



MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



LİNEER REGRESYON MODELLERİNDE ETKİLİ GÖZLEMLERİN TESPİTİ İÇİN YENİ BİR YÖNTEM

GÖKÇE DELİORMAN

YÜKSEK LİSANS TEZİ
Matematik Anabilim Dalı
Matematik Programı

DANIŞMAN
Doç. Dr. Deniz İNAN

İSTANBUL, 2020

TEŞEKKÜR / ÖNSÖZ

Çalışmalarım boyunca beni her zaman destekleyen, emeğini ve katkısını daima hissettiğim saygıdeğer danışman hocam Sayın Doç. Dr. Deniz İNAN'a, beni her zaman yüreklendiren sevgili arkadaşım Cevahir Yıldırım'a ve bu süreçte hep yanımda olan canım aileme teşekkürlerimi sunarım.

Gökçe Deliorman

Ekim 2020



İÇİNDEKİLER

TEŞEKKÜR / ÖNSÖZ	i
ÖZET	iv
ABSTRACT	v
SEMBOLLER	vi
KISALTMALAR.....	viii
ŞEKİL LİSTESİ.....	ix
TABLO LİSTESİ.....	x
1 GİRİŞ.....	1
2 TEMEL KAVRAMLAR.....	5
3 ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE KULLANILAN KLASİK YÖNTEMLER	12
3.1 Cook Uzaklığı.....	12
3.2 DFFITS	12
3.3 Olabilirlik Değişim İstatistiği (Likelihood Displacement) (OD).....	13
4 PARÇACIK SÜRÜ OPTİMİZASYONU	14
5 KESİKLİ PARÇACIK SÜRÜ OPTİMİZASYONU	17
6 ETKİLİ ALT KÜMELERİN TESPİTİNDE KESİKLİ PARÇACIK SÜRÜ OPTİMİZASYONU KULLANILMASI	18
6.1 Parçacığın Yapısı.....	20
6.2 Amaç Fonksiyonu	20
7 SİMÜLASYON	22
7.1 Simülasyon 1.....	23
7.2 Simülasyon 2.....	23
7.3 Simülasyon 3.....	23
8 GERÇEK VERİ SETİ UYGULAMASI.....	26
8.1 Hills Yarışları Veri Seti	26
8.2 CYG OB1 Yıldız Kümesi Verisi	30
9 SONUÇ.....	35
10 KAYNAKLAR	36
EK1: Hills Yarışları Verisi	39
EK 2: CYG OB1 Yıldız Kümesi Verisi	40



ÖZET

LİNEER REGRESYON MODELLERİNDE ETKİLİ GÖZLEMLERİN TESPİTİ İÇİN YENİ BİR YÖNTEM

Etkili gözlemler, veri kümesinde yer alan diğer gözlemlerden farklılık gösteren, regresyon katsayılarında büyük değişikliğe sebep olan gözlemlerdir. Veri setinde bu tür gözlemlerin varlığı istatistiksel analizlerin geçerliliğini ve duyarlılığını azaltmaktadır. Literatürde etkili gözlemleri belirlemek için kullanılan birçok yöntem vardır. Ancak bu yöntemlerin çoğu dağılımsal varsayımlar gerektirir ve maskeleye ve süpürme etkilerinden oldukça etkilenirler. Özellikle veri setinde birden fazla etkili gözlem olması durumunda bu yöntemlerin çoğu yetersiz kalmaktadır. Bu tezin amacı, meta-sezgisel Kesikli Parçacık Sürü Optimizasyon algoritmasını kullanarak etkili gözlem kümelerini tespit etmek için yeni bir teşhis yöntemi geliştirmektir. Önerilen bu yaklaşım, herhangi bir dağılımsal varsayım gerektirmemekte ve ayrıca klasik yöntemler gibi maskeleye ve süpürme etkilerinden etkilenmemektedir. Önerilen yöntemin performansı çeşitli simülasyonlar ve gerçek veri setleri ile analiz edilmiştir.

Anahtar kelimeler: Etkili Alt Kümeler, Kesikli Parçacık Sürü Optimizasyonu, Meta-sezgisel Algoritmalar, Lineer Regresyon

ABSTRACT

A NEW METHOD FOR DETECTION OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION MODELS

Influential observations are observations that differ from other observations in the data set and cause a large change (effect) in the regression coefficients. The presence of such observations in the data set reduces the validity and sensitivity of statistical analyzes. There are many methods used to determine influential observations in the literature. However, most of these methods require distributional assumptions and are highly affected by masking and swamping effects. Most of these methods are inadequate, especially if there is more than one influential observation in the data set. The aim of this thesis is to develop a new diagnostic method to detect influential observation sets using the meta-heuristic Binary Particle Swarm Optimization algorithm. This proposed approach does not require any distributional assumptions and is also not affected by the masking and swamping effects such as the classical methods. The performance of the proposed method has been analyzed via various simulations and real data set applications.

Keywords: Influential Subsets, Binary Particle Swarm Optimization, Meta-heuristic Algorithms, Linear Regression

SEMBOLLER

AKO	Artık kareler ortalaması
AKT	Artık kareler toplamı
c_1	Bilişsel bileşen
c_2	Sosyal bileşen
d_i	Standartlaştırılmış artıklar
D_i	Cook uzaklığı
h_{ii}	Şapka matrisinin köşegen elemanları
H	Şapka matrisi
I	$n \times n$ 'lik birim matris
k	Bağımsız değişken sayısı
K	Sürüdeki parçacık sayısı
$L(\beta, \sigma^2)$	Lineer regresyon modelinin log-likelihood fonksiyonu
m	Potansiyel etkili gözlemlerin sayısı
n	Gözlem sayısı
p	Parametre sayısı
P	Parçacığın konum matrisi
r_i	İçsel student artıklar
r_i^*	Harici (dışsal) student artıklar
RKT	Regresyon kareler toplamı
R_{adj}^2	Düzeltilmiş (adjusted) belirleyicilik katsayısı
R^2	Belirleyicilik katsayısı
s	Veri setinde kirletilen gözlem sayısı

t	İterasyon sayısı
TKT	Toplam kareler toplamı
V	Parçacığın hız matrisi
var	Varyans
v_{max}	Parçacığın maksimum hızı
Y	Bağımlı değişken vektörü
X	Bağımsız değişkenler matrisi
w	Eylemsizlik ağırlığı
β	Bilinmeyen katsayılar vektörü
$\hat{\beta}$	β 'nın en küçük kareler tahmin edicisi
β_j	Regresyon katsayıları
ε	Rastgele hatalar vektörü
$\hat{\sigma}^2$	Artık varyansı

KISALTMALAR

BPSO	Kesikli Parçacık Sürü Optimizasyonu
EKK	En Küçük Kareler
k.o.	Kirlenme Oranı (contamination rate)
MO	Maskeleme Oranı (Masking rate)
OD	Olabilirlik Değişimi (Likelihood Displacement)
PSO	Parçacık Sürü Optimizasyonu
SO	Süpürme Oranı (Swamping rate)

ŞEKİL LİSTESİ

Şekil 2-1: Basit lineer regresyon modelinde kaldıraç noktası örneği.....	8
Şekil 2-2: Basit lineer regresyon modelinde etkili gözlemin regresyon doğrusuna etkisi	10
Şekil 2-3: Basit lineer regresyon modelinde maskeleye ve süpürme etkisi	11
Şekil 6-1 BPSO'nun akış diyagramı.....	19
Şekil 6-2: Bir parçacıkta bulunan potansiyel etkili gözlemler	20
Şekil 8-1: Gözlemlerin Cook uzaklığı grafiği	28
Şekil 8-2: Gözlemlerin DFFITS grafiği	28
Şekil 8-3: Gözlemlerin OD grafiği	29
Şekil 8-4: CYG OB1 verisi saçılım grafiği	32
Şekil 8-5: CYG OB1 verisiyle elde edilen regresyon doğruları	34

TABLO LİSTESİ

Tablo 7-1: Simülasyon 1 sonuçları	23
Tablo 7-2: Simülasyon 2 sonuçları	24
Tablo 7-3: Simülasyon 3 sonuçları	24
Tablo 7-4: Veride tek etkili gözlem olması durumunda teşhis yöntemlerinin SO ve MO sonuçları	24
Tablo 8-1: Hills Yarışları verisi için geleneksel yöntemlerin kesim değerleri.....	26
Tablo 8-2: Hill Yarışları veri setinde etkili gözlemlerin tespitinde kullanılan tüm yöntemlerin sonuçları	27
Tablo 8-3: Hills Yarışları verisi için kurulan üç farklı regresyon modeli	30
Tablo 8-4: CYG OB1 verisi için geleneksel yöntemlerin kesim değerleri.....	31
Tablo 8-5: Önerilen yöntemin etkili olarak işaretlediği gözlemler ve etki dereceleri	32
Tablo 8-6: CYG OB1 verisinde etkili gözlemlerin tespitinde kullanılan tüm yöntemlerin sonuçları	33

1 GİRİŞ

Regresyon analizi deęiřkenler arasındaki iliřkiyi modellemek ve incelemek ii kullanılan istatistiksel bir tekniktir (Montgomery, Peck, & Vinning, 2012). Mühendislik, fizik, kimya ve biyoloji bilimleri, sosyal bilimler gibi birçok farklı uygulama alanlarında kullanılır.

Regresyon analizinin önemli amaçlarından birisi regresyon modelindeki bilinmeyen parametreleri kestirmektir. Bu süreç modelin verilere uydurulması olarak da adlandırılabilir. Oluřturulacak regresyon modelinde iyi parametre kestirimleri iin bazı varsayımların saęlanması gerekir. Varsayımların geçerlilięini ve model yeterlilięini kontrol etme süreci regresyon analizinde önemli bir adımdır. Modeldeki bilinmeyen β parametrelerini kestirmek iin genellikle En Küçük Kareler (EKK) yöntemi kullanılır. Ancak EKK yönteminin veri kümesi ierisinde yer alan bazı gözlemlerden oldukça etkilendięi ve bu gözlemlerin etkisini yansıttıęı bilinmektedir. Tek başına veya dięer birkaç gözlemle birlikte, çeřitli tahminlerin hesaplanan deęerleri üzerinde bariz řekilde büyük bir etkiye sahip olan gözlemlere etkili gözlem denir (Belsley, Kuh, & Welsch, 1980). Bu tür gözlemlerin veriden ıkarılması regresyon katsayılarını ve kestirim deęerlerini önemli derecede deęiřtirir.

Regresyon analizinde bir noktanın etkisini deęerlendirmek iin genel prosedür, o gözlem göz ardı edildięinde meydana gelen deęiřiklikleri belirlemektir. Bir veri kümesinde bu tür gözlemlerin bulunmasının nedenleri yanlış ölçümler, veri giriřindeki hatalar, ölçüm cihazındaki sorunlar veya nadir olaylar olabilir. Bir gözlemin etkili olarak iřaretlenmesi, o gözlemin hatalı olduęu anlamına gelmemektedir. Etkili olarak tespit edilen bir nokta son derece önemli bir gözlem olup daha önce bilinmeyen etkenlere iliřkin yeni bilgiler elde etmeye yarayabilir (Rawlings, Pantula, & Dickey, 2001). Örneęin regresyon modelinde kullanılan verilerin hasta bilgileri olduęunu düşünürsek, etkili olarak tespit ettięimiz bir gözlem belki de o hastayla ilgili önemli biyolojik bilgilere ulařmamızı saęlayacak ve hastanın tedavi kararında etkili olacaktır. Aynı řekilde bir mühendisinin makine arızasını etkileyen deęiřkenlerle regresyon modeli kurduęunu düşünelim. Bu modelde etkili olarak tespit edilen gözlemler makinenin bozulmadan bakım yapılması kararında yardımcı olup,

makine ömrünün uzamasını sağlayabilir. Böylece ileriye dönük daha verimli sonuçlar elde edilebilir.

Etkili gözlemlerin tespiti ile ilgili literatürde birçok çalışma önerilmiştir. Yapılan literatür taraması sonucunda Cook uzaklığı, DFFITS ve Olabilirlik Değişim istatistiği (Likelihood Displacement (OD)) gibi tek gözlem silmeye dayalı yöntemler çok yaygın olarak kullanılmaktadır. Ancak bu yöntemlerin büyük sorunlarından biri maskeleye ve süpürme etkilerinden oldukça etkilenmeleridir. Ayrıca bu teşhis yöntemleri birçok dağılımsal varsayım da gerektirir. Bu yöntemlerin genellikle bir karar kuralı vardır. Veri kümesinde bulunan bir gözlemin etkili gözlem olarak belirlenmesi için bu gözlemin kesim değerinden büyük olması koşulu söz konusu olabilir ve bu kesim noktaları genellikle örneklem hacmine bağlıdır. Tüm bu durumlar etkili gözlem alt kümelerini yanlış tespit etmemize yol açabilir.

Etkili gözlemleri incelemek için ilk çalışmalar Cook (1977) tarafından yapılmıştır. Cook uzaklığı veriden tek gözlemin silinmesiyle regresyon katsayılarının EKK tahmin edicisindeki değişimin dağılımına dayanır (Cook R. D., 1977). Aslında bu değişiklik herhangi bir dağılımsal özelliğe sahip değildir. Bu nedenle etkili gözlemlerin tespitinde Cook uzaklığını kullanmak yanlış sonuçlara neden olabilir (Kim, 2017).

Etkili gözlemleri tespit edebilmek için kullanılan bir diğer yöntem DFFITS, Belsley, Kuh ve Welsch (1980) tarafından bulunmuştur. Bu yöntem i . gözlem veriden kaldırıldığında $\hat{y}$ üzerindeki değişimin ölçüsüdür. Bu teşhis aracı için önerilen kesim noktası örneklem büyüklüğüne ve açıklayıcı değişken sayısına bağlıdır, tüm durumlar için uygun değildir. Bu yüzden tüm durumlar için doğru bir kesim noktası belirlemek oldukça zordur, önerilen kesim noktaları sadece birer yol göstericidir (Montgomery, Peck, & Vinning, 2012).

OD istatistiği Cook ve Weisberg (1982) tarafından önerilen bir diğer teşhis aracıdır. χ^2 dağıldığı varsayılır. Ancak bu varsayım büyük örneklerde doğru değildir. Aksine örneklem büyüklüğü arttıkça OD değerleri 0'a yakınsar (Wang, Chow, & Wei, 1995).

Son yıllarda, optimizasyon problemlerini çözmek için geleneksel ve yaklaşık algoritmalar olmak üzere birçok optimizasyon algoritmaları önerilmektedir. Geleneksel algoritmalar birçok problemde iyi bir performansla sahiptir, ancak daha büyük boyutlu kombinatoriyal ve doğrusal olmayan optimizasyon problemlerinin çözümlerinde verimli değildirler. Ayrıca, bu algoritmalar genellikle birçok durumda doğrulanması kolay olmayabilecek birkaç varsayım gerektirir (Neapolitan, 2004). Bu tür sınırlamaların üstesinden gelmek için daha uyarlanabilir ve esnek algoritmalara ihtiyaç vardır. Herhangi bir problem hakkında çok az veya hiç varsayımda bulunmayan ve geniş çözüm alanları araştırabilen meta-sezgisel algoritmalar, günümüzde optimizasyon problemlerini çözmek için geliştirilmiştir. Parçacık Sürü Optimizasyonu (PSO), yapay arı kolonisi algoritmaları, karınca kolonisi optimizasyonu, tabu arama ve genetik algoritmalar gibi doğadan esinlenerek geliştirilmiş çeşitli algoritmalar optimizasyon problemlerini çözmek için sıklıkla kullanılmaktadırlar. (Beheshti & Shamsuddin, 2013).

PSO, Eberhart ve Kennedy (1995) tarafından önerilmiş kuş sürülerinin yiyecek arama davranışlarından ilham alan yaygın olarak kullanılan algoritmalarından biridir. PSO, tıp bilimleri, iletişim ağları, bulanık mantık sistemleri, ekonomi, işletme, finans, mühendislik gibi birçok farklı bilimlerde geniş uygulama alanlarına sahiptir. PSO'nun bu kadar çok kullanılmasının sebebi algoritmanın basitliği, uygulama kolaylığı, farklı uygulama alanlarına adapte edilebilmesidir (Poli, 2008). PSO'nun diğer önemli özelliklerinden biri de küresel bir optimum çözümü belirlemek için, çözüm alanının farklı bölgelerindeki farklı noktaları aynı anda araştırmasıdır. Ayrıca düşük parametre gereksinimi, başlangıç noktalarına daha az bağımlılık, verimli küresel arama yaklaşımı gibi avantajlarının olması diğer meta-sezgisel algoritmaların önüne geçmesini sağlamıştır (Beheshti & Shamsuddin, 2013). PSO'nun ikili uzaydaki optimizasyon problemlerini çözmek amacıyla Kennedy ve Eberhart (1997) Kesikli Parçacık Sürü Optimizasyonunu (BPSO) geliştirmiştir. Özellik seçimi (Cervante, Xue, Zhang, & Shang, 2012), zamanlama (Pedrasa, Spooner, & MacGill, 2009), yönlendirme (Ayob, ve diğerleri, 2010) ve lot boyutlandırma (Taşgetiren & Liang, 2003) gibi ikili uzayda tanımlanan birçok optimizasyon problemi için BPSO kullanılmış ve etkili performans göstermiştir.

Bu tezin amacı, herhangi bir dağılımsal varsayım gerektirmeyen, etkili gözlem alt kümelerinin varlığında bile iyi performans gösteren ve maskeleye ve süpürme

etkilerinden etkilenmeyen yeni bir tespit yöntemi geliştirmektir. BPSO'nun kombinatoriyal problemlerin çözümündeki başarısı dikkate alınarak, gözlemlerin etkili olup olmadığı bir kombinatoriyal optimizasyon problemi olarak düşünülmüştür. Bu amaçla BPSO bir teşhis aracı olarak kullanılmıştır. Önerilen yöntemin performansı, gerçek veri setleri üzerinde ve çeşitli simülasyon çalışmaları ile analiz edilmiştir.

Bu tezin ikinci bölümünde regresyon analizi ile ilgili temel kavramlar verilmiştir. Üçüncü bölümde etkili gözlemleri tespit etmekte kullanılan klasik tanı yöntemleri tanıtılmıştır. Dördüncü ve beşinci bölümlerde PSO ve BPSO algoritmaları açıklanmıştır. Altıncı bölümde etkili alt kümelerin tespitinde önerilen BPSO tabanlı yaklaşımın kullanılması anlatılmıştır. Yedinci bölümde üç farklı simülasyon çalışması, sekizinci bölümde ise iki tane gerçek veri setleri üzerinden uygulamalar yapılmıştır. Uygulamalar ve simülasyon çalışmaları kullanılarak, geleneksel yöntemler ile önerilen BPSO yaklaşımının performansı karşılaştırılmıştır. Son olarak dokuzuncu bölümde sonuç kısmına yer verilmiştir.

2 TEMEL KAVRAMLAR

Regresyon analizinin genel tanımı bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi matematiksel modellerle açıklamaktır (Alpar, 2013). Tek bir bağımsız değişken içeren regresyon modeli basit lineer regresyon, birden fazla bağımsız değişken içeren regresyon modeli ise çoklu lineer regresyon modeli olarak adlandırılır. Bağımlı değişkenin k sayıdaki bağımsız değişken ile açıklandığı çoklu lineer regresyon modeli

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2.1)$$

şeklindedir. β_j ($j = 0, 1, \dots, k$) parametreleri regresyon katsayıları olarak bilinir. β_j parametresi x_i ($i \neq j$) bağımsız değişkenleri sabit tutulduğunda x_j bağımsız değişkenindeki bir birimlik değişime karşılık gelen y yanıt değişkenindeki beklenen değişimi göstermektedir. Bu modelinin matris gösterimi ise $Y = X\beta + \varepsilon$ biçiminde yazılır. Burada n gözlem sayısı, k bağımsız değişken sayısı ve $k + 1 = p$ parametre sayısı olmak üzere;

Y : $n \times 1$ boyutlu bağımlı değişken vektörü,

X : $n \times p$ boyutlu tam ranklı bağımsız değişkenler matrisi,

β : $p \times 1$ boyutlu bilinmeyen katsayılar vektörü,

ε : $n \times 1$ boyutlu rastgele hatalar vektörüdür.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} \quad X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & X_{21} & \dots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}_{n \times p} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{p \times 1} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

β parametrelerini tahmin etmek için genellikle EKK yöntemi kullanılır. EKK tahminleri hata kareler toplamının minimizasyonu ile elde edilir. Aşağıda verilen $S(\beta)$ amaç fonksiyonunda modeldeki hata terimlerinin kareleri toplamı minimize edilerek β 'nın EKK tahmin edicisi $\hat{\beta} = (X'X)^{-1}X'Y$ vektörü elde edilir.

$$S(\beta) = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon \quad (2.2)$$

EKK yönteminin kullanılabilmesi için aşağıdaki varsayımların sağlanması gerekir.

- i) Hata terimleri sıfır ortalamalı, σ^2 sabit varyanslı normal dağılan rastgele değişkenlerdir.

$$\varepsilon_i \sim N(0, \sigma^2)$$

- ii) Hata terimleri ilişkisizdir.

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

- iii) Yanıt değişkeni ile X bağımsız değişkenler arasındaki ilişki doğrusaldır.
iv) X bağımsız değişkenleri rastgele değişken değildir.

Sabit varyans varsayımı ile hataların ilişkisiz olma varsayımlarının sağlanması durumunda EKK tahmin edicileri diğer tüm yansız tahmin ediciler arasında minimum varyans olma özelliğine sahiptir. Bu özellikler Gauss-Markov teoremi ile tanımlanmıştır. Verilen varsayımların sağlanmaması durumunda EKK yöntemiyle tahmin edilen parametreler güvenilir olmayabilir. Varsayımların kontrolü üç başlık altında incelenebilir. Bunlar hata terimine dayalı varsayımlar, model varsayımları ve olağandışı gözlemlere dayalı varsayımlardır (Faraway, 2014).

- 1) Hata terimi varsayımları: i ve ii. varsayımlar hataya ilişkin varsayımlardır. Normallik varsayımı parametrelerin anlamlılık testleri ve güven aralığı tahminlerinin oluşturulması için gereklidir. Sadece normallik varsayımı ihlal edilip diğer varsayımların karşılanması durumunda, EKK tahmin edicileri, tüm doğrusal yansız tahmin ediciler arasında en küçük varyansa sahip olmaya devam eder. Hataların ilişkili olması durumunda, EKK tahmin edicileri yansız olmaya devam eder, ancak artık en iyi tahmin ediciler değildir. Elde edilen regresyon katsayıları, modelden bir değişkenin çıkartılması ve eklenmesine karşın çok duyarlı hale gelir. Bu da tahminlerin önyargılı olmasına neden olur ve anlamlılık testlerini geçersiz kılar. Sabit varyans varsayımı EKK yönteminde önemli bir yere sahiptir. Sıradan EKK tahmin edicilerinin minimum varyans özelliği doğrudan bu varsayıma bağlıdır. Varsayım, bağımlı değişken üzerindeki her gözlemin eşit miktarda bilgi içerdiğini ima eder. Ancak heterojen varyanslar bazı gözlemlerin diğerlerinden daha fazla bilgi içerdiğini ifade etmektedir. Bu varsayımların sağlanmamasının nedeni olağandışı gözlemlerin varlığı olabilir. Bu nedenle sorunların giderilmesi

için bu tür gözlemlerin dikkatlice incelenmesi gereklidir. Hatalar gözlenemez olduğundan artıklar üzerinden varsayımları kontrol edilir.

2. Model varsayımları: $E(y) = X\beta$ model formunun doğru olduğu varsayılır.
3. Olağan dışı gözlemlere dayalı varsayımlar: Bazen bir veya birkaç gözlem model ile uyumlu olmayabilir. Bu tür gözlemler model seçimini ve parametre tahminlerini ciddi şekilde etkilemektedir.

Bu varsayımların geçerliliğine her zaman dikkat edilmeli ve ayrıca model yeterliliğinin belirlenmesi gereklidir.

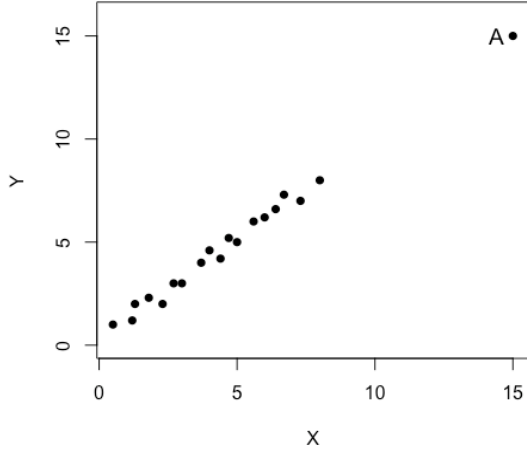
Model yeterliliğinin belirlenmesine ilişkin ölçülerden biri de artıkların incelenmesidir. Ham artıklar gözlenen y_i ile tahmini $\hat{y}_i$ değeri arasındaki farkı gösterir. Burada tahmini $\hat{Y}$ değeri

$$\hat{Y} = X\hat{\beta} \quad (2.3)$$

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

$$\hat{Y} = HY$$

şeklindedir. Burada $H = X(X'X)^{-1}X'$ $n \times n$ boyutlu şapka (hat) matrisidir. Şapka matrisi simetrik ($H' = H$) ve eş güçlüdür (idempotent) ($HH = H$). Şapka matrisinin köşegen elemanları (h_{ii}), x gözlem değerleri ile verinin merkezi arasındaki mesafenin bir ölçüsüdür. Büyük bir h_{ii} değeri, i -inci gözlemin merkezden uzak olduğunu gösterir. Şekil 2.1'de kaldırıcı noktasına örnek olarak bir saçılım grafiği verilmiştir. A noktası şapka matrisinin büyük köşegen elemanı olup kaldırıcı noktasıdır.



Şekil 2-1: Basit lineer regresyon modelinde kaldıraç noktası örneği

Bir gözlemin ham artık değeri $e_i = y_i - \hat{y}_i$ olarak tanımlanır. Artıklar regresyon modeli ile açıklanamayan değişikliği gösterir. Bu değişikliği ölçmede kullanılan ölçütlerden biride artık kareler ortalamasıdır (AKO). AKO aşağıdaki gibi hesaplanır.

$$AKO = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^n e_i^2}{n-p} = \frac{AKT}{n-p} \quad (2.4)$$

Burada AKT artık kareler toplamıdır.

Modelin genel anlamda yeterliliği ile ilgili olarak bir ölçü de R^2 ile gösterilen belirtme katsayısıdır.

$$R^2 = \frac{RKT}{TKT} = 1 - \frac{AKT}{TKT} \quad (2.5)$$

Burada RKT regresyon kareler toplamı, TKT ise toplam kareler toplamıdır. Belirtme katsayısı 0 ile 1 arasında değişir. 1'e yakın değerleri y 'deki değişkenliğin büyük kısmının regresyon modeliyle açıklandığını belirtmektedir. Düzeltilmiş R^2 (R_{adj}^2) ise

$$R_{adj}^2 = 1 - \frac{RKT/(n-p)}{TKT/(n-1)} \quad (2.6)$$

şeklinde tanımlanır. R_{adj}^2 sadece modele eklenen değişken artık kareler ortalamasını düşürdüğünde artar. Regresyon modellerinin değerlendirilmesi ve karşılaştırılması R_{adj}^2 ile kolaylıkla yapılabilir.

Artıklar model yeterliliğini incelenmesinde ve temel varsayımların saptanmasında önemli rol oynarlar. Artık incelenmesinde çeşitli artıklar türleri kullanılır. Genellikle artıkların ilk önce ortak bir varyansa sahip olması için standartlaştırılması önerilir. Standartlaştırılmış artıklar (d_i) her bir artık değerinin artık varyansının kareköküne bölünmesiyle elde edilir. Standartlaştırılmış artıklar sıfır ortalama ve birim varyansa sahiptir.

$$d_i = \frac{e_i}{\sqrt{AKO}} \quad (2.7)$$

Standart artıklarda pay ve payda bağımlı olduğundan her bir artık kendi standart sapmasına oranlanarak student türü artık (r_i) elde edilir.

$$r_i = \frac{e_i}{\sqrt{AKO(1-h_{ii})}} \quad (2.8)$$

Burada i . artığın varyansı $var(e_i) = AKO(1 - h_{ii})$ şeklindedir. Bu durum artığın içsel ölçeklendirmesi gibi düşünülebilir (Montgomery, Peck, & Vinning, 2012).

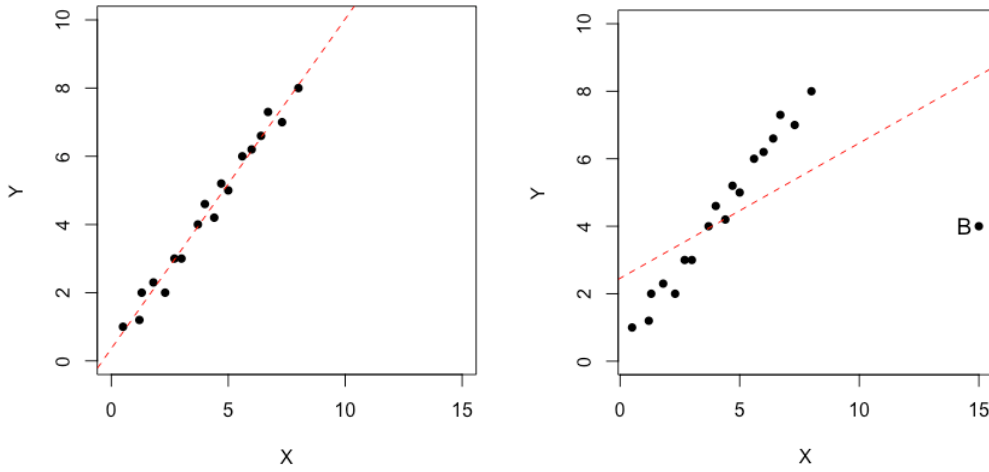
Harici (dışsal) student türü artıklar (r_i^*) için e_i 'den bağımsız bir σ^2 tahmincisi gereklidir. Hataların normal dağıldığı varsayımın altında $s_{(i)}$ hesaplanırken i . gözleme ilişkin artığın dahil edilmemesi $s_{(i)}$ ile e_i 'nin bağımsız olmasını sağlar. Bu artıklara jackknife artıklar da denir (Atkinson A. C., 1983).

$$r_i^* = \frac{e_i}{\sqrt{s_{(i)}^2(1-h_{ii})}} \quad (2.9)$$

$s_{(i)}$ burada $var(e)$ 'nin i . gözlemin çıkarılmasıyla elde edilen tahminidir ve aşağıdaki gibidir.

$$s_{(i)}^2 = \frac{(n-p)AKO - e_i^2/(1-h_{ii})}{n-p-1} \quad (2.10)$$

Büyük artık değerine sahip gözlemlere aykırı değer denir (Alpar, 2013). Başka bir tanıma göre aykırı değer, veri kümesinde yer alan diğer gözlemlerden önemli ölçüde farklılık gösteren aşırı bir gözlemdir (Montgomery, Peck, & Vinning, 2012). Model tarafından iyi tahmin edilemezler. Etkili gözlemler ise model katsayıları üzerinde önemli değişikliklere neden olurlar ve regresyon modelini kendilerine doğru çekerler (Montgomery, Peck, & Vinning, 2012). Bir gözlemin hem aykırı değer hem de etkili gözlem olması mümkündür. Şekil 2.2'deki ilk grafikte içerisinde etkili gözlem olmayan bir veri seti ile elde edilen regresyon doğrusu verilmiştir. İlk grafiğe etkili bir B noktası eklenip yeni bir regresyon doğrusu çizilmiştir. Etkili gözlemin etkisi, yani B noktasının regresyon doğrusunu kendine doğru çektiği ikinci grafikten açıkça görülmektedir.

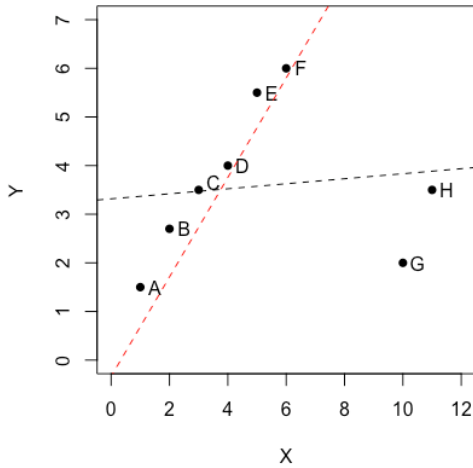


Şekil 2-2: Basit lineer regresyon modelinde etkili gözlemin regresyon doğrusuna etkisi

Etkili gözlemlerin belirlenmesi çok önemlidir çünkü model kurma hatasına, yanlış parametre tahminlerine veya yanlış analiz sonuçlarına sebep olabilirler (Liu, Shah, & Jiang, 2004). Bu tip gözlemlerin tespitinde kullanılan birçok yöntem herhangi bir gözlemin veriden çıkarılması durumunda yarattığı etkinin ölçülmesi prensibine dayanır. Ancak veri setinde birden fazla etkili gözlem olması durumunda tek gözlem silmeye dayalı yöntemleri genellikle başarısız olurlar. Tek gözlem silmeye dayalı yöntemlerin en büyük sorunu maskeleye ve süpürme problemidir. Maskeleye problemi, etkili bir gözlemin diğer etkili gözlemlerin varlığından dolayı tespit edilememesi durumudur.

Süpürme etkisi ise etkili olmayan bir gözlemin etkili olarak belirlenmesidir (Lawrance, 1995).

Şekil 2.3'te basit lineer regresyon modelinde maskeleme ve süpürme etkileri gözlemlenebilen bir saçılım grafiği verilmiştir. G ve H noktaları regresyon doğrusunu kendilerine doğru çekmişlerdir. Özellikle H noktasının artık değeri oldukça azalmıştır, bu gözlem etkili gözlem olmasına rağmen belirlenememiştir. G noktası H noktasını maskeleymiştir. A noktası ise modele uygun olduğu halde regresyon doğrusundan uzak kalmış, artık değeri büyümüştür. Bu sebeple bu gözlem sorunsuz olmasına karşın, etkili gözlem olarak belirlenmiş yani süpürme etkisine maruz kalmıştır. Sonuç olarak etkili gözlemleri tespit etmek için sadece artık değerlerini kullanmak başarılı sonuçlar vermemektedir.



Şekil 2-3: Basit lineer regresyon modelinde maskeleme ve süpürme etkisi

3 ETKİLİ GÖZLEMLERİN BELİRLENMESİNDE KULLANILAN KLASİK YÖNTEMLER

Bu bölümde Cook uzaklığı istatistiği, DDFITS ve OD olmak üzere etkili gözlemlerin tespiti için kullanılan geleneksel tanı yöntemleri gösterilecektir. Bu yöntemlerin hepsi i . gözlemin veri kümesinden silinmesiyle oluşan değişikliğe dayanır. Bu kavram kullanılarak çeşitli tanı yöntemleri geliştirilmiştir. Bu yöntemlerin amacı, regresyon modelindeki veya verilerdeki zayıflıkları belirlemektir.

3.1 Cook Uzaklığı

Cook (1977) tarafından önerilen Cook uzaklığı $\hat{\beta}$ ile i . gözlemin veri kümesinden silinmesiyle elde edilen $\hat{\beta}_{(i)}$ vektörü arasındaki karesel uzaklığın bir ölçüsüdür. Bu tanı istatistiği D_i ile gösterilir ve genel formu

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} \quad (3.1)$$

şeklindedir. D_i daha farklı olarak aşağıdaki gibi de yazılabilir.

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}} \quad i = 1, 2, \dots, n \quad (3.2)$$

Cook uzaklığının kritik değeri olarak $F_{(0.50, p, n-p)}$ alınabilir. D_i değerleri %50'lik güven elipsoidinin dışında ise bu gözlemin etkili gözlem olduğu söylenebilir (Cook R. D., 1977). D_i değerlerinin F dağılımına uygun olduğu varsayılmaktadır. Ancak D_i değerleri bir F dağılımı göstermez, bu karşılaştırma sadece D_i 'yi bilinen bir ölçeğe dönüştürmek için kullanılır (Cook R. D., 1977). Özellikle büyük örneklerde ve çoklu etkili gözlemlerin varlığında Cook uzaklığının kullanılması etkili gözlemler hakkında doğru bilgiler vermeyebilir (Kim, 2017).

3.2 DFFITS

Belsley, Kuh ve Welsch (1980) tarafından önerilen DFFITS $\hat{y}$ ile i . gözlemin veriden silinmesiyle elde edilen $\hat{y}_{(i)}$ arasındaki değişimi gösterir.

$$DFFITS_i = \frac{\hat{y} - \hat{y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}} = \sqrt{\left(\frac{h_{ii}}{1-h_{ii}}\right)} r_i^* \quad i = 1, 2, \dots, n \quad (3.3)$$

DFFITS değerleri hem student türü artıklardan hem de kaldıraç değerlerinde oldukça etkilenmektedir. Belsley, Kuh ve Welsch (1980) $|DFFITS_i| > 2\sqrt{p/n}$ ise i . gözlemi etkili gözlem olarak kabul edilmesini önermişlerdir. Velleman ve Welsch (1981) ise 1 veya 2'den büyük $DFFITS_i$ değerlerinin dikkatle incelenmesi gerektiğini belirtmişlerdir. Ancak önerilen bu kesim değerleri örneklem büyüklüğüne bağlı olup her durum için doğru bir kesim değeri oluşturmak oldukça zordur (Montgomery, Peck, & Vinning, 2012).

3.3 Olabilirlik Değişim İstatistiği (Likelihood Displacement) (OD)

Cook ve Weisberg (1982) tarafından önerilen OD istatistiği, etkili gözlemleri tespit etmek için yaygın olarak kullanılır. Bu istatistik, $\hat{\beta}$ ile i . gözlemin veriden çıkarılmasıyla elde edilen $\hat{\beta}_{(i)}$ 'nin log-likelihood fonksiyonlarını karşılaştırmayı amaçlamaktadır.

$$LD_i = 2|L(\hat{\beta}) - L(\hat{\beta}_{(i)})|, \quad i = 1, 2, \dots, n \quad (3.4)$$

Lineer regresyon modelinin log-likelihood fonksiyonu $(L(\beta, \sigma^2))$ aşağıdaki gibidir:

$$\begin{aligned} L(\beta, \sigma^2) &= \sum_{i=1}^n \left(-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (y - x\beta_{(i)})^2 \right) \\ &= -\frac{n}{2} (\ln(2\pi\sigma^2)) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y - x\beta_{(i)})^2 \end{aligned} \quad (3.5)$$

LD_i değerlerinin p serbestlik dereceli χ^2 dağılımı ile karşılaştırılması ve $LD_i > \chi_p^2$ ise i . gözlemin etkili gözlem olduğu önerilmiştir (Cook & Weisberg, 1982). Ancak büyük örneklem için χ^2 dağılımı uygun değildir, aksine örneklem büyüklüğü n arttıkça olasılık 0'a yaklaşmaktadır (Wang, Chow, & Wei, 1995). OD istatistiğinin bir başka dezavantajı diğer tek gözlem silmeye dayalı yöntemlerine benzer olarak maskeleye ve süpürme etkilerinden etkilenmesidir.

4 PARÇACIK SÜRÜ OPTİMİZASYONU

Popülasyon tabanlı meta-sezgisel optimizasyon yöntemlerinden biri olan Parçacık Sürü Optimizasyonu Eberhart ve Kennedy tarafından bulunmuştur (Eberhart & Kennedy, 1995). Bu algoritma hem sürü halinde yaşayan hayvanların sosyal davranışından hem de sürüdeki kuşların kendi davranışlarından ilham almıştır. Her bir kuş (potansiyel çözüm) bir “parçacık” belirtir. Parçacık popülasyonuna “sürü” denir. PSO’da her bir parçacık rastgele hızlar ve konumlarla çözüm aramaya başlar ve her iterasyonda parçacık konumları en iyi çözümü bulmak için önceden tanımlanmış uygunluk (amaç) fonksiyonuna göre güncellenir. Bu algoritmada, düzgün bir performans elde etmek için global ve yerel arama olmak üzere iki yaklaşım uygulanır. Her bir parçacığın uçuş boyunca en iyi pozisyonuna (en iyi uygunluk değeri) P_{best} , sürünün arama alanında şimdiye kadar sahip olduğu en iyi pozisyonuna ise global en iyi yani G_{best} denir. PSO’nun optimum çözüm belirlemek için problem uzayının farklı alanlardaki farklı noktaları aynı anda araştırması önemli özelliklerindendir.

PSO yaklaşımının adımları aşağıdaki gibidir.

1. Parametreler tanımlanır.

w : Eylemsizlik ağırlığı

c_1 : Bilişsel katsayı

c_2 : Sosyal katsayı

t : İterasyon sayısı

K : Parçacık sayısı

v_{max} : Maksimum hız

n : Parçacık boyutu

2. Rastgele üretilen başlangıç pozisyon ve hızları sırasıyla aşağıdaki gibidir.

$$P = (\vec{p}_1, \dots, \vec{p}_i, \dots, \vec{p}_K)$$

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ p_{i1} & p_{i2} & \cdots & p_{in} \\ \vdots & \vdots & \vdots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{Kn} \end{bmatrix}_{K \times n}$$

$$V = (\vec{v}_1, \dots, \vec{v}_l, \dots, \vec{v}_K)$$

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ v_{i1} & v_{i2} & \cdots & v_{in} \\ \vdots & \vdots & \vdots & \vdots \\ v_{K1} & v_{K2} & \cdots & v_{Kn} \end{bmatrix}_{K \times n}$$

$i = 1, 2, \dots, K$ olmak üzere, $\vec{p}_i$ popülasyonun i . parçacığıdır $\vec{v}_i$ 'de bu parçacığa karşılık gelen hız vektörüdür.

3. Sürüdeki tüm parçacıkların uygunluk değerleri uygunluk fonksiyonuna göre hesaplanır.

$$F(P) = (F(\vec{p}_1), \dots, F(\vec{p}_l), \dots, F(\vec{p}_K))$$

4. Her parçacığın en iyi pozisyon matrisi P_{best} aşağıdaki gibi bulunur. İlk iterasyonda $P_{best} = P$ dir.

$$P_{best} = (\vec{p}_1^{best}, \vec{p}_2^{best}, \dots, \vec{p}_i^{best}, \dots, \vec{p}_K^{best})$$

5. Global en iyi $\vec{p}_g^{best} = \max(or \min)_{(\vec{p} \in P)(F(\vec{p}))}$ şeklinde korunur.
6. Hız ve pozisyonlar sırasıyla aşağıdaki denklemlere göre güncellenir.

$$\vec{v}_i^{t+1} = w \times \vec{v}_i^t + \underbrace{c_1 r_1 (\vec{p}_1^{best} - \vec{p}_i^t)}_{bilisşel \ bileşen} + \underbrace{c_2 r_2 (\vec{p}_g^{best} - \vec{p}_i^t)}_{sosyal \ bileşen} \quad (4.1)$$

$$\vec{p}_i^{t+1} = \vec{p}_i^t + \vec{v}_i^{t+1} \quad (4.2)$$

Burada r_1 ve r_2 $[0,1]$ arasında düzgün dağılımdan üretilen rastgele değişkenlerdir. Bu rastgele değişkenler algoritmaya rastlantısal bir özellik kazandırmıştır. w parametresi parçacıkların bir önceki hızlarına bağlı olarak hız güncellemesini sağlar ve çözüm uzayında ani hız değişimi yapmasını engeller. Daha yüksek bir w değeri, küresel keşfi teşvik ederken (yeni alanlar araştırmayı), daha küçük bir w değeri ise yerel aramayı kolaylaştırır. Denklem 4.2’de parçacığın yeni hızının bir önceki hızına göre hesaplanması verilmiştir. Sosyal bileşen geçmiş performanslardaki en iyi pozisyon bilgisini içerir. Böylece parçacıklar en iyi pozisyonlarla değiştirilmiş olur. Bilişsel bileşen ise parçacığın yakınında en iyi pozisyonun bulunmasını sağlar ve böylece sürünün en iyi konumuna doğru yönelmesi sağlanır. Denklem 4.3’te ise konumun hıza bağlı olarak güncellenmesi verilmiştir.

7. Maksimum iterasyon sayısına kadar yukarıdaki adımlar tekrarlanır. Durdurma kriteri sağlandığında elde edilen G_{best} değeri çözüm olarak belirlenir.

5 KESİKLİ PARÇACIK SÜRÜ OPTİMİZASYONU

PSO başlangıçta sürekli uzaylardaki optimizasyon problemlerini çözmek için kullanılmıştır. Fakat birçok kesikli (ayrık, ikili) uzay optimizasyon problemlerinin mevcut olmasından dolayı Kennedy ve Eberhart (1997) tarafından ikili uzaydaki kombinatoriyal optimizasyon problemlerini çözmek için BPSO geliştirilmiştir. Özellik seçimi (Cervante, Xue, Zhang, & Shang, 2012), zamanlama (Pedrasa, Spooner, & MacGill, 2009), yönlendirme (Ayob, ve diğerleri, 2010) ve lot boyutlandırma (Taşgetiren & Liang, 2003) gibi ikili uzayda tanımlanan birçok optimizasyon problemi için BPSO kullanılmış ve etkili performans göstermiştir.

BPSO yaklaşımında i . parçacık elemanları $\vec{p}_i = (p_{i1}, p_{i2}, \dots, p_{ij}, \dots)$ $p_{ij} \in \{0,1\}$ olacak şekilde kesikli olarak tanımlanır. PSO ve BPSO arasındaki temel fark hızların tanımlanmasıdır. Her parçacığın hızı, 1 değerini alan parçacığa karşılık gelen elemanların olasılıkları açısından tanımlanır. Sigmoid fonksiyonu $\vec{v}_i$ 'nin elemanlarını (0,1) aralığına dönüştürmek için kullanılır. t . iterasyonda her bir parçacığın hızı ve pozisyonu sırasıyla aşağıdaki denklemler gibi güncellenir.

$$\vec{v}_i^{t+1} = w \times \vec{v}_i^t + \underbrace{c_1 r_1 (\vec{p}_{best_i}^t - \vec{p}_i^t)}_{\text{bilişsel bileşen}} + \underbrace{c_2 r_2 (G_{best} - \vec{p}_i^t)}_{\text{sosyal bileşen}}$$

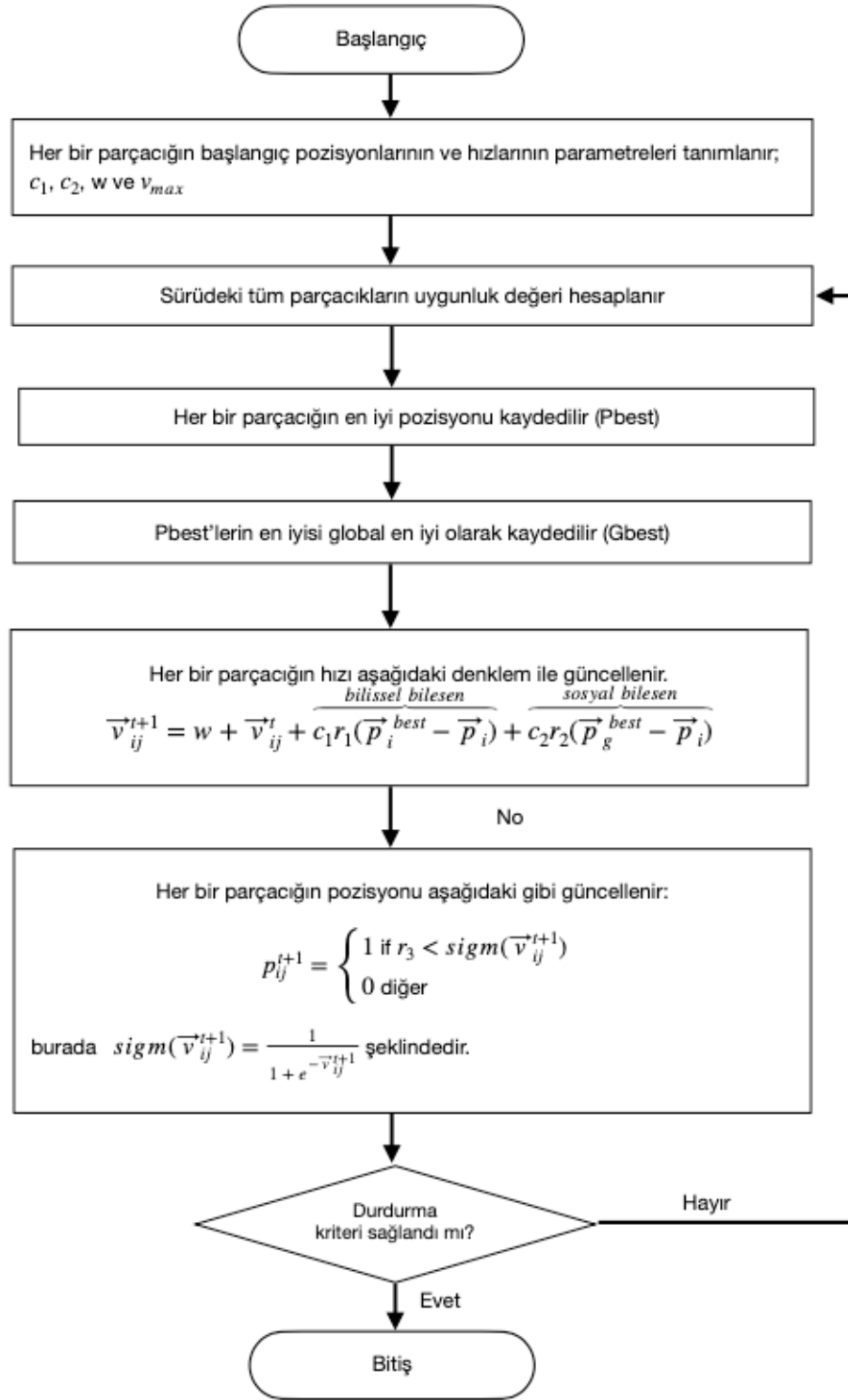
$$\vec{p}_i^{t+1} = \begin{cases} 1, & r_3 < \text{sigm}(\vec{v}_i^{t+1}) \\ 0, & \text{diğer} \end{cases} \quad (5.1)$$

Burada $\vec{v}_i$ i . parçacığın hız vektörü, $\text{sigm}(\vec{v}_i^{t+1}) = \frac{1}{1+e^{-\vec{v}_i^{t+1}}}$ sigmoid dönüşüm fonksiyonu, r_1, r_2 ve r_3 [0,1] arasında düzgün dağılımlı rastgele sayılardır. BPSO'nun adımları Şekil 6.1'de gösterilmiştir.

6 ETKİLİ ALT KÜMELERİN TESPİTİNDE KESİKLİ PARÇACIK SÜRÜ OPTİMİZASYONU KULLANILMASI

Daha önce de belirtildiği gibi, tek gözlem silmeye dayalı yöntemlerin birçok dezavantajı vardır. Özellikle etkili gözlem alt kümelerin varlığında yanıltıcı sonuçlar doğururlar. Gözlemlerin etkili olup olmadığı düşünüldüğünde, etkili altkümelerin tanımlanması bir kombinatoriyal optimizasyon problemi olarak düşünülebilir.

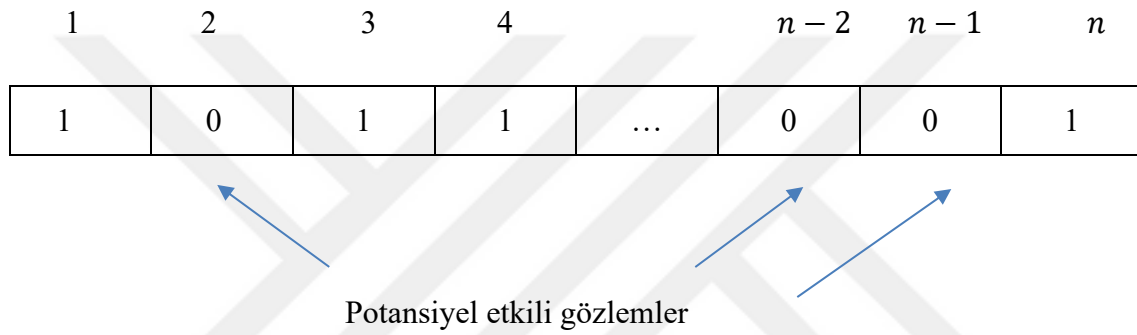
BPSO, kısa çalışma süresi, basitliği, düşük parametre gereksinimleri, gelişmiş yakınsama, verimli küresel arama, başlangıç noktalarına daha az bağımlılık gibi özellikleri nedeniyle diğer meta-sezgisel algoritmalara kıyasla çeşitli avantajlara sahiptir (Sancar & Inan, 2020). Bu özellikler göz önüne alındığında, BPSO yaklaşımının lineer regresyon modellerinde etkili gözlemlerin eşzamanlı tanımlanması için yeni bir yöntem olarak uygun olduğu düşünülmektedir.



Şekil 6-1 BPSO'nun akış diyagramı

6.1 Parçacığın Yapısı

BPSO yaklaşımında her bir $\vec{p}_i = (p_{i1}, p_{i2}, \dots, p_{ij}, \dots)$ parçacığının pozisyonu 0 veya 1 olacak şekilde ikili değer olarak belirlenmektedir. Her bir parçacık $1 \times n$ boyutludur. Burada n gözlem sayısıdır. $p_{ij} = 0$ olan bileşenler potansiyel etkili gözlemleri gösterir, $p_{ij} = 1$ ise i . parçacıkta etkili olmayan gözlemleri belirtir. En son durumda $G_{best} = 0$ olan gözlemler etkili gözlem olarak işaretlenir. Şekil 6.2'deki $\vec{p}_i = (1, 0, 1, 1, \dots, 0, 0, 1)$ parçacığında 2., $n - 2$. ve $n - 1$. gözlemler potansiyel etkili gözlemlerdir.



Şekil 6-2: Bir parçacıkta bulunan potansiyel etkili gözlemler

6.2 Amaç Fonksiyonu

Bu çalışmada etkili gözlem alt kümelerinin tespiti için aşağıdaki OD tabanlı amaç fonksiyonu kullanılmıştır. Amaç fonksiyonunu tanımlamak en iyi çözümü elde etmek için önemlidir. Denklem 6.1'de etkili alt kümeleri tanımlamak için kullanılan amaç fonksiyonu verilmiştir.

$$LD_{0_s} = \frac{2[L(\hat{\beta}, \sigma^2) - L(\hat{\beta}_{(-0_s)}, \sigma_{(-0_s)}^2)]}{m+1} \quad (6.1)$$

Burada 0_s potansiyel etkili gözlem olarak belirlenen gözlemlerdir, m de potansiyel etkili gözlemlerin sayısı yani 0_s 'lerin sayısıdır. $(\hat{\beta}, \sigma^2)$ tüm veri setinden elde edilen

parametre tahminleri kümesi, $(\hat{\beta}_{(-0_s)}, \sigma^2_{(-0_s)})$ ise m tane gözlemin veriden silinmesiyle elde edilen parametre tahminidir.

OD istatistiği lineer regresyon modellerinde etkili gözlemleri tanımlamak için iyi bir yöntem olabilir ancak diğer yöntemlere benzer olarak maskeleme ve süpürme etkilerinden etkilenir. Bu nedenle gözlemlerin tek tek silinmesi yerine etkili gözlem kümesinin veriden çıkartılarak OD fonksiyonunu hesaplamanın daha uygun olacağı düşünülmüştür. Başka bir deyişle buradaki amaç, BPSO aracılığıyla etkili gözlemler olarak tanımlanan gözlem grubu silindiğinde OD üzerinde en büyük etkiye sahip olan gözlem grubunu belirlemektir. Her bir gözlem OD değerlerinde bir farklılık yapma eğiliminde olduğundan hesaplanan OD değerleri $m + 1$ 'e bölünerek, etkili olmayan veri noktalarının etkili olarak tanımlanması engellenecek ve etkili veri noktalarının doğru tanımlanması sağlanacaktır. Amacımız maksimizasyon olduğundan 0'lar kümesi etkili gözlemler olarak tanımlanan optimum gözlemler olduğunda, önerilen uygunluk fonksiyonunun maksimum değere ulaşacağı düşünülmektedir. Ayrıca, 6.1. denkleminde görüldüğü gibi OD istatistiği $m + 1$ 'e bölünerek, bir gözlem grubu için hesaplanan sonuçlar hiçbir veri noktasını etkili olarak belirlememiş olsa dahi önerilen BPSO tabanlı yaklaşımın çalışması sağlanmış, oluşabilecek belirsizliğin önüne geçilmiştir. Aynı zamanda, arama alanında optimal etkili gözlem kümesini aynı anda arayacağından, maskeleme ve süpürme etkilerinden önemli ölçüde etkilenmeyecektir.

7 SİMÜLASYON

Bu bölümde BPSO tabanlı yaklaşımın performansı simülasyon çalışması ile değerlendirilmiştir. Önerilen bu yaklaşım, maskeleye oranı (MO) ve süpürme oranı (SO) olmak üzere iki değerlendirme kriterine göre daha önce bahsedilen geleneksel teşhis yöntemleri ile karşılaştırılmıştır. Bu iki kriterin sıfıra yakın çıkması bir başarı olarak değerlendirilecektir. Veri üretimi ve model uyumu her simülasyon için 100 kez tekrar edilmiş ve her bir yöntem için maskeleye ve süpürme olasılıklarının ortalaması kaydedilmiştir. Simülasyon çalışmasında veri üretimi için R paket programı (RStudio, 2015) kullanılmıştır. Veri üretirken üç farklı örneklem büyüklüğü ($n=50, 100, 200$), üç farklı kirletme oranı ($k.o.= 0.02, 0.04, 0.06$) ve üç farklı hata dağılımı ($t(3)$ ağır kuyruklu, normal, log normal) kullanılmıştır. İlk olarak içerisinde etkili gözlem içermeyen temiz data üretilmiştir. Ardından örneklem büyüklüğü ile kirletme oranı çarpımı ($s = k.o. \times n$) kadar etkili gözlem üretilmiştir. Veride kirletilecek olan gözlemlerin ilk s tanesi kirletilmiştir. Böylece etkili gözlemlerin hangi gözlemler olduğu önceden bilinip, sonuçların doğruluğu değerlendirilebilmiştir. Veri kirletme işleminde etkili gözlemlerin yönlerinin rastgele oluşturulabilmesi için Bernouilli dağılımı kullanılmıştır. Kirletilecek her bir gözlem için Bernouilli dağılımından bir değer üretilmiş olup eğer bu değer 1 ise sağ kuyruktan aşırı bir değerle, 0 ise sol kuyruktan aşırı bir değerle değiştirilmiştir. Böylece farklı yönlerde etkili gözlemlerin üretilmesi sağlanmıştır.

BPSO'nun ilk parçacık matrisi binom dağılımından rastgele üretilmiştir. İlk rastgele hız matrisi ise maksimum hızı 4 olacak şekilde düzgün dağılımdan elde edilmiştir. BPSO algoritmasının parametreleri aşağıdaki gibi belirlenmiştir.

$$w = 1,$$

$$c_1 = c_2 = 0.5,$$

$$v_{max} = 4,$$

$$K = 30,$$

$$t = 300$$

Yapılan üç farklı simülasyon çalışması aşağıda verildiği gibidir.

7.1 Simülasyon 1

- $y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon$
- $X_1 \sim N(2,1), X_2 \sim N(-1,0.5), X_3 \sim Uni(0,1), X_4 \sim B(1,0.5)$ ve $\varepsilon \sim N(0,0.5)$
- $\beta = (2,3,6,1,2)$
- x_1, x_2, x_3 bağımsız değişkenleri ile hata terimi kirletilmiştir.

7.2 Simülasyon 2

- $y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon$
- $X_1 \sim N(3,0.5), X_2 \sim N(2,0.5), \varepsilon \sim \exp(N(0,0.5)) - \exp(N(0.5, \sigma^2))$
- $\beta = (2,1,1.2)$
- x_1 ve x_2 bağımsız değişkenleri ile hata terimi kirletilmiştir

7.3 Simülasyon 3

- $y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon$
- $X_1 \sim N(2,1), X_2 \sim N(-1,0.5), X_3 \sim Uni(0,1), X_4 \sim B(1,0.5)$ ve $\varepsilon \sim t(3)$
- $\beta = (2,3,6,1,2)$
- x_1, x_2, x_3 bağımsız değişkenleri ile hata terimi kirletilmiştir.

Tablo 7-1: Simülasyon 1 sonuçları

k.o.	Teşhis Yöntemleri	n=50		n=100		n=200	
		MO	SO	MO	SO	MO	SO
0.02	Cook	0.000	0.000	0.283	0.000	0.962	0.000
	DFFITS	0.000	0.063	0.000	0.054	0.000	0.050
	OD	0.000	0.000	0.727	0.000	0.999	0.000
	BPSO	0.000	0.006	0.000	0.000	0.000	0.015
0.04	Cook	0.726	0.000	0.997	0.000	1.000	0.000
	DFFITS	0.042	0.067	0.110	0.058	0.143	0.053
	OD	0.937	0.000	1.000	0.000	1.000	0.000
	BPSO	0.000	0.031	0.000	0.001	0.000	0.028
0.06	Cook	0.988	0.000	1.000	0.000	1.000	0.000
	DFFITS	0.311	0.070	0.409	0.061	0.444	0.056
	OD	1.000	0.000	1.000	0.000	1.000	0.000
	BPSO	0.008	0.031	0.007	0.007	0.029	0.048

Tablo 7-2: Simülasyon 2 sonuçları

k.o.	Teşhis Yöntemleri	n=50		n=100		n=200	
		MO	SO	MO	SO	MO	SO
0.02	Cook	0.013	0.000	0.307	0.000	0.804	0.000
	DFFITS	0.000	0.071	0.003	0.062	0.008	0.059
	OD	0.093	0.000	0.662	0.000	0.948	0.000
	BPSO	0.000	0.044	0.001	0.003	0.002	0.015
0.04	Cook	0.401	0.000	0.860	0.000	0.995	0.000
	DFFITS	0.032	0.071	0.047	0.065	0.062	0.062
	OD	0.754	0.000	0.972	0.000	0.999	0.000
	BPSO	0.002	0.026	0.002	0.003	0.002	0.022
0.06	Cook	0.770	0.000	0.982	0.000	0.999	0.000
	DFFITS	0.104	0.072	0.142	0.067	0.163	0.064
	OD	0.945	0.000	0.996	0.000	0.999	0.000
	BPSO	0.001	0.043	0.001	0.007	0.004	0.033

Tablo 7-3: Simülasyon 3 sonuçları

k.o.	Teşhis Yöntemleri	n=50		n=100		n=200	
		MO	SO	MO	SO	MO	SO
0.02	Cook	0.007	0.000	1.000	0.000	0.084	0.000
	DFFITS	0.000	0.066	0.316	0.064	0.099	0.043
	OD	0.043	0.001	1.000	0.003	0.902	0.000
	BPSO	0.003	0.071	0.050	0.063	0.007	0.014
0.04	Cook	0.684	0.000	0.879	0.000	0.540	0.000
	DFFITS	0.156	0.060	0.299	0.046	0.068	0.051
	OD	0.817	0.001	0.917	0.000	0.746	0.000
	BPSO	0.003	0.090	0.024	0.030	0.007	0.019
0.06	Cook	0.858	0.000	0.934	0.000	0.971	0.000
	DFFITS	0.366	0.554	0.485	0.041	0.566	0.034
	OD	0.900	0.000	0.942	0.000	0.968	0.000
	BPSO	0.022	0.137	0.088	0.045	0.171	0.027

Tablo 7-4: Veride tek etkili gözlem olması durumunda teşhis yöntemlerinin SO ve MO sonuçları

Teşhis Yöntemleri	n=200 k.o.=0.05		n=500 k.o.=0.002		n=800 k.o.=0.125		n=1000 k.o.=0.001	
	MO	SO	MO	SO	MO	SO	MO	SO
Cook	0.050	0.000	0.550	0.000	0.980	0.000	1.000	0.000
DFFITS	0.000	0.063	0.000	0.067	0.000	0.069	0.000	0.070
OD	0.600	0.000	1.000	0.000	1.000	0.000	1.000	0.000
BPSO	0.000	0.003	0.000	0.004	0.000	0.004	0.000	0.004

Tablo 7-1, Tablo7-2 ve Tablo 7-3'te yer alan $n=50$ ve $k.o.=0.02$ olduđu durumlar üretilen veride tek etkili gözlemin bulunması anlamına gelmektedir. Cook uzaklığı, DFFITS, OD ve önerilen BPSO yaklaşımının hepsi bu durumda iyi performans göstermektedir. Ancak gözlem sayısı n 'nin ve kirletme oranı $k.o.$ 'nın artmasıyla geleneksel yöntemlerin performansları azalmaktadır. Etkili alt kümelerin varlığında, özellikle Cook uzaklığı ve OD istatistikleri, etkili gözlemlerin neredeyse hiçbirini tespit edememiştir. $k.o.$ arttırılmadan sadece örneklem büyüklüğünün yöntemlerin performansı üzerindeki saf etkisini gözlemlemek için artan örneklem büyüklüklerinde tek etkili gözlem durumu Simülasyon 1 için ayrıca incelenmiştir. Sonuçlar Tablo 7-4'te sunulmaktadır. Beklendiği gibi, Cook uzaklığı ve OD'nin performansı artan örneklem büyüklüğü ile birlikte önemli ölçüde bozulmaktadır. DFFITS maskeleye sorunundan diğerleri kadar etkilenmese de SO açısından önerilen yönteme göre daha kötü sonuçlar vermektedir. Önerilen yöntemin performansının örneklem büyüklüğüne bağlı olmadığı ve diğer yöntemlere göre etkili gözlemleri oldukça iyi tespit edebildiği görülmektedir. Ancak, örneklem büyüklüğündeki artış önerilen yöntemin hesaplama süresini arttırmıştır. Bu sorunu gidermek için, her bir iterasyondaki Gbest parçacıkları kaydedilmiş ve her bir gözlem için kaydedilen tüm Gbest parçacıklarının yüzde kaçında etkili gözlem olarak belirlendiği hesaplanmıştır. Böylelikle daha az iterasyonla oldukça gerçekçi sonuçlar elde edilebilmiştir.

8 GERÇEK VERİ SETİ UYGULAMASI

Bu bölümde önerilen BPSO yaklaşımının performansı Hills Yarışları ve CYG OB1 Yıldız Kümesi gerçek veri setleri aracılığıyla değerlendirilmiştir. MO ve SO değerlendirme kriterlerine ek olarak R_{adj}^2 ve σ^2 değerleri ile de model karşılaştırmaları yapılmıştır.

8.1 Hills Yarışları Veri Seti

Bu uygulamada 1984 yılında İskoç'ta düzenlenen Hills yarışları veri seti kullanılmıştır. 35 gözlemden oluşan veride yanıt değişkeni kayıt süresidir. Veride rota sırasında kazanılan toplam yükseklik (fit=0,0003048 km) ve gidilen mesafe (mil=1,609344 km) olmak üzere iki açıklayıcı değişken bulunmaktadır (Atkinson A. , 1986).

Hills veri setindeki etkili gözlemler geleneksel yöntemlerle ve önerilen BPSO yaklaşımıyla belirlenmiş ve her bir yöntemin sonuçları Tablo 8-2'de sunulmuştur. Etkili gözlemler olarak bulunan gözlemler tabloda koyu olarak yazılmıştır. Tablo 8-2'nin son sütunu önerilen yöntem için etki değerlerini göstermektedir. Geleneksel yöntemler için kesim değerleri ise Tablo 8-1'de verilmektedir.

Tablo 8-1: Hills Yarışları verisi için geleneksel yöntemlerin kesim değerleri

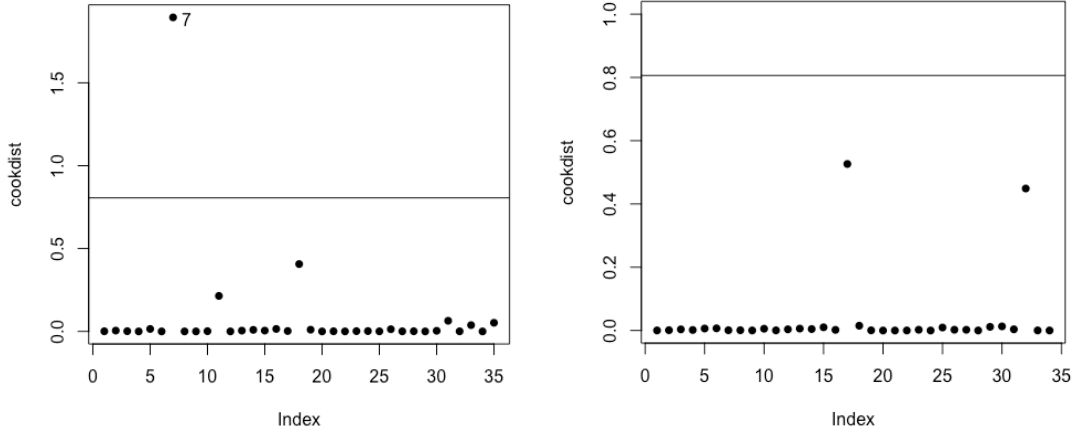
Teşhis Yöntemleri	Kesim değerleri	Kesim noktaları
Cook uzaklığı	$F_{\{\alpha;p,n-p\}}$	0.8057312
DFFITS	$2\sqrt{p/n}$	± 0.58554
OD	$\chi_{(\alpha,p)}^2$	7.81944034

Tablo 8-2: Hill Yarışları veri setinde etkili gözlemlerin tespitinde kullanılan tüm yöntemlerin sonuçları

	Cook	DFFITS	OD	BPSO
1	8.134912e-04	0.048655302	0.10830752	0.006666667
2	4.898145e-03	-0.119841802	0.09131054	0.003333333
3	1.379338e-03	-0.063417046	0.10241931	0.320000000
4	6.721553e-05	-0.013977456	0.11090565	0.483333333
5	1.476689e-02	-0.209658654	0.06830629	0.010000000
6	4.924224e-05	0.011963432	0.11097150	0.793333333
7	1.894671e+00	2.700267176	7.62445375	1.000000000
8	4.547035e-05	-0.011496146	0.11093833	0.003333333
9	9.953011e-05	-0.017009519	0.11060455	0.010000000
10	1.389065e-03	0.063623681	0.10448092	0.256666667
11	2.141871e-01	0.792567587	0.70430502	0.313333333
12	1.003818e-04	-0.017082009	0.11066742	0.503333333
13	4.714510e-03	-0.117835726	0.07457506	0.003333333
14	9.295217e-03	0.165701135	0.07776703	0.710000000
15	4.858681e-03	-0.119507194	0.08202771	0.166666667
16	1.491993e-02	-0.211443081	0.04819088	0.000000000
17	2.382076e-03	-0.083306655	0.10837558	0.196666667
18	4.063631e-01	1.837974703	24.37872690	1.000000000
19	1.030265e-02	-0.175164254	0.05349975	0.000000000
20	3.881560e-05	0.010621505	0.11101489	0.006666667
21	8.494582e-04	0.049719433	0.10838628	0.000000000
22	1.788072e-04	-0.022799788	0.11040184	0.000000000
23	1.662892e-03	-0.069649493	0.10118278	0.000000000
24	2.070977e-03	0.077702579	0.10484469	0.003333333
25	4.842645e-04	0.037531204	0.10922193	0.006666667
26	1.322339e-02	-0.198161215	0.07150468	0.000000000
27	5.202016e-04	-0.038904505	0.10817766	0.000000000
28	1.031178e-03	-0.054811489	0.10438671	0.396666666
29	3.856609e-06	-0.003347895	0.11116876	0.000000000
30	3.698622e-03	-0.104034328	0.09530843	0.000000000
31	6.414452e-02	-0.441461308	0.15944553	0.006666667
32	5.411661e-04	-0.039678286	0.10872144	0.000000000
33	3.780872e-02	0.334357794	0.15858742	1.000000000
34	6.434007e-06	-0.004324246	0.11116570	0.023333333
35	5.213423e-02	-0.393344183	0.18904368	0.756666667

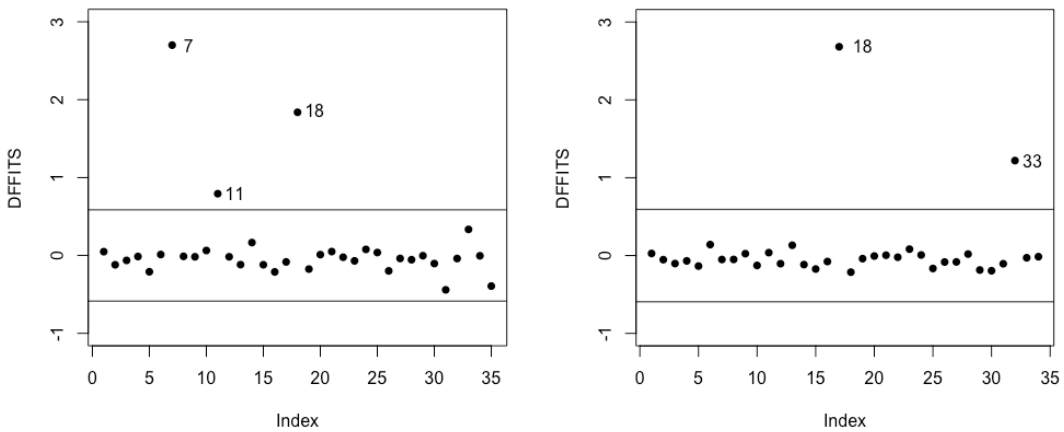
Tablo 8-2’den de görüldüğü gibi Cook istatistiği sadece Bens of Jura Fell’i (7. gözlem) etkili gözlem olarak tanımlamıştır. 7. gözlemi veri setinden çıkartıp diğer gözlemler için tekrar Cook uzaklığı hesaplandığında bu istatistik hiçbir gözlemi etkili gözlem olarak belirleyememiştir. Ek olarak 18 ve 33 numaralı gözlemlerin Cook uzaklık değerleri

diğerlerinden daha büyük olduđu Şekil 8-1’den açıkça görünmektedir. Bu durum Cook uzaklığının hem kesim noktası belirlenmesindeki eksikliğinin hem de veride çoklu etkili gözlem olması durumunda doğru sonuçlar elde edemeyişinin göstergesidir.



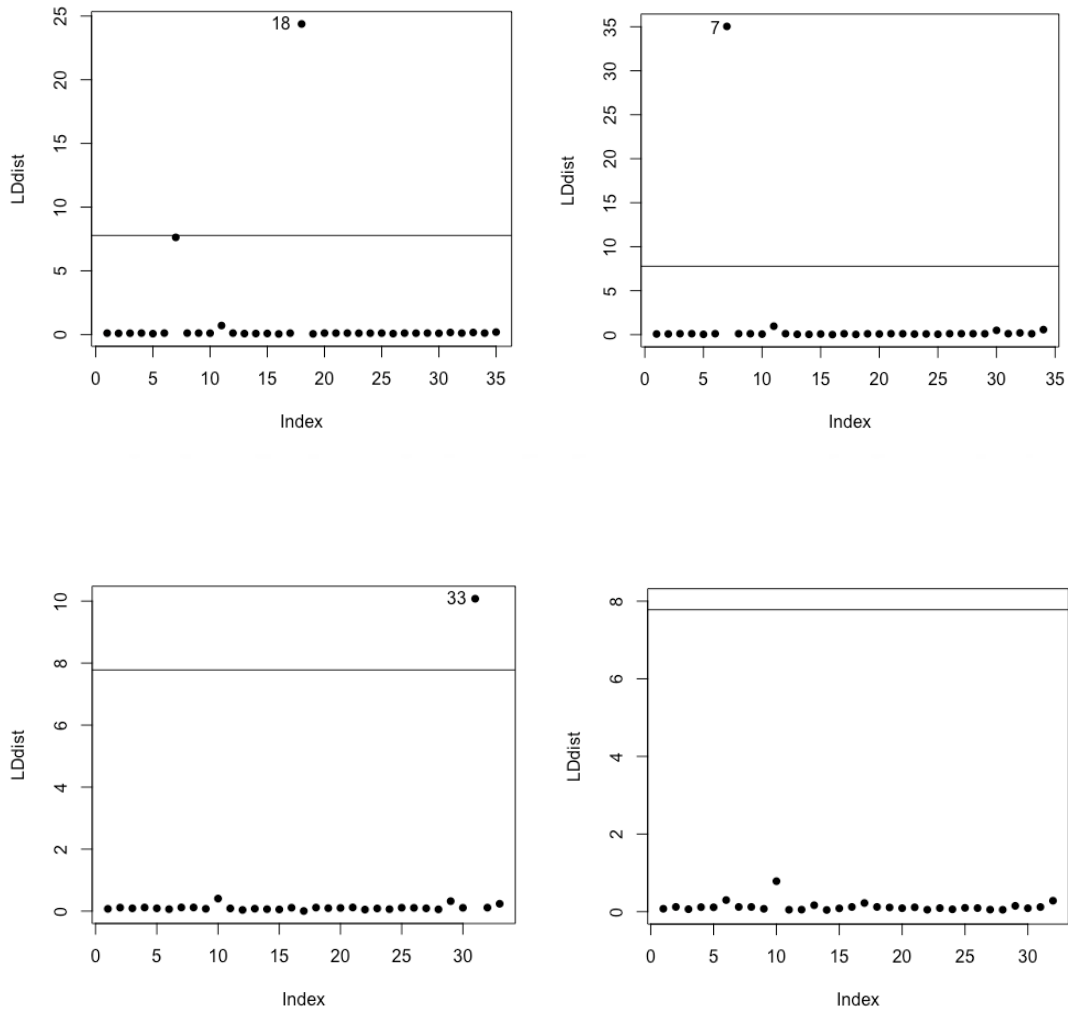
Şekil 8-1: Gözlemlerin Cook uzaklığı grafiği

Jura Fell (7. gözlem), Lairig Ghru Fun Run (11. gözlem) ve Knock Hill (18. gözlem) DFFITS tarafından etkili gözlemler olarak belirlenmişlerdir. 7. gözlemin kaldırıldığı veri seti için DFFITS 18. ve 33. gözlemleri etkili gözlem olarak işaretlemektedir. Buna ek olarak, 11. gözlem artık etkili olarak işaretlenmemektedir. Bu durum, 33. gözlemin maskelendiği ve 11. gözlemde ise süpürme etkisinin olduğunun bir göstergesidir.



Şekil 8-2: Gözlemlerin DFFITS grafiği

OD istatistiği sadece Knock Hill'i (18. gözlem) etkili gözlem olarak tespit etmiştir. Bu gözlem veri kümesinden kaldırıldığında OD 7. gözlemi de etkili gözlem olarak belirlemiştir. Hem 7. hem de 18. gözlemlerin veri setinden birlikte kaldırıldığı durumda ise OD 33. gözlemi etkili olarak işaretlemiştir. Bu 3 gözlemi birlikte veri setinden çıkardığımızda da OD hiçbir gözlemi etkili olarak belirlememiştir. Bu durum, OD istatistiğinin veri setinde birden çok etkili gözlem bulunmasında etkili gözlemleri tespit etme eksikliğinin bir göstergesidir.



Şekil 8-3: Gözlemlerin OD grafiği

Önerilen yöntem Jura Fell Bens (7. gözlem), Knock Hill (18. gözlem) ve Two Breweries (33. gözlem) gözlemlerini etkili gözlem olarak belirlemiştir. Buna ek olarak, sonuçlarda bu üç gözlem dışında dikkate değer yüksek etki derecesine sahip başka gözlemler bulunmamaktadır.

Tablo 8-3'te bu veri setiyle kurulan üç farklı regresyon modeli verilmiştir ve etkili gözlemlerin regresyon sonuçlarına olan etkisi açıkça gözlemlenebilmektedir. 1. model herhangi bir gözlemin veriden silinmediği tüm veri setinden elde edilen modeldir. 2. model DFFITS tarafından etkili olarak işaretlenen gözlemlerin (7, 11, 18) kaldırıldığı durumdaki model ve 3. model ise önerilen yöntemin etkili gözlem olarak bulduğu alt kümeyi (7, 18, 33) veri kümesinden silerek elde edilen regresyon modelidir. Bu modeller karşılaştırıldığında 3. modelin düzeltilmiş R^2 değeri daha yüksektir ve hata varyansı 2. modele göre %40 azalmıştır. Tüm bu sonuçlar ve literatürde bulunan Atkinson (1986) yorumu incelendiğinde, önerilen yöntemin sonuçları doğrulanmaktadır. Önerilen yöntem diğer yöntemlerden farklı olarak tüm etkili gözlemleri aynı anda tespit edebilmektedir. Aynı zamanda maskeleye ve süpürme etkilerinden etkilenmemektedir.

Tablo 8-3: Hills Yarışları verisi için kurulan üç farklı regresyon modeli

	Silinen gözlemler	Regresyon Modeli	Düzeltilmiş R^2	σ^2
1	Hiçbiri	$y = -0.1491383 + 0.1035808x_1$ $+ 0.0001841x_2$	0.914	0.0599
2	7,11,18	$y = -0.1774611 + 0.1137933 x_1$ $+ 0.0001284 x_2$	0.9721	0.0105
3	7, 18, 33	$y = -0.1362572 + 0.1105251x_1$ $+ 0.0001097x_2$	0.9855	0.0062

8.2 CYG OB1 Yıldız Kümesi Verisi

İkinci uygulamada CYG OB1 yıldız kümesi verisi kullanılmıştır. Veri iki değişkenden ve 47 gözlemden oluşmaktadır. Yanıt değişkeni yıldızların ışık yoğunluğunun logaritması (logI), bağımsız değişken ise yıldızların yüzey sıcaklığının logaritmasıdır (logst).

Hertzsprung-Russell (H-R) diyagramı yıldız evrim teorisinin temelini oluşturur. Diyagram esasen yüzey sıcaklıklarına karşı çizilen yıldızların enerji çıkışının bir grafiğidir. Vanisma Greve (1972)'ye göre kalibre edilen Yıldız Kümesi CYG OB1'in H-R diyagramından elde edilen veri kümesi EK-2'de verilmektedir.

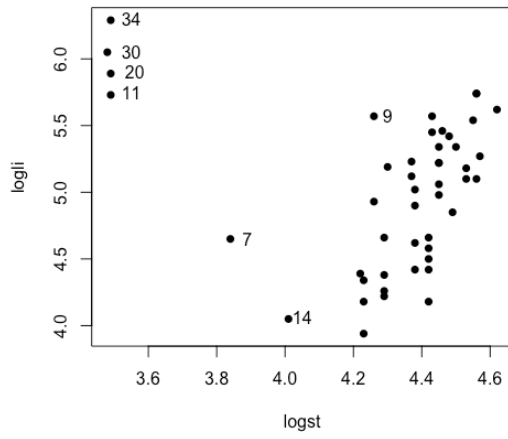
CYG OB1 veri setindeki etkili gözlemler geleneksel yöntemler ve önerilen yöntemle belirlenmiş, sonuçlar Tablo 8-6'te sunulmuştur. Veride tek bağımsız değişken bulunduğu Şekil 8-3'te verinin serpm diyagramı verilmiştir. Tablo 8-5'da önerilen yöntemin etkili gözlem olarak bulduğu gözlemler ve etki değerleri yer almaktadır. Geleneksel yöntemler için kesim değerleri Tablo 8-4'te sunulmuştur. Cook uzaklığı ve OD istatistiği herhangi bir gözlemi etkili gözlem olarak belirlemekte başarısız olmuşlardır. Bu iki yöntemin veride çoklu etkili gözlem bulunması durumunda maskeleme sorunun olduğunu kanıttır. DFFITS 14, 20, 30 ve 34 numaralı gözlemleri etkili gözlem olarak tespit etmiştir. Ancak Şekil 8-3'ten gözlemlenebildiği gibi 11. gözlem maskelenmiş ve DFFITS tarafından bulunamamıştır. Önerilen yaklaşım ise 11, 20, 30, 34, 7, 9 ve 14 numaralı gözlemleri etkili olarak işaretlemiş ve ayrıca maskeleme ve süpürme etkilerinden etkilenmemiştir. Aynı zamanda önerilen yöntemin etkili olarak bulduğu gözlemler Şekil 8-3 ile uyumludur. 11, 20, 30 ve 34 numaralı gözlemlerin etki değerleri %95'in üzerinde, 7. gözlemin etki değeri %91'e eşit ve geriye kalan 9 ve 14 numaralı gözlemlerin etki değerleri %90'ın altında olarak bulunmuştur. Diğer gözlemlerin etki değerleri ise %85'in altında olup önemsizmemektedir.

Tablo 8-4: CYG OB1 verisi için geleneksel yöntemlerin kesim değerleri

Teşhis Yöntemleri	Kesim değerleri	Kesim noktaları
Cook uzaklığı	$F_{\{\alpha;p,n-p\}}$	0.7039344
DFFITS	$2\sqrt{p/n}$	± 0.4125685
OD	$\chi^2_{(\alpha,p)}$	5.991465

Tablo 8-5: Önerilen yöntemin etkili olarak işaretlediği gözlemler ve etki dereceleri

Gözlem numaraları	Etki dereceleri
7	0.910
9	0.885
11	0.960
14	0.880
20	0.995
30	1.000
34	0.990



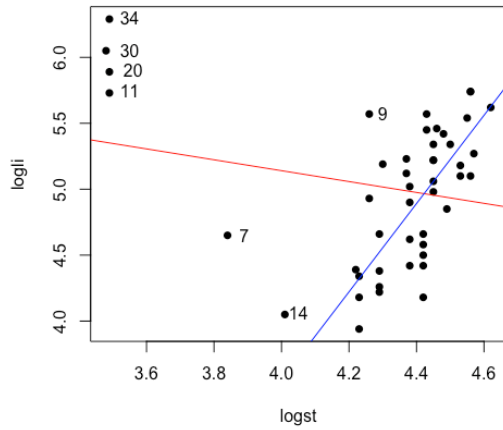
Şekil 8-4: CYG OB1 verisi saçılım grafiği

Tablo 8-6: CYG OBI verisinde etkili gözlemlerin tespitinde kullanılan tüm yöntemlerin sonuçları

	Cook uzaklığı	DFFITs	OD	BPSO
1	2.144696e-03	0.064898025	0.04820537	0.555
2	4.366058e-02	0.299794676	0.05108715	0.010
3	3.796996e-04	-0.027259555	0.05471988	0.780
4	4.366058e-02	0.299794676	0.05108715	0.010
5	1.052847e-03	0.045423963	0.05193309	0.745
6	1.165485e-02	0.152395064	0.03087269	0.230
7	4.458315e-02	-0.298787936	0.08688781	0.910
8	8.754312e-03	0.131479394	0.04642517	0.060
9	1.037046e-02	0.143895305	0.02415350	0.885
10	6.410336e-04	0.035428120	0.05374746	0.305
11	6.731445e-02	0.365092909	0.15547081	0.960
12	9.791721e-03	0.139567767	0.03079769	0.445
13	1.090914e-02	0.147272727	0.03434047	0.000
14	8.997550e-02	-0.438768531	0.16238430	0.880
15	2.024341e-02	-0.203193189	0.01080737	0.005
16	6.007649e-03	-0.108973115	0.03842780	0.005
17	4.599398e-02	-0.313893857	0.06979265	0.000
18	2.486654e-02	-0.225560155	0.01711307	0.695
19	2.818711e-02	-0.241311168	0.01730407	0.005
20	1.361546e-01	0.522608792	0.26737381	0.995
21	1.435823e-02	-0.170068134	0.01573823	0.005
22	2.242922e-02	-0.214379731	0.01081474	0.000
23	1.200820e-02	-0.154903260	0.02579385	0.505
24	3.795101e-04	-0.027250022	0.05531441	0.330
25	5.010681e-05	0.009899299	0.05603264	0.015
26	3.778877e-03	-0.086254232	0.04444110	0.025
27	4.547379e-03	-0.094740111	0.03922351	0.085
28	2.560500e-04	-0.022382300	0.05524942	0.000
29	1.669607e-02	-0.183563370	0.01745295	0.180
30	2.336906e-01	0.690666047	0.44563413	1.000
31	1.173372e-02	-0.153221943	0.02266862	0.000
32	2.310193e-03	0.067303052	0.05295711	0.380
33	3.073919e-03	0.077728775	0.04762804	0.015
34	4.132486e-01	0.935330295	0.83525734	0.990
35	1.872580e-02	-0.194849300	0.01487989	0.055
36	4.291566e-02	0.295605066	0.05882376	0.000
37	1.810543e-03	0.059571816	0.05296401	0.010
38	3.073919e-03	0.077728775	0.04762804	0.000
39	3.793181e-03	0.086335328	0.04966537	0.000
40	1.520594e-02	0.174762084	0.02242647	0.620

41	4.879765e-03	-0.098149221	0.03947507	0.005
42	4.866670e-04	0.030862059	0.05478881	0.000
43	8.421488e-03	0.129099215	0.04020419	0.005
44	6.476989e-03	0.113148044	0.03943233	0.025
45	2.394800e-02	0.219550106	0.03637114	0.015
46	2.882967e-05	0.007508708	0.05613842	0.005
47	8.750756e-03	-0.131844820	0.03200104	0.350

Şekil 8-5'te iki tane regresyon doğrusu bulunmaktadır. Kırmızı renkle gösterilen doğru tüm CYG OB1 verisi üzerinden elde edilen regresyon doğrusudur. Mavi renk ile gösterilen doğru ise önerilen yöntemin bulduğu etkili alt kümenin veriden çıkarılmasıyla elde edilen verinin regresyon doğrusudur. Etkili gözlemlerin regresyon doğrusunu kendilerine doğru çektikleri basit lineer regresyon modelinde açıkça görünmektedir.



Şekil 8-5: CYG OB1 verisiyle elde edilen regresyon doğruları

9 SONUÇ

Literatürde sıklıkla kullanılan etkili gözlemleri tespit yöntemleri, özellikle etkili gözlem alt kümelerinin varlığında ve büyük örneklerde etkin bir şekilde çalışmamaktadır. Bu yöntemlerin kesim noktalarını belirlemek için bazı dağılımsal varsayımlar gereklidir. Ancak, bu varsayımlar büyük ölçüde örneklem büyüklüğüne ve açıklayıcı değişkenlerin sayısına bağlıdır yani her durum için uygun değildir. Simülasyon ve gerçek veri setleri uygulamalarından da anlaşılacağı üzere geleneksel yöntemler pek çok durumda güvenilir sonuçlar vermemektedir. Ancak etkili gözlemlerin güvenilir bir şekilde belirlenmesi ve bu gözlemlerin kaynağının tespit edilmesi istatistiksel analizde çok önemli bir konudur. Etkili gözlem olarak teşhis edilen gözlemler silinmek için tespit edilmemelidir. Bazen etkili gözlemler, farklılıklarıyla daha önemli bilgi veriyor olabilirler ve bazen bu gözlemlerin kendilerinin çalışma konusu yapılması gerekebilir.

Literatürdeki yöntemlerin eksiklikleri dikkate alınarak bu çalışmada yeni bir yöntem önerilmiştir. Bu çalışmanın ana fikri, lineer regresyon modellerinde optimal etkili gözlemler kümesinin belirlenmesi için BPSO'ya dayalı yeni bir eşzamanlı yaklaşım önerisidir. Klasik yöntemlerden farklı olarak, bu yöntem herhangi bir dağılımsal varsayım gerektirmez.

Simülasyon ve gerçek veri seti uygulamaları, birden fazla etkili gözlem durumunda bile yöntemin düzgün çalıştığını ve maskeleye ve süpürme etkilerinden diğerleri kadar etkilenmediğini de göstermektedir. Diğer yöntemlerden farklı olarak, örneklem büyüklüğü önerilen yöntemin performansını etkilememektedir. Ancak bu durumda önerilen yöntemin bir dezavantajı, hesaplama süresinin artması olmuştur. Bu sorun yeni bir yaklaşım olan BPSO'nun hafızası kullanılarak çözülmüştür. Bu yeni yaklaşım ile her bir gözlem için bir etki derecesi belirlenerek hesaplama süresi kısaltılmıştır. Gözlemler için etki derecesi hesaplamak daha az iterasyonla da başarılı sonuçlar elde etmemizi sağlamıştır.

10 KAYNAKLAR

- Ab Wahab, M. N., Nefti-Meziani, S., & Atyabi, A. (2015). A comprehensive review of swarm optimization algorithms. *PloS one*, 10(5).
- Acarlar, I., & Gamgam, H. (2015). Çoklu regresyonda etkili gözlemlerin belirlenmesinde iki yeni yaklaşım. *Doktora Tezi*. Ankara: Gazi Üniversitesi. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp> adresinden alındı
- Alpar, R. (2013). *Uygulamalı çok değişkenli istatistiksel yöntemler*. Detay yayıncılık.
- Atkinson, A. (1986). [Influential observations, high leverage points, and outliers in linear regression]: comment: aspects of diagnostic regression analysis. *Statistical Science*, 1(3), 397-402.
- Atkinson, A. C. (1983). Diagnostic regression analysis and shifted power transformations. *Technometrics*, 25(1), 23-33.
- Ayob, M. N., Yusof, Z. M., Adam, A., Abidin, A. F., Ibrahim, I., Ibrahim, Z., . . . Hani, M. K. (2010). A particle swarm optimization approach for routing in VLSI. *2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks*, 49-53.
- Beheshti, Z., & Shamsuddin, S. M. (2013). A review of population-based meta-heuristic algorithms. *Int. J. Adv. Soft Comput. Appl*, 5(1), 1-35.
- Belsley, D. A., Kuh, E., & Welsch, R. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.
- Beyaztaş, U., & Alin, A. (2014). Sufficient jackknife-after-bootstrap method for detection of influential observations in linear regression models. *Statistical Papers*, 55(4), 1001-1018.
- Cervante, L., Xue, B., Zhang, M., & Shang, L. (2012). Binary particle swarm optimisation for feature selection: A filter based approach. *2012 IEEE Congress on Evolutionary Computation*, 1-8.
- Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical science*, 379-393.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1), 15-18.
- Cook, R. D. (1977). The Editor Technometrics. *Technometrics*, 19(3), 349-350.

- Cook, R. D. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Darıca, C., & Gamgam, H. (2014). Regresyonda etkili gözlemleri belirleme yöntemleri ve karşılaştırmaları. Ankara, Türkiye: Gazi Üniversitesi. <https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp> adresinden alındı
- Eberhart, R., & Kennedy, J. (1995). Particle swarm optimization. *Proceedings of the IEEE international conference on neural networks* (s. 1942-1948). Citeseer.
- Faraway, J. J. (2014). *Linear models with R*. Chapman & Hall/CRC.
- Kennedy, J., & Eberhart, R. C. (1997). A discrete binary version of the particle swarm algorithm. *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation* (s. 4104-4108). IEEE.
- Kim, M. G. (2017). A cautionary note on the use of Cook's distance. *Communications for Statistical Applications and Methods*, 317-324.
- Lawrance, A. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 181-189.
- Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning . *Computers & chemical engineering*, 28(9), 1635-1647.
- Marandi, A., Afshinmanesh, F., Shahabadi, M., & Bahrami, F. (2006). Boolean particle swarm optimization and its application to the design of a dual-band dual-polarized planar antenna. *2006 IEEE International Conference on Evolutionary Computation*, 3212-3218.
- Montgomery, D. C., Peck, E. A., & Vinning, G. G. (2012). *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons.
- Neapolitan, R. E. (2004). *Foundations of algorithms using Java pseudocode*. Jones & Bartlett Learning.
- Nurunnabi, A. H. (2014). Procedures for the identification of multiple influential observations in linear regression. *Journal of Applied Statistics*, 41(6), 1315-1331.
- Pedrasa, M. A., Spooner, T. D., & MacGill, I. F. (2009). Scheduling of demand side resources using binary particle swarm optimization. *IEEE Transactions on Power Systems*, 24(3), 1173-1181.

- Poli, R. (2008). Analysis of the publications on the applications of particle swarm optimisation. *Journal of Artificial Evolution and Applications*, 2008.
- Puntanen, S., & Styan, G. P. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician*, 43(3), 153-161.
- Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool*. Springer Science & Business Media.
- Ren, Y., & Bai, G. (2010). Determination of optimal SVM parameters by using GA/PSO. *JCP*, 5(8), 1160-1168.
- RStudio. (2015). Integrated development for R. RStudio. Inc. Boston, MA, USA, 42. <https://rstudio.com/> adresinden alındı
- Sancar, N., & Inan, D. (2020). Identification of influential observations based on binary particle swarm optimization in the cox PH model. *Communications in Statistics-Simulation and Computation*, 1-24.
- Shen, Q., Jiang, J.-H., Jiao, C.-X., Shen, G.-l., & Yu, R.-Q. (2004). Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *European Journal of Pharmaceutical Sciences*, 22(2-3), 145-152.
- Taşgetiren, M. F., & Liang, Y.-C. (2003). A binary particle swarm optimization algorithm for lot sizing problem. *Journal of Economic and Social Research*, 5(2), 1-20.
- Velleman, P. F., & Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*(4), 234-242.
- Wang, W., Chow, S. C., & Wei, W. W. (1995). On likelihood distance for outliers detection. *Journal of biopharmaceutical statistics*, 5(3), 307-322.
- Zhang, C., & Hu, H. (2005). Using PSO algorithm to evolve an optimum input subset for a SVM in time series forecasting. *2005 IEEE international conference on systems, man and cybernetics*, 4, 3793-3796.

EKLER

EK1: Hills Yarışları Verisi

Gözlem numaraları	Yarış adı	Record times (dk)	uzaklık (mil)	yükseklik (fit)
1	Greenmantle	16.083	2.5	650
2	Carnethy	48.350	6	2500
3	Craig Dunain	33.650	6	900
4	Ben Rha	45.600	7.5	800
5	Ben Lomond	62.267	8	3070
6	Goatfell	73.217	8	2866
7	Bens of Jura Fell	204.617	16	7500
8	Cairnpaple	36.367	6	800
9	Scotly	29.750	5	800
10	Traprain Law	39.750	6	650
11	Lairig Ghru Fun Run	192.667	28	2100
12	Dollar	43.050	5	2000
13	Lomonds of Fife	65.000	9.5	2200
14	Cairn Table	44.133	6	500
15	Eildon Two	26.933	4.5	1500
16	Cairngorm	72.250	10	3000
17	Seven Hills of Edinburgh	98.417	14	2200
18	Knock Hill	78.650	3	350
19	Black Hill	17.417	4.5	1000
20	Creag Beag	32.567	5.5	600
21	Kildoon Hill	15.905	3	300
22	Meall Ant-Suidhe	27.900	3.5	1500
23	Half Ben Nevis	47.633	6	2200
24	Cow Hill	17.933	2	900
25	North Berwick Law	18.683	3	600
26	Creag Dubh	26.217	4	2000
27	Burnswark	34.433	6	800
28	Largo Law	28.567	5	950
29	Criffel	50.500	6.5	1750
30	Achmony	20.950	5	500
31	Ben Nevis	85.583	10	4400
32	Knockfarrel	32.383	6	600
33	Two Breweries	170.250	18	5200
34	Cockleroi	28.100	4.5	850
35	Moffat Chase	159.833	20	5000

EK 2: CYG OB1 Yıldız Kümesi Verisi

Gözlem numarası	logst(x)	logli(y)
1	4.37	5.23
2	4.56	5.74
3	4.26	4.93
4	4.56	5.74
5	4.30	5.19
6	4.46	5.46
7	3.84	4.65
8	4.57	5.27
9	4.26	5.57
10	4.37	5.12
11	3.49	5.73
12	4.43	5.45
13	4.48	5.42
14	4.01	4.05
15	4.29	4.26
16	4.42	4.58
17	4.23	3.94
18	4.42	4.18
19	4.23	4.18
20	3.49	5.89
21	4.29	4.38
22	4.29	4.22
23	4.42	4.42
24	4.49	4.85
25	4.38	5.02
26	4.42	4.66
27	4.29	4.66
28	4.38	4.90
29	4.22	4.39
30	3.48	6.05
31	4.38	4.42
32	4.56	5.10
33	4.45	5.22
34	3.49	6.29
35	4.23	4.34
36	4.62	5.62
37	4.53	5.10
38	4.45	5.22
39	4.53	5.18
40	4.43	5.57
41	4.38	4.62
42	4.45	5.06
43	4.50	5.34
44	4.45	5.34
45	4.55	5.34
46	4.45	4.98
47	4.42	4.50

ÖZGEÇMİŞ

GÖKÇE DELİORMAN

Kişisel Bilgiler

Doğum Tarihi : 12.09.1992

Doğum Yeri : İSTANBUL

İletişim Bilgileri

E-posta adresi : gokcedeliorman@gmail.com

Tel (cep) : 05558737100

Adres : Muhtar Sokak No: 3 Kavacık - Beykoz / Istanbul

Öğrenim Bilgisi

Lise : Dede Korkut Anadolu Lisesi
(2006-2010)

Lisans : Anadolu Üniversitesi
(2010-2015) Fen Fakültesi/Matematik Bölümü

Yüksek Lisans : Marmara Üniversitesi
(2017-) Fen Bilimleri Enstitüsü/ Matematik (YL Tezli)

Görevler

- Matematik Öğretmeni** : Özel Onadım Eğitim Bilimleri Ortaokulu
(2016-2018)
- Matematik Öğretmeni** : Oğuzkaan Koleji
(2018-2019)
- Araştırma Görevlisi** : Beykoz Üniversitesi
(2019-)

Eserler

Uluslararası hakemli dergilerde yayımlanan makaleler:

1. DELIORMAN GÖKÇE, INAN DENİZ Binary Particle Swarm Optimization as a Detection Tool for Influential Subsets in Linear Regression. JOURNAL OF APPLIED STATISTICS, Doi:10.1080/02664763.2020.1779196 (Yayın No: 6251790)
2. ÇEVİK MEHTAP, DARTAN GÜLER, ÜLKER MELİKE, BEZCİ KIVANÇ, DELIORMAN GÖKÇE, SÜSLEYİCİ DUMAN BELGİN, ÇAGATAY PENBE, AKSU MEHMET BURAK, CAN ZEHRA SEMRA, LAÇIN TUNÇ, YURDUN TÜRKAN, KESKİN YASAR (2020). Evaluation of Cytotoxicity and Mutagenicity of Wastewater from Istanbul: Data from Hospitals and Advanced Wastewater Treatment Plant. BULLETIN OF ENVIRONMENTAL CONTAMINATION AND TOXICOLOGY, Doi:10.1007/s00128-020-02853-6 (Yayın No: 6238225)

Uluslararası bilimsel toplantılarda sunulan vebildiri kitaplarında (proceedings) basılan bildiriler:

1. INAN DENİZ, DELIORMAN GÖKÇE (2019). Binary Particle Swarm Optimization as a Detection Tool for Influential Subsets in Linear Regression. 11 th. International Statistics Congress ISC2019 (Özet Bildiri/Sözlü Sunum) (Yayın No:5332820)