistant) tahmin ediciler olarak bilinir. Yüksek bozulma noktalı tahmin ediciler hem x hem de y yönündeki aykırı değerlerin varlığı durumunda güvenilir parametre tahminleri verir.

3.1.1 En Küçük Medyan Kareler Yöntemi (EMK)

Rousseeuw ve Leory (1987) tarafından, en küçük medyan kareler yönteminin, aykırı değerlerin ortaya çıkarılması için kullanılan sağlam bir yöntem olduğu iddia edilmektedir. Yöntem, artık kareler toplamı yerine artık karelerin medyanını en küçük yapan β`nın değeri olarak aşağıdaki biçimde tanımlanır:

$$Min_{\beta_j} med_i \ (y_i-\beta)^2$$

Bu yöntem, aykırı değerlerin belirlemesinde kullanılan bir veri çözümlemesi olarak düşünülebilir.

EMK tahmin edicisinin ölçek tahmini de sağlam olmalıdır. Bu amaçla aşağıdaki tahmin edicinin kullanımı önerilmektedir:

$$\hat{\sigma}= C_1 \sqrt{\text{mede}_i^2}$$

Burada,

e_i – EMK tahmine göre elde edilmiş artık değerleridir.

C_1 – sabit normal hata dağılımında tutarlılığı sağlamak için kullanılır.

3.1.2 En Küçük Kesilmiş Kareler Yöntemi

En Küçük Medyan Kareler, asimptotik etkinlik bakımından zayıf bir performansa sahiptir. En Küçük Kesilmiş Kareler Tahmin Edicisi $\hat{\beta}$, sıralı artık karelerin toplamını en küçük yapan β değeri olarak Rousseeuw (1987) tarafından aşağıdaki biçimde önerilmiştir:

$$Min_{\beta} \quad \sum_{i=1}^{h} e_{(i)}^2 \qquad (3.12)$$

Burada, $e_{(1)}^2 \leq e_{(2)}^2 \leq \cdots \leq e_{(n)}^2$ sıralı artık kare değerleridir ve h = [n/2] +1`dir.

h = [n/2] +1 iken, EMK ile aynı bozulma noktası elde edilmektedir. Tekrarlı medyanla aynı bozulma noktası elde etmek için,

$$h = [n/2] + [(p + 1) / 2]$$

alınmalıdır.

Rousseeuw ve Leory (1987), h değerinin, h = [n (1 - α)] + 1 olarak seçilmesini önermişlerdir.

EKKK`in hesaplanması için

$$\{y_{(1)}, \dots, y_{(h)}\}; \{y_{(2)}, \dots, y_{(h+1)}\}; \dots; \{y_{(n-h+1)}, \dots, y_{(n)}\}$$

örneklemleri belirlenir. Herbiri h gözlem içeren bu gözlemlerin ortalamaları,

$$\bar{y}^{(1)} = \frac{1}{h}\sum_{i=1}^{h} y_{(i)} \ , \dots , \quad \bar{y}^{(n-h+1)} = \frac{1}{h}\sum_{i=1}^{h} y_{(i)}$$

bulunur ve bunlara karşılık gelen kareler toplamları (KT) aşağıdaki biçimde elde edilir:

$$KT^{(1)} = \sum_{i=1}^{h}(y_{(i)} - \bar{y}^{(1)})^2 \ , \dots \ , KT^{(n-h+1)} = \sum_{i=n-h+1}^{h}(y_{(i)} - \bar{y}^{(1)})^2$$

EKKK tahmini, en küçük $KT^{(j)}$ değerine karşılık gelen $\bar{y}^{(j)}$ olarak bulunur.

$$\bar{y}^{(j)} = \frac{h\bar{y}^{(j-1)} - y_{(j-1)} + y_{(j+h-1)}}{h}$$

Karşılık gelen ortalama kareler toplamını hesaplanmak için:

$$KT^{(j)} = KT^{(j-1)} - y_{(j-1)}^2 + y_{(j+h-1)}^2 - h\bar{y}_{(j)}^2 + h\bar{y}_{(j-1)}^2$$

eşitliği kullanılır.

Rousseeuw ve Leroy (1987) tarafından, EKKK`de EMK`de olduğu gibi regresyon, ölçek, afin eşdeğişkendir. Ayrıca, $1/2$ (n + p – 1) gözlemden fazlası $y_i = x_i \beta$ eşitliğini sağladığında EKKK tahmin edicisi de tam uyum özelliğini sağlamaktadır.

3.1.3 S - Yöntemi

Kavruk (2005) tarafından, hem EMK hem de EKKK yöntemi, artıkların dağılımının sağlam ölçümünün minimumu olarak tanımlanmaktadır. Candan (1995) tarafından, S tahmin ediciler, regresyon gibi çok değişkenli durumlarda meydana gelen yüksek bozulma noktasından dolayı önerilir. Rousseuw ve Leroy (1987) tarafından, bir boyutlu durumda önemli bir tahmin edici sınıfı da S`dir. Bir S tahmin edicisi, artıkların dağılımını en küçük yapan β değeri olarak aşağıdaki gibi önermiştir:

$$Min_{\beta}\ s(y_1 - \beta, \dots, y_n - \beta),$$

$$S(\beta) = s(e_1(\beta), \dots, e_n(\beta)) \qquad (3.13)$$

Artık dağılımının β`ya göre en küçük yapılmasından elde edilmektedir

$$Min_{\beta}\ S(\beta)$$

Burada, S tahmin edicileri,

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

eşitliğinin çözümünden elde edilmektedir.

Rousseeuw ve Leroy (1987) tarafından, ρ artık fonksiyonunun aşağıdaki verilen şekildeki ifadenin kullanılması önerilmiştir:

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

Bir Y = {y_1, **...** , y_n} örnekleminde EMK, EKKK ve S tahmin edicilerinin bozulma noktası,

$$\varepsilon_n^*(T,Y)=\frac{\left[\frac{n+1}{2}\right]}{n}$$

dir.

Bu oran $n \rightarrow \infty$ iken 0.50`ye yakınsar.

Rousseeuw ve Leroy (1987) tarafından, S tahmin edicilerinin de diğer yüksek bozulma noktalı tahmin ediciler gibi tam uyum özelliğini sağladığı belirtilmiştir.

S tahmin edicileri yüksek bozulma noktasına sahip olmasına karşın hataların normal dağıldığı durumda düşük etkinliğe sahiptir. S tahminlerinin etkinliğini artırmanın bir yolu da bir adım M tahmininin ya da yeniden ağırlıklı EKK tahmininin kullanılmasıdır.

3.1.4 MM Yöntemi

Tüm yüksek bozulma noktasına sahip tahmin ediciler, yüksek etkinliğe sahip MM tahmin edicilerin hesaplamasında başlangıç noktası olarak kullanılabilir. Yohai (1987) tarafından, MM yöntemi, istatistiksel etkinliğinin (hataların normal dağıldığı varsayımı altında) yüksek ve aynı zamanda yüksek bozulma noktasına sahip bir yöntem olduğu gösterilmiştir.

Yohai (1988) tarafından, bu tahmin edicinin çökmeye dayanıklı ve yüksek etkinliğe sahip üç aşamadan oluşan bir tahmin edici olduğu kanıtlanmıştır. İlk aşama S tahmin edicisini kullanır ve artıkları elde eder. İkinci aşamada, artıkları kullanarak M tahmin edicisi hesaplanır. Son aşama olarak, çok yüksek artıklara 0 ağırlık veren bir fonksiyon ile M tahmin edicisi hesaplanır.

MM tahminleri üç aşamalı olarak aşağıdaki gibi tanımlanabilir:

Aşama 1:

Yüksek bozulma noktasına sahip (mümkünse 0.50), bir başlangıç tahmini seçilir.

Aşama 2:

$$e_i(\hat{\beta}^o) = y_i - \hat{\beta}^o x_i \qquad , \qquad 1 \leq i \leq n$$

artıkları hesaplanır. $S_n = s_n(e(\beta^o))M$ ölçek tahmin, varsayımlarını sağlayan bir ρ_0 fonksiyonu kullanılarak,

b / a = 0.5

eşitliğini sağlayan bir b sabiti için

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}\rho(e_i(\beta)/s_n) = b$$

eşitliğinden hesaplanır. Burada,

$$a = max\rho_0(y)$$

dir.

Bu ilk ölçek tahmininin 0.5 bozulma noktasına sahip olması için, b/a = 0.5 olması gerektiğini Huber (1981) tarafından tanıtlanmıştır.

Aşama 3:

ρ_1, ρ_0 için verilen koşulları sağlayan diğer bir fonksiyon olmak üzere:

$$\rho_1(y) \leq \rho_0(y) \qquad ve \quad Sup\rho_1(u) = sup\rho_0(u) = a$$

olmak üzere, MM tahmini $\hat{\beta}$,

$$\sum_{i=1}^{n} \Psi_1\left(\frac{e_i(\beta)}{s_n}\right)x_i=0$$

eşitliğinin bir çözümü olarak tanımlanır, ve bu tahmin

$$S(\hat{\beta}_1) \leq S(\hat{\beta}_0)$$

eşitsizliğini sağlamaktadır. Burada,

$$S(\beta)= \sum_{i=1}^{n} \rho_1(e_i(\beta)/s_n) \qquad (3.14)$$

dir ve $\rho_1(0/0)$, 0 olarak tanımlanmaktadır. Bu tahminlerin hesaplamasında kullanılan iteratif ağırlıklı EKK algoritmasının değişik bir biçimi Yohai (1987) tarafından önerilmiştir.

Bu algoritma aşağıdaki gibi ifade edilebilir:

$$Z_i = (y_i , x_i), 1 \leq i \leq n$$

veri kümesi olmak üzere, T_0, bu örneklemden hesaplanan yüksek bozulma noktalı EMK olsun.

Her $t \in R^p$ için ağırlık fonksiyonları,

$$W_i(t) = \Psi_1 (e_i (t) / s_n) / (e_i (t) / s_n)$$

biçimde de tanımlanır.

Yüksek bozulma noktasına sahip EMK´in değişik yöntemler için başlangıç noktası olarak kullanımı yaygındır.

3.2 Sağlam M Tahmin Yöntemleri

Inal ve arkadaşları (2006) tarafından, sağlam tahmin yöntemleri içinde en yaygın kullanılanlardan biri M tahminlerdir. M tahminler maksimum olasılık tahmininin genelleştirilmiş biçimidir**.** M tahmini olarak çok sayıda yöntem sunulmuştur. Bu yöntemlerin her biri farklı kayıp, etki ve ağırlık fonksiyonu ile tanımlanır. Parametre tahmininde bilinmeyen parametrelerin gerçek değeri ile tahmin edilen değerleri arasındaki fark kayıp fonksiyonu ile ifade edilir.

Huber, Andrew ve Tukey olmak üzere bu konuda birçok M- tahmini önerilmiştir. Her bir M- tahmine ait ρ(t), ψ(t) değerlerinin hesaplanışı sırasıyla aşağıda verilmiştir (Akbilgiç ve Keskintürk 2008).

Huber`in Minimax Tahmini

$$\rho(t) = \begin{cases} \frac{t^2}{2} & , \quad |t| \leq b \\ \frac{b^2}{2} & , \quad |t| > b \end{cases}$$

$$\psi(t) = \begin{cases} t & , \quad |t| < c \\ bsign(t) & , \quad |t| \geq c \end{cases}$$

Andrew`in Sinüs Dalgası Fonksiyonu

$$\rho(t) = \begin{cases} \frac{t^2}{2} & , \quad |t| \leq b \\ \frac{b^2}{2} & , \quad |t| > b \end{cases}$$

$$\psi(t) = \begin{cases} \sin(t), & -\pi \leq t < \pi \\ 0 \quad , & değer\ durumlar \end{cases}$$

Tukey'in İkili Ağırlıklar (Bi – Weight)

$$\rho(t) = \begin{cases} \frac{t^2}{2} & , \quad |t| \leq c \\ \frac{a^2}{2} & , \quad |t| > c \end{cases}$$

$$\psi(t) = \begin{cases} \mathrm{t}(1 - (\mathrm{t} / \mathrm{c})^2)^2, & -\pi \leq t < \pi \\ 0 \quad , & değer\ durumlar \end{cases}$$

M- tahmini için ρ(t) = $t^2/2$ ve ψ(t) = t olarak belirlenirse EKK tahmini elde edilir. Yukarıdaki eşitliklerde yer alan a ve c birer sabittir ve genellikle ayarlama sabitleri olarak adlandırılır (Akbilgiç ve Keskintürk 2008).

Hampel´ın yeniden azalan tahmileri

Hampel üç parçalı M tahmin edicileri aşağıdaki Ψ fonksiyonu yardımıyla hesaplanabilir:

$$\psi(t) = \begin{cases} x \quad , & 0 \leq |x| \leq a \\ a\ signn\ (x), & a \leq |x| \leq b \\ \frac{a(r - |x|)}{r - b}\ sign(x)\ , & b \leq |x| \leq r \\ 0\ , & r \leq |x| \end{cases}$$

DÖRDÜNCÜ BÖLÜM

4. UYGULAMA

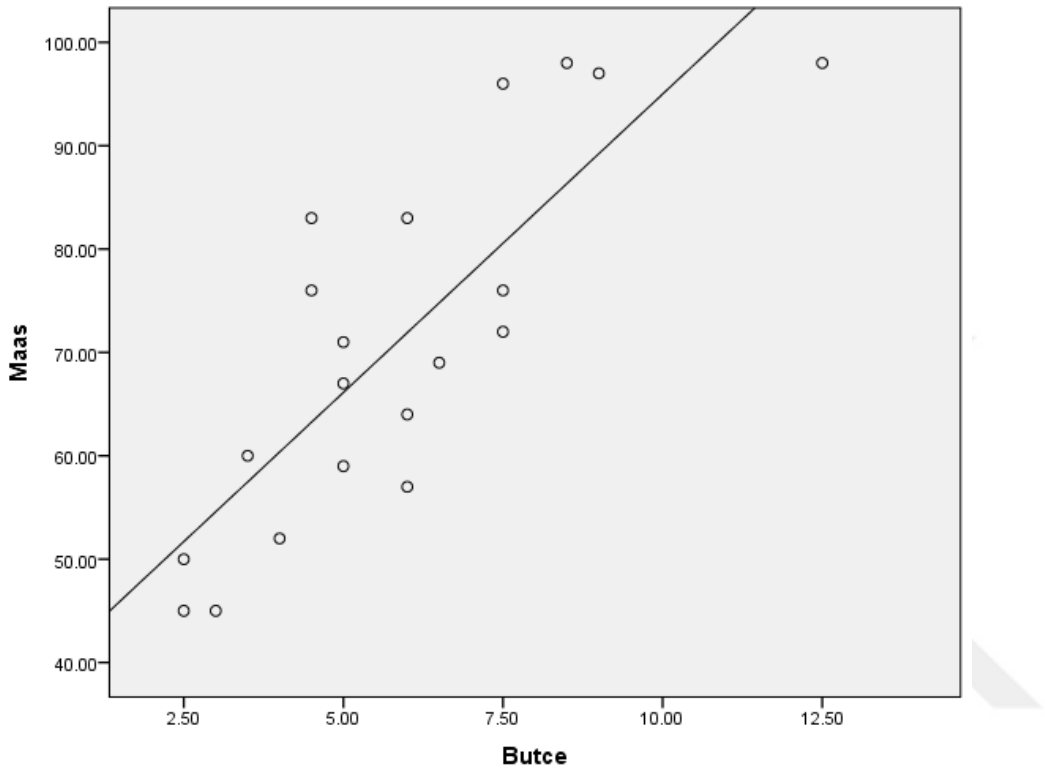
Bu bölümde sağlam regresyon tekniklerinden Huber, Hampel, Andrew ve Tukey´ın M-tahminlerini karşılaştırmak amacıyla, önce gerçek bir veri kümesi, sonra gerçek veri kümesi değiştirilerek elde edilen yeni veri kümesi üzerinde M parametre tahminleri elde edilmiştir. Çalışmada IBM SPSS Statistic V23 ile Huber, Tukey, Hampel, Andrew türü M tahminleri hesaplanmıştır.

Gerçek veri olarak bir şirketin farklı bölümlerinde görev yapan yöneticilerin aldıkları maaşlar ve yönettikleri bütçeler dikkate alınarak aşağıdaki çizelge hazırlanmıştır (Aydın 2014).

Çizelge 4.1 Bütçe verileri.

Gözlem No	Maaş (Y) (100TL)	bütçe (X) (100.000TL)
1	60	3.5
2	67	5
3	50	2.5
4	83	6
5	96	7.5
6	76	4.5
7	64	6
8	52	4
9	83	4.5
10	59	5
11	45	2.5
12	98	12.5
13	97	9
14	72	7.5
15	57	6
16	71	5
17	45	3
18	98	8.5
19	76	7.5
20	69	6.5
Toplam	1418	116.5

Uygulamamızda çizelge 4.1’ de verilen veri kümesinin saçılım grafiği aşağıda verilmektedir:



Şekil 4.1 Yönetilen bütçe miktarlarına karşılık alınan maaşların saçılımı

Şekil 4.1’ deki saçılım grafiğinden veri kümesinin doğrusal regresyon modeli ile iyi bir biçimde temsil edilebileceği açık olarak görülmektedir.

Çizelge 4.1’ deki veri kümesinde, aykırı değerin olduğu ve olmadığı durumlar göz önüne alınarak sağlam M-tahmin yöntemleri ile elde edilen değişik tahminler incelenmiştir.

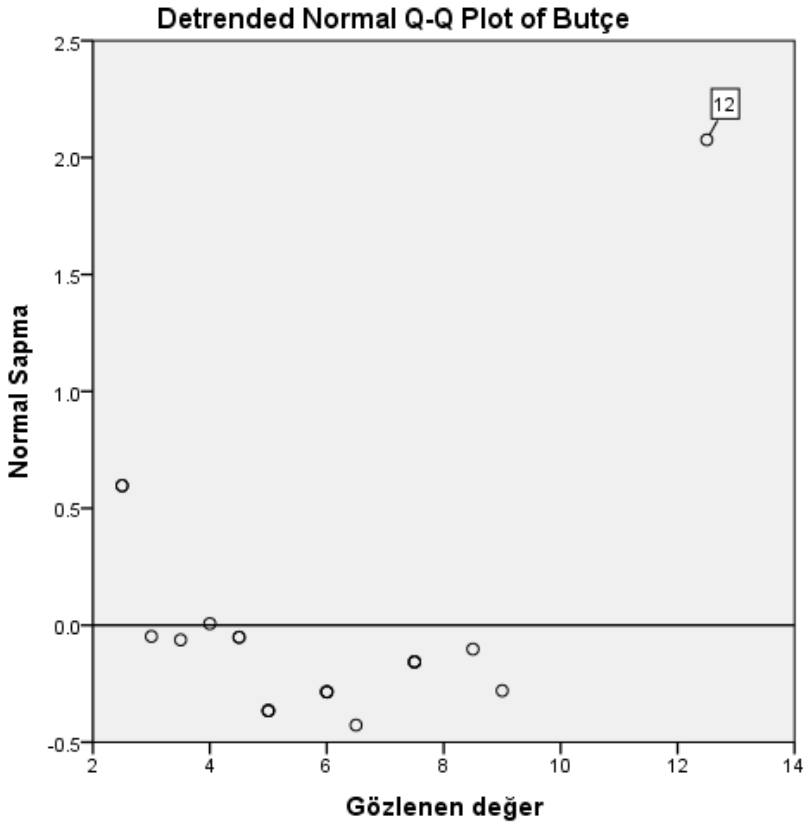
1. **Durum:** Tek aykırı değer olduğunda

Aşağıda sağlam regresyon M tahminleri çizelge’de verilmektedir.

Çizelge 4.2 Tek aykırı değer durumunda sağlam regresyon tahminlerinin karşılaştırılması

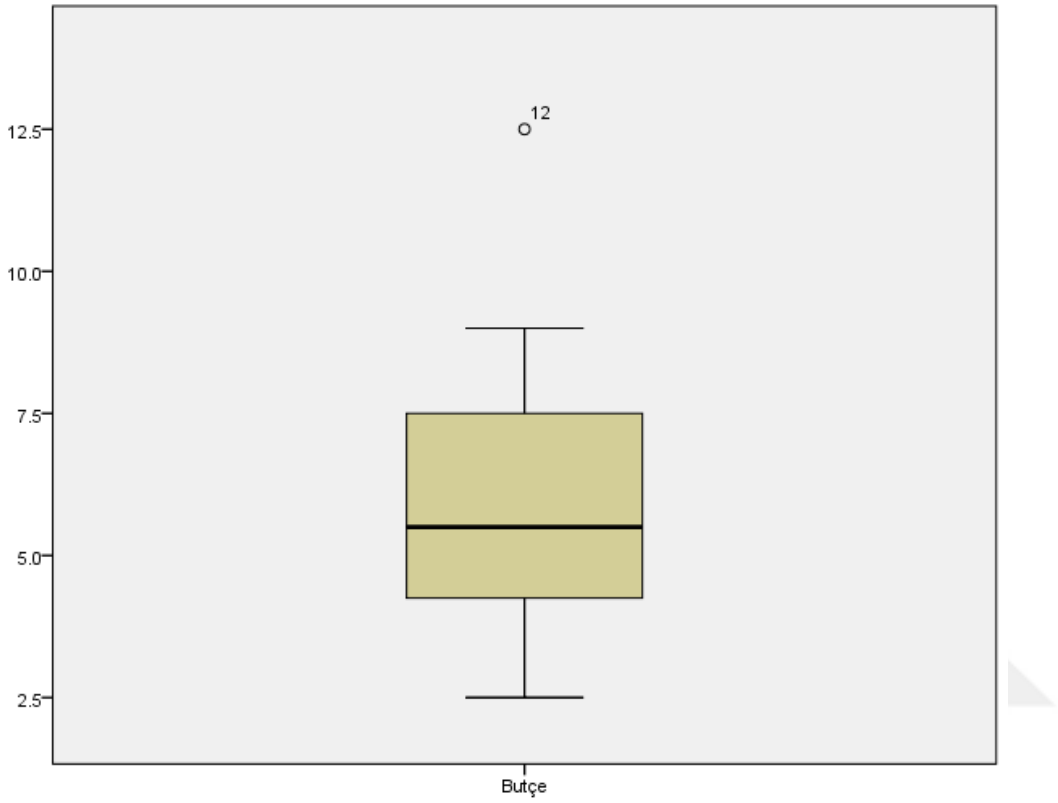
Yöntem	Maaş (Y)	Bütçe
Huber M	69.831	5.601
Tukey M	70.067	5.483
Andrew M	70.071	5.475
Hampel M	70.168	5.602

Veri kümesinin tek aykırı değer olduğu durumunda saçılım grafiği aşağıda verilmektedir:



Şekil 4.2 Tek aykırı değer durumunda veri kümesinin saçılımı

Hesaplanan artıkların saçılım grafiğinden 12. no'lu gözlemin aykırı değer olduğu şekil 4.2'de açık biçimde görülmektedir.



Şekil 4.3 Butçe Box-plot grafiği

Yukarıdaki grafik incelendiğinde veri kümesinde tek aykırı değer olduğu söylenebilir.

2. **Durum:** İki aykırı değer olduğunda

Aşağıdaki çizelge'de veri kümesinde, iki aykırı değer olduğu durumu incelemek amacıyla, bağımsız değişkende rasgele bir değişim yapılarak 7. gözlem için bağımsız değişken değer 14 olarak tanımlandı.

Çizelge 4.3 Veri kümesinde bağımsız değişkende 7. gözlem değerinin değiştirilmesi

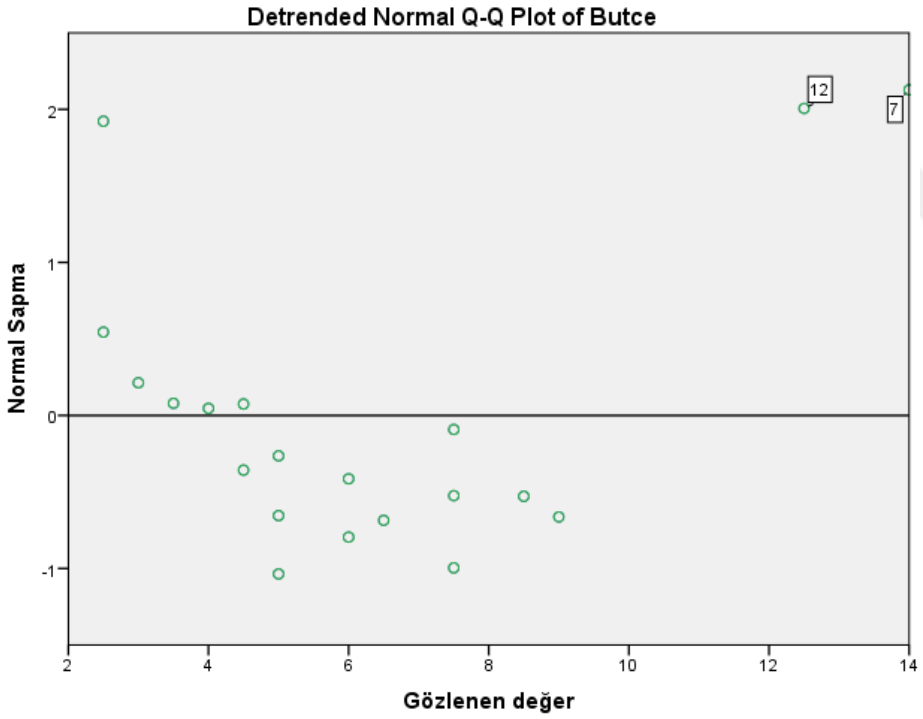
Gözlem No	Maaş (Y) (100TL)	bütçe (X) (100.000TL)
1	60	3.5
2	67	5
3	50	2.5
4	83	6
5	96	7.5
6	76	4.5
7	64	14
8	52	4
9	83	4.5
10	59	5
11	45	2.5
12	98	12.5
13	97	9
14	72	7.5
15	57	6
16	71	5
17	45	3
18	98	8.5
19	76	7.5
20	69	6.5
Toplam	1418	124.5

Aşağıda iki aykırı değer olduğunda sağlam parametre M tahminleri çizelge'de verilmektedir.

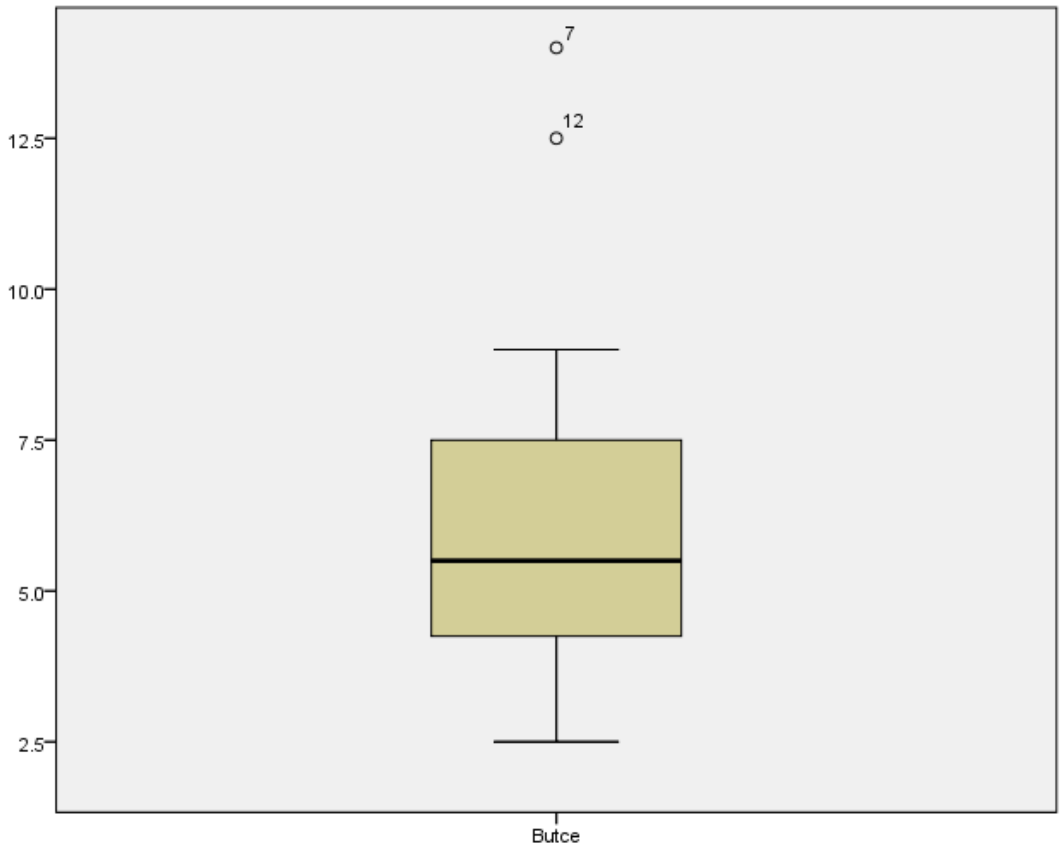
Çizelge 4.4 İki aykırı değer durumunda sağlam regresyon tahminlerin karşılaştırılması

Yöntem	Maaş (Y)	Bütçe
Huber M	69.831	5.77
Tukey M	70.067	5.52
Andrew M	70.071	5.5
Hampel M	70.168	5.79

İki aykırı değer durumunda plot ve saçılım grafik aşağıda verilmektedir



Şekil 4.4 İki aykırı değer durumunda veri kümesinin saçılımı



Şekil 4.5 İki aykırı değer durumunda Box-plot grafiği

Çizelge 4.4 (İki aykırı değer durumu) ve çizelge 4.2 (tek aykırı değer durumu) dikkatle incelendiğinde sonuçların birbirine çok yakın olduğu görülebilir. Yani, Huber M ve Hampel M- tahmin edicileri arasında önemli bir fark yoktur. Aynı şekilde Tukey M ile Andrew M- tahmin edicileri de birbirine çok yakın olduğu görülebilir. Ayrıca çalıştığımız veri kümesi için Tukey ve Andrew M tahmin değerleri Huber ve Hampel M tahminlerinden daha anlamlı olduğu görülmektedir.

3. **Durumu:** Aykırı değer olmadığında

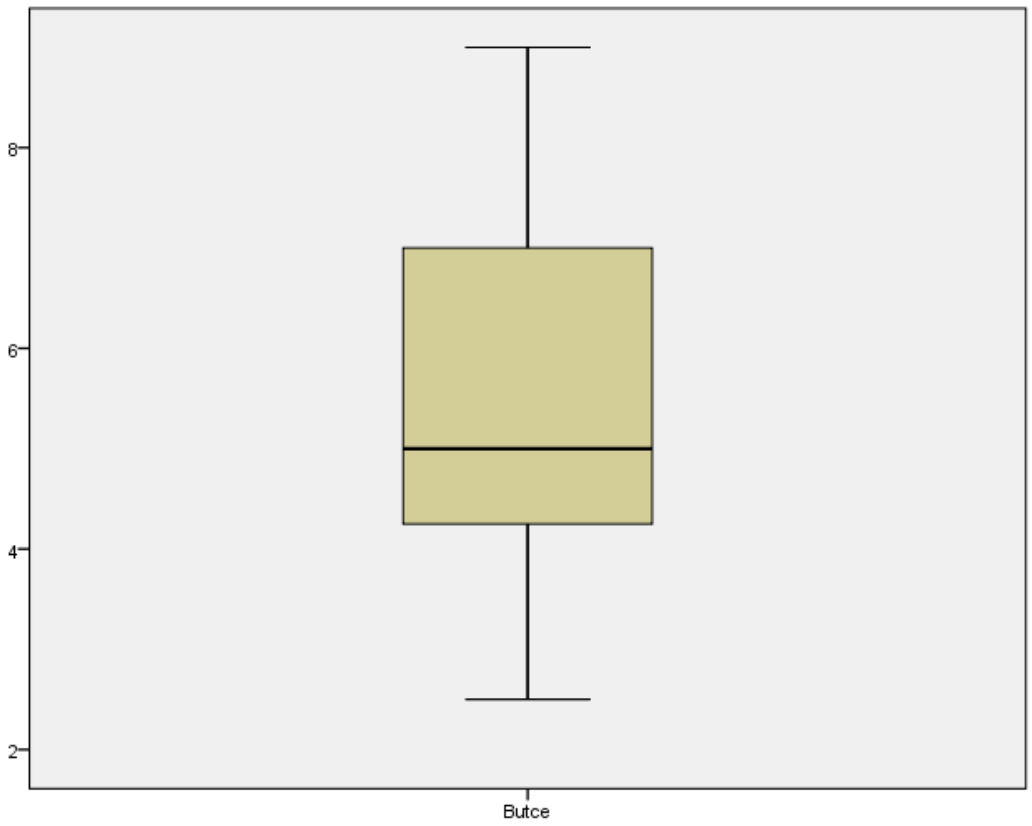
Aykırı değer olmadığında sağlam tahminleri birbiriyle karşılaştırmak amacıyla 12. gözlem değeri veri kümesinden çıkarılmıştır. Aşağıdaki çizelge oluşturan yeni veri kümesi kullanılarak yeniden elde edilen tüm sonuçlar aşağıda verilmiştir.

Çizelge 4.5 Veri kümesinden 12. gözlem değerinin çıkarılması

Gözlem No	Maaş (Y) (100TL)	bütçe (X) (100.000TL)
1	60	3.5
2	67	5
3	50	2.5
4	83	6
5	96	7.5
6	76	4.5
7	64	6
8	52	4
9	83	4.5
10	59	5
11	45	2.5
12		
13	97	9
14	72	7.5
15	57	6
16	71	5
17	45	3
18	98	8.5
19	76	7.5
20	69	6.5
Toplam	1320	104

Çizelge 4.6 Sağlam regresyon tahminlerinin karşılaştırılması

Yöntem	Maaş (Y)	Bütçe
Huber M	68.46	5.44
Tukey M	68.49	5.41
Andrew M	68.6	5.41
Hampel M	68.5	5.43



Sekil 4.6 Aykırı değer olmadığı durumunda Box-plot grafiği

Şekil 4.6, incelendiğinde veri kümesinde aykırı değer olmadığı görülmektedir. Aykırı değer yokken çizelge 4.6 incelendiğinde Huber, Tukey, Andrew ve Hampel tahminlerin birbirine çok yakın olduğu görülmektedir

BEŞİNCİ BÖLÜM

5. SONUÇ VE TARTIŞMA

Birinci bölümde sağlam yöntemler ile ilgili bilgiler derlenerek doğrusal regresyon tahminleri için sağlam tahminler incelendi.

İkinci bölümde, kullanılan tahmin yöntemleri ve tanısal ölçümlere önemli sağlamlık unsurlarından olan bozulama noktası tanımlandı. Üçüncü bölümde yüksek bozulma noktasına sahip yöntemlerin özellikleri tartışıldı.

Dördüncü bölümde, kullanılan veri kümesi için Sağlam Regresyon tekniklerinden Huber, Hampel, Andrew ve Tukey´ın M-tahminleri uygulandı**.** Bu yöntemleri, aykırı değerin olduğu ve olmadığı durumlar göz önüne alınarak birbiriyle karşılaştırıldı.

Uygulamada gerçek bir veri kümesi kullanıldı. İlk olarak veri kümesinin aykırı değer olup olmadığı test edildi ve tek aykırı değer olduğu görüldü. Aynı veri kümesinde rasgele yapılan bir değişiklikle iki aykırı değer olan veri kümesi yaratıldı. Öte yanda orjinal veri kümesinden 12. gözlem değeri çıkartılarak aynı sağlam teknikler karşılaştırıldı.

Veri kümesindeki tek veya iki aykırı değer olması durumunda Tukey ve Andrew'in M-tahminleri diğer tahminlerinden daha iyi sonuçlar vermektedir. Aykırı değerin olmadığı durumda ise sağlam yöntemler arasında büyük bir fark olmadığı görülebilir.

Veri kümesindeki aykırı değer sayısı artığında Hampel ve Huber yöntemlerinin sağlam olmadığı gözlemlenebilir. Aykırı değer olmadığı veya tek aykırı değer olması durumunda Tukey ve Andrew'in M tahminleri çok yakın sonuçlar verdiği görülmektedir.

KAYNAKLAR

[1] Aydın, Dursun. *Uygulamalı Regresyon Analizi, Kavramlar ve R hesaplamları*, Ankara. **2014.**

[2] Alpu Özlem, Şamkar Hatice, Altan Ekrem. *Sağlam Ridge Regresyon Analizi ve Bir Uygulama*. Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt:25, Sayı:2, ss.137-148. **2010.**

[3] Akbilgiç, Oğuz; Keskintürk, Timur, *Yapay Sinir Ağları Ve Çoklu Regresyon Analizinin Karşılaştırılması*. İstanbul Üniversitesi. **2008.**

[4] Candan, M. *Doğrusal Regresyon Çözümlemesinde Sağlam Kestiriciler*, Hacettepe Üniversitesi. ANKARA. **1995.**

[5] Ertaş Hasan. *Çoklu Lineer Regresyonda Sapan Değerlerin Belirlenmesi İçin Tanılama Ölçüleri*. Çukurova Üniversitesi. Adana. **2011.**

[6] Gamgam, H., Altunkaynak, B*., Regresyon Analizi*, Seçkin Yayınevi, Ankara, **2015.**

[7] Hampel, F., *A General Qualitative Definition Of Robustness*, The Annals of Mathematical Statistics. **1971.**

[8] Huber, J., *Robust Statistics*, John Wiley and Sons Inc. NY **1981.**

[9] Huber, J. *Robust Statıstıcs Pricedures*, J. W. Arrowsmith Ltd. Bristol 3. England **1985.**

[10]Huber, Peter; Ronchetti, Evezio. *Robust Statistics*, John Wiley and Sons Inc. NY **2009.**

[11] İnal, Cevat; Yetkin, Mevlüt. *Robust Yöntemlere Uyuşumsuz Ölçülerin Berlirlenmesi*. Selçuk Üniversitesi. Konya. **2006.**

[12]Kavruk, Neslihan Tuba, *Doğrusal Regresyonda Sağlam Güven Aralıkları*. Hacettepe Üniversitesi. Ankara. **2005.**

[13]Maronna, R. A. And Yohai, V. J. *The Breakdown Point of Simultaneous General M Estimates of Regression and Scale*. **1991.**

[14]Myers, H. *Classical And Modern Regression With Applıcatıon*, Duxbury Press, Boston, Massachusetts. **1986.**

[15]Pekgör Ahmet, Genç Aşır, *Doğrusal Olmayan Regresyonda Bozulma Noktası Ve Bir Uygulama*, Selçuk Üniversitesi Fen Fakültesi İstatistik Bölümü Kampüs/ Konya. **2010.**

[16]Ross, G. J., *Nonlınear Models*, Icpam, França. **1987.**

[17]Rousseeuw, P., *Leastmedıan Of Squares Regressıon*. Journal of American Satistical Association. **1984.**

[18]Rousseeuw, P. J., Leroy, A. *Robust Regression And Outlier Detection*, John Wiley, Ny. **1987.**

[19]Ryan, P., *Modern Regression Methods,* John Wiley and Sons Ny. **1997.**

[20]Yohaı, V. And Zamar, R. *High Breakdown-Poınt Estımates Of Regressıon By Means Of Mınımızatıon Of An Effıcıent Scale*. American Statistical Association. **1988.**

[21]Yohaı, V. *High Breakdown-Poınt And Hıgh Effıcıency Robust Estımates For Regressıon*, The Annals Statıstıcs. **1987.**

- İNTERNET KAYNAKLARININ YAZIMI

[22]https://scholar.google.com.tr/scholar?q=do%C4%9Frusal+regresyon+modeli&hl=en&as_sdt=0&as_vis=1&oi=scholart&sa=X&ved=0CBgQgQMwAGoVChMIs4aziYeFxgIVhw8sCh08KQZY

[23]http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6828704

[24]http://www.turkistatistik.org/wp-content/uploads/2015/03/IstKon7_bildiriler_kitabi.pdf

[25]https://en.wikipedia.org/wiki/Redescending_M-estimator

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı Soyadı: Joaquim Jorge da Costa Khalau

Doğum Yeri: Mozambik

Medeni Hali: Bekar

E-posta: Joaquim.khalau@gmail.com

Adress: Çankırı caddesi No 5. Ankara

Eğitim

Lisans: Universidade Pedagogica – Mozambik

Yüksek Lisans: -

Doktora: -

Yabancı Dil ve Düveyi: Portekizce; İngilizce

İş Deneyimi: Öğretmenlik – Sao Tomas University - Mozambik

Deneyim Alanları: -

Tezden Üretilmiş Projeler ve Bütçesi: -

Tezden Üretilmiş Yayınlar: -

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile katıldığı Toplantılar: -

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 06/12/2016

Tez Başlığı / ~~Konusu~~: DOĞRUSAL REGRESYONDA BOZULMA NOKTASINA SAHİP TAHMİNLER

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç ve e)Kaynakça kısımlarından oluşan toplam 40 sayfalık kısmına ilişkin, 10/11/2016 tarihinde ~~şahsım~~/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 7 'tür.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar ~~hariç~~/dâhil
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

06/12/2016
Tarih ve İmza

Adı Soyadı: JOAQUİM JORGE DA COSTA KHALAU
Öğrenci No: N12124224
Anabilim Dalı: İSTATİSTİK
Programı: YÜKSEK LİSANS
Statüsü: ☒ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

Prof. Dr. Süleyman Günay
(Unvan, Ad Soyad, İmza)