ISBN No: 978-975-19-4105-3 Volume: 05 Number: 01 Page: 35-48 ISSN No: 1303-6319 Cilt: 05 Sayı: 01 Sayfa: 35-48

ÇOKLU REGRESYON UÇDEĞERLERİNİN TEŞHİSÇİSİ OLARAK ÇANTA ÇİZİTİ

Enis SINIKSARAN[††] **M. Hakan SATMAN**[**]

ÖZET

Çanta çiziti tek değişkenli veriler için kullanılan ve uçdeğerlerin tespitinde de faydalanılan kutu çizitinin iki değişkenli versiyonudur. Dolayısıyla tek bağımsız değişkenli regresyon uçdeğerlerinin tespitinde kullanılabilir. Ancak bağımsız değişken sayısı birden fazla olduğunda, çanta çizitinin dolaysız kullanılma şansı yoktur. Öte yandan regresyon kalıntıları ile bağımlı değişkenin tahmin değerlerinin belirlediği iki boyutlu uzayın tüm veriyi belirli nispette temsil etme yeteneği literatürde kanıtlanmıştır. Bu çalışmada temel olarak bu sonuçtan hareket edilmiştir. Çoklu regresyon modelinden elde edilen kalıntı ve tahmin değerlerinin belirlediği uzayda çanta çizitinin regresyon uçdeğerlerini belirlemedeki performansı bazı klasik verilerle ve çeşitli senaryolarda Monte Carlo simülasyonlarıyla araştırılmıştır. Yaklaşım bir çok senaryoda başarılı bulunmuştur.

Anahtar Kelimeler: Çanta Çiziti, Gizleme, Uç Değer, Yanlış Alarm.

1. GİRİŞ

Regresyon uçdeğerleri, verinin çoğunluğu ile belirlenen regresyon düzleminden uzak olan noktalardır. Bu uçdeğerlerin teşhis edilmeleri, parametre tahminleri ve dolayısıyla istatistiksel çıkarsamalar üzerindeki olumsuz etkileri nedeniyle büyük önem taşır. Regresyon uçdeğerlerinin tanımlanma problemine ve klasik teşhisçilere ilişkin temel bilgiler Belshey vd. (1980), Cook ve Weisberg (1980) ve Rousseeuw ve Leroy (1987)'de bulunabilir.

Bir regresyon verisinde, tek ya da çok az sayıda uçdeğer söz konusu ise Cook mesafesi, Dfbetas ve COVRATIO gibi klasik en küçük kareler teşhisçileri işe yarayabilir. Ancak veri içindeki küçük bir grup uçdeğer birlikte hareket ettiğinde, klasik teşhisçiler bu uçdeğerleri teşhis edemezler (masking) ya da gerçekte uçdeğer olmayan gözlemleri uçdeğer olarak tanımlayabilirler (swamping). Çalışmanın bundan sonraki bölümlerinde masking, "gizleme" olarak swamping ise "yanlış alarm" olarak isimlendirilecektir. Araştırmacılar, gizlemeyi, yanlış alarma kıyasla daha ciddi bir hata olarak görseler de, iyi bir uçdeğer saptama algoritması her iki hatayı da en küçükte tutmalıdır. İstatistik literatüründe bu iddiayı taşıyan pek çok algoritma mevcuttur. Hadi ve Simonoff (1993) ile Wisnowski vd., (2001)'de bu algoritma ve performanslarına ilişkin, özlü bilgilere ulaşılabilir.

Literatürde, uçdeğer saptama algoritmalarının performansları değerlendirilirken, genellikle iki yola başvurulmaktadır:

[††] Yard. Doç. Dr., İstanbul Üniversitesi İktisat Fakültesi Ekonometri Bölümü, İstanbul, Türkiye, e-mail: esiniksaran@istanbul.edu.tr
[**] Araş. Gör., İstanbul Üniversitesi İktisat Fakültesi Ekonometri Bölümü, İstanbul, Türkiye, e-mail: mhsatman@istanbul.edu.tr

i. Telephone Data (Rouseeuw ve Leroy, 1987), Stackloss Data (Brownlee, 1965), Hawkins, Bradu ve Kass Data (Hawkins vd., 1984), Modified Wood Gravity Data (Rouseeuw ve Leroy, 1987) ya da Hadi-Simonoff Data (Hadi ve Simonoff, 1993) gibi uçdeğer analizleri ve dayanıklı (robust) yöntemlere ilişkin çalışmalarda sıkça başvurulan, bir kısmı gerçek veri bir kısmı ise hipotetik veri olan ve klasik teşhisçilerin genellikle başarısız olduğu verilerdeki başarı yüzdeleri.
ii. Gözlem ve değişken sayısı, uçdeğerin tipi, uçdeğerlerin verinin genelinden uzaklığı gibi pek çok faktöre bağlı olarak değişen senaryolarda gerçekleştirilen Monte Carlo simülasyonlarındaki başarı yüzdeleridir. Kianifard ve Swallow (1990) ve Wisnowski vd., (2001)'nın çalışmaları bu konuda iyi birer örnektir.

Rousseuw, Ruts ve Tukey'in (1999) tarafından geliştirilen çanta çiziti (bagplot), tek değişken için kullanılan kutu çizitinin (boxplot) iki değişkene uyarlanmasıdır. Söz konusu çalışmada 3 boyutlu çanta çizitinin, çizim olanakları da tartışılmıştır. k tane bağımsız değişkenli çoklu regresyon için, bağımlı değişken de hesaba katıldığında, uçdeğerleri gösterecek bir çanta çiziti k + 1 boyutlu olacaktır. Çizim olanağı 3 boyutla sınırlı olduğuna göre k > 2 olduğunda, regresyon uçdeğerlerinin tespitinde çanta çizitinden yararlanmak şansı kalmayacaktır. Bu noktada Sebert vd., (1998)'nın çalışmasının sonuçlarından yararlanılabilir. Bu çalışmada, çoklu regresyon verisine en küçük kareler uygulanmakta, buradan elde edilen kalıntı ve bağımlı değişkenin tahmin değerleri standartlaştırıldıktan sonra, bu ikililerin belirlediği noktalar arasındaki Öklit mesafeleri hesaplanmaktadır. Bu mesafelere dayanarak, gözlemler bir dendogram yardımıyla sınıflandırılmakta ve Mojena kuralı yardımıyla belirli bir noktada kesilen dendogram verideki grupları ve dolayısıyla uçdeğerleri ortaya çıkarmaktadır. Sebert vd. (1998) tasarlanan bu sürecin gerek klasik verilerde, gerekse Monte Carlo simülasyonlarında başarılı olduğunu kanıtlamışlardır. Wisnowski (2001)'deki Monte Carlo simülasyonları da bu sonucu destekler görünmektedir. Sebert vd., (1998) çalışmasındaki süreç standartlaştırılmış tahmin değerleri ile kalıntı değerlerinin belirlediği 2 boyutlu uzaya (çalışmanın bundan sonraki bölümlerinde bu uzay TK uzayı olarak adlandırılacaktır) dayanmaktadır. Bu çalışmada çoklu regresyon uçdeğerlerini tespitte TK uzayında çanta çizitinin performansı araştırılmıştır. Bunu yaparken yukarıda belirtilen geleneğe bağlı kalınmıştır. Bir başka ifade ile, yöntemin başarısı hem bazı klasik veriler, hem de Monte Carlo simülasyonları ile araştırılmıştır. Sezgisel olarak, bu 2 boyutlu uzayın çoklu regresyonu temsil etme yeteneğinin değişken sayısı ve gözlem sayısı artarken azalması beklenmelidir. Bunun yanı sıra uçdeğerlerin oranı, verinin genelinden uzaklıkları, hangi tipte olduğu gibi faktörler de şüphesiz yöntemin başarısını etkileyecektir. Bu nedenle Monte Carlo simülasyonlarında söz konusu faktörler göz önüne alınmıştır.

Çalışmanın 2. Bölümünde çanta çiziti tanıtılmış ve bazı klasik veriler için tasarlanan yöntemin başarısı araştırılmıştır. Bölüm 3'te çeşitli senaryolar için yöntemin performansı Monte Carlo simülasyonları ile değerlendirilmiş, Bölüm 4'te ise çalışmanın sonuçları kısaca tartışılmıştır. Makaledeki hesaplamalarda çanta çizitinin (bagplot) çizimini sağlayacak gözlemlerin saptanmasında, Rousseeuw ve Ruts (1999)'ın bir Fortran programı olan BAGPLOT'tan yararlanılmıştır. Simülasyonlar ve grafikler, Fortran ile Mathematica programı arasında yaratılan bir arayüz ile gerçekleştirilmiştir.

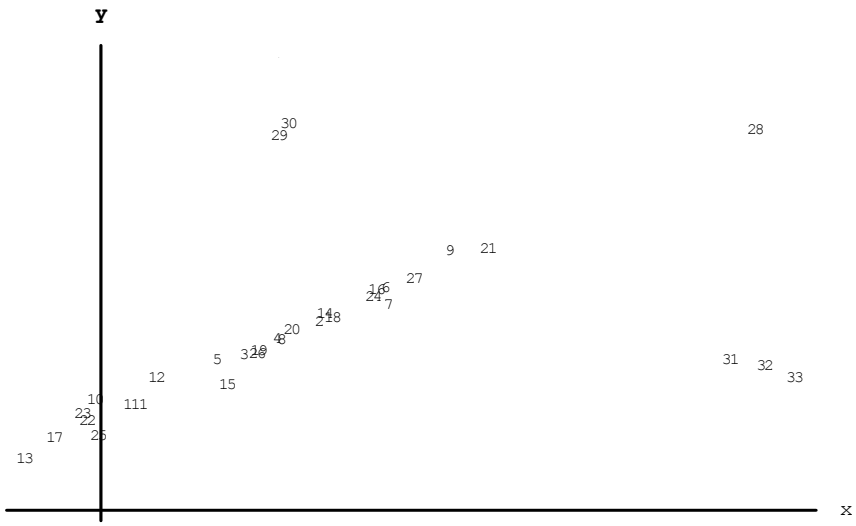
2. ÇANTA ÇİZİTİ VE TK UZAYI

Regresyon uçdeğerleri, verinin geneliyle olan ilişkileri ile tanımlanır. Basitçe ifade edilirse, veride çoğunluğun oluşturduğu lineer örüntüye uymayan gözlemler uçdeğer olarak nitelenir. Bu uçdeğerlerden bağımlı değişken yönünde uzağa düşenler *y*-uzayı uçdeğeri ya da dikey uçdeğer, bağımsız değişken/değişkenler (*x*-uzayı olarak da isimlendirilir) yönünde uzağa düşenler ise, *x*-uzayı uçdeğeri olarak isimlendirilir. *x*-uzayı uçdeğerleri regresyon düzlemini kendilerine doğru çektikleri, diğer bir deyişle kaldıraç etkisine sahip oldukları için kötü kaldıraçlar olarak da isimlendirilir. Kötü olarak nitelenme sebepleri, bağımsız değişkenlerin parametre tahminlerini ve yapılacak çıkarsamaları olumsuz etkilemeleridir. Dikey uçdeğerler de bir ölçüde regresyon düzlemini etkilese de, olumsuz etki bağımsız değişken parametrelerinin tahminlerinden daha çok sabit terimin tahmini üzerine olduğu için, *x*-uzayı uçdeğerleri kadar tehlikeli görülmezler. Öte yandan verinin genelinden uzak olsa da genelin belirlediği lineer örüntüye uyan gözlemler de söz konusu olabilir. Bu tür gözlemler de regresyon düzlemine kaldıraç etkisi yaparlar, ancak bu etki lineer örüntü yönünde olacağı için, tahminler ve çıkarsamalar olumlu yönde etkilenecektir. Bu nedenle bu tür gözlemler iyi kaldıraç olarak nitelenir.

Tablo 1'de 33 gözlemden oluşan hipotetik bir veri ve Şekil 1'de bu verinin serpilme çiziti görülmektedir. Veride 6 gözlem dışındaki 27 temiz gözlem bir lineer örüntü sergilemektedir. Lineer örüntüye uymayan 5 uçdeğer söz konusudur. Bunlardan 29 ve 30. gözlemler dikey uçdeğer, 31,32 ve 33. gözlemler ise kötü kaldıraçtır. 28. gözlem ise verinin genelinden uzak da olsa, temiz gözlemlerin oluşturduğu lineer örüntüye uyduğu için iyi kaldıraç olarak nitelenebilir.

Tablo 1. 33 gözlemli hipotetik veri

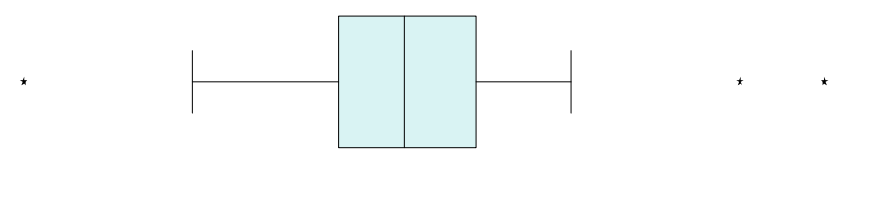
x	6,68	13,7	10,66	11,98	9,59	16,32	16,4	12,15	
y	9,96	17,81	14,67	16,27	14,34	21,07	19,58	16,06	
x	3,15	14,21	11,32	12,6	20,44	4,49	4,31	15,83	
y	6,91	18,29	15,09	17,05	24,82	8,52	9,17	20,19	
x	18,85	4,8	6,23	7,23	1,99	13,91	10,02	15,97	
y	24,63	10,46	10,08	12,52	4,92	18,71	11,92	21	
x	4,91	11,25	17,51	31,02	12,08	12,51	30,05	31,4	32,64
y	7,06	14,85	21,87	36,12	35,62	36,7	14,22	13,71	12,63



Şekil 1. 33 gözlemli hipotetik verinin serpilme çiziti ve uçdeğerler

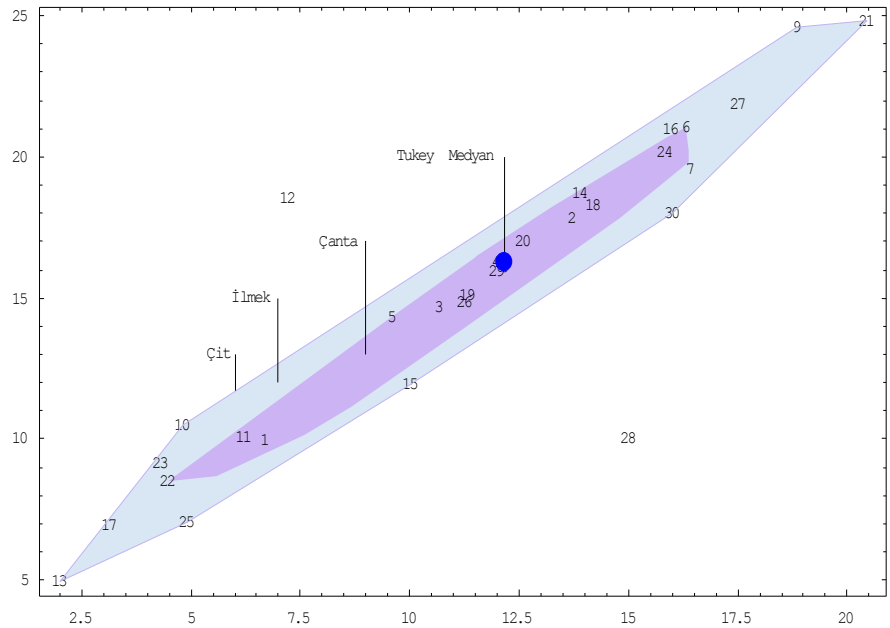
Tek bağımsız değişken söz konusu olduğunda bu örnekte olduğu gibi uçdeğerleri saptamak zor olmaz. Ancak bağımsız değişken sayısı arttırıldığında, verinin bütününün bir serpilme çizitinde gösterilmesi olanağı ortadan kalkacağı için, uçdeğerlerin tek bir grafikle görsel tespiti de mümkün olmayacaktır. Ancak yukarıda tek bağımsız değişken için yapılan uçdeğer tanımlamaları çoklu regresyon için de geçerli olacaktır.

Rousseeuw, Ruts ve Tukey (1999) tarafından tasarlanan çanta çiziti, sorgulayıcı veri analizinde (exploratory data analysis) kullanılan kutu çizitinin (boxplot) iki değişkene uyarlanmasıdır. Şekil 2'de tek değişkenli 55 gözlemden oluşan bir verinin kutu çiziti yer almaktadır. Bilindiği gibi kutu çizitinde, kutu kısmı (şekildeki boyalı bölüm) merkezde (en derin değer olan medyan civarında) kalan % 50'lik kısmı gösterir. Kutunun sol tarafı birinci kartil, sağ tarafı ise üçüncü kartil ile belirlenir. Kutunun ortasındaki çizgi ise medyanı gösterir. Kutunun uzunluğu kartil aralığı olarak isimlendirilirken, kutunun yanlarından çıkan bıyıklar (whiskers) kartil aralığının 1,5 katı kadar mesafedeki gözlemlere kadar uzatılır. Bu mesafede gözlem yoksa, bıyıklar bu mesafe içine düşen kutuya en uzak gözleme kadar uzatılır. Dolayısıyla bıyıkların dışına düşen gözlemler uçdeğer olarak düşünülür. Şekil 2 ile temsil edilen veride solda 1, sağda ise 2 uçdeğer görülmektedir. Bazı yazarlar kartil aralığının 3 katı mesafeden öte olanları uzak uçdeğer, kartil aralığının 1,5–3 katı mesafede olanları ise yakın uçdeğer olarak nitelerler. Kutu çiziti, yalnızca uçdeğerler değil, gerek medyan civarındaki % 50'nin, gerekse genel olarak verinin dağılımı, simetrisi hakkında oldukça bilgi verici bir araçtır.



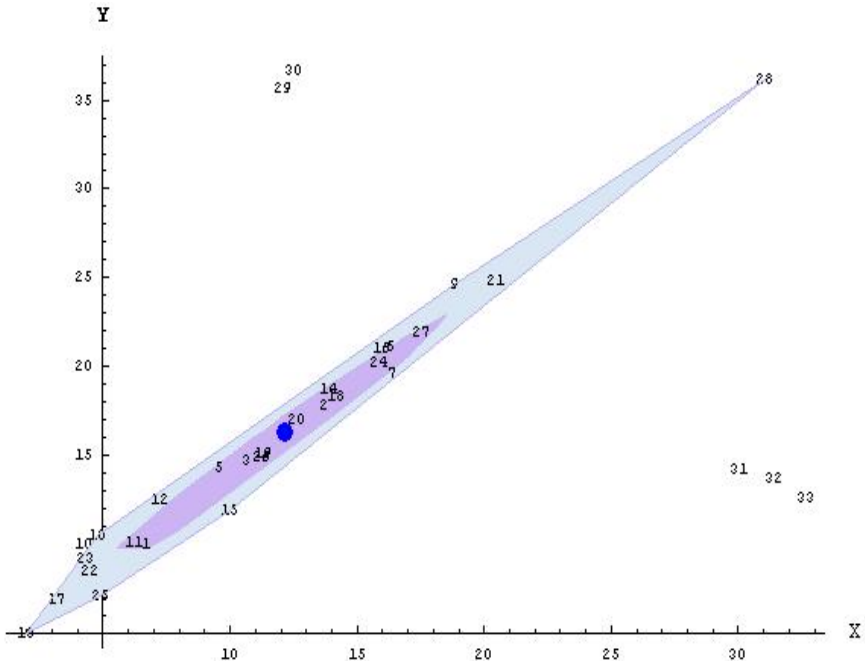
Şekil 2. 55 gözlemden oluşan bir verinin kutu çiziti ve uçdeğerler

Şekil 3'te 30 gözlemden oluşan iki değişkenli bir verinin çanta çiziti görülmektedir. Bir çanta çiziti, gözlemlerin merkeze (Tukey medyanına) en yakın % 50'sini içeren bir çanta (bag), temiz gözlemleri, uçdeğerlerden ayıran bir çit (fence) ve çanta ile çit arasında kalan bölgeyi temsil eden ilmekten (loop) oluşur. Analoji kurmak açısından; kutu çizitindeki medyanın, çanta çizitinde Tukey medyanına, kutunun, çantaya, bıyıkların ise çite denk düştüğü söyleyenebilir. Çanta çizitine bakıldığında, dikey uçdeğer olan 12 ve 28 numaralı gözlemlerin çitin dışında kaldığı, çanta çiziti tarafından uçdeğer olarak belirlendiği söylenebilir.

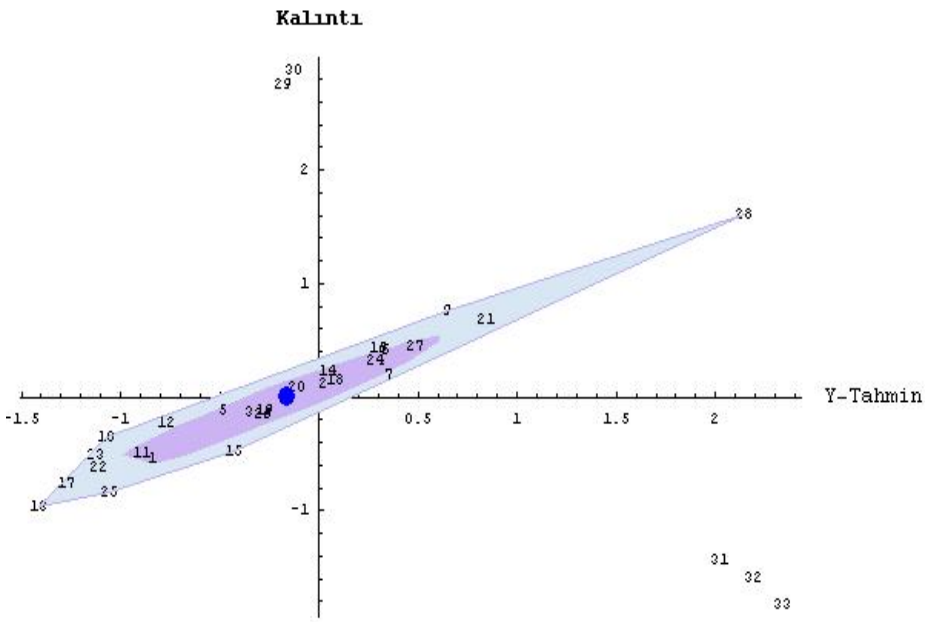


Şekil 3. 30 gözlemden oluşan bir verinin çanta çiziti ve uçdeğerler

Şekil 4'te ise, Tablo 1'de verilen hipotetik verinin çanta çiziti görülmektedir. Dikey uçdeğerler olan 29 ve 30. gözlemler ile kötü kaldıraçlar 31,32 ve 33. gözlemler çit dışında yer alırken, iyi kaldıraç olan 28. gözlem çit içine düşmüştür. Çanta çizitinin genel şekli ise Rousseeuw vd., (1999)'nın öngördüğü gibi verinin korelasyon yapısına uygun bir şekil almıştır. Şekil 5'te ise aynı verinin TK uzayındaki (yatay eksenin bağımlı değişkenin standartlaştırılmış En Küçük Kareler Yöntemi (EKKY) tahmin değerleri, dikey eksenin ise EKKY kalıntıları ile belirlendiği 2 boyutlu koordinat sistemi) çanta çiziti yer almaktadır. Görüldüğü gibi TK uzayı, Şekil 4'te serpilme diyagramı verilen ham verinin yapısını çok büyük ölçüde yansıtmakta ve çanta çizitinin genel karakteri benzerlik göstermektedir. Ayrıca çanta çiziti her iki uzayda da aynı uçdeğerleri teşhis etmiştir.

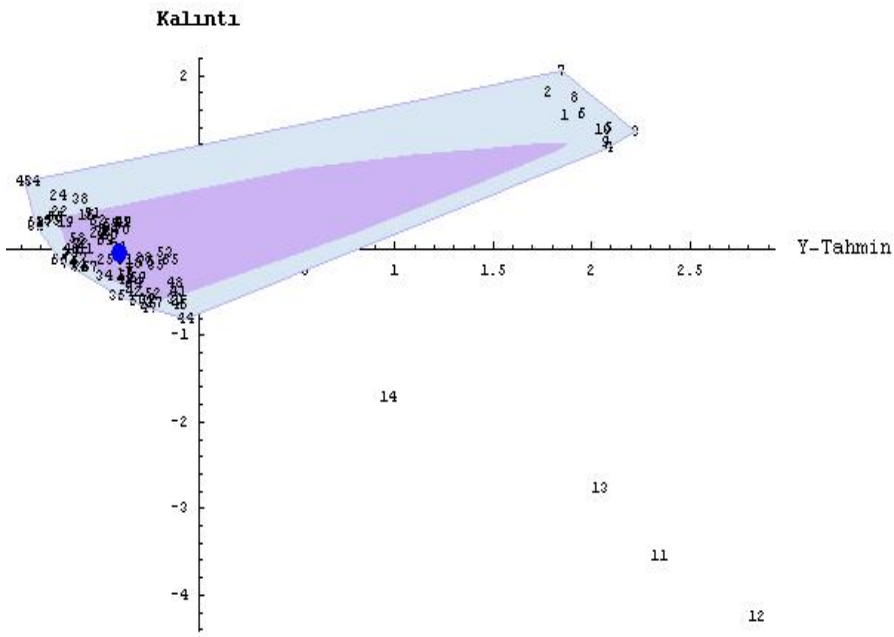


Şekil 4. Hipotetik verinin çanta çiziti ve uçdeğerler



Şekil 5. Hipotetik verinin TK uzayında çanta çiziti ve uçdeğerler

Bu noktada, akla bağımsız değişken sayısı arttırıldığında TK uzayının verinin genelini temsil yeteneğinin, dolayısıyla uçdeğerlerin görsel olarak teşhis edilme olanağının ne ölçüde etkileneceği sorusu gelebilir. Bu sorunun yanıtı, makalenin ana konusudur. Sorunun yanıtı öncelikle klasik verilerde aranmıştır. İlk örnek olarak 3 bağımsız değişkenli ve 75 gözlemli Hawkins, Bradu ve Kass verisi (Hawkins vd., 1984) ele alınmıştır. Bu veride ilk 10 gözlem *xy*-uzayı uçdeğerleri iyi huylu kaldıraçlar iken, 11, 12, 13 ve 14. gözlemler *x*-uzayı uçdeğeri, kötü kaldıraçlardır. Geriye kalan 61 gözlem ise temizdir. Şekil 6'da TK uzayındaki çanta çiziti yer almaktadır. Her şeyden önce, TK uzayının verinin genel yapısını oldukça doğru bir biçimde yansıttığına, temiz gözlemler, iyi kaldıraçlar ve kötü kaldıraçların ayrı kümeler olarak yer aldığına dikkat edilmelidir. Çanta çiziti de hipotetik veri örneğinde olduğu gibi, iyi kaldıraçları çitin içine alırken, kötü kaldıraçları dışarıda bırakmıştır.



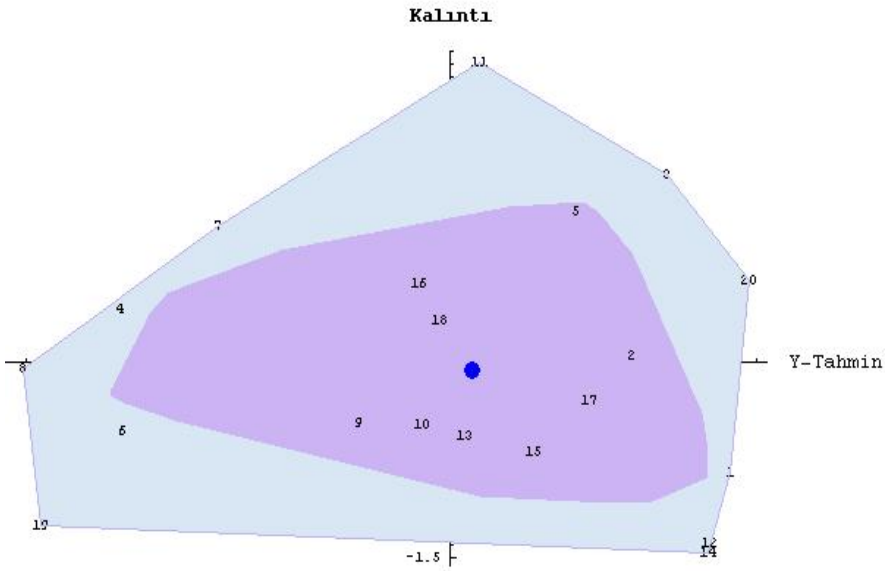
Şekil 6. Bradu Kass verisinin TK uzayında çanta çiziti ve uçdeğerler

Şüphesiz her veri için, sonuçların bu kadar mükemmel çıkması beklenmemelidir. Klasik verilerden 5 bağımsız değişkenli ve 20 gözlemli Modified Wood Gravity (Rouseeuw ve Leroy, 1987) ve 3 bağımsız değişkenli ve 21 gözlemli Stackloss verileri (Brownlee, 1965) ele alındığında; ilk veride 4, 6, 8 ve 19. gözlemler, ikinci de ise 1, 2, 3, 4 ve 21. gözlemler uçdeğerdir. Şekil 7 ve Şekil 8'de bu iki verinin TK uzayında çanta çiziti yer almaktadır. TK uzayı bu iki veri için temiz gözlemleri, uçdeğerlerden daha önceki örneklerdeki kadar net ayırt edememiştir. Çanta çizitlerine bakıldığında ise uçdeğerlerin, çitin dışında yer almasalar da hiç birisinin çantanın içine düşmediği görülmektedir. Modified Wood Gravity veride uçdeğerlerden ikisi çiti oluşturan gözlemlere katılırken, ikisi ilmek bölgesine düşmüştür. Stackloss veride ise, uçdeğerler çiti oluşturan gözlemlerden olup, hiçbiri ilmek bölgesine düşmemiştir.

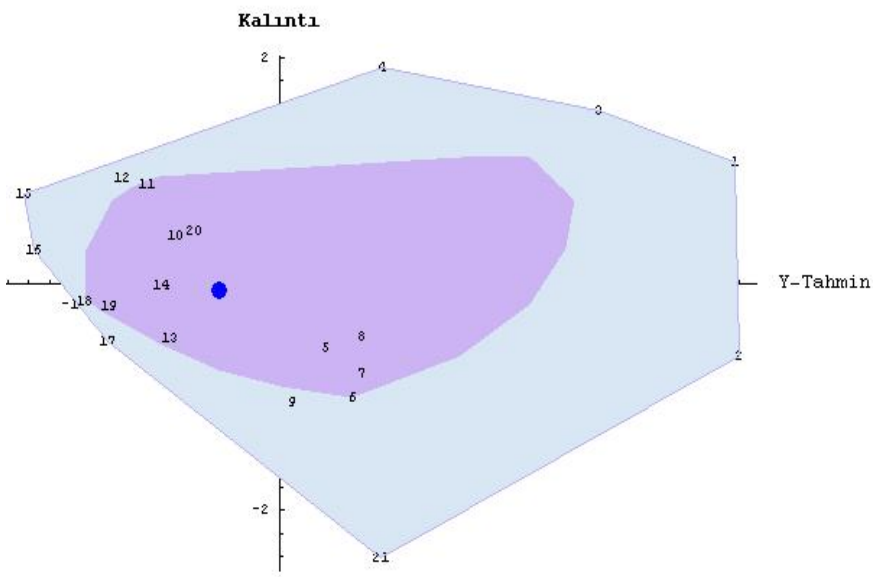
Hipotetik veri ve 3 klasik veriden elde edilen sonuçlara bakılarak şunlar söylenebilir: TK uzayı regresyon verisi genel yapısını belirli ölçüde yansıtmakta, bazen uçdeğerleri

tiplerine de bağlı olmak üzere net biçim ortaya koymaktadır. Bu uzayda çizilen çanta çiziti de uçdeğerleri çitin dışında belirleyerek teşhis edebilmekte, bazı durumlarda uçdeğerler çitin dışında yer almasa da, en azından çanta bölgesine de düşmemektedir.

Şüphesiz TK uzayının ve çanta çizitinin performansına ilişkin genellemeler yapmak için, bu kadar örnek yeterli olmayacağı için, bundan sonraki kısımda sunulacak olan Monte Carlo simülasyonlarına ihtiyaç duyulmuştur.



Şekil 7. Modified Wood Gravity verisinin TK uzayında çanta çiziti ve uçdeğerler



Şekil 8. Stackloss verisinin TK uzayında çanta çiziti ve uçdeğerler

3. MONTE CARLO SİMÜLASYONLARININ SONUÇLARI

Simülasyonlarda $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ regresyon modeli göz önüne alınmış ve modelde sabit terimin olduğu varsayılmıştır. Temiz veri, $\boldsymbol{X}$ matrisinin $\mathbf{1} = [1,1,...,1]'$ vektörünün oluşturduğu ilk sütunu dışındaki sütunları; ortalamaları 7, standart sapması 16 ve kovaryansları 0 olan çok değişkenli normal dağılımdan; hata vektörü $\boldsymbol{\varepsilon}$ ise, standart normal dağılımdan elde edilmiştir. Parametre vektörü $\boldsymbol{\beta} = [\beta_0 , \beta_1 ,..., \beta_k]' = [5 , 5 ,..., 5]'$ şeklinde seçilirken, y değerleri de $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ denkleminden elde edilmiştir. Bu şekilde üretilen regresyon verisi klasik varsayımları karşılamaktadır. Simülasyonlar Tablo 2'de verilen faktörler ve düzeyler için gerçekleştirilmiştir. 5 faktörün toplam 13 farklı düzeyi olduğu için, $3\times3\times3\times2\times2 = 108$ farklı senaryo söz konusudur. Her bir senaryo ise 500 kez iterasyona sokulmuştur.

Tablo 2. Simülasyon senaryoları

Faktörler	Düzeyler
Sapma tipi	***x*,*y*, *xy*-uzayı**
Bağımsız değişken sayısı	**2,3,5**
Gözlem sayısı	**40,60,100**
Uçdeğerlerin yüzdesi	**% 5, % 10**
Sapma miktarı	**1σ, 2σ**

Bir çanta çizitinde çanta, ilmek ve çit dışı olmak üzere 3 bölge söz konusu olduğu için her bölgeye düşen uçdeğerler ayrı ayrı sayılmış ve buna bağlı olarak gizleme ve yanlış alarm yüzdeleri hesaplanmıştır. Tablolardaki bir bölgeye ait gizleme yüzdesi, o bölgenin verideki uçdeğerleri yakalayamama yüzdesi olarak anlaşılmalıdır. Şüphesiz bir çanta çiziti için beklenen, uçdeğerlerin çantaya ve ilmeğe değil, daha çok çit dışına düşmesidir. Dolayısıyla gizlenme yüzdelerinin çantada ve ilmekte yüksek, çit dışında düşük olması çanta çizitinin başarısı olarak değerlendirilmelidir.

Simülasyon sonuçları x, y ve xy-uzayındaki sapmalar, çanta, ilmek ve çit dışı için ayrı ayrı sunulmuştur. Dolayısıyla tüm sonuçlar aşağıdaki 9 tablo yardımıyla görülebilir. Örnek olarak yorumlanırsa, Tablo 3'te, x-uzayında uçdeğerlerin tespitinde çantanın performansı sunulmaktadır. Burada örneğin, 40 birimlik bir örneklemde % 5 oranında kirlenmede ve temiz verilerden 1σ uzaklık söz konusu olduğunda, gizleme oranı % 98 olarak görülmektedir. Öyleyse bu senaryoda, uçdeğerlerin ancak $1 - 0{,}98 = 0{,}02$'si çantaya düşmektedir. Çantanın gözlemlerden Tukey medyanı civarındaki % 50'lik kısmı içermesi, dolayısıyla uçdeğerleri içermemesi bekleneceği için, gizleme oranın yüksekliği çantanın başarısı anlamına gelecektir. Bu anlamda Tablo 3, 4 ve 5'e bakıldığında çanta, uçdeğerleri içermeme konusunda oldukça başarılı görülmektedir. Bu başarı y ve xy-uzayındaki sapmalarda daha da yüksektir. Kirlenme yüzdesinin ve değişken sayısının artması, bekleneceği gibi, başarıyı az da olsa olumsuz yönde etkilemektedir. Yanlış alarm, aslında temiz olan gözlemleri dışarıda bırakma konusunda, yüzdelerin gizleme yüzdeleri kadar başarılı görülmemesini de doğal karşılamak gerekir. Çünkü çanta çizitinin tasarımı, Çantanın gözlemlerden en derin % 50'yi içermesine dayanmaktadır. Dolayısıyla bu en derin bölgeye düşmeyen öte yandan uç değer de olmayan bazı gözlemler doğallıkla çanta dışında yer alacak ve yanlış alarm olarak kaydedilecektir. Dikkat edilirse yanlış alarm oranları, bu söylenilenlerle tutarlı olmak üzere % 50 civarındadır.

Tablo 6, 7 ve 8'de ise ilmeğin performansı görülmektedir. Sonuçlar beklentilere oldukça uygundur. Şüphesiz uçdeğerleri içermeme konusunda, çanta kadar başarılı olmasının beklenmemesi gereken bu bölge, çanta da olduğu gibi *y* ve *xy*-uzayındaki sapmalarda daha başarılıdır. Ayrıca yine çanta da olduğu gibi kirlenme yüzdesinin ve değişken sayısının artması, performansı olumsuz yönde etkilemektedir.

Çanta çizitinde uçdeğerleri içermesi beklenen bölge şüphesiz çitin dışı olacaktır. Bu anlamda bu bölgenin performansı, diğer 2 bölgeye kıyasla en çok dikkate değer olandır. Tablo 9, 10 ve 11'de bu sonuçlar yer almaktadır. Doğaldır ki bu bölgedeki başarı, gerek gizleme gerekse yanlış alarm yüzdelerinin düşüklüğü ile orantılı olacaktır. Sonuçlardan bu bölgenin *y* ve *xy*-uzayındaki sapmalarda hemen her senaryoda oldukça başarılı olduğunu, *x*-uzayındaki sapmalarda ise özellikle düşük kirlenme yüzdelerinde ve küçük örneklemlerde başarılı olduğunu göstermektedir. Bu yöndeki kirlenmede, kirlenme yüzdesinin bağımsız değişken sayısının ve gözlem sayısının arttırılması başarıyı belirli ölçülerde olumsuz yönde etkilemektedir.

Tablo 3. *x*-uzayında sapmada çantanın performansı

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,492**	**1,000**	**0,501**	**1,000**	**0,511**
2	**5**	**2**	**1,000**	**0,490**	**1,000**	**0,497**	**1,000**	**0,511**
2	**10**	**1**	**1,000**	**0,506**	**1,000**	**0,522**	**1,000**	**0,539**
2	**10**	**2**	**1,000**	**0,517**	**0,973**	**0,526**	**0,980**	**0,538**
3	**5**	**1**	**1,000**	**0,478**	**1,000**	**0,496**	**1,000**	**0,508**
3	**5**	**2**	**0,990**	**0,487**	**1,000**	**0,505**	**0,996**	**0,508**
3	**10**	**1**	**0,960**	**0,508**	**0,950**	**0,521**	**0,926**	**0,532**
3	**10**	**2**	**0,870**	**0,504**	**0,857**	**0,516**	**0,822**	**0,517**
5	**5**	**1**	**0,980**	**0,485**	**0,973**	**0,501**	**0,960**	**0,509**
5	**5**	**2**	**0,950**	**0,482**	**0,960**	**0,498**	**0,900**	**0,504**
5	**10**	**1**	**0,800**	**0,493**	**0,817**	**0,514**	**0,718**	**0,505**
5	**10**	**2**	**0,670**	**0,489**	**0,640**	**0,489**	**0,674**	**0,505**

Tablo 4. *y*-uzayında sapmada çantanın performansı

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,485**	**1,000**	**0,498**	**1,000**	**0,515**
2	**5**	**2**	**1,000**	**0,489**	**1,000**	**0,506**	**1,000**	**0,509**
2	**10**	**1**	**1,000**	**0,515**	**1,000**	**0,524**	**1,000**	**0,542**
2	**10**	**2**	**1,000**	**0,510**	**1,000**	**0,534**	**1,000**	**0,531**
3	**5**	**1**	**1,000**	**0,485**	**1,000**	**0,499**	**1,000**	**0,510**
3	**5**	**2**	**1,000**	**0,488**	**1,000**	**0,503**	**1,000**	**0,510**
3	**10**	**1**	**1,000**	**0,512**	**1,000**	**0,530**	**1,000**	**0,536**
3	**10**	**2**	**1,000**	**0,508**	**1,000**	**0,534**	**1,000**	**0,536**
5	**5**	**1**	**1,000**	**0,483**	**1,000**	**0,505**	**1,000**	**0,512**
5	**5**	**2**	**1,000**	**0,484**	**1,000**	**0,499**	**1,000**	**0,509**
5	**5**	**1**	**1,000**	**0,524**	**1,000**	**0,532**	**1,000**	**0,538**
5	**10**	**2**	**1,000**	**0,504**	**1,000**	**0,533**	**1,000**	**0,535**

Tablo 5. *xy*-uzayında sapmada çantanın performansı

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,479**	**1,000**	**0,491**	**1,000**	**0,509**
2	**5**	**2**	**1,000**	**0,482**	**1,000**	**0,501**	**1,000**	**0,511**
2	**10**	**1**	**1,000**	**0,506**	**1,000**	**0,527**	**1,000**	**0,539**
2	**10**	**2**	**1,000**	**0,516**	**1,000**	**0,521**	**1,000**	**0,540**
3	**5**	**1**	**1,000**	**0,485**	**1,000**	**0,491**	**1,000**	**0,514**
3	**5**	**2**	**1,000**	**0,485**	**1,000**	**0,507**	**1,000**	**0,513**
3	**10**	**1**	**1,000**	**0,519**	**1,000**	**0,524**	**1,000**	**0,536**
3	**10**	**2**	**1,000**	**0,515**	**1,000**	**0,520**	**1,000**	**0,540**
5	**5**	**1**	**1,000**	**0,486**	**1,000**	**0,497**	**1,000**	**0,509**
5	**5**	**2**	**1,000**	**0,484**	**1,000**	**0,503**	**1,000**	**0,512**
5	**10**	**1**	**1,000**	**0,511**	**1,000**	**0,522**	**1,000**	**0,535**
5	**10**	**2**	**1,000**	**0,515**	**1,000**	**0,529**	**1,000**	**0,539**

Tablo 6. *x*-uzayında sapmada ilmeğin performansı

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,497**	**1,000**	**0,488**	**1,000**	**0,485**
2	**5**	**2**	**1,000**	**0,495**	**1,000**	**0,493**	**0,996**	**0,487**
2	**10**	**1**	**0,940**	**0,487**	**0,970**	**0,468**	**0,916**	**0,455**
2	**10**	**2**	**0,790**	**0,473**	**0,800**	**0,466**	**0,752**	**0,456**
3	**5**	**1**	**0,920**	**0,512**	**1,000**	**0,500**	**0,996**	**0,489**
3	**5**	**2**	**1,000**	**0,501**	**0,947**	**0,489**	**0,916**	**0,488**
3	**10**	**1**	**0,580**	**0,477**	**0,673**	**0,473**	**0,564**	**0,465**
3	**10**	**2**	**0,630**	**0,484**	**0,467**	**0,479**	**0,560**	**0,480**
5	**5**	**1**	**0,820**	**0,503**	**0,813**	**0,490**	**0,796**	**0,487**
5	**5**	**2**	**0,560**	**0,511**	**0,653**	**0,493**	**0,548**	**0,492**
5	**10**	**1**	**0,300**	**0,497**	**0,393**	**0,478**	**0,362**	**0,492**
5	**10**	**2**	**0,425**	**0,502**	**0,437**	**0,504**	**0,466**	**0,489**

Tablo 7. *y*-uzayında sapmada ilmeğin performansı

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,498**	**1,000**	**0,495**	**1,000**	**0,481**
2	**5**	**2**	**1,000**	**0,496**	**1,000**	**0,489**	**1,000**	**0,486**
2	**10**	**1**	**0,995**	**0,477**	**0,997**	**0,467**	**1,000**	**0,452**
2	**10**	**2**	**0,990**	**0,477**	**0,990**	**0,459**	**0,998**	**0,446**
3	**5**	**1**	**1,000**	**0,501**	**1,000**	**0,491**	**1,000**	**0,486**
3	**5**	**2**	**1,000**	**0,497**	**1,000**	**0,488**	**1,000**	**0,486**
3	**10**	**1**	**0,995**	**0,479**	**0,997**	**0,460**	**0,998**	**0,460**
3	**10**	**2**	**0,970**	**0,483**	**0,983**	**0,460**	**0,990**	**0,458**
5	**5**	**1**	**0,990**	**0,508**	**1,000**	**0,487**	**1,000**	**0,484**
5	**5**	**2**	**1,000**	**0,501**	**1,000**	**0,494**	**1,000**	**0,486**
5	**10**	**1**	**0,940**	**0,462**	**0,980**	**0,461**	**0,984**	**0,455**
5	**10**	**2**	**0,955**	**0,486**	**0,977**	**0,459**	**0,990**	**0,460**

Tablo 8. ***xy*-uzayında sapmada ilmeğin performansı**

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**1,000**	**0,515**	**1,000**	**0,482**	**1,000**	**0,487**
2	**5**	**2**	**1,000**	**0,509**	**1,000**	**0,491**	**1,000**	**0,484**
2	**10**	**1**	**1,000**	**0,483**	**1,000**	**0,463**	**0,998**	**0,457**
2	**10**	**2**	**0,980**	**0,475**	**0,997**	**0,470**	**1,000**	**0,454**
3	**5**	**1**	**1,000**	**0,503**	**1,000**	**0,499**	**1,000**	**0,483**
3	**5**	**2**	**1,000**	**0,503**	**1,000**	**0,484**	**1,000**	**0,482**
3	**10**	**1**	**0,990**	**0,472**	**1,000**	**0,463**	**1,000**	**0,459**
3	**10**	**2**	**0,975**	**0,473**	**0,997**	**0,473**	**0,998**	**0,454**
5	**5**	**1**	**0,990**	**0,499**	**1,000**	**0,494**	**1,000**	**0,486**
5	**5**	**2**	**0,990**	**0,505**	**0,993**	**0,492**	**1,000**	**0,484**
5	**10**	**1**	**0,980**	**0,474**	**0,990**	**0,471**	**1,000**	**0,460**
5	**10**	**2**	**0,955**	**0,478**	**0,990**	**0,462**	**0,998**	**0,454**

Tablo 9. ***x*-uzayında sapmada çit dışının performansı**

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi.	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi.	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi.
2	**5**	**1**	**0,000**	**0,011**	**0,000**	**0,010**	**0,000**	**0,005**
2	**5**	**2**	**0,000**	**0,015**	**0,000**	**0,010**	**0,004**	**0,002**
2	**10**	**1**	**0,060**	**0,019**	**0,030**	**0,011**	**0,084**	**0,006**
2	**10**	**2**	**0,210**	**0,017**	**0,227**	**0,010**	**0,268**	**0,006**
3	**5**	**1**	**0,080**	**0,011**	**0,000**	**0,005**	**0,004**	**0,003**
3	**5**	**2**	**0,010**	**0,013**	**0,053**	**0,006**	**0,088**	**0,004**
3	**10**	**1**	**0,460**	**0,015**	**0,377**	**0,006**	**0,510**	**0,003**
3	**10**	**2**	**0,500**	**0,013**	**0,677**	**0,005**	**0,618**	**0,003**
5	**5**	**1**	**0,200**	**0,013**	**0,213**	**0,008**	**0,244**	**0,004**
5	**5**	**2**	**0,490**	**0,008**	**0,387**	**0,009**	**0,552**	**0,004**
5	**10**	**1**	**0,900**	**0,009**	**0,790**	**0,008**	**0,920**	**0,004**
5	**10**	**2**	**0,905**	**0,009**	**0,923**	**0,007**	**0,860**	**0,006**

Tablo 10. ***y*-uzayında sapmada çit dışının performansı**

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**0,000**	**0,017**	**0,000**	**0,007**	**0,000**	**0,004**
2	**5**	**2**	**0,000**	**0,015**	**0,000**	**0,006**	**0,000**	**0,005**
2	**10**	**1**	**0,005**	**0,023**	**0,003**	**0,012**	**0,000**	**0,006**
2	**10**	**2**	**0,010**	**0,023**	**0,010**	**0,016**	**0,022**	**0,004**
3	**5**	**1**	**0,000**	**0,014**	**0,000**	**0,009**	**0,000**	**0,004**
3	**5**	**2**	**0,000**	**0,015**	**0,000**	**0,009**	**0,000**	**0,005**
3	**10**	**1**	**0,005**	**0,012**	**0,003**	**0,012**	**0,002**	**0,004**
3	**10**	**2**	**0,030**	**0,032**	**0,017**	**0,006**	**0,010**	**0,006**
5	**5**	**1**	**0,010**	**0,009**	**0,000**	**0,008**	**0,000**	**0,004**
5	**5**	**2**	**0,000**	**0,015**	**0,000**	**0,007**	**0,000**	**0,005**
5	**10**	**1**	**0,060**	**0,024**	**0,020**	**0,007**	**0,016**	**0,007**
5	**10**	**2**	**0,045**	**0,018**	**0,023**	**0,011**	**0,010**	**0,005**

Tablo 11. ***xy*-uzayında sapmada çit dışının performansı**

			n = 40		n = 60		n = 100	
Değişken Sayısı	Uçdeğer Yüzdesi	Sapma Miktarı	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi	Gizleme Yüzdesi	Yanlış Alarm Yüzdesi
2	**5**	**1**	**0,000**	**0,006**	**0,020**	**0,007**	**0,000**	**0,005**
2	**5**	**2**	**0,000**	**0,009**	**0,000**	**0,008**	**0,000**	**0,004**
2	**10**	**1**	**0,000**	**0,014**	**0,000**	**0,010**	**0,002**	**0,004**
2	**10**	**2**	**0,020**	**0,012**	**0,003**	**0,010**	**0,000**	**0,006**
3	**5**	**1**	**0,000**	**0,012**	**0,000**	**0,010**	**0,000**	**0,003**
3	**5**	**2**	**0,000**	**0,012**	**0,000**	**0,009**	**0,000**	**0,005**
3	**10**	**1**	**0,010**	**0,016**	**0,000**	**0,014**	**0,000**	**0,005**
3	**10**	**2**	**0,025**	**0,020**	**0,003**	**0,008**	**0,002**	**0,006**
5	**5**	**1**	**0,010**	**0,015**	**0,000**	**0,008**	**0,000**	**0,005**
5	**5**	**2**	**0,010**	**0,012**	**0,007**	**0,006**	**0,000**	**0,004**
5	**10**	**1**	**0,020**	**0,018**	**0,010**	**0,014**	**0,000**	**0,005**
5	**10**	**2**	**0,045**	**0,017**	**0,010**	**0,013**	**0,002**	**0,008**

4. TARTIŞMA VE SONUÇ

Regresyon analizi varsayımlarının ihlal edilmemesi ve yapılacak çıkarsamaların güvenilirliği açısından verilerdeki uçdeğerlerin saptanması çok önemlidir. Bu sorunu çözmek amacıyla pek çok başarılı algoritma ve yöntem geliştirilmiştir. Ancak bu yöntemlerden hiç biri, her koşul altında % 100 başarılı değildir. Bazı algoritmalar düşük kirlenme yüzdelerinde başarılıyken, bazılarında durum tersidir. *x*-uzayındaki sapmalarda başarılı olan bir algoritma *y*-uzayındaki sapmalarda başarısız olabilmektedir. Gizleme oranlarında oldukça başarılı olan bir algoritma, çok yüksek yanlış alarm oranı nedeniyle tercih edilmemekte ya da yine başarılı bir algoritmanın kullanımı çok uzun hesaplama süreleri nedeniyle sınırlı kalabilmektedir. SPSS, S-Plus ya da EViews gibi yaygın istatistik yazılımlarında ise, yeterli sayıda uç değer teşhis yöntemine yer verildiğini söylemek zordur.

Bu çalışmada incelenen, TK uzayında çanta çiziti yöntemi de, regresyon uçdeğerlerinin teşhisinde her koşul altında % 100 başarılı değildir. Ancak literatürde sıkça başvurulan bazı klasik veriler ve yapılan Monte Carlo simülasyonlarına dayanarak yöntemin, görsellik avantajı da düşünülürse, en azından diğer algoritmaların sonuçlarını kontrol etmek açısından tamamlayıcı olarak kullanılabileceği söylenebilir.

5. KAYNAKLAR

Belshey, D.A., Kuh, E., Welsh, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity.Wiley, New York.

Brownlee, K.A., 1965. Statistical Theory and Methodology. Wiley, 2nd Edition, NewYork.

Cook, R.D. ve Weisberg, S., 1980. Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. Techonometrics, 22,495-508.

Hadi, A.S. ve Simonoff, J.S., 1993. Procedures for the Identification of Multiple Outliers in Linear models. J.Amer.Statist.Assoc., 88, 1264-1272.

Hawkins, D.M, Bradu, D. ve Kass, G.V., 1984. Location of Several Outliers in Multiple Regression Data Using Elemental Sets. Technometrics, 26, 197-208.

Kianifard, F. ve Swallow, W., 1990. A Monte Carlo Comparison of Some Procedures for Identifying Outliers in Linear Regression. Commun. Statist, Part A Theory Methods 19, 1913-1938.

Rousseeuw, P.J. ve Leroy, A.M., 1987. Robust Regression and Outlier Detection, Wiley, Newyork.

Rousseeuw, P.J. ve Ruts, I., 1999. BAGPLOT, Çevrimiçi http://www.agoras.ua.ac.be/.

Rousseeuw, P.J. ve Ruts, I. ve Tukey, J.W., 1999. The Bagplot:A Bivarite Boxplot, the American Statistician, Vol.53, No.4, 382-387.

Sebert, D.M., Montgomery, D.C. ve Rollier, D., 1998. A Clustering Algorithm for Identifying Multiple Outliers in Linear Regression. Computational Statistics&Data Analysis, 27, 461- 484.

Wisnowski, J.W., Montgomery, D.C. ve Simpson, J.R., 2001. A Comparative Analysis of Multiple Outlier Detection Procedures in the Linear Regression Model. Computational Statistics&Data Analysis, 36, 351-382.

THE BAGPLOT AS A DIAGNOSTIC TOOL FOR MULTIPLE REGRESSION OUTLIERS

ABSTRACT

The Bagplot is a bivariate generalization of the univariate boxplot which is also used in determining the outliers. Hence it can be used in diagnosing the outliers for the simple linear regression. It cannot be used, however, when the number of variables exceeds one. On the other hand, in statistics literature, it has been shown that the predicted value versus residual plot can represent the whole data in some instances. The main motivation of this paper is that point. The performance of the Bagplot in the predicted value versus residual plot is investigated for some classical and simulated data sets. The approach is found successful for many scenarios.

Key Words: Bagplot, Masking, Outlier, Swamping.