YARI PARAMETRİK REGRESYON MODELİNDE ETKİLİ GÖZLEM ANALİZİ

ANALYSIS OF INFLUENTIAL OBSERVATION IN SEMIPARAMETRIC REGRESSION MODEL

SEMRA TÜRKAN

Hacettepe Üniversitesi
Lisansüstü Eğitim – Öğretim ve Sınav Yönetmeliğinin
İSTATİSTİK Anabilim Dalı için Öngördüğü
DOKTORA TEZİ
olarak hazırlanmıştır.

2012

Fen Bilimleri Enstitüsü Müdürlüğü'ne,

Bu çalışma jürimiz tarafından **İSTATİSTİK ANABİLİM DALI 'nda DOKTORA TEZİ** olarak kabul edilmiştir.

Başkan	:…..... Prof. Dr. Mehmet Akif BAKIR
Üye (Danışman)	:…..... Prof. Dr. Öniz TOKTAMIŞ
Üye	:…..... Prof. Dr. Hüseyin TATLIDİL
Üye	:…..... Doç. Dr. Meral Candan ÇETİN
Üye	:…..... Doç. Dr. Serpil Gökçe CULA

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından ...../...../2012 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ...../...../2012 tarihinde kabul edilmiştir.

Prof. Dr. Fatma SEVİN DÜZ
Fen Bilimleri Enstitüsü Müdürü

YARI PARAMETRİK REGRESYON MODELİNDE ETKİLİ GÖZLEM ANALİZİ

Semra Türkan

ÖZ

Tez çalışmasının amacı, doğrusal regresyon modelinde etkili gözlemleri belirlemek için önerilen ölçütlerin yarı parametrik regresyon modeli için geliştirilmesi ve geliştirilen ölçütlerin etkili gözlemleri belirlemede başarılı olup olmadıklarının gerçek bir veri kümesi kullanılarak ve simülasyon çalışması yapılarak incelenmesidir.

Bu çalışmada yarı parametrik regresyon modelinde parametreleri tahmin etmek için kullanılan yöntemler tanıtılmış, sonra doğrusal regresyon modelinde etkili gözlemleri belirlemek için son yıllarda sıkça kullanılan Cook Uzaklığı, Hadi'nin Ölçütü, Pena'nın Ölçütü ve COVRATIO gibi ölçütler incelenmiş ve bu ölçütlerden Hadi'nin Ölçütü, Pena'nın Ölçütü ve COVRATIO ölçütü yarı parametrik regresyon modeli için geliştirilmiştir.

Gerçek bir veri kümesi ve yapay bir veri kümesi kullanılarak önerilen ölçütlerin etkili gözlemleri ortaya çıkarmadaki başarıları araştırılmıştır. Simülasyon çalışması yapılarak farklı örneklem büyüklüklerinde önerilen ölçütlerin etkili gözlemleri ortaya çıkarmadaki başarıları incelenmiştir ve karşılaştırılmıştır.

Anahtar Kelimeler: Yarı parametrik regresyon modeli, etkili gözlemler, Cook Uzaklığı, Hadi'nin Ölçütü, Pena'nın Ölçütü, COVRATIO

Danışman: Prof.Dr. Öniz Toktamış, Hacettepe Üniversitesi, İstatistik Bölümü.

ANALYSIS OF INFLUENTIAL OBSERVATION IN SEMIPARAMETRIC REGRESSION MODEL

Semra Türkan

ABSTRACT

The aim of thesis is to develope diagnostics, used to detect influential observations in linear regression model, for semiparametric regression model and examine whether the developed diagnostics are successful or not in determining influential observations using a real data set and simulation.

In this study, methods used for estimation of parameters in semiparametric regression are introduced, then, diagnostics such as Cook's Distance, Hadi's diagnostic, Pena's diagnostic and COVRATIO which are frequently used in recent years are examined and Hadi's diagnostic, Pena's diagnostic and COVRATIO are developed for semiparametric regression model.

It is investigated whether developed diagnostics detect or not the influential observations by using real data and artificial data. It is examined and compared the success of developed diagnostics in determining influential observations in different sample sizes via simulation study.

Key Words: Semiparametric regression model, influential observations, Cook's Distance, Hadi's diagnostic, Pena's diagnostic, COVRATIO

Advisor: Prof Dr. Öniz Toktamış, Hacettepe University, Depertmant of Statistics.

TEŞEKKÜR

Çalışmamın her aşamasında bilgisi ve manevi desteği ile her zaman yanımda olan, katkı ve eleştirileri ile çalışmama yön veren değerli danışmanım Sayın Prof.Dr. Öniz TOKTAMIŞ'a teşekkür ederim. Ayrıca değerli katkı ve görüşleri için Prof.Dr. Hüseyin TATLIDİL'e ve Prof. Dr. M.Akif BAKIR'a teşekkür ederim.

Tez çalışması süresince olumlu katkı ve görüşleri ve değerli yönlendirmeleri için Doç.Dr. Meral ÇETİN'e ve Doç.Dr. Serpil CULA'ya teşekkür ederim.

Çalışmamın her aşamasında manevi desteği ile bana güç veren değerli arkadaşım Arş. Gör. Esra POLAT'a, bilgisini ve manevi desteğini hiç esirgemeyen değerli arkadaşım Öğr. Gör. Gamze ÖZEL'e içtenlikle teşekkür ederim.

Tez çalışmam süresince gösterdikleri sevgi ve sabırla her zaman yanımda olan AİLEM'e teşekkür ederim.

İÇİNDEKİLER DİZİNİ

Sayfa

ŞEKİLLER DİZİNİ

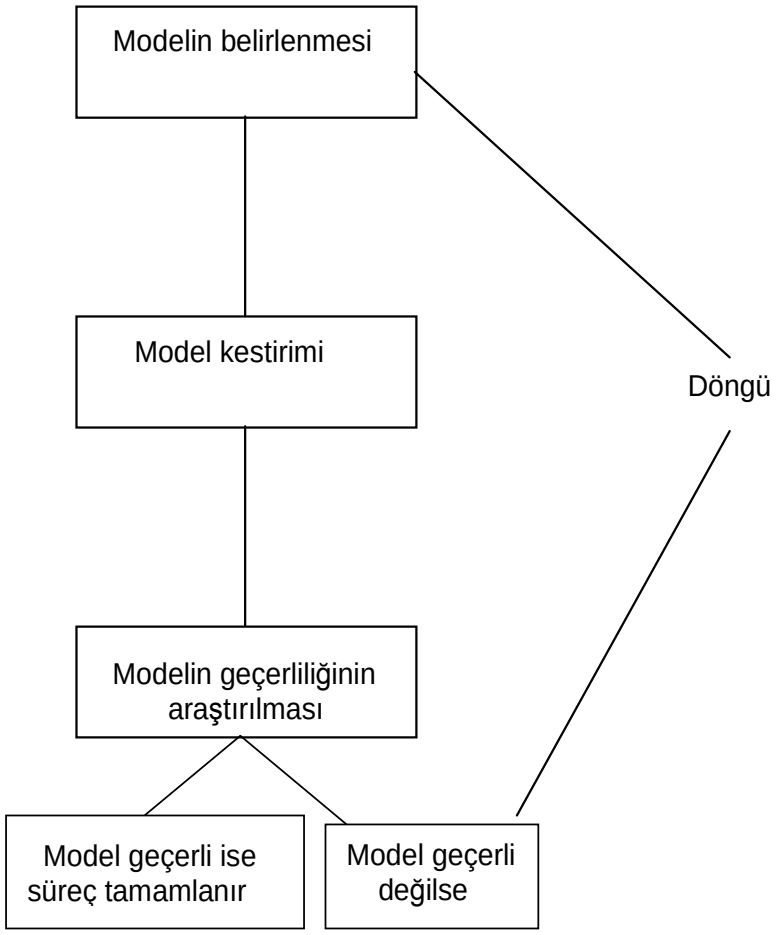
Sayfa

ÇİZELGELER DİZİNİ

Sayfa

1.GİRİŞ

Veriyi modelleme, modern istatistiksel analizin ayrılmaz bir parçasıdır. İstatistiksel modelleme süreci aşağıdaki adımlardan oluşmaktadır. Birinci adım modelin belirlenmesi, ikinci adım modelin kestirimi, üçüncü adım ise modelin geçerliliğinin araştırılmasıdır. Modelleme sürecinin adımları Şekil 1.1'de aşağıdaki gibi gösterilebilir:



Şekil 1.1. İstatistiksel modelleme süreci

Şekil 1.1'de modelin belirlenmesi aşamasında, ilgilenilen veri kümesine uygun olabilecek modeller seçilir. Model kestirimi aşamasında, seçilen modeldeki bilinmeyen parametrelerin kestirimleri veri kümesinden yararlanılarak elde edilir ve bulunan kestirim değerleri modelde yerine konulur. Modelin geçerliliğinin araştırılması aşamasında belirlenen ve kestirilen modelin veriye uygun olup olmadığı araştırılır. Kestirilen modelin veriye uygun olup olmadığı araştırılırken etki

analizi yapılır. Etki analizi yardımı ile analiz sonuçlarını etkileyen gözlemler (aykırı değer, etkili gözlem ve büyük kaldıraç değeri) olup olmadığı araştırılır (Dillane, 2005;. Türkan, 2008). Doğrusal regresyon analizinde etki analizi ile ilgili birçok kitap ve makale vardır. Cook (1977), Chatterjee ve Hadi (1986), Hadi (1992) ve Pena (2005)'in makaleleri, Belsley vd. (1980) ve Cook ile Weisberg (1982)'in kitapları bunlardan bazılarıdır. Yapılan çalışmalarda, doğrusal regresyon modelinde analiz sonuçlarını etkileyen gözlemleri ortaya çıkarmak için çeşitli ölçütler geliştirilmiştir. Bu ölçütlerden en çok kullanılanı Cook uzaklığı ölçütüdür. Son yıllarda Hadi'nin ölçütü ve Pena'nın ölçütü de etkili gözlemleri belirlemede sıkça kullanılmaktadır. Parametrik olmayan regresyon modellerinde ve yarı parametrik regresyon modellerinde etki analizi ile ilgili doğrusal regresyon modelinde olduğu kadar çok sayıda çalışma yoktur. Özellikle yarı parametrik regresyon modellerinde etki analizi ile ilgili çalışmalar son on yılda başlamıştır. Parametrik olmayan ve yarı parametrik regresyon modellerinin kestiriminde kullanılan pürüzlülük ceza yaklaşımında hata kareler toplamına bir ceza fonksiyonu eklenir. Ceza fonksiyonunun eklenmesindeki amaç esnek eğimli uyumlar ile sabit eğimli uyumlar arasında bir uzlaşma sağlamaktır. Bu uzlaşma düzleştirme parametresi ile belirlenir ve düzleştirme parametresinin seçimi pratikte zor bir problemdir (Tabakan, 2009).

Bu tez çalışmasında hem parametrik hem de parametrik olmayan bileşenleri içeren yarı parametrik regresyon modelinde etki analizi incelenmiştir. Yarı parametrik regresyon modeli, doğrusal parametrik bileşen ve parametrik olmayan bileşenlerin her ikisini de içerdiğinden, bu modele kısmi doğrusal model (partially linear model) de denir. Yarı parametrik regresyon modeli son yıllarda çok kullanılan istatistiksel bir modeldir. Çünkü bu model hem parametrik kısmı, hem de parametrik olmayan kısmı birleştirdiğinden doğrusal regresyon modeline göre çok daha esnektir. Yarı parametrik regresyon modelinde etki analizi ile ilgili literatürde sınırlı sayıda çalışma vardır. Bu tezin amacı, yarı parametrik regresyon modelinde analiz sonuçlarını etkileyen gözlemleri ortaya çıkarmak için doğrusal regresyon modelindekine benzer ölçütler geliştirmektir.

Bu çalışma altı bölümden oluşmaktadır. Birinci bölüm olarak ele alınan giriş bölümünde tezin konusu, önemi, bu konuda yapılan önceki çalışmalar, tezin içeriği ve izlenecek düzen ana hatları ile verilmiştir.

İkinci Bölüm'de parametrik, parametrik olmayan ve yarı parametrik regresyon modelleri hakkında genel bilgiler verilmiş ve regresyonda düzleştirme kavramı ve pürüzlülük ceza yaklaşımı açıklanmıştır. Bu bölümde düzleştirme parametresinin seçimi için kullanılan yöntemler üzerinde de durulmuştur.

Üçüncü Bölüm'de doğrusal regresyon modelinde etki analizi üzerinde durulmuş, regresyon modelinde analiz sonuçlarını etkileyen gözlemleri ortaya çıkarmak için önerilen ölçütlerden yaygın olarak kullanılan Cook uzaklığı, Hadi'nin ölçütü, Pena'nın ölçütü ve COVRATIO ölçütü tanıtılmıştır.

Dördüncü Bölüm'de yarı parametrik regresyon modelinde yerel polinom düzleştiricisi kullanıldığında etkili gözlemleri ortaya çıkarmak için önerilen Cook uzaklığı ölçütü incelenmiştir. Ayrıca bu bölümde daha önce yarı parametrik regresyon modellerinde uygulanmamış olan doğrusal modellerdeki etkili gözlemleri ortaya çıkarmak için önerilen Hadi'nin ölçütü, Pena'nın ölçütü ve COVRATIO ölçütü, yarı parametrik regresyon modellerinden yerel polinom düzleştiricisi için geliştirilmiştir.

Beşinci Bölüm'de ilk olarak gerçek bir veri kümesi kullanılarak yarı parametrik regresyon modelinde yerel polinom düzleştiricisi için literatürde var olan Cook uzaklığı ölçüt değerleri ve bu tez çalışmasında geliştirilen Hadi, Pena ve COVRATIO ölçüt değerleri elde edilmiştir. İncelenen veri kümesi için geliştirilen ölçütlerin etkili gözlemleri ortaya çıkarmada başarılı olup olmadıkları araştırılmıştır. Daha sonra hem büyük kaldıraç değeri hem aykırı değer olan gözlemlerin olduğu yapay bir veri kümesi türetilmiştir. Doğrusal regresyonda bu gözlemleri ortaya çıkarmada özellikle büyük örneklemlerde diğer ölçütlere göre daha başarılı olan Pena ölçütünün yarı parametrik regresyon modeli için de aynı sonucu verip vermediği araştırılmıştır. Yapılan simülasyon çalışması ile farklı örneklem büyüklüklerinde oluşturulan etkili gözlemlerin önerilen ölçütler tarafından

belirlenme yüzdeleri elde edilmiştir. Ölçütler etkili gözlemleri belirleme yüzdelerine göre karşılaştırılmıştır.

Altıncı Bölüm'de ise bir önceki bölümde elde sonuçlar tartışılmıştır. Yapılan çalışmalar sonucunda yarı parametrik regresyon modelinde polinom düzleştiricisi kullanıldığında büyük örneklemlerde hem büyük kaldıraç değeri hem de aykırı değer olan gözlemleri ortaya çıkarmada geliştirilen Pena ölçütünün diğer ölçütlere göre daha başarılı olduğu görülmüştür. Diğer durumlarda önerilen ölçütler, veri kümelerinde oluşturulan etkili gözlemleri belirlemede birbirine benzer sonuçlar vermiştir.

2. REGRESYON MODELLERİ

Birçok bilim dalı için iki değişken arasındaki ilişki önemli bir konudur. İki değişken arasındaki ilişkiyi açıklamak için çok sayıda model geliştirilmiştir. Regresyon modeli, değişkenler arasındaki ilişkiyi açıklamak için geliştirilen bir modeldir. Regresyon modeli ile bağımsız değişkenlerin verilen özel değerleri için bağımlı değişkenin ortalama değeri tahmin edilmeye çalışılır. Regresyon modeli genel olarak

$$y_i = E(y_i | \mathbf{x}_i) + \varepsilon_i, \qquad i = 1,...,n \tag{2.1}$$

şeklinde yazılabilir. Bu eşitlikte y_i, i.bağımlı değişkeni, $\mathbf{x}_i$, kx1 boyutlu i. gözlem için bağımsız değişkenler vektörünü, $E(y_i | \mathbf{x}_i)$, $\mathbf{x}_i$ bilindiğinde y_i'nin koşullu beklenen değerini ve ε_i hatayı göstermektedir.

Bu bölümde bağımlı ve bağımsız değişkenler arasındaki fonksiyonel bağımlılığı belirleyen parametrik regresyon, parametrik olmayan regresyon ve yarı-parametrik regresyon yöntemleri tanıtılmış, düzleştirme kavramı, pürüzlülük ceza yaklaşımı ve düzleştirme yöntemleri incelenmiştir.

2.1. Parametrik Regresyon

Parametrik regresyon yaklaşımında, bağımlı ve bağımsız değişkenler arasındaki ortalama ilişki, matematiksel yapısı bilinen bir fonksiyonla ifade edilir ve bu fonksiyonda yer alan parametreler veriden tahmin edilip, model denkleminde yerine konularak model belirlenir. Yaygın olarak kullanılan modellerden birisi doğrusal modeldir. Doğrusal regresyon modelinde, bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkinin parametrelere göre doğrusal olduğu varsayılır. Bilindiği gibi Eş. (2.1)'deki model klasik doğrusal regresyon modeli olarak matris vektör biçiminde

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

şeklinde yazılır. Eş. (2.2)'de **y**, nx1 boyutlu gözlemler vektörünü, **X**, nxp boyutlu tasarım matrisini, **β**, px1 boyutlu parametreler vektörünü ve **ε**, nx1 boyutlu hatalar vektörünü göstermektedir. Doğrusal regresyon modelini kestirebilmek için Gauss-Markov varsayımları olarak bilinen bazı varsayımların sağlanması gerekir. Bu varsayımlar, ε_i'lerin ortalamasının sıfır ve tüm i'ler için ε'nun sabit varyansa sahip olması, x_i'nin sabit olması, x_i ile ε_i arasında ilişki olmaması, ε_i'ler arasında ilişki olmamasıdır. Gauss-Markov varsayımları altında en küçük kareler kestiricisi $\hat{\boldsymbol{\beta}}$, en iyi doğrusal yansız tahmin edicidir (Best Linear Unbiased Estimator; BLUE). Bu varsayımlar sağlanmadığında regresyon analizi sonuçları güvenilir olmayacaktır. Varsayımlar sağlansa bile veri noktalarının saçılım grafiği doğrusal bir yapı göstermiyorsa, doğrusal regresyon modeli ilgilenilen veri kümesi için uygun olmayacaktır. Regresyonda doğrusal olmamanın üstesinden gelmenin en yaygın yolu, yüksek dereceli polinomların kullanılmasıdır. Bu modeller, polinom regresyon modelleri olarak bilinmektedir. p. dereceden polinom regresyon modeli,

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i \qquad (2.3)$$

şeklinde ifade edilir. Eş. (2.3)'deki modelde polinom derecesi arttıkça veriyi temsil edebilecek uygun polinomun seçilmesi zor ve zaman alıcıdır.

Doğrusal regresyon ve polinom regresyon modelleri parametrelere göre doğrusal modellerdir. Değişkenler arasındaki ilişkinin fonksiyonel yapısını temsil eden, parametrelere göre doğrusal olmayan modeller de bulunmaktadır. Bu modeller fonksiyonel yapı bilindiği için parametrik modellerdir. Ancak bu modellerde parametre kestirimleri iteratif yöntemler ile elde edildiği için zordur. Bu durumda parametrik olmayan regresyon yöntemleri kullanılabilir (Shi, 2009).

2.2. Parametrik Olmayan Regresyon

Regresyon modelinin fonksiyonel yapısının bilindiğini varsaymak yerine daha iyi bir yaklaşım, uygun fonksiyonel yapıyı verilerden kestirmektir. Fonksiyonel yapıyı verilerden tahmin etmek için genel tahminler yerine yerel tahminler kullanılır. Parametrik olmayan regresyon yöntemi ile, bağımlı değişken ve bağımsız

değişken arasındaki doğrusal olmayan ilişkinin fonksiyonel yapısı belirtilmeden veriye uygun bir model, yerel kestirimlerin bir dizisi olarak elde edilir (Keele, 2008). Bir y bağımlı değişkeni ile yalnız bir x bağımsız değişkeninin bulunduğu basit parametrik olmayan regresyon modeli,

$$y_i = m(x_i) + \varepsilon_i, \quad i=1,\ldots,n \tag{2.4}$$

şeklinde gösterilir. Bu eşitlikte m(.), x ve y arasındaki fonksiyonel ilişkiyi gösterir ve belirgin bir şekle sahip olmayan bilinmeyen bir fonksiyondur. Parametrik olmayan regresyon yönteminin amacı, m(.) fonksiyonunun düzgün (smooth) ve sürekli olduğu, hata terimleri ε_i'lerin ortalaması 0 ve varyansı σ^2 olan özdeş bir dağılıma sahip olduklarını varsayarak, m(.) fonksiyonunu verilerden tahmin etmektir.

Parametrik olmayan basit regresyon modeli, genel olarak "saçılım grafiği düzleştiricisi" (scatterplot smoothing) olarak da adlandırılır. Çünkü, parametrik olmayan basit regresyon yöntemi ile saçılım grafiğindeki noktalar olabildiğince düzgün bir eğri ile temsil edilmeye çalışılır.

İki ya da daha fazla bağımsız değişken olduğunda, parametrik olmayan regresyon modelini uydurmak ve geometrik olarak göstermek zordur. Bu sorunu çözmek için daha kısıtlayıcı modeller geliştirilmiştir. Toplamsal regresyon modeli, bu modellerden birisidir ve

$$y_i = \alpha + m_1(x_{i1}) + m_2(x_{i2}) + ,...,+ m_k(x_{ik}) + \varepsilon_i, \quad i=1,\ldots,n \tag{2.5}$$

şeklinde gösterilir. Eş. (2.5)'de m_j (.) fonksiyonlarına kısmi regresyon fonksiyonları denir ve düzgün oldukları varsayılır (Fox, 2002).

Parametrik olmayan regresyon modeli, regresyon fonksiyonunda doğrusallık varsayımı yerine regresyon fonksiyonunun düzgünlük varsayımını dikkate aldığı için doğrusallık varsayımını esnetmiş olur. Doğrusallık varsayımının esnetilmesi daha çok hesaplama gerektirir ve bazı durumlarda yorumlanması zor sonuçlar

elde edilir. Ancak regresyon fonksiyonu daha doğru bir şekilde kestirilir (Keele, 2008; Fox, 2002).

Literatürde önerilen birçok parametrik olmayan regresyon modeli ya da düzleştiriciler vardır: Bunlar çekirdek (Kernel) düzleştiriciler, eğrisel çizgi (Spline Smoothing) düzleştiriciler, cezalandırılmış eğrisel çizgi düzleştiriciler vb. Eğrisel çizgi ve cezalandırılmış eğrisel çizgi düzleştiricilerinin temeli, pürüzlülük ceza yaklaşımına dayanır. Bu nedenle bundan sonraki alt bölümde regresyonda düzleştirme kavramı ve pürüzlülük cezası yaklaşımı hakkında genel bir bilgi verilmiştir.

2.2.1. Düzleştirme Kavramı ve Pürüzlülük Cezası Yaklaşımı

Regresyon analizinin amacı, bilinmeyen regresyon fonksiyonu için uygun bir tahmin elde etmektir. Birçok durumda yüzlerce noktadan oluşan eğrileri çok karmaşık bir şekle sahip olması nedeniyle parametrik modellerle temsil etmek olanaksız hale gelir.

Parametrik olmayan regresyon fonksiyonu eğrisinin düzgün bir şekilde elde edilmesi işlemine düzleştirme (smoothing) denir. Bu nedenle parametrik olmayan regresyon fonksiyonunu tahmin etmek için kullanılan yöntemlere, düzleştirme yöntemleri de denilmektedir.

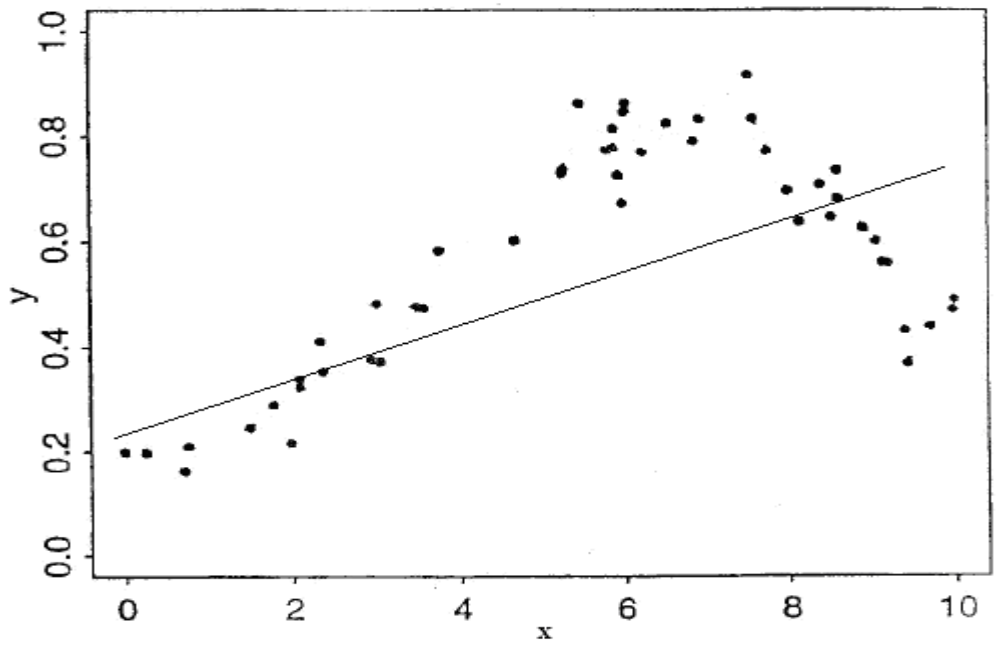
m regresyon fonksiyonunu tahmin etmek için kullanılan klasik yöntemlerden biri doğrusal regresyon yaklaşımıdır. Bu yöntemle, (x_i, y_i), i=1,…,n veri dizisi için artık kareler toplamı (AKT(m))

$$AKT(m)=\sum_{i=1}^{n}(y_i - m(x_i))^2$$

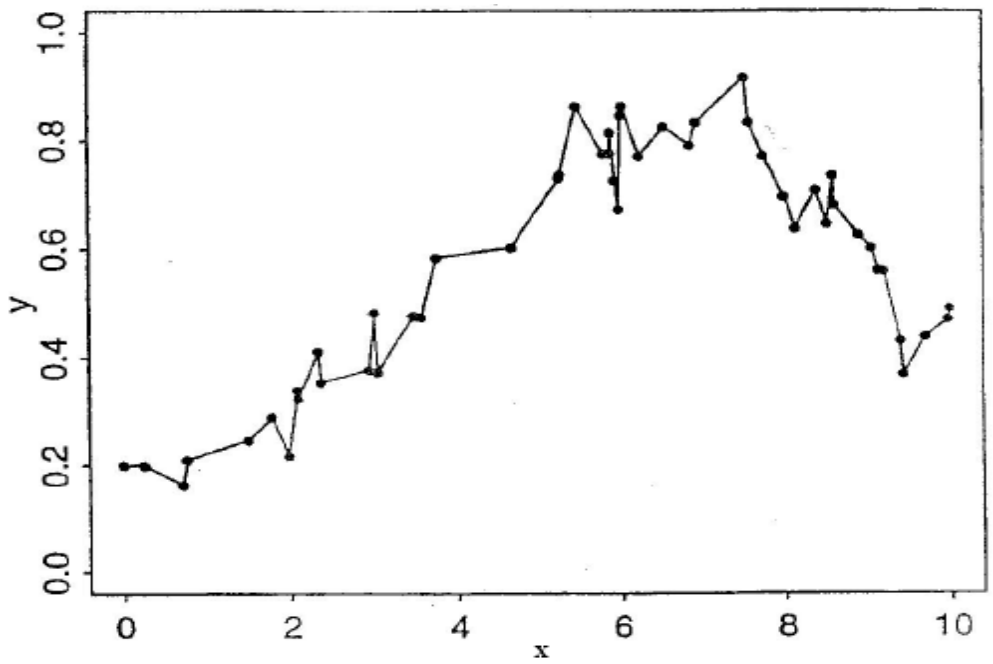
olup, minimum yapılarak m fonksiyonu tahmin edilir. Verilerin saçılım grafiği bir doğru ile temsil edilebilecek gibi ise, m fonksiyonunu tahmin etmek için doğrusal regresyon yaklaşımı uygun olacaktır. Verilerin gerçek yapısı doğrusal değil ise bu yaklaşım başarısız olur. Saçılım grafiği ve uydurulan doğrunun verildiği böyle bir

örnek, Şekil 2.1'de görülmektedir. Bu kestirici için doğru ile birleştirilen noktalarda sabit eğimlilik sağlanır.

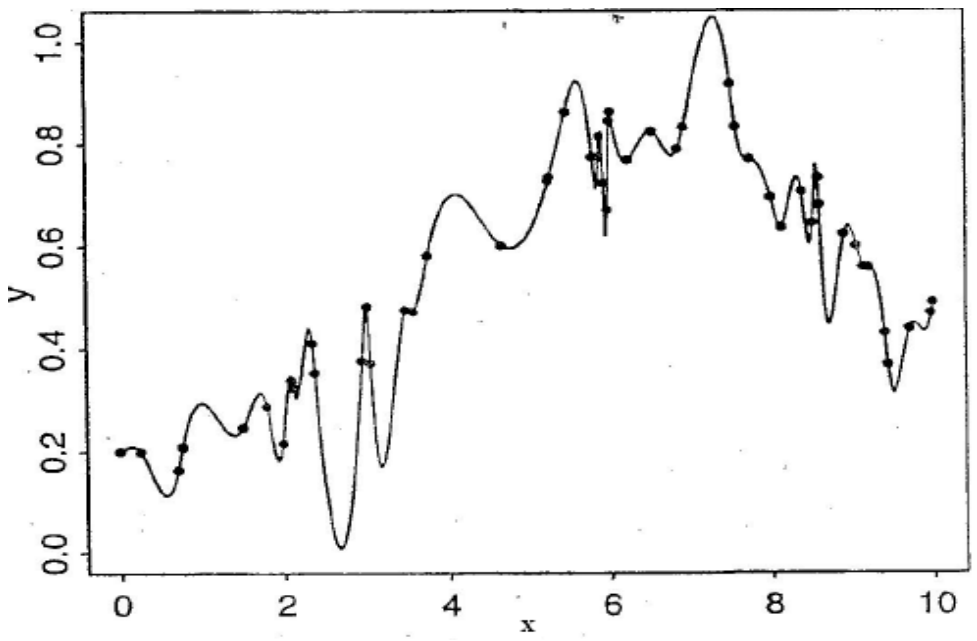
Birçok kestirim problemi için doğrusal modeller verilen veri kümesine uymaz ve bu durumlarda, doğrusal olmayan uyum eğrilerinin oluşturulması zorunluluğu ortaya çıkar (Omay, 2007). Şekil 2.2'de m fonksiyonunu elde etmek için verilen noktalar, doğrular ile birleştirilip, interpolasyon ile m fonksiyonunun tahmini yapılmıştır. Aynı veri noktaları için Şekil 2.3'de m fonksiyonu üzerine düzgünlük kısıtı konularak elde edilen başka bir kestirim verilmektedir. Şekil 2.3'de verilen eğri sürekli ikinci mertebeden türevlere sahiptir ve tüm veri noktalarından geçer. Şekil 2.2 ve Şekil 2.3'deki tahminlerin ikisi de veri noktaları için iyi bir uyum göstermektedir. Tahminler, Şekil 2.2 ve Şekil 2.3'de görüldüğü gibi çok pürüzlüdür. Bu kestiriciler için eğriler ile birleştirilen noktalarda esnek eğimlilik vardır. Her iki kestirim için de AKT(m)=0'dır. Ancak eğri uydurmada tek amaç iyi bir uyum elde etmek değildir. Çoğu kez, buna aykırı düşen veri uydurmadaki başka bir amaç, çok hızlı değişim göstermeyen bir eğri kestirimi elde etmektir. Pürüzlülük cezası yaklaşımındaki ana fikir, hızlı olarak dalgalanan bir eğrinin eğilimini ölçmek ve daha sonra eğri kestiriminde sabit eğimli uyumlar ile esnek eğimli uyumlar arasında gerekli uzlaşmayı sağlayacak biçimde tahmin problemini ortaya koymaktır.



Şekil 2.1. Veri noktalarının saçılım grafiği ve uydurulan doğru (Green ve Silverman, 2000)



Şekil 2.2. Veri noktaları doğrusal çizgiler ile birleştirilerek elde edilen tahmin (Green ve Silverman, 2000)



Şekil 2.3. Veri noktaları eğriler ile birleştirildiğinde elde edilen tahmin (Green ve Silverman, 2000)

Bir $[a,b]$ aralığında tanımlanan m eğrisinin ne kadar pürüzlü ya da ne kadar dalgalı olduğunu ölçmenin birçok yolu vardır. İkinci mertebeden türevi alınabilen m eğrisinin pürüzlülüğünü ölçmenin en yaygın yolu m fonksiyonunun ikinci

mertebeden türevinin karesinin integrali olan $\int_a^b \{m''(t)\}^2 dt$ integralini hesaplamaktır. Bu ölçüme göre sadece doğrusal m(x) fonksiyonları sıfır pürüzlülüğe sahipken, sürekli fonksiyonlar sınıfındaki tüm fonksiyonlar pozitif pürüzlülüğe sahiptirler. İkinci mertebeden türevin karesinin integrali, önemli hesaplama avantajlarına sahip olduğu için en yaygın kullanılan pürüzlülük ölçüsüdür.

Bir $[a,b]$ aralığında tanımlanan ikinci mertebeden türevi alınabilen herhangi bir m fonksiyonu ve düzleştirme parametresi λ>0 için cezalandırılmış kareler toplamı,

$$S(m) = \sum_{i=1}^{n} \{y_i - m(x_i)\}^2 + \lambda \int_a^b \{m''(t)\}^2 dt \qquad (2.6)$$

şeklinde tanımlanır. Eş. (2.6)'daki ilk terim hata kareler toplamıdır, ikinci terim ise pürüzlülük cezası olarak adlandırılır. Cezalandırılmış en küçük kareler tahmin edicisi $\hat{m}$'yı elde etmek için S(m) fonksiyonu minimum yapılır. Bu işlem, pürüzlülük cezası kısıtı altında y ile m(x) arasındaki farkların kareleri toplamının minimum yapılması problemidir. Pürüzlülük cezası iki kısımdan oluşur. Birincisi düzleştirme parametresi olarak bilinen λ'dır. İkincisi m(x) fonksiyonunun ikinci mertebeden türevinin karesinin integralidir. İkinci mertebeden türev, bir fonksiyonun eğiminin değişim oranını ya da eğriliğini ölçer. Karesi alınmış $m''(t)$ fonksiyonunun tanım aralığındaki integrali, esasen parametrik olmayan tahminin tanım aralığı boyunca eğriliğinin bir ölçüsünü verir. Bu integralin değerinin büyük olması, m(x) fonksiyonunun pürüzlü (dalgalı) olduğunu; küçük olması m(x) fonksiyonunun daha düzgün olduğunu göstermektedir (Green ve Silverman, 2000; Keele, 2008). Negatif değer almayan λ düzleştirme parametresi ise, m fonksiyonunun ne kadar düzleştiğinin bir ölçüsü olan ikinci mertebeden türeve verilen ağırlığın büyüklüğünü kontrol eder. λ parametresi 0 dan ∞'a değerler alır. λ=0 olması esnek eğilimli bir model uyumunu, $\lambda \to \infty$ olması doğrusal regresyon modeline uygunluğu gösterir (Keele, 2008). Eğer λ büyük ise, S(m)'deki asıl bileşen pürüzlülük cezası terimi olacaktır ve S(m)'yi minimum yapan m fonksiyonunun tahmini çok az eğrilik gösterecektir. Eğer λ küçük ise S(m)'deki asıl bileşen hata kareler toplamı

olacaktır ve bu durumda m fonksiyonunun tahmini, verinin saçılım grafiğindeki noktalara yakın bulunacaktır (Green ve Silverman, 2000).

2.3. Düzleştirme Yöntemleri

Bağımlı değişken y ve bağımsız değişken x arasındaki ilişkiyi veren bilinmeyen m(x) fonksiyonunun düzgün bir şekilde tahmin edilmesi için kullanılan yöntemlere düzleştirme yöntemleri denir. Düzleştirme yöntemleri ile parametrik olmayan regresyon yöntemleri aynı anlamda kullanılmaktadır. Bu çalışmada

- Çekirdek (Kernel)
- Yerel (Local) polinom
- Eğrisel çizgi (Spline Smoothing)
- Cezalandırılmış eğrisel çizgi
- Cezalandırılmış eğrisel çizgide karışık doğrusal model yaklaşımı

düzleştirme yöntemleri incelenmiştir.

2.3.1. Çekirdek (Kernel) düzleştirme yöntemi

Düzleştirme yöntemlerinin en basit olanı çekirdek düzleştirme yöntemidir. Tek değişkenli bir dağılımdan alınan x_1, x_2, … , x_n gözlemlerinin bağımsız, özdeş bilinmeyen bir dağılıma sahip olduğu varsayılsın. Bu gözlemlerin alındığı kitlenin dağılımının olasılık yoğunluk fonksiyonu f için ilk çekirdek kestiricisi Rosenblatt (1956) tarafından,

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=}^{n} K\left(\frac{x_i - x}{h}\right) \qquad (2.7)$$

şeklinde önerilmiştir. Olasılık yoğunluk fonksiyonunun çekirdek kestiriminde yer alan K fonksiyonu, çekirdek fonksiyonu olarak adlandırılır. Her bir x noktasındaki çekirdek kestirimi Eş. (2.7)'de görüldüğü gibi bir ağırlıklı ortalamadır ve çekirdek fonksiyonu K, ağırlıkların hesaplanmasında yardımcı olmakta bant genişliği h ise çekirdek fonksiyonunun dağılımını belirleyen bir ölçekleme etkeni rolünü

oynamaktadır. Olasılık yoğunluk fonksiyonunun çekirdek kestirimi, K ve h nın farklı seçimlerine göre değişmektedir Çekirdek fonksiyonu ve bant genişliği araştırmacı tarafından seçilmektedir (Gökmen, 2002; Cula, 1998).

Çekirdek fonksiyonları genellikle sıfır merkezli, simetrik bir olasılık yoğunluk fonksiyonu olarak seçilmektedir. Uygulamada çekirdek fonksiyonunun seçimi bant genişliği seçimi kadar önemli değildir. Birçok çekirdek fonksiyonu vardır. Uygulamada kullanılan bazı çekirdek fonksiyonları Çizelge 2.1'de verilmektedir:

Çizelge 2.1. Çekirdek Fonksiyonları

Gaussian çekirdek	$K(u)=\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}u^2}, \quad u\in[-\infty;\infty]$
Tebiçimli çekirdek	$K(u)=\frac{1}{2}, \quad u\in[-1,1]$
Üçgensel çekirdek	$(1-\lvert u\rvert), \quad u\in[-1,1]$
Epanechnikov çekirdek	$\frac{3}{4}(1-u^2), \quad u\in[-1,1]$
İki Ağırlıklı çekirdek	$\frac{15}{16}(1-u^2)^2, \quad u\in[-1,1]$
Üç Ağırlıklı çekirdek	$\frac{35}{32}(1-u^2)^3, \quad u\in[-1,1]$

Regresyon fonksiyonunun parametrik olmayan kestirim yöntemlerinden biri çekirdek kestirim yöntemidir. Nadaraya ve Watson tarafından önerilen çekirdek kestirimi, bu yöntemlerden biridir.

Koşullu beklenen değer olarak tanımlanan regresyon fonksiyonu,

$$m(x)=E(Y\lvert X=x)=\int y\, f(y\lvert x)dy=\frac{\int y\, f(x,y)dy}{f(x)} \qquad (2.8)$$

şeklinde yazılabilir. Bu koşullu beklenen değer için doğal bir kestirim, pay ve paydanın kestirimlerinin ayrı ayrı elde edilerek Eş. (2.8)'de yerine konulmasıyla

bulunabilir. Pay için genellikle çarpımsal çekirdek yoğunluk fonksiyonu kullanılarak f(x,y) bileşik yoğunluk fonksiyonu kestirilebilmektedir. Çarpımsal çekirdek kestiricisi,

$$\hat{f}_{h_1,h_2}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1}K\left(\frac{x_i - x}{h_1}\right)\frac{1}{h_2}K\left(\frac{y_i - y}{h_2}\right) \tag{2.9}$$

şeklindedir. Buna göre Eş. (2.8)'de verilen ifadenin payının kestirimi,

$$\int y\,\hat{f}(x,y)\,dy = \int y\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1}K\left(\frac{x_i - x}{h_1}\right)\frac{1}{h_2}K\left((\frac{y_i - y}{h_2})dy\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1}K\left(\frac{x_i - x}{h_1}\right)\int\frac{y}{h_2}K\left(\frac{y_i - y}{h_2}\right)dy \tag{2.10}$$

olur. Eş. (2.10)'daki integralde $\left(\frac{y_i - y}{h_2}\right) = t$ dönüşümü yapılarak ve çekirdek fonksiyonunun sağladığı $\int t\,K(t)dt = 0$ ve $\int K(t)dt = 1$ özellikleri göz önüne alınarak Eş. (2.10),

$$\int y\,\hat{f}(x,y)\,dy = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{h_1}K\left(\frac{x_i - x}{h_1}\right) \tag{2.11}$$

olarak bulunur. Buna göre Nadaraya ve Watson tarafından verilen regresyon fonksiyonunun çekirdek kestirimi,

$$\hat{m}_h(x) = \frac{\hat{f}_{h_1,h_2}(x,y)}{\hat{f}_h(x)}$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}y_i\frac{1}{h_1}K\left(\frac{x_i - x}{h_1}\right)}{\hat{f}_h(x)} \tag{2.12}$$

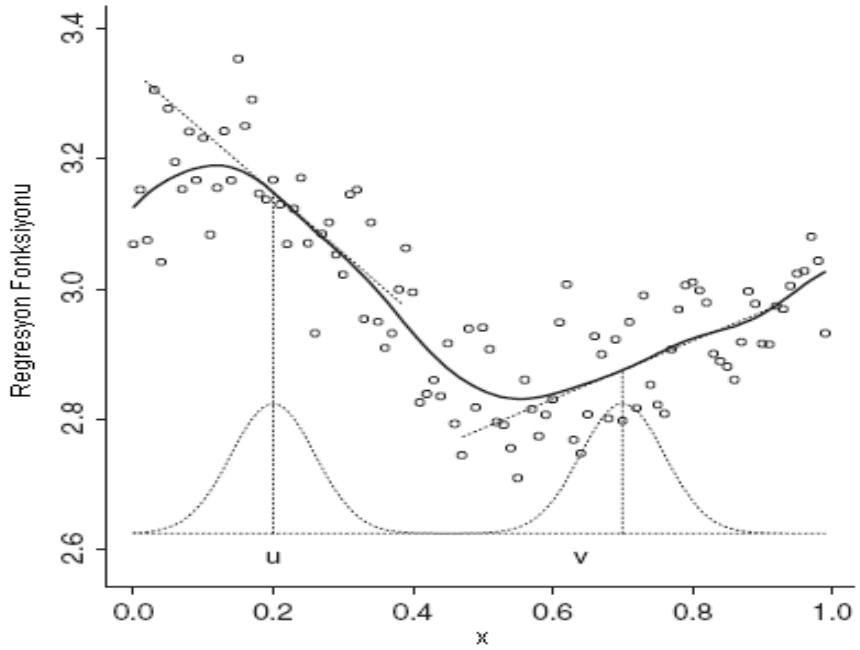
şeklinde elde edilir. h_1=h değişmez bant genişliği kullanılarak Nadaraya-Watson kestiricisi,

$$\hat{m}_h(x) = \frac{\frac{1}{nh}\sum_{i=1}^{n} y_i K\left(\left(\frac{x_i - x}{h}\right)\right)}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\left(\frac{x_i - x}{h}\right)\right)} = \frac{1}{n}\sum_{i=1}^{n} W_i(x) y_i \qquad (2.13)$$

şeklinde yazılabilir. Burada $W_i(x)$ ağırlık fonksiyonudur (Härdle vd., 2004; Demir, 2005). Nadaraya ve Watson tarafından önerilen regresyon fonksiyonunun sabit bir x noktasındaki çekirdek kestirimi, her bir veri noktasında merkezileştirilmiş eşit ölçekli çekirdeklerin o noktada aldıkları değerlerin ağırlıklı bir ortalamasıdır.

2.3.2. Yerel polinom düzleştiricisi

Bir saçılım grafiğini düzleştirmek için kullanılan en yaygın yöntemlerden biri yerel polinom düzleştiricisidir. Yerel polinom regresyonu, çekirdek regresyonuna benzer ancak ağırlıklandırılmış ortalama olan kestirim değerleri yerine ağırlıklandırılmış en küçük kareler kestirim değerlerini kullanır. Buna göre herhangi bir x noktasındaki düzleştirme, ağırlıkların çekirdek fonksiyonunun yüksekliğine karşılık geldiği ağırlıklandırılmış en küçük kareler doğrusu uydurularak elde edilir. Şekil 2.4, yerel doğrusal düzleştirici ile nasıl düzleştirme yapıldığını göstermektedir.



Şekil 2.4 Yerel doğrusal düzleştirici (Ruppert vd., 2003)

Şekil 2.4'de, simülasyon ile elde edilen 100 veri noktası yuvarlaklar ile gösterilmiştir. u'nun komşuluğunda yerel bir doğru uydurulup, çekirdek fonksiyonu yükseklikleri ile ağırlıklandırılmış en küçük kareler yöntemi kullanılarak, x=u noktasındaki tahmin elde edilir. x=v noktasındaki kestirim de benzer şekilde elde edilir. Şekil 2.4'deki grafiğin üst kısmında bulunan kesikli çizgiler, u ve v noktasındaki ağırlıklandırılmış en küçük kareler ile elde edilen regresyon doğrusunu göstermektedir. Çekirdek fonksiyonu da grafiğin alt kısmında kesikli çizgiler ile verilmiştir. Eğer bu işlem birçok x değeri için uygulanırsa Şekil 2.4'de kalın çizgi ile gösterilen yerel doğrusal tahmin eğrisi elde edilir.

Şekil 2.4'de yerel doğrusal kestiriciler kullanılmıştır. Ancak, herhangi bir dereceden polinomlar da kullanılabilir. Yerel doğrusal ve daha yüksek dereceden yerel polinom kestirimleri elde etmek için ilk olarak bilinmeyen koşullu beklenen değer fonksiyonu m(.) için Taylor açılımı bulunur. Buna göre x noktası komşuluğundaki x_i için Taylor açılımı kullanılarak $m(x_i)$ fonksiyonu yaklaşık olarak,

$$m(x_i) \approx m(x) + m^{(1)}(x)(x_i - x) + \ldots + m^{(p)}(x)(x_i - x)^p \frac{1}{p!} \qquad (2.14)$$

şeklinde yazılır. Eş. (2.14)'deki model, $\beta_j = \frac{m^{(j)}(x)}{j!}$ yazılarak

$$m(x_i) = \beta_0 + \beta_1(x_i - x) + \beta_2(x_i - x)^2 + \ldots + \beta_p(x_i - x)^p \qquad (2.15)$$

şeklinde ifade edilebilir. Görüldüğü gibi $m(x_i)$, basit polinom modelidir. Bu ise yerel olarak ağırlıklandırılmış polinom regresyonun kullanılmasını akla getirir.

Bir x noktasının komşu noktalarındaki kestirim, p.dereceden polinom modelinin artık kareler toplamı, çekirdek ağırlıkları $K\{h^{-1}(x_i - x)\}$ ile ağırlıklandırılarak elde edilir. Çekirdek fonksiyonu K(.) genellikle simetrik pozitif fonksiyon olarak alınır. Çekirdek fonksiyonundaki h bant genişliği, yerel polinom düzleştiriciler için düzleştirme parametresidir.

Yerel polinom regresyonunun x noktasındaki ağırlıklandırılmış en küçük kareler kestiricisini elde etmek için

$$\sum_{i=1}^{n}\left\{y_i - \beta_0 - \quad \ldots \quad -\beta_p(x_i - x)^p\right\}^2 K\left(\frac{x_i - x}{h}\right) = (\mathbf{y} - \mathbf{X}_x\boldsymbol{\beta})^T \mathbf{W}_x (\mathbf{y} - \mathbf{X}_x\boldsymbol{\beta}) \qquad (2.16)$$

ifadesini minimum yapan $\hat{\boldsymbol{\beta}}$ değerleri bulunur. Bunun için Eş. (2.16)'daki ifadenin $\boldsymbol{\beta}$'ya göre türevi alınıp sıfıra eşitlenerek $\hat{\boldsymbol{\beta}}$ değerleri

$$\hat{\boldsymbol{\beta}}_x = (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{y} \qquad (2.17)$$

biçiminde elde edilir.. Eş. (2.17)'de $\mathbf{X}_x$ ve $\mathbf{W}_x$ matrisleri

$$\mathbf{X}_x = \begin{bmatrix} 1 & (x_1 - x) & (x_1 - x)^2 & \cdots & (x_1 - x)^p \\ 1 & (x_2 - x) & (x_2 - x)^2 & \cdots & (x_2 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & (x_n - x)^2 & \cdots & (x_n - x)^p \end{bmatrix}$$

$$\mathbf{W}_x = \begin{bmatrix} K_h(x_1 - x) & 0 & \cdots & 0 \\ 0 & K_h(x_2 - x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_h(x_n - x) \end{bmatrix}$$

şeklindedir. Eş. (2.17)'deki ifade Eş.(2.15)'de yerine konularak m(x) fonksiyonunun yerel polinom regresyon kestiricisi,

$$\hat{m}(x) = \mathbf{t}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{y} \qquad (2.18)$$

şeklinde elde edilir. Bu ifadede $\mathbf{t}^T = (1,0,...,0)^T_{1x(p+1)}$ 'dir (Kim vd., 2001; Ruppert vd., 2003).

Uygulamada, Eş. (2.17)'de elde edilen $\hat{\boldsymbol{\beta}}$ değerleri, $m^{(j)}(x)$'in bir kestirimi olan $\hat{m}_j(x) = j!\,\hat{\beta}_j$, j=1,..,p, eşitliğinde yerine konularak $m^{(j)}(x)$'in kestirimi bulunur. Bu değerler de Eş. (2.14)'de yerlerine konularak $\hat{m}(x)$ tahmini elde edilir.

Yerel polinom regresyonunda p=0 olduğunda yerel sabit kestirici ya da Nadaraya-Watson kestiricisi elde edilir. Eş. (2.16)'da p=0 alındığında m(x) fonksiyonunun Nadaraya-Watson kestiricisi

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i K_h(x_i - x)}{\sum_{i=1}^{n} K_h(x_i - x)} \tag{2.19}$$

şeklinde bulunur (Sheather, 2009). Eş. (2.19)'da $K_h(x_i - x) = \frac{1}{n} K\left(\frac{x_i - x}{h}\right)$'dir.

Eş. (2.16)'da p=1 olduğunda m(x) fonksiyonunun yerel doğrusal kestiricisi

$$\hat{m}(x) = n^{-1} \sum_{i=1}^{n} \frac{(s_2(x) - s_1(x)(x_i - x))K_h(x_i - x)y_i}{s_2(x)s_0(x) - s_1(x)^2} \tag{2.20}$$

şeklinde elde edilir. Eş. (2.20)'de $s_r(x) = n^{-1}\sum_{i=1}^{n}(x_i - x)^r K_h(x_i - x)$ 'dir.

Yerel polinom regresyonda daha yüksek dereceden polinom seçimi temel eğriye daha iyi bir yaklaşımın elde edilmesini ve buna bağlı olarak yanın azalmasını sağlar. Ancak, bu da tahminlerde daha fazla değişkenliğin olmasına neden olur. Loader 1992'de yerel polinom regresyonunda polinomun derecesinin seçimi ile ilgili olarak "İyi bir kestirim değeri elde etmek için düşük dereceli polinom seçilmesi ve bant genişliğinin seçimine yoğunlaşılması yeterlidir. Polinom derecesinin seçiminde en yaygın kullanılan polinom dereceleri yerel doğrusal (p=1 olduğu) ve yerel karesel (p=2 olduğu) dir. Yerel sabit kestirim (p=0 olduğu) yana duyarlıdır. Yerel doğrusal kestirim özellikle uç noktalarda genel olarak daha iyi sonuç verir. Yerel karesel kestirim yanı azaltır ancak özellikle uç noktalarda varyansın artmasına neden olabilir. Yerel kübik ve daha yüksek dereceli kestirimler nadiren iyi sonuçlar verir" görüşünü savunmuştur.

Ruppert, Wand ve Carroll (2003) yaptıkları çalışmalara dayanarak, regresyon fonksiyonu monoton artan bir fonksiyon ise yerel doğrusal kestirimi diğer durumlarda yerel karesel kestirimi önermişlerdir (Sheather, 2009).

2.3.3. Eğrisel çizgi düzleştirme (spline smoothing) yöntemi

Eğrisel çizgi (spline) fonksiyonlarının amacı, tanımlanan aralığı gözlem değerleri yardımıyla alt aralıklara bölerek, her bir alt aralıkta farklı bir polinom ile bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellendirerek istenilen dereceden türevi olan sürekli bir fonksiyon elde etmektir. (x_i, y_i), $i = 1,...,n$, gözlem değerleri olsun. Herhangi bir [a,b] aralığında $a = x_o < x_1 < ... < x_n < x_{n+1} = b$ koşulunu sağlayan gerçel x_i reel sayıları verilmiş olsun. $(a, x_1),(x_1, x_2),...,(x_i, x_{i+1}),...,(x_{n-1}, x_n),(x_n, b)$ alt aralıklarının her birinde örneğin $x_i \le x < x_{i+1}$ alt aralığında p.dereceden bir $m_i(x)$ polinomu tanımlansın. Burada x_i noktaları "düğüm noktaları (knots)" olarak adlandırılır. Polinomlardan oluşan ve eğrisel çizgi adı verilen m(x) fonksiyonu, alt aralıklarda tanımlanan p. dereceden polinomların birleşimi olarak tanımlanır (Aydın, 2005; Green ve Silverman, 2000).

$m(x)=m_i(x)$, $x_i \le x \le x_{i+1}$, i=1,…,n

ve

$m_i(x_{i+1})= m_{i+1}(x_{i+1})$, i=1,…,n

olmak üzere $m_i(x)$, $x_i \le x < x_{i+1}$, i=0,…,n, p.dereceden bir polinomdur. p=1 olduğunda parçalı doğrusal çizgi, p=2 olduğunda karesel eğrisel çizgi ve p=3 olduğunda kübik eğrisel çizgi fonksiyonu olarak adlandırılır.

[a,b] aralığı üzerinde tanımlı bir m fonksiyonu aşağıdaki iki koşulu sağlıyorsa kübik eğrisel çizgi fonksiyonudur. Birincisi, $(a, x_1),(x_1, x_2),...,(x_i, x_{i+1}),...,(x_{n-1}, x_n),(x_n, b)$ aralıklarının her birinde m kübik polinomdur; ikincisi m'nin kendisi, birinci ve ikinci mertebeden türevleri her bir x_i düğüm noktalarında, böylece [a,b] aralığında süreklidir. Kübik eğrisel çizgi fonksiyonu,

$$m(x) = m_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x_i \le x \le x_{i+1} \qquad (2.21)$$

biçiminde ifade edilir. Eş. (2.21)'de a_i, b_i, c_i ve d_i sabit değerlerdir. x_0=a ve x_{n+1}=b dır. [a,b] aralığında bir kübik eğrisel çizgi, a ve b uç noktalarında ikinci ve üçüncü mertebeden türevlerinde sıfır değerini alırsa buna *doğal* kübik eğrisel çizgi (natural cubic spline) denir. Bu koşullar doğal sınır koşulları olarak adlandırılır ve $d_0 = c_0 = d_n = c_n = 0$ olduğunu yani m(x) fonksiyonunun [a, x_1] ve [x_n, b] aralığında doğrusal olduğunu ifade eder (Aydın, 2005; Green ve Silverman, 2000).

Eş. (2.21), hem matematiksel irdeleme hem de hesaplama açısından doğal kübik eğrisel çizgi fonksiyonu için uygun gösterim değildir. Bu eşitlik, her bir düğüm noktasında doğal kübik eğrisel çizgi fonksiyonunun değerini ve ikinci mertebeden türevini veren bir fonksiyon olarak belirtilebilir. Bu gösterim ikinci mertebeden türev- değer gösterimi (value-second derivative representation) olarak adlandırılır. m fonksiyonunun $x_1 < \ldots < x_n$ düğümlerinde doğal kübik eğrisel çizgi fonksiyonu olduğu varsayılsın ve

$$m_i = m(x_i) \quad \text{ve} \quad \gamma_i = m''(x_i), \quad i = 1,\ldots,n$$

olarak tanımlansın. Doğal kübik eğrisel çizgi tanımına göre, x_1 ve x_n noktalarında m fonksiyonun ikinci türevi sıfırdır yani $\gamma_1 = \gamma_n = 0$ dır. $\mathbf{m} = (m_1,\ldots,m_n)^T$ ve $\boldsymbol{\gamma} = (\gamma_2,\ldots,\gamma_{n-1})^T$ olsun. **m** ve $\boldsymbol{\gamma}$ vektörleri, m eğrisini tam olarak belirlenmesini sağlar. m eğrisi, herhangi bir x noktasında m 'nin değeri ve türevleri için, **m** ve $\boldsymbol{\gamma}$ vektörleri ile açık olarak formüle edilebilir. Ancak her **m** ve $\boldsymbol{\gamma}$ vektörleri doğal kübik eğrisel çizgi belirtmez. Verilen düğüm noktalarında **m** ve $\boldsymbol{\gamma}$ vektörlerinin kübik eğrisel çizgi belirtmesi için gerek ve yeter koşul **Q** ve **R** gibi iki bant matrisine bağlıdır. Bir matrisin sıfır olmayan elemanlarının hepsi az sayıda köşegen elemanları üzerinde toplanmışsa, bu matrise "bant matris" adı verilir. Matrisin sıfır olmayan köşegen elemanlarının sayısı matrisin bant genişliği olarak adlandırılır. $h_i = x_{i+1} - x_i$ olmak üzere **Q** ve **R** matrislerinin elemanları aşağıdaki gibi tanımlanır:

$$q_{ij}=\begin{cases}\dfrac{1}{h_{j-1}}, & i=j-1\\ -\left(\dfrac{1}{h_{j-1}}+\dfrac{1}{h_j}\right), & i=j\\ 0, & |i-j|\geq 2\\ \dfrac{1}{h_j}, & i=j+1\end{cases} \qquad i=1,2,\ldots,n \text{ ve } j=2,\ldots,n-1$$

$$r_{ij}=\begin{cases}\dfrac{1}{6}h_{j-1}, & i=j-1\\ \dfrac{1}{3}(h_{j-1}+h_j), & i=j\\ 0, & |i-j|\geq 2\\ \dfrac{1}{6}h_j, & i=j+1\end{cases} \qquad i=2,\ldots,n-1 \text{ ve } j=2,\ldots,n-1$$

Buna göre **R** ve **Q** matrisleri aşağıdaki gibi gösterilebilir: **R** ,(n-2)x(n-2) boyutlu simetrik matris, **Q**, nx(n-2) boyutlu matrisdir:

$$\mathbf{R}=\begin{bmatrix}\frac{1}{3}(h_1+h_2) & \frac{1}{6}h_2 & 0 & . & . & . & 0\\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \frac{1}{6}h_3 & 0 & . & . & .\\ 0 & \frac{1}{6}h_3 & \frac{1}{3}(h_3+h_4) & \frac{1}{6}h_4 & 0 & . & .\\ . & . & . & . & . & . & .\\ . & . & . & . & . & . & 0\\ . & . & . & 0 & \frac{1}{6}h_{n-3} & \frac{1}{3}(h_{n-3}+h_{n-2}) & \frac{1}{6}h_{n-2}\\ . & . & . & . & 0 & \frac{1}{6}h_{n-2} & \frac{1}{3}(h_{n-2}+h_{n-1})\end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{h_1} & -(\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & 0 & 0 & . & . & . & 0 \\ 0 & \frac{1}{h_2} & -(\frac{1}{h_2}+\frac{1}{h_3}) & \frac{1}{h_3} & 0 & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & 0 & \frac{1}{h_{n-3}} & -(\frac{1}{h_{n-3}+h_{n-2}}) & \frac{1}{h_{n-2}} & 0 \\ 0 & . & . & . & 0 & 0 & \frac{1}{h_{n-2}} & -(\frac{1}{h_{n-2}+h_{n-1}}) & \frac{1}{h_{n-1}} \end{bmatrix}$$

R ve **Q** matrisleri üç köşegen matrislerdir, başka bir deyişle, indisleri $|i-j| \geq 2$ koşulunu sağlayan elemanları sıfırdır (Aydın, 2005). **R** matrisi her bir i için $|r_{ii}| > \sum_{i \neq j} |r_{ij}|$ ise köşegen dominanttır. Bu özellik **R** matrisinin pozitif tanımlı bir matris olduğunu göstermektedir. Buna göre **R** matrisinin tersi vardır ve

$$\mathbf{K} = \mathbf{Q}\,\mathbf{R}^{-1}\mathbf{Q}^{T} \tag{2.22}$$

olacak biçimde bir **K** matrisi tanımlanabilir. Bu eşitlik, aşağıda iki önemli sonucu veren teoremi ifade etmek için kullanılır.

Teorem 1: m ve γ vektörleri, yalnız ve yalnız aşağıdaki koşul sağlanırsa doğal kübik eğrisel çizgi belirtir:

$$\mathbf{Q}^{T}\mathbf{m} = \mathbf{R}\gamma \tag{2.23}$$

Eğer Eş. (2.23) sağlanırsa, pürüzlülük cezası için,

$$\int_a^b m''(x)^2 dx = \gamma^{T}\mathbf{R}\gamma = \mathbf{m}^{T}\mathbf{K}\mathbf{m} \tag{2.24}$$

eşitliği elde edilir (Green ve Silverman, 2000).

Teorem 1'den yararlanarak parametrik olmayan m fonksiyonu, doğal kübik eğrisel çizgi fonksiyonları yardımı ile tahmin edilebilir. Buna göre Eş. (2.6)'daki cezalandırılmış kareler toplamı, kübik eğrisel çizgi fonksiyonlarından yararlanarak,

$$S(m)= \sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda \int_a^b \{m''(x)\}^2 dx$$

$$= (\mathbf{y}-\mathbf{m})^T(\mathbf{y}-\mathbf{m}) + \lambda \mathbf{m}^T \mathbf{K} \mathbf{m} \qquad (2.25)$$

şeklinde yazılabilir. Eş. (2.25)'deki ifadenin **m** vektörüne göre türevi alınıp sıfıra eşitlenirse $\hat{\mathbf{m}}$

$$\hat{\mathbf{m}} = (\mathbf{I} + \lambda \mathbf{K})^{-1} \mathbf{y} \qquad (2.26)$$

olarak elde edilir. Bu eşitlikte **K** matrisi yarı pozitif tanımlı ve λ>0 olduğundan $\lambda \mathbf{K}$ matrisi de yarı pozitif tanımlıdır. Eş. (2.26)'daki $(\mathbf{I} + \lambda \mathbf{K})^{-1}$ matrisi düzleştirme matrisi olarak adlandırılır ve $\mathbf{S}_\lambda$ ile gösterilir. Buna göre Eş. (2.26),

$$\hat{\mathbf{m}}_\lambda = \mathbf{S}_\lambda \mathbf{y} \qquad (2.27)$$

şeklinde yazılabilir. Burada $\hat{\mathbf{m}}_\lambda$, tahmin edilen eğrisel çizgi düzleştiricisidir.

Gözlem değerleri w_i gibi ağırlıklarla ağırlıklandırıldığında ağırlıklandırılmış eğrisel çizgi düzleştirici, cezalı ağırlıklandırılmış en küçük kareler yaklaşımı ile elde edilebilir. Ağırlıklandırılmış eğrisel çizgi düzleştirme yönteminin esası, $m \in C^2[a,b]$ uzayındaki tüm m fonksiyonları arasında,

$$S(m) = (\mathbf{y}-\mathbf{m})^T \mathbf{W} (\mathbf{y}-\mathbf{m}) + \lambda \mathbf{m}^T \mathbf{K} \mathbf{m} \qquad (2.28)$$

eşitliği ile belirtilen "cezalı ağırlıklandırılmış en küçük kareler toplamını" minimum yapan m fonksiyonunu elde etmektir. Burada **W** köşegen elemanları w_i olan köşegen bir matristir. Eş. (2.28)'deki ifadenin **m** vektörüne göre türevi alınıp sıfıra eşitlenirse, ağırlıklandırılmış eğrisel çizgi düzleştiricisi,

$$\hat{\mathbf{m}} = (\mathbf{W} + \lambda\mathbf{K})^{-1}\mathbf{W}\,\mathbf{y}$$
$$= \mathbf{S}_{\lambda}\mathbf{y} \qquad (2.29)$$

olarak elde edilir. Eş. (2.29)'daki $\mathbf{S}_{\lambda} = (\mathbf{W} + \lambda\mathbf{K})^{-1}\mathbf{W}$ dır (Hastie ve Tibshirani, 1990).

2.3.4. Cezalandırılmış (Penalized) eğrisel çizgi regresyonu

Eğrisel çizgi yaklaşımında kestirimin pürüzlülüğü çok sayıda düğüm noktasının olmasından kaynaklanmaktadır. Bu sorun düğüm noktalarının etkileri üzerine bir kısıtlama getirilerek çözülebilir. K tane düğüm noktası için (K büyük bir değer) m(x) fonksiyonu, aşağıdaki gibi modellenebilir:

$$y = m(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{1k}(x - \kappa_k)_+ + \varepsilon = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (2.30)$$

Eş. (2.30)'da $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_{11}, ..., \beta_{1K}]^T$ olup β_{1k}, k.düğüme ilişkin parametreyi ve κ_1, $\kappa_2, \ldots, \kappa_K$ düğüm noktalarını göstermektedir. Eş. (2.30)'daki modelde düğüm noktalarının seçimi önemli bir konudur. Literatürde düğüm noktalarının seçimi ile ilgili çok sayıda çalışma vardır. Bunun için bazı formül ve algoritmalar önerilmiştir. Düğüm noktasının seçimi için en yaygın kullanılan formül,

$$\kappa_k = \left(\frac{k+1}{K+2}\right)$$

dir. Bu formülde K=min ($\frac{1}{4}$x(tekrar etmeyen x_i'lerin sayısı), 35)'dir (Yao ve Lee, 2008).

Eş. (2.30)'daki **X** matrisi

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \vdots & \vdots & \cdots & \\ 1 & x_n & (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_K)_+ \end{bmatrix}$$

biçimindedir. Burada

$$(x-\kappa_k)_+ = \begin{cases} 0, & x \le \kappa_k \\ (x-\kappa_k), & x > \kappa_k \end{cases}$$

dir. Eş. (2.30)'daki $\boldsymbol{\beta}$, $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$ hata kareler toplamını minimum yapan değerdir. Burada düğüm noktalarının sayısının çok olması kestirimin kıvrımlı olmasına neden olur. Bu sorunu gidermek için izlenecek yol, düğüm noktalarına ilişkin β_{ik} katsayıları üzerine kısıt koymaktır. $\beta_{1\,k}$, k =1,…K, katsayıları üzerine konulan kısıtlar bu durumu düzeltebilir. Bu kısıtlar,

- $\max|\beta_{1i}| < C$
- $\sum_{i=1}^{m}|\beta_{1i}| < C$
- $\sum_{i=1}^{m}\beta_{1i}^2 < C$

biçimindedir. C uygun bir şekilde seçildiğinde, bu kısıtların her biri daha düz bir kestirim elde edilmesini sağlayacaktır. Ancak üçüncü kısıtın uygulanması diğerlerinden daha kolaydır. (K+2)x(K+2) boyutlu bir **D** matrisi

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{2x2} & \mathbf{0}_{2xK} \\ \mathbf{0}_{Kx2} & \mathbf{I}_{KxK} \end{bmatrix}$$

olarak tanımlanırsa, minimizasyon problemi, $\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta} \le C$ kısıt altında $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2$ hata kareler toplamının minimizasyon problemine dönüşür. Lagrange çarpanları yöntemi kullanılarak minimum yapılacak fonksiyon,

$$S(\beta)=\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\lambda\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta} \qquad (2.31)$$

şekline dönüşür ($\lambda \ge 0$). Eş. (2.31)'deki cezalandırılmış artık kareler toplamının $\boldsymbol{\beta}$ vektörüne göre türevi alınıp sıfıra eşitlendiğinde,

$$\hat{\boldsymbol{\beta}}_{\lambda} = (\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{2.32}$$

biçiminde elde edilir. Kestirim değerleri ise

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda} \tag{2.33}$$

dir. Eş. (2.33)'de düzleştirme matrisi $\mathbf{S}_{\lambda} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^{\mathbf{T}}$ 'dir.

p. dereceden eğrisel çizgi regresyon modeli ise

$$\mathbf{y} = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk}(x - \kappa_k)_+^{\,p} + \boldsymbol{\varepsilon} = \mathbf{X}_{\mathbf{p}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.34}$$

şeklinde ifade edilir. Burada $\mathbf{X}_p$ matrisi

$$\mathbf{X}_{\mathbf{p}} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p & (x_1-\kappa_1)_+^p & \cdots & (x_1-\kappa_K)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \\ 1 & x_n & \cdots & x_n^p & (x_n-\kappa_1)_+^p & \cdots & (x_n-\kappa_K)_+^p \end{bmatrix}$$

biçimindedir. p. dereceden cezalandırılmış eğrisel çizgi regresyonunda minimum yapılacak fonksiyon

$$S(\beta) = \left\| \mathbf{y} - \mathbf{X}_p\boldsymbol{\beta} \right\|^2 + \lambda\,\boldsymbol{\beta}^{\mathbf{T}}\mathbf{D}\boldsymbol{\beta} \tag{2.35}$$

şekline dönüşür. Bu fonksiyonun $\boldsymbol{\beta}$'ya göre türevi alınıp sıfıra eşitlenirse $\hat{\boldsymbol{\beta}}_{\lambda}$

$$\hat{\boldsymbol{\beta}}_{\lambda} = (\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}\mathbf{X}_{\mathbf{p}} + \lambda\mathbf{D})^{-1}\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}\mathbf{y} \tag{2.36}$$

olarak bulunur. p.dereceden cezalandırılmış eğrisel çizgi regresyonun kestirim değerleri aşağıdaki gibi elde edilir:

$$\hat{\mathbf{y}} = \mathbf{X}_{\mathbf{p}}(\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}\mathbf{X}_{\mathbf{p}} + \lambda\mathbf{D})^{-1}\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}\mathbf{y} = \mathbf{X}_{\mathbf{p}}\hat{\boldsymbol{\beta}}_{\lambda} \tag{2.37}$$

Bu eşitlikte düzleştirme matrisi $\mathbf{S}_{\lambda} = \mathbf{X}_{\mathbf{p}}(\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}\mathbf{X}_{\mathbf{p}} + \lambda\mathbf{D})^{-1}\mathbf{X}_{\mathbf{p}}{}^{\mathbf{T}}$'dir (Ruppert vd., 2003).

2.3.5. Cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı ile düzleştirme

Cezalandırılmış eğrisel çizgi regresyon yöntemi ile karışık doğrusal modeller arasında yakın bir ilişki vardır. Cezalandırılmış eğrisel çizgi regresyon modeli, karışık doğrusal modele benzetilerek ifade edilebilir.

Karışık doğrusal modelde amaç, hem sabit etkilere hem de rasgele etkilere ilişkin parametre kestirimlerinin elde edilmesidir. Bu amaçla önerilen yöntemlerden birisi, hem sabit hem de rasgele etkilere ilişkin parametrelerin birlikte tahmin edilebildiği en iyi doğrusal yansız kestirim (best linear unbiased predictor, BLUP) yöntemidir. BLUP yöntemi kullanılarak parametrik olmayan regresyon fonksiyonu m(.), karışık doğrusal modele benzetilip tahminler elde edilebilir. Bundan sonraki alt bölümlerde cezalandırılmış eğrisel çizgi regresyonunda BLUP yöntemi ile düzleştirme yönteminin daha iyi anlaşılması için karışık doğrusal model ve bu modelde parametre ve varyans oranlarının tahminleri hakkında kısaca bilgi verilecektir.

2.3.5.1. Karışık doğrusal model

Karışık doğrusal modeller, doğrusal modellerin genişletilmiş hali olarak düşünülebilir. Sabit doğrusal modellerde sadece özel seçimli etkiler yer alırken, karışık doğrusal modellerde, hem özel seçimli etkiler hem de rasgele seçimli etkiler bir arada bulunur. Karışık doğrusal model matris formunda aşağıdaki gibi ifade edilebilir:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \qquad (2.38)$$

Eş. (2.38)'de $\mathbf{y}$, nx1 boyutlu gözlemlenebilir yanıt vektörü, $\mathbf{X}$, nxp boyutlu sabit etkilere ilişkin tasarım matrisi, $\boldsymbol{\beta}$, px1 boyutlu sabit etkilere ilişkin parametreler vektörü, $\mathbf{Z}$, nxq boyutlu rasgele etkilere ilişkin tasarım matrisi, $\mathbf{u}$, qx1 boyutlu rasgele etkilere ilişkin parametreler vektörü, $\boldsymbol{\varepsilon}$, nx1 boyutlu rasgele hatalar vektörüdür.

u ve **ε**'nun beklenen değerleri, **u** ve **ε** arasındaki varyans-kovaryans matrisi, sırasıyla

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{0}_{qx1}\\ \mathbf{0}_{nx1}\end{bmatrix} \quad \text{ve} \quad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\mathbf{D}_{qxq} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}_{nxn}\end{bmatrix} \tag{2.39}$$

şeklindedir. Buna göre **y**, yanıt değişkeninin beklenen değeri ve varyansı,

$$\mathrm{E}(\mathbf{y})=\mathbf{X\beta} \tag{2.40}$$

$$\mathrm{V}(\mathbf{y})=\mathbf{ZDZ}^{\mathbf{T}}+\mathbf{R}=\mathbf{V}_{nxn} \tag{2.41}$$

olarak bulunur (McCulloch ve Searle, 2001; Türkan, 2008).

2.3.5.2. Karışık doğrusal modelde tahmin

Karışık doğrusal modelde hem sabit etkilere hem de rasgele etkilere ilişkin parametre kestirimleri elde edilir. Başka bir deyişle sabit etkiler vektörü **β**'nın, rasgele etkiler vektör **u**'nun ve varyans-kovaryans matrisi **V**'deki varyans parametrelerinin tahminleri elde edilir.

Parametrik olmayan regresyonda karışık doğrusal model yaklaşımı ile düzleştirme yapıldığında parametrelerin kestirimlerini elde etmek için hem sabit hem de rasgele etkilere ilişkin parametrelerin birlikte kestirilebildiği en iyi doğrusal yansız kestirici (best linear unbiased predictor, BLUP) kullanılmıştır.

Henderson'ın karışık model eşitliklerinden yararlanarak BLUP elde edilebilir. Henderson'ın karışık model eşitliklerinin elde edilebilmesi için **u** verilmişken **y'**nin ve **u**'nun, sırasıyla,

$$\mathbf{y}\,|\,\mathbf{u}\sim \mathrm{N}(\mathbf{X\beta}+\mathbf{Zu},\mathbf{R}) \quad \text{ve} \quad \mathbf{u}\sim \mathrm{N}(\mathbf{0},\mathbf{D})$$

biçiminde normal dağıldığı varsayılır. Buna göre **u** ve **y**'nin bileşik yoğunluk fonksiyonu, aşağıdaki gibidir:

$$f(\mathbf{u};\mathbf{y}) = f(\mathbf{y}|\mathbf{u})f(\mathbf{u})$$

$$= \frac{\exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})+\mathbf{u}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{u}\}}{(2\pi)^{\frac{1}{2}(N+q)}|\mathbf{R}|^{\frac{1}{2}}|\mathbf{D}|^{\frac{1}{2}}} \tag{2.42}$$

Eş. (2.42)'deki ifadeyi maksimum yapan parametre kestirimlerini bulmak için Eş. (2.42)'deki ifadenin logaritması alınırsa

$$L = \log(f(\mathbf{u};\mathbf{y})) \approx (\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})^{\mathsf{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})+\mathbf{u}^{\mathsf{T}}\mathbf{D}^{-1}\mathbf{u} \tag{2.43}$$

olarak elde edilir. Bu ifadenin $\boldsymbol{\beta}$ ve $\mathbf{u}$ 'ya türevleri bulunup sıfıra eşitlenirse bulunan eşitlikler matris gösterimi ile

$$\begin{bmatrix} \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

biçiminde yazılabilir. ($\boldsymbol{\beta}$, $\mathbf{u}$)'nun tahmini BLUP ile

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{Z}+\mathbf{D}^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} = (\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{C}+\mathbf{B})^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \tag{2.44}$$

olarak elde edilir. Burada $\mathbf{C} \equiv [\mathbf{X} \quad \mathbf{Z}]$ ve $\mathbf{B} \equiv \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^{-1} \end{bmatrix}$ dir.

BLUP kestirim yöntemi ile elde edilen kestirim değerleri,

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} = \mathbf{C}(\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{C}+\mathbf{B})^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{R}^{-1}\mathbf{y} \tag{2.45}$$

olarak bulunur (Ruppert vd., 2003).

Karışık doğrusal modellerde kovaryans matrisinin tahmini ile ilgili çeşitli yöntemler geliştirilmiştir. Normallik varsayımı gerektirmeyen en küçük normlu karesel yansız kestirim yöntemi (Minimum quadratic unbiased estimation, MINQUE) ve en küçük varyanslı karesel yansız kestirim yöntemi (Minimum variance unbiased estimation,

MIVQUE) ve dağılım varsayımı gerektiren olabilirlik yöntemi (Maximum likelihood, ML) ve sınırlandırılmış en çok olabilirlik yöntemi kovaryans matrislerinin tahmini için önerilen yöntemlerdir. Ancak, hesaplama algoritmalarındaki gelişmeler ile kovaryans matrisindeki parametrelerin tahmini için en çok olabilirlik ya da kısıtlanmış en çok olabilirlik yöntemleri en yaygın kullanılan yöntemlerdir.

$\mathbf{y} \sim N(\mathbf{X\beta}, \mathbf{V})$ varsayımı altında **y**'nin olabilirlik fonksiyonunun logaritması aşağıdaki gibidir:

$$L(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X\beta})^{T}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})\} \quad (2.46)$$

($\boldsymbol{\beta}$,**V**)'nin en çok olabilirlik tahmini(ML), Eş. (2.46)'daki ifade maksimum yapılarak bulunur. Eş. (2.46)'da olabilirlik fonksiyonun $\boldsymbol{\beta}$'ya göre türevi bulunup sıfıra eşitlenirse,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{y} \quad (2.47)$$

olarak elde edilir. $\hat{\boldsymbol{\beta}}$ en iyi doğrusal yansız tahmin edicidir. Eş. (2.47)'deki $\hat{\boldsymbol{\beta}}$, Eş. (2.46)'da yerine yazılırsa **V** matrisi için profil olabilirlik fonksiyonu elde edilir. Profil olabilirlik fonksiyonun logaritması alınıp profil log-olabilirlik (profile log-likelihood) fonksiyonu,

$$L_{P}(\mathbf{V}) = -\frac{1}{2}\{n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{T}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\}$$

$$= -\frac{1}{2}\{\log|\mathbf{V}| + \mathbf{y}^{T}\mathbf{V}^{-1}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{V}^{-1})\mathbf{y}\} - \frac{n}{2}\log(2\pi) \quad (2.48)$$

olarak elde edilir. **V** matrisindeki parametrelerin en çok olabilirlik tahminleri bu parametrelere göre Eş. (2.48)'deki ifade maksimum yapılarak bulunur.

Sınırlandırılmış en çok olabilirlik yöntemine (Restricted Maximum Likelihood, REML) göre **V**'deki parametrelerin tahminlerinin elde edilmesi ise daha karmaşıktır. REML tahminleri, **y**'nin elemanlarının doğrusal bileşimlerinin olabilirlik fonksiyonun ($\boldsymbol{\beta}$'ya bağlı olmayan) maksimum yapılmasıyla elde edilir.

Sınırlandırılmış en çok olabilirlik fonksiyonun logaritması alınırsa sınırlandırılmış log-olabilirlik (restricted log-likelihood) fonksiyonu,

$$L_R(\mathbf{V}) = L_P(\mathbf{V}) - \frac{1}{2}\log\left|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right| \tag{2.49}$$

olarak bulunur. **V** matrisindeki parametrelerin(varyans oranları) sınırlandırılmış en çok olabilirlik tahminleri bu parametrelere göre Eş. (2.49)'daki ifade maksimum yapılarak bulunur.

Küçük örneklemler için REML tahminlerinin ML tahminlerinden daha doğru olması beklenir ancak büyük örneklemler için iki tahmin yöntemi arasında çok az bir fark vardır (Ruppert vd., 2003).

2.3.5.3. Parametrik olmayan regresyonda cezalandırılmış eğrisel çizgiler (penalized splines) için BLUP tahminleri

Eş. (2.30)'daki cezalandırılmış eğrisel çizgi modeli karışık doğrusal modele benzetilerek gösterilebilir. Bu durumda

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} (x_1-\kappa_1)_+ & \cdots & (x_1-\kappa_k)_+ \\ \vdots & \vdots & \vdots \\ (x_n-\kappa_1)_+ & \cdots & (x_n-\kappa_k)_+ \end{bmatrix}$$

olarak tanımlansın. Buna göre Eş. (2.30)'daki modelde $\beta_{ik} = u_k$ alınarak,

$$y = m(x) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k\,(x-\kappa_k)_+ + \varepsilon \tag{2.50}$$

şeklinde yazılabilir. $\boldsymbol{\beta}$, x değerlerine ilişkin regresyon katsayıları vektörünü ve **u** düğüm noktalarına ilişkin regresyon katsayıları vektörünü göstermek üzere, Eş. (2.50)'deki model, karışık doğrusal model olarak matris formunda

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{2.51}$$

şeklinde yazılabilir. Eş. (2.51)'deki karışık doğrusal modelde, **u** ve **ε**'nun beklenen değerlerinin ve **u** ile **ε** arasındaki varyans-kovaryans matrisinin sırasıyla,

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \quad \text{ve} \quad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_u^2\mathbf{I} & 0\\ 0 & \sigma_\varepsilon^2\mathbf{I}\end{bmatrix} \tag{2.52}$$

biçiminde olduğu varsayılır. Buna göre **β** ve **u** parametreler vektörlerinin birlikte tahmini en iyi doğrusal yansız kestirici (BLUP) kullanılarak elde edilebilir. Eş. (2.52)'den **u** bilindiğinde **y**'nin

$$\mathbf{y}\,|\,\mathbf{u} \sim \mathrm{N}(\mathbf{X\beta} + \mathbf{Zu}, \sigma_\varepsilon^2\mathbf{I}), \quad \mathbf{u} \sim \mathrm{N}(0, \sigma_u^2\mathbf{I})$$

şeklinde normal dağıldığı varsayılır. **u** ve **y**'nin bileşik yoğunluk fonksiyonu, aşağıdaki gibidir:

$$\begin{aligned} \mathrm{f}(\mathbf{u};\mathbf{y}) &= \mathrm{f}(\mathbf{y}\,|\,\mathbf{u})\mathrm{f}(\mathbf{u}) \\ &\cong \exp\{-\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu})^{\mathrm{T}}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu}) + \frac{\sigma_\varepsilon^2}{\sigma_u^2}\mathbf{u}^{\mathrm{T}}\mathbf{u}\} \end{aligned} \tag{2.53}$$

Eş. (2.53)'den $\mathrm{f}(\mathbf{u};\mathbf{y})$'nun log-olabilirlik fonksiyonu

$$\mathrm{L} = \log(\mathrm{f}(\mathbf{u};\mathbf{y})) \approx (\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu})^{\mathrm{T}}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu}) + \lambda\mathbf{u}^{\mathrm{T}}\mathbf{u} \tag{2.54}$$

olarak elde edilir. Burada düzleştirme parametresi $\lambda = \sigma_\varepsilon^2/\sigma_u^2$'dır. Eş. (2.54)'deki ifadenin **β** ve **u** 'ya göre türevleri bulunup sıfıra eşitlenirse (**β**, **u**)'nun en iyi doğrusal yansız kestiricisi (BLUP),

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix} = (\mathbf{C}^{\mathrm{T}}\mathbf{C} + \lambda\mathbf{M})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y} \tag{2.55}$$

olarak elde edilir. Burada $\mathbf{C} = [\mathbf{X} \quad \mathbf{Z}]$ ve $\mathbf{M} = \begin{bmatrix} \mathbf{0}_{2x2} & \mathbf{0}_{2xK} \\ \mathbf{0}_{Kx2} & \mathbf{I}_{KxK} \end{bmatrix}$ dir.

Eş.(2.50)'deki model için kestirim değerleri ise

$$\hat{\mathbf{m}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} = \mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y} \qquad (2.56)$$

olarak bulunur. Eş. (2.56)'daki ifade parametrik olmayan m fonksiyonunun en iyi doğrusal yansız kestiricisi, bir çeşit Ridge Regresyonu'dur. (Ruppert vd, 2003). Cezalandırılmış eğrisel çizgi regresyonunun karışık doğrusal model olarak gösterimi oldukça yararlıdır. Çünkü karışık doğrusal model olarak ifade edilebilen cezalandırılmış eğrisel çizgi regresyonunda kestirimler karışık doğrusal model teorisi ve karışık doğrusal model için geliştirilen bilgisayar programları kullanılarak elde edilebilir. Ayrıca karışık doğrusal modelde kestirilen varyansların oranı ($\sigma_\varepsilon^2/\sigma_u^2$) eğrisel çizgi regresyonundaki düzleştirme parametresine karşılık gelmektedir. Bu da düzleştirme parametresinin seçiminin önemli olduğu eğrisel çizgi regresyonunda düzleştirme parametresinin belirlenmesinde önemli kolaylık sağlar.

Yukarıda bahsedilen düzleştirme yöntemlerinden cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı dışında diğer düzleştirme yöntemlerinde düzleştirme parametresinin seçimi önemlidir ve düzleştirme parametresinin seçimi için birçok yöntem geliştirilmiştir. Bir sonraki alt bölümde bu yöntemlerden Çapraz Geçerlilik (CV), Genelleştirilmiş Çapraz Geçerlilik, Mallows'un C_p ölçütü ve Akaike Bilgi Ölçütü hakkında genel bir bilgi verilecektir.

2.3.6. Düzleştirme Parametresinin Seçimi

Parametrik olmayan regresyon modellerinde veri kümesinde düzleştirme miktarını kontrol eden düzleştirme parametresi kullanılır (Hurvich vd., 1998). Eğer düzleştirme parametresi çok büyük seçilirse veriler fazla düzleştirilmiş, çok küçük seçilirse veriler az düzleştirilmiş olur. Bu nedenle düzleştirme parametresinin seçimi oldukça önemlidir. Düzleştirme parametresinin seçimi için önerilen yöntemler genel olarak iki sınıfta toplanabilir: Klasik yaklaşımlar ve yerleştirme (plug-in) yaklaşımı (Aydın, 2005; Tabakan, 2009). Çapraz Geçerlilik (CV),

Genelleştirilmiş Çapraz Geçerlilik, Mallows'un C_p ölçütü ve Akaike Bilgi Ölçütü düzleştirme parametresinin seçimi için kullanılan klasik yaklaşımlardır. Klasik yaklaşımların yanı sıra risk tahmin yöntemleri olarak adlandırılan "Klasik Pilotları Kullanan Risk Tahmini" ve "Yerel Risk Tahmini" yöntemleri de bulunmaktadır. Ayrıca düzleştirme parametresinin seçimi için kullanılan diğer bir yaklaşım olabilirlik yaklaşımı olarak bilinen karışık doğrusal model yaklaşımıdır.

Aşağıdaki alt bölümlerde düzleştirme parametresinin seçiminde yaygın olarak kullanılan klasik yaklaşımlar hakkında genel bilgiler verilecektir.

2.3.6.1. Çapraz geçerlilik (Cross-validation, CV)

Bir modelin artık kareler toplamı, o modelin ilgilenilen veri kümesini iyi açıklayıp açıklamadığının bir ölçüsüdür. Çünkü artık (residual), yanıt değişkeninin gözlenen değeri ile kestirilen değeri arasındaki farka eşittir ve

$$e_i = y_i - \hat{y}_i \tag{2.57}$$

şeklinde tanımlanır. Ancak, artık kareler toplamı model seçimi için yeterli bir ölçüt değildir. Çünkü sorun, y_i'nin kestiriminde y_i gözleminin de kullanılmasıdır. Bu problemin basit bir çözümü vardır. Bu çözüm, i.gözlem veri kümesinden çıkarıldıktan sonra kalan diğer gözlem değerleri kullanılarak y_i gözleminin tahmin edilmesidir. Bu şekilde elde edilen kestirim değeri $\hat{y}_{i,-i}$ ile gösterilir. i. gözlem çıkarıldıktan sonra elde edilen artık değeri e_{-i},

$$e_{-i} = y_i - \hat{y}_{i,-i} \tag{2.58}$$

şeklinde elde edilir. Kestirilmiş artık kareler toplamı ise (predicted residual sum of squares, PRESS)

$$\text{PRESS} = \sum_{i=1}^{n} e_{-i}^2 \tag{2.59}$$

şeklinde tanımlanır. PRESS, çapraz geçerlilik istatistiği olarak da adlandırılır. Parametrik regresyonda çapraz geçerlilik ölçütü

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 \tag{2.60}$$

şeklindedir. Eş. (2.60)'daki $y_i - \hat{y}_{i,-i}$ değeri, aşağıdaki gibi bulunur:

$$\begin{aligned} y_i - \hat{y}_{i,-i} &= y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{-i} \\ &= y_i - \mathbf{x}_i^T \left[\hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i (y_i - \hat{y}_i)}{1 - h_{ii}} \right] \\ &= y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \frac{\mathbf{x}_i^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i (y_i - \hat{y}_i)}{1 - h_{ii}} \\ &= y_i - \hat{y}_i + \frac{h_{ii}(y_i - \hat{y}_i)}{1 - h_{ii}} \\ &= \frac{(y_i - \hat{y}_i)}{1 - h_{ii}} \end{aligned} \tag{2.61}$$

Eş. (2.61)'de h_{ii}, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ şapka matrisinin köşegen elemanıdır ve $\hat{\mathbf{y}} = \mathbf{H}\,\mathbf{y}$ olarak yazılabilir. Buna göre çapraz geçerlilik ölçütü

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} \left(\frac{(y_i - \hat{y}_i)}{1 - h_{ii}} \right)^2 \tag{2.62}$$

şeklinde elde edilir.

Parametrik olmayan regresyonda çapraz geçerlilik ölçütü parametrik regresyondakine benzer olarak,

$$CV(\lambda) = \sum_{i=1}^{n} (y_i - \hat{m}_{-i}(x_i;\lambda))^2 \tag{2.63}$$

şeklinde ifade edilir.

$\hat{m}(x_i;\lambda)$, x_i noktasında λ düzleştirme parametresiyle parametrik olmayan regresyon tahminini göstersin. Parametrik olmayan regresyonda kestirim değerleri vektörü aşağıdaki gibidir:

$$\begin{bmatrix} \hat{m}(x_1;\lambda) \\ \vdots \\ \hat{m}(x_n;\lambda) \end{bmatrix} = \mathbf{S}_\lambda \mathbf{y} \tag{2.64}$$

ve

$$\hat{m}(x_i;\lambda) = \sum_{j=1}^{n} S_{\lambda,ij} y_j \tag{2.65}$$

Eş. (2.65)'de $S_{\lambda,ij}$, $\mathbf{S}_\lambda$'nın (i,j). elemanıdır. Birçok düzleştirici aşağıdaki biçimde yazılabilir:

$$\hat{m}_{-i}(x_i;\lambda) = \frac{\sum_{i \neq j} S_{\lambda,ij} y_j}{\sum_{i \neq j} S_{\lambda,ij}} \tag{2.66}$$

(Ruppert vd., 2003). Eş. (2.66) her zaman sağlanamasa da, genellikle yaklaşık olarak sağlanır. Eş.(2.66)'da $\hat{m}_{-i}(x_i;\lambda)$, veri kümesinden (x_i,y_i) gözlemi çıkarıldıktan sonra elde edilen parametrik olmayan regresyon tahmin değeridir. Ayrıca bu gösterim çapraz geçerlilik ölçütünde kullanılan $\hat{m}_{-i}(x_i;\lambda)$'nin tanımı olarak düşünülebilir. Tüm düzleştiriciler, her zaman $y_i \equiv 1$ ise $\hat{y}_i \equiv 1$ özelliğine sahiptir ve bu özellik tüm düzleştiriciler için,

$$\sum_{i=1}^{n} S_{\lambda,ij} = 1 \text{ , tüm i'ler için}$$

olduğunu gösterir. Buna göre $\sum_{i\neq j} S_{\lambda,ij} = 1 - S_{\lambda,ii}$'dir ve parametrik olmayan regresyonda çapraz geçerlilik ölçütü parametrik regresyondakine benzer olarak

$$
\begin{aligned}
CV(\lambda) &= \sum_{i=1}^{n} \left(y_i - \hat{m}_{-i}(x_i;\lambda)\right)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \frac{\sum_{i\neq j} S_{\lambda ij} y_j}{1 - S_{\lambda,ii}} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{y_i - \hat{m}(x_i;\lambda)}{1 - S_{\lambda,ii}} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{\{(\mathbf{I} - \mathbf{S}_{\boldsymbol{\lambda}})\mathbf{y}\}_i}{1 - S_{\lambda,ii}} \right)^2 \\
&= \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - S_{\lambda,ii}} \right)^2
\end{aligned}
\qquad (2.67)
$$

şeklinde yazılabilir. Çapraz geçerlilik ölçütüne göre en uygun düzleştirme parametresi (λ), CV(λ) fonksiyonunu minimum yapan değerdir (Ruppert vd., 2003).

2.3.6.2. Genelleştirilmiş çapraz geçerlilik (Generalized cross-validation, GCV)

Genelleştirilmiş çapraz geçerlilik ölçütü, çapraz geçerlilik eşitliğinde $S_{\lambda,ii}$'nin yerine $S_{\lambda,ii}$'nin ortalaması

$$\frac{1}{n}\sum_{i=1}^{n} S_{ii} = \frac{1}{n} \text{İz}(\mathbf{S}_{\lambda}) \qquad (2.68)$$

yazılarak elde edilir. Buna göre genelleştirilmiş çapraz geçerlilik ölçütü

$$GCV(\lambda) = \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{m}(x_i;\lambda)}{1 - \text{İz}(\mathbf{S}_{\lambda})/n} \right\}^2$$

$$= \frac{\left\| (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y} \right\|^2}{\left\{1 - n^{-1} \mathrm{Iz}(\mathbf{S}_\lambda)\right\}^2} \tag{2.69}$$

şeklinde ifade edilir. Genelleştirilmiş çapraz geçerlilik ölçütüne göre en uygun düzleştirme parametresi, $GCV(\lambda)$ fonksiyonunu minimum yapan değerdir.

2.3.6.3. Mallows'un C_p ölçütü

Mallows (1973) tarafından önerilen C_p kriteri

$$C_p = HKT_p + 2\hat{\sigma}_\varepsilon^2 p \tag{2.70}$$

şeklindedir. Burada p modeldeki parametrelerin sayısıdır. HKT_p aday model için hata kareler toplamıdır. Parametrik olmayan regresyon modeli için Mallows'un C_p kriteri geliştirilmiştir. Buna göre,

$$\mathbf{y} = \mathbf{m} + \boldsymbol{\varepsilon}, \quad \mathrm{Kov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I} \tag{2.71}$$

parametrik olmayan regresyon modelini göz önüne alalım. Bu modelde $\mathbf{m}$ vektörünün tahmini $\hat{\mathbf{m}}$

$$\hat{\mathbf{m}} = \mathbf{S}_\lambda \mathbf{y} \tag{2.72}$$

olmak üzere ortalama hata kareler toplamı (mean summed squared error) aşağıdaki gibidir:

$$OHKT(\hat{\mathbf{m}}) = E\left\| \hat{\mathbf{m}} - \mathbf{m} \right\|^2 = \sum_{i=1}^{n} \left\{E\hat{m}(x_i) - m(x_i)\right\}^2 + \mathrm{Var}\left\{\hat{m}(x_i)\right\}$$

$$= \sum_{i=1}^{n} \left\{E(\mathbf{S}_\lambda \mathbf{y})_i - m_i\right\}^2 + \mathrm{Var}\left\{(\mathbf{S}_\lambda \mathbf{y})_i\right\}$$

$$= \sum_{i=1}^{n} \left\{E(\mathbf{S}_\lambda \mathbf{y})_i - m_i\right\}^2 + \mathrm{Kov}\left\{(\mathbf{S}_\lambda \mathbf{y})_{ii}\right\}$$

$$= \left\| (\mathbf{S}_\lambda - \mathbf{I})\mathbf{m} \right\|^2 + \text{İz}\{\text{Kov}(\mathbf{S}_\lambda \mathbf{y})\} \quad (2.73)$$

Değişen varyanslılığın olmadığı varsayımı altında $\text{Kov}(\mathbf{y}) = \sigma_\varepsilon^2 \mathbf{I}$ olmak üzere ortalama hata kareler toplamı,

$$\text{OHKT}(\hat{\mathbf{m}}) = \left\| (\mathbf{S}_\lambda - \mathbf{I})\mathbf{m} \right\|^2 + \sigma_\varepsilon^2 \text{İz}\{(\mathbf{S}_\lambda \mathbf{S}_\lambda^\text{T})\} \quad (2.74)$$

şeklinde elde edilir. Hata kareler toplamı (HKT)

$$\text{HKT}(\lambda) = \left\| \hat{\mathbf{m}} - \mathbf{y} \right\|^2 \quad (2.75)$$

olup, hata kareler toplamının beklenen değeri,

$$\begin{aligned}
\text{E}(\text{HKT}(\lambda)) &= \text{E}\left\| \hat{\mathbf{m}} - \mathbf{y} \right\|^2 = \text{E}\left\| (\mathbf{S}_\lambda - \mathbf{I})\mathbf{y} \right\|^2 \\
&= \text{E}\{\mathbf{y}^\text{T} (\mathbf{S}_\lambda - \mathbf{I})^\text{T} (\mathbf{S}_\lambda - \mathbf{I})\mathbf{y}\} \\
&= \mathbf{m}^\text{T} (\mathbf{S}_\lambda - \mathbf{I})^\text{T} (\mathbf{S}_\lambda - \mathbf{I})\mathbf{m} + \sigma^2 \text{İz}\{(\mathbf{S}_\lambda - \mathbf{I})^\text{T} (\mathbf{S}_\lambda - \mathbf{I})\} \\
&= \left\| (\mathbf{S}_\lambda - \mathbf{I})\mathbf{m} \right\|^2 + \sigma^2 \{\text{İz}(\mathbf{S}_\lambda \mathbf{S}_\lambda) - 2\text{İz}(\mathbf{S}_\lambda) + \text{n}\} \\
&= \left\| (\mathbf{S}_\lambda - \mathbf{I})\mathbf{m} \right\|^2 + \sigma^2 (\text{sd}_{\text{Artık}}) \quad (2.76)
\end{aligned}$$

dir. Eş. (2.76)'da $\text{sd}_{\text{Artık}} = \{\text{İz}(\mathbf{S}_\lambda \mathbf{S}_\lambda) - 2\ \text{İz}(\mathbf{S}_\lambda) + \text{n}\}$'dir (Ruppert vd. , 2003). Hata kareler toplamının beklenen değeri, ortalama hata kareler toplamına göre aşağıdaki gibi yazılabilir:

$$\text{E}(\text{HKT}(\lambda)) = \text{OHKT}(\hat{\mathbf{m}}) + \sigma^2 (\text{n} - 2(\text{sd}_{\text{kestirim}})) \quad (2.77)$$

Eş. (2.77)'de $\text{sd}_{\text{kestirim}}$, $\hat{\mathbf{m}}$'yı kestirmek için kullanılan parametrelerin sayısını göstermektedir ve $\text{sd}_{\text{kestirim}} = \text{İz}(\mathbf{S}_\lambda)$'dır. Buna göre parametrik olmayan regresyonda Mallows'un C_p kriteri parametrik regresyondakine benzer olarak

$$C_p(\lambda) = \text{HKT}(\lambda_p) + 2\hat{\sigma}_\varepsilon^2 \ (\text{sd}_{\text{kestirim}}) \quad (2.78)$$

şeklinde elde edilir. Burada $\hat{\sigma}_{\varepsilon}^{2}$

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{HKT(\lambda)}{n - 2İz(S_{\lambda}) + İz(S_{\lambda}S_{\lambda}^{T})} \tag{2.79}$$

dir. σ_{ε}^{2}'nin tahmini için λ, CV ya da GCV ölçütlerini minimum yapan değer olarak seçilir (Ruppert vd., 2003).

2.3.6.4. Akaike bilgi ölçütü (Akaike information criterion)

Parametrik problemler için klasik Akaike Bilgi Ölçütü (AIC), beklenen Kullback-Leibler bilgisinin yaklaşık olarak yansız tahmin edicisi olarak geliştirilmiştir. Akaike (1973) tarafından önerilen AIC bilgi ölçütü,

$$AIC(\lambda) = \log\{HKT(\lambda)\} + 2\,(sd_{kestirim}) / n \tag{2.80}$$

şeklindedir. Hurvich vd. (1998) doğrusal parametrik olmayan düzleştiricilerde düzleştirme parametresinin seçimi için AIC'yi önermişlerdir. Hurvich vd. (1998) tarafından geliştirilen bu ölçüt,

$$AIC_{c}(\lambda) = \log\frac{\left\|(\mathbf{S}_{\lambda} - \mathbf{I})\mathbf{y}\right\|^{2}}{n} + 1 + \frac{2\left\{İz(\mathbf{S}_{\lambda}) + 1\right\}}{n - İz(\mathbf{S}_{\lambda}) - 2} \tag{2.81}$$

şeklindedir. $AIC_{c}(\lambda)$ ölçütüne göre en uygun düzleştirme parametresi, $AIC_{c}(\lambda)$ ölçütünü minimum yapan değerdir (Hurvich vd., 1998; Lee, 2003).

Parametrik olmayan yöntemler esnek yöntemlerdir. Çünkü bu yöntemler bağımlı değişken ve bağımsız değişkenler arasındaki fonksiyonel ilişkinin şekli ile ilgili herhangi bir varsayım gerektirmez. Ancak bazen bağımlı değişken ile bazı bağımsız değişkenler arasındaki ilişkinin doğrusal olduğu bilinirken, bağımlı değişken ile diğer bağımsız değişkenler arasındaki ilişkinin şeklinin bilinmediği durumlar ile karşılaşılabilir. Bu durumda bilinen doğrusal ilişkiyi göz ardı eden parametrik olmayan bir modelin kestirilmesi doğru olmayan sonuçlara neden

olabilir. Çünkü doğrusal ilişki bazı önemli teorik sonuçlar içerebilir. Bu sorunu ortadan kaldırmak için yarı parametrik (semiparametric) modeller geliştirilmiştir. Yarı parametrik modeller hem parametrik hem de parametrik olmayan bileşenleri içerdiğinden doğrusallığı ve doğrusal olmamayı aynı anda ele almaktadır (Shi, 2009).

2.4. Yarı Parametrik Regresyon

Yarı parametrik regresyon modelinde bağımlı değişkenin bir ya da daha çok açıklayıcı değişkenle doğrusal olarak ilişkili olduğu, ancak eklenilen bazı açıklayıcı değişken ya da değişkenlerle arasındaki ilişkinin doğrusal olmayan bir ilişki içinde oldukları varsayılır (Speckman, 1988). Yarı parametrik modellerin en basit şeklinde bağımsız değişkenlerden biri (x diye adlandıracağımız) ile y bağımlı değişkeni arasında doğrusal bağımlılık yokken, diğer bağımsız değişkenler arasında doğrusal bağımlılık vardır. Bu durumda aşağıdaki yarı parametrik regresyon modeli kullanılır:

$$y_i = \mathbf{z_i^T}\boldsymbol{\beta} + m(x_i) + \varepsilon_i, \quad (1 \le i \le n) \tag{2.82}$$

Eş.(2.82)'de $\mathbf{z_i^T}\boldsymbol{\beta}$ modelin doğrusal parametrik bileşeni, $m(x_i)$ ise doğrusal olmayan parametrik olmayan bileşendir. $\mathbf{z_i}^T$, (1 x p) boyutlu parametrik kısma karşılık gelen i.gözlem vektörünü; $\boldsymbol{\beta}$, (p x 1) boyutlu bilinmeyen parametreler vektörünü; $m(x_i)$, modelin parametrik olmayan kısmına karşılık gelen ikinci mertebeden türevi alınabilen fonksiyonlar uzayının bir elemanını; x_i, i. gözlem değerini; ve ε_i, i. gözlem için hata terimini göstermektedir. Bu modelde amaç, parametrik kısmındaki parametreler vektörü $\boldsymbol{\beta}$'yı ve parametrik olmayan m(x) fonksiyonunu $\{y_i, z_i, x_i\}$ veri kümesinden tahmin etmektir. Bu model matris-vektör formunda,

$$\mathbf{y = Z\beta + m + \varepsilon} \tag{2.83}$$

şeklinde ifade edilir. Eş. (2.83)'de,

- **y**, (nx1) boyutlu gözlemler vektörünü,
- **Z,** parametrik kısma karşılık gelen (nxp) bağımsız değişkenler matrisini,
- $\boldsymbol{\beta}$, parametrik kısma karşılık gelen (px1) boyutlu parametreler vektörünü,
- $\mathbf{m} \in C^2[a,b]$, parametrik olmayan kısma karşılık gelen (nx1) boyutlu bilinmeyen fonksiyonlar vektörünü,
- $\boldsymbol{\varepsilon}$, (nx1) boyutlu ortalaması 0 ve varyansı σ^2 olan bağımsız ve aynı dağılımlı hata terimleri vektörünü göstermektedir.

Yarı parametrik regresyon modeli toplamsal modellerin (additive model) özel bir durumudur. Yarı parametrik modeller, doğrusal parametrik bileşen ve parametrik olmayan bileşenlerin her ikisini içerdiğinden, bu modellere kısmi doğrusal modeller (partially linear model) de denir. Bu modeller standart regresyon modellerine göre çok daha esnektir.

Yarı parametrik regresyon modeli, ilk olarak günlük ortalama hava sıcaklığı ile elektrik satışları arasındaki ilişkiyi incelemek için Engle vd. (1985) tarafından kullanılmıştır. Green vd. (1985), Green ve Yandell (1985), Heckman (1986) yarı parametrik regresyonda doğrusal olmayan bileşenin kestirimi için kullanılan eğrisel düzleştirme (spline smoothing) ile birleştirilen en küçük kareler yaklaşımını önermişlerdir. Chen (1988) yarı parametrik regresyon modelinde parametrik bileşen ve parametrik olmayan bileşenin aynı anda tahmin edildiği en küçük kareler yaklaşımını önermiştir. Speckman (1988) ve Robinson (1988) yarı parametrik regresyonda parametre kestirimleri için çekirdek düzleştirme yöntemini kullanmışlardır. Wahba (1990), Green ve Silverman (1994) yarı parametrik regresyon modelinde parametreleri eğrisel çizgi düzleştirme yöntemi ile tahmin etmişlerdir. Akdeniz ve Tabakan (2009) yarı parametrik regresyon modelinde parametrelerin kısıtlanmış Rigde tahmin edicilerini, Duran, Akdeniz ve Hu (2011) yarı parametrik regresyon modelinde Liu tipi tahmin edicileri incelemişlerdir.

Yarı parametrik regresyon modelinde parametreleri tahmin etmek için kullanılan yaklaşımların hepsi parametrik olmayan regresyon yöntemlerine dayanır. Yarı parametrik regresyon modelinde $\boldsymbol{\beta}$ ve $\mathbf{m}$ vektörlerini ve $\boldsymbol{\mu} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{m}$ ortalama

vektörünü tahmin etmek için çok sayıda yaklaşım vardır. Eğrisel çizgi düzleştirme yöntemi, çekirdek düzleştirme yöntemi, yerel doğrusal regresyon düzleştirme vb. gibi. Green vd. (1985), Engle vd. (1985), Wahba (1990) ve Green ve Silverman (1994) yarı parametrik regresyon modelinde parametre tahmini için eğrisel düzleştirme yöntemini, Speckman (1988) ve Robinson(1988) çekirdek düzleştirme yöntemini, Hamilton ve Truong (1997) yerel doğrusal regresyon düzleştirme yöntemini kullanmışlardır.

2.4.1. Yarı parametrik regresyon modeli için cezalandırılmış en küçük kareler yöntemi

Eş. (2.82)'deki yarı parametrik regresyon modelini veriye uydurmak için

$$\sum_{i=1}^{n}\left\{y_i - \mathbf{z}_i^T \boldsymbol{\beta} - m(x_i)\right\}^2$$

biçimindeki artık kareler toplamını minimum yapan m fonksiyonu ve $\boldsymbol{\beta}$ parametreleri elde edilmeye çalışılır. Ancak m üzerinde bir kısıt olmadıkça bu yaklaşım başarısız olur. Bir an için x_i 'lerin farklı olduğunu varsayalım. $\boldsymbol{\beta}$'nın herhangi bir değeri için m , $m(x_i) = y_i - \mathbf{z}_i^T \boldsymbol{\beta}$ interpolasyonu ile elde edilebilir, ancak $\boldsymbol{\beta}$ bilinmiyorsa bu yaklaşımla belirlenemez (Tabakan,2009). Bu sorun, $\boldsymbol{\beta}$ ve m 'in değerine karar vermek yerine cezalandırılmış kareler toplamı minimum yapılarak çözülür (Green ve Silverman, 1994). Yarı parametrik regresyon modeli için cezalandırılmış en küçük kareler toplamı, aşağıdaki gibidir:

$$S(\boldsymbol{\beta},\mathbf{m}) = \sum_{i=1}^{n}\left\{y_i - \mathbf{z}_i^T \boldsymbol{\beta} - m(x_i)\right\}^2 + \lambda \int_a^b \left\{m''(x)\right\}^2 dx \qquad (2.84)$$

Cezalandırılmış en küçük kareler toplamının minimum yapılması, kübik eğrisel çizgi fonksiyonlarından yararlanan eğrisel çizgi düzleştirme yöntemi esasına dayanır.

$x_1,...,x_n$ düğüm noktaları $x_1 < ... < x_n$ koşulu sağlandığında Teorem 1'den yararlanarak yarı parametrik regresyon modeli için cezalandırılmış en küçük kareler toplamı,

$$S(\boldsymbol{\beta}, \mathbf{m}) = \sum_{i=1}^{n} \left\{ y_i - \mathbf{z}_i^T \boldsymbol{\beta} - m(x_i) \right\}^2 + \lambda \int_a^b \{ m''(x) \}^2 dx$$
$$= (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{m})^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{m}) + \lambda \mathbf{m}^T \mathbf{K} \mathbf{m} \qquad (2.85)$$

şeklinde yazılır. Eğer $x_1 < ... < x_n$ koşulu sağlanmazsa yani x_i 'ler farklı ve sıralı değilse, N tekrarlanma matrisi (incidence matrix) yardımıyla sıralı hale getirilir. $x_1,...,x_n$ düğüm noktalarının farklı ve sıralı değerleri $s_1,...,s_q$ ile gösterilsin. $x_1,...,x_n$ ve $s_1,...,s_q$ arasındaki bağlantı n x q tekrarlanma matrisi yardımıyla elde edilir. Tekrarlanma matrisinin elemanları, $x_i = s_j$ ise $N_{ij} = 1$ diğer durumlarda $N_{ij} = 0$ 'dır. Buna göre, Eş. (2.85)'deki cezalandırılmış en küçük kareler toplamı,

$$S(\boldsymbol{\beta}, \mathbf{m}) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{N}\mathbf{m})^T (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta} - \mathbf{N}\mathbf{m}) + \lambda \mathbf{m}^T \mathbf{K} \mathbf{m} \qquad (2.86)$$

şeklinde yazılır. Veriler sıralı ise Eş. (2.85) ve veriler sıralı değil ise Eş. (2.86)'daki cezalı en küçük kareler toplamlarını minimum yapan $\boldsymbol{\beta}$ ve $\mathbf{m}$ vektörlerinin kestirimleri, bu eşitliklerin $\boldsymbol{\beta}$ ve $\mathbf{m}$ 'e göre türevi alınıp sıfıra eşitlenmesi ile bulunur. Eş. (2.86)'daki cezalı en küçük kareler toplamının $\boldsymbol{\beta}$ ve $\mathbf{m}$ 'e göre türevi alınıp sıfıra eşitlenirse elde edilen denklemler matris formunda

$$\begin{bmatrix} \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{N} \\ \mathbf{N}^T\mathbf{Z} & \mathbf{N}^T\mathbf{N} + \lambda\mathbf{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}^T \\ \mathbf{N}^T \end{bmatrix} \mathbf{y} \qquad (2.87)$$

şeklinde yazılabilir. Eş. (2.87)'de $\boldsymbol{\beta}$=0 olduğunda modelin parametrik kısmını içeren $\mathbf{Z}$ ve $\boldsymbol{\beta}$ göz ardı edilebilir ve

$$(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})\mathbf{m} = \mathbf{N}^T\mathbf{y} \qquad (2.88)$$

yazılabilir. $\hat{\mathbf{m}} = \mathbf{Nm}$ vektörünü elde etmek için **y** vektörüne uygulanan ve verilen bir $\lambda > 0$ sabitine bağlı düzleştirme matrisi,

$$\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T \tag{2.89}$$

olarak elde edilir. Eğer x_i 'ler farklı ve sıralı ise **N=I** 'dır ve Eş. (2.89)'daki düzleştirme matrisi,

$$\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1} \tag{2.90}$$

şekline dönüşür (Green ve Silverman, 1994; Tabakan, 2009).

Eş. (2.87)'den

$$\mathbf{Z}^T\mathbf{Z}\boldsymbol{\beta} = \mathbf{Z}^T(\mathbf{y} - \mathbf{Nm}) \tag{2.91}$$

$$(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})\mathbf{m} = \mathbf{N}^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \tag{2.92}$$

denklemleri yazılabilir. **m** biliniyorsa Eş. (2.91)'den $\boldsymbol{\beta}$'nın tahmini,

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{Nm}) \tag{2.93}$$

olup, $\boldsymbol{\beta}$ biliniyorsa Eş. (2.92)'den **m** vektörünün tahmini,

$$\hat{\mathbf{m}} = \mathbf{Nm} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) \tag{2.94}$$

olarak elde edilir.

Eş. (2.93) ve Eş.(2.94)' deki denklemler dönüşümlü biçimde kullanılarak, $\boldsymbol{\beta}$ ve **m'**nin tahminleri sırasıyla tekrarlı olarak elde edilir. Bu işleme güncelleme algoritması (backfitting algorithm) adı verilir. Bu algoritma Hastie ve Tibshirani (1990) tarafından önerilmiştir. Bu algoritma yarı parametrik regresyondaki

parametrik ve parametrik olmayan terimlerin tahminlerinin aynı anda elde edilebilmesi için yeterince esnek ve oldukça basit bir iteratif algoritmadır (Keele, 2008). n. adımda $\boldsymbol{\beta}$ ve $\mathbf{m}$ vektörlerinin tahmini bu algoritma ile aşağıdaki gibi elde edilir:

$$\hat{\boldsymbol{\beta}}^{(n)} = (\mathbf{Z}^{\mathsf{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}(\mathbf{y} - \hat{\mathbf{m}}^{(n)}) \tag{2.95}$$

$$\hat{\mathbf{m}}^{(n)} = \mathbf{S}_{\lambda}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}^{(n-1)}), \quad n=1,2,... \tag{2.96}$$

Eş. (2.95) 'deki algoritma, $\boldsymbol{\beta}$, cezalı en küçük kareler tahminine yakınsayıncaya kadar tekrarlı olarak devam eder.

2.4.2. Yarı parametrik regresyon modelinde Green ve Silverman yaklaşımı

$\boldsymbol{\beta}$ ve $\mathbf{m}$ vektörlerini tahmin etmek için güncelleme algoritması yerine, Green ve Silverman (1994) tarafından önerilen İterasyona gerek duymayan alternatif bir yaklaşım kullanılabilir. Bu yaklaşıma göre,

$$[\mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{Z}]\boldsymbol{\beta} = \mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{y} \tag{2.97}$$

denkleminden $\boldsymbol{\beta}$ vektörünün tahmini,

$$\hat{\boldsymbol{\beta}} = [\mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{Z}]^{-1}\mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_{\lambda})\mathbf{y} \tag{2.98}$$

olarak elde edilir. $\mathbf{m}$ vektörünün tahmini ise

$$\hat{\mathbf{m}} = \mathbf{S}_{\lambda}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \tag{2.99}$$

dir. Burada elde edilen tahmin ediciler kısmi eğrisel çizgi olarak adlandırılır (Green ve Silverman, 2000; Akdeniz ve Tabakan, 2009).

Bu yaklaşıma göre $\boldsymbol{\mu}$ ortalama vektörü,

$$\begin{aligned}\boldsymbol{\mu} &= \mathbf{Z}\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{S}_\lambda(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{S}_\lambda \mathbf{y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}\hat{\boldsymbol{\beta}} \\ &= \mathbf{S}_\lambda \mathbf{y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}\left[\mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}\right]\mathbf{Z}^{\mathsf{T}}(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y} \\ &= [\mathbf{S}_\lambda + \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^{\mathsf{T}}\mathbf{Z})^{-1}\tilde{\mathbf{Z}}^{\mathsf{T}}]\mathbf{y} \\ &= \mathbf{H}\,\mathbf{y} \end{aligned} \qquad (2.100)$$

olarak elde edilir. Burada $\tilde{\mathbf{Z}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{Z}$'dır. **H** matrisi ise doğrusal regresyondaki şapka matrisine benzer bir matristir.

2.4.3. Yarı parametrik regresyon modelinde çekirdek düzleştirme yöntemi

Çekirdek düzleştirme yöntemi Speckman (1988) ve Robinson (1988) tarafından önerilmiştir. Eğrisel çizgi düzleştirme yönteminden farklı olarak çekirdek düzleştirme yönteminde $\boldsymbol{\beta}$ ve **m** vektörleri ayrı ayrı tahmin edilir. Parametre tahminleri iki adımda bulunur.

Adım 1: Sabit $\boldsymbol{\beta}$ için Eş. (2.82)'deki model parametrik olmayan regresyon modeli gibi düşünülebilir. Buna göre Eş. (2.82),

$$y_i - \mathbf{z}_i^T\boldsymbol{\beta} = m(x_i) + e_i, \quad (1 \le i \le n) \qquad (2.101)$$

şeklinde yazılır. Bu modelde m, parametrik olmayan çekirdek kestirim yöntemi ile tahmin edilebilir. Bu durumda m fonksiyonunun çekirdek tahmini, $\hat{m}(x;\beta)$, aşağıdaki gibi bulunur.

$$\hat{m}(x;\beta) = \sum_{i=1}^{n} K_h(x, x_i)(y_i - \mathbf{z}_i^T\boldsymbol{\beta}) \qquad (2.102)$$

Burada Nadaraya ve Watson çekirdek kestiricisinin ağırlıkları, $K_h(x, x_i)$,

$$K_h(x, x_i) = \frac{K\left(\left(\frac{x_i - x}{h}\right)\right)}{\sum_{i=1}^{n} K\left(\left(\frac{x_i - x}{h}\right)\right)}$$

olup, bant genişliği h>0 dır (Speckman, 1988). Çekirdek tahminlerini elde etmek için,

$$\hat{m}_1(x) = \sum_{i=1}^{n} K_h(x, x_i)\, y_i \qquad (2.103)$$

$$\hat{m}_2(x) = \sum_{i=1}^{n} K_h(x, x_i)\, \mathbf{z_i}^T \qquad (2.104)$$

fonksiyonları tanımlansın. Bu fonksiyonlardan yararlanarak $\hat{m}(x;\boldsymbol{\beta})$

$$\hat{m}(x;\boldsymbol{\beta}) = \hat{m}_1(x) - \hat{m}_2(x)\boldsymbol{\beta} \qquad (2.105)$$

şeklinde yazılabilir.

Adım 2: Eş. (2.82)'deki modelde m(x) yerine $\hat{m}(x;\boldsymbol{\beta})$ yazılırsa

$$\begin{aligned} y_i &= \mathbf{z_i}^T \boldsymbol{\beta} + \hat{m}(x;\boldsymbol{\beta}) + \varepsilon_i, \quad (1 \le i \le n) \\ &= \mathbf{z_i}^T \boldsymbol{\beta} + \hat{m}_1(x) - \hat{m}_2(x)\boldsymbol{\beta} + \varepsilon_i \\ y_i - \hat{m}_1(x) &= [\mathbf{z_i}^T - \hat{m}_2(x)]\boldsymbol{\beta} + \varepsilon_i \\ \tilde{y}_i &= \tilde{\mathbf{z}}_i^T \boldsymbol{\beta} + \varepsilon_i \end{aligned} \qquad (2.106)$$

olarak elde edilir. Eş. (2.106)'da $\tilde{y}_i = y_i - \hat{m}_1(x)$ ve $\tilde{\mathbf{z}}_i^T = [\mathbf{z_i}^T - \hat{m}_2(x)]$ 'dir. Eş. (2.106)'daki modelde en küçük kareler yöntemi kullanılarak, $\boldsymbol{\beta}$'nın çekirdek kestirimi, $\hat{\boldsymbol{\beta}}_h$,

$$\hat{\boldsymbol{\beta}}_h = (\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}^T \tilde{\mathbf{y}} \qquad (2.107)$$

şeklinde bulunur. Bu eşitlikte $\tilde{\mathbf{Z}}$ ve $\tilde{\mathbf{y}}$ aşağıdaki gibidir:

$$\tilde{\mathbf{Z}} = (\tilde{z}_1,...,\tilde{z}_n)^T = (\mathbf{I} - \mathbf{W}(h))\mathbf{Z} \quad (2.108)$$

$$\tilde{\mathbf{y}} = (\tilde{y}_1,...,\tilde{y}_n)^T = (\mathbf{I} - \mathbf{W}(h))\mathbf{y} \quad (2.109)$$

Eş. (2.109)'da $\mathbf{W}(h)$ köşegen elemanları, yukarıda tanımlanan Nadaraya ve Watson çekirdek kestiricisinin ağırlıkları, $(K_h(x_i,x_j))$, olan köşegen bir matristir. $\mathbf{m} = (m(x_1),...,m(x_n))^T$ vektörünün tahmini,

$$\hat{m}(x;\hat{\beta}) = \hat{m}_1(x) - \hat{m}_2(x)\hat{\boldsymbol{\beta}} \quad (2.110)$$

eşitliğinden yararlanarak ve daha önce elde edilen $\hat{\boldsymbol{\beta}}_h$ değeri yerine yazılarak

$$\hat{\mathbf{m}}_h = \mathbf{W}(h)(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_h) \quad (2.111)$$

olarak bulunur. Yarı parametrik regresyon modelinin ortalama vektörü $\boldsymbol{\mu}_h$ ise,

$$\begin{aligned}\boldsymbol{\mu}_h &= \mathbf{Z}\hat{\boldsymbol{\beta}}_h + \hat{\mathbf{m}}_h = \mathbf{Z}\hat{\boldsymbol{\beta}}_h + \mathbf{W}(h)(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_h) \\ &= \mathbf{W}(h)\mathbf{y} + (\mathbf{I} - \mathbf{W}(h))\mathbf{Z}\hat{\boldsymbol{\beta}}_h \\ &= \mathbf{W}(h)\mathbf{y} + (\mathbf{I} - \mathbf{W}(h))\mathbf{Z}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T(\mathbf{I} - \mathbf{W}(h))\mathbf{y} \\ &= [\,\mathbf{W}(h) + \tilde{\mathbf{Z}}\,(\tilde{\mathbf{Z}}^T\,\mathbf{Z})^{-1}\tilde{\mathbf{Z}}^T(\mathbf{I} - \mathbf{W}(h))\,]\mathbf{y} \\ &= \mathbf{H}_h\,\mathbf{y} \end{aligned} \quad (2.112)$$

şeklinde elde edilir. $\hat{\mathbf{m}}$ yerine $\hat{\mathbf{m}}_h$ kullanıldığında Green ve Silverman tarafından önerilen $\hat{\boldsymbol{\beta}}$ kestirimine benzer olarak,

$$\hat{\boldsymbol{\beta}}_{h'} = [\mathbf{Z}^T\,(\mathbf{I} - \mathbf{W}(h))\,\mathbf{Z}]^{-1}\mathbf{Z}^T\,(\mathbf{I} - \mathbf{W}(h))\mathbf{y} \quad (2.113)$$

kestirimi elde edilir (Hong, 1998). Bu kestirim, Eş.(2.111)'de yerine konularak yarı parametrik regresyon modelinin kestirimi , $\hat{\mathbf{m}}_h$, elde edilir

2.4.4. Yarı parametrik regresyon modelinde yerel polinom regresyon (local polinomial) düzleştirme yöntemi

Yarı parametrik regresyon modelinde tahminler yerel doğrusal düzleştirici kullanılarak da elde edilebilir. Yerel doğrusal tahminler ya da daha genel olarak yerel polinom tahminler ilk olarak 1977'de Stone tarafından çalışılmıştır. Bu tahminler çekirdek kestirimine dayalı tahminler ile karşılaştırıldığında daha iyi özelliklere sahiptir. Çünkü bu tahminler uç noktalardaki tahminlerden kötü bir şekilde etkilenmeyen asimtotik yan ve varyans özelliğine sahip olmaları açısından dikkat çekicidir. Fan (1993), çekirdek ve eğrisel tahminleri içeren doğrusal tahmin edicilerin içerisinde yerel doğrusal tahminlerin en iyi yakınsama oranına ulaştığını göstermiştir. Hamilton ve Truong (1997) bu özelliklerinden dolayı yerel doğrusal düzleştiricileri yarı parametrik regresyon modeli için uyarlamışlardır.

$\mathbf{S}_x$, nxn boyutlu doğrusal düzleştirme matrisini göstermek üzere yarı parametrik model için **Z** matrisi ve **y** vektörü

$$\tilde{\mathbf{Z}} = (\mathbf{I} - \mathbf{S}_x)\mathbf{Z} \quad \text{ve} \quad \tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S}_x)\mathbf{y}$$

şeklinde tanımlansın. Speckman (1988)'nın yaklaşımına dayanarak **β** ve **m** vektörlerinin herhangi bir x noktasındaki yerel polinom tahmin edicileri

$$\hat{\boldsymbol{\beta}}_x = (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{y}} \tag{2.114}$$

$$\hat{\mathbf{m}}_x = \mathbf{S}_x(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}_x) \tag{2.115}$$

şeklinde elde edilir. Eş. (2.115)'deki 1xn boyutlu satır vektörü $\mathbf{S}_{xj}$,

$$\mathbf{S}_{xj} = \mathbf{t}^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x \tag{2.116}$$

şeklindedir ve $\mathbf{S}_{xj}$, j=1,…,n, nxn boyutlu $\mathbf{S}_x$ düzleştirme matrisinin j.satırını göstermektedir (Hamilton ve Truong, 1997). Eş. (2.116)'da $\mathbf{t}^T=(1,0,\ldots,0)_{1x(p+1)}$ olup $\mathbf{X}_x$ ve $\mathbf{W}_x$ matrisleri

$$\mathbf{X}_x=\begin{bmatrix}1 & (x_1-x) & (x_1-x)^2 & \cdots & (x_1-x)^p\\ 1 & (x_2-x) & (x_2-x)^2 & \cdots & (x_2-x)^p\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 1 & (x_n-x) & (x_n-x)^2 & \cdots & (x_n-x)^p\end{bmatrix}$$

$$\mathbf{W}_x=\begin{bmatrix}K_h(x_1-x) & 0 & \cdots & 0\\ 0 & K_h(x_2-x) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & K_h(x_n-x)\end{bmatrix}$$

dır. p=1 olduğunda $\boldsymbol{\beta}$ ve **m** vektörlerinin yerel doğrusal tahmin edicisi elde edilir.

2.4.5. Yarı parametrik regresyonda karışık doğrusal model yaklaşımı

Yarı parametrik regresyon modelinde parametrik olmayan m fonksiyonu karışık doğrusal model yaklaşımı ile tahmin edilebilir. Eş. (2.82)'de parametrik olmayan kısmı gösteren m(x) fonksiyonu p.inci dereceden eğrisel çizgi modeli olarak aşağıdaki gibi ifade edilsin:

$$m(x,\gamma)=\gamma_0+\gamma_1x+\ldots+\gamma_1x^p+\sum_{k=1}^{K}u_k(x-\kappa_k)_+^{\,p} \qquad (2.117)$$

Eş. (2.117)'deki gösterimden yararlanarak Eş. (2.82)'deki yarı parametrik regresyon modeli matris formunda

$$\mathbf{y}=\mathbf{Z}\boldsymbol{\beta}+\mathbf{m}(\mathbf{X},\gamma)+\boldsymbol{\varepsilon} \qquad (2.118)$$

şeklinde gösterilir. Bu modelde $(\boldsymbol{\beta},\gamma)$'nın eğrisel çizgi yöntemine göre tahmini

$$S(\boldsymbol{\beta},\mathbf{m})=\sum_{i=1}^{n}\left\{y_i - \mathbf{z}_i^T\boldsymbol{\beta} - m(x_i,\gamma)\right\}^2 + \lambda \mathbf{u}^T\mathbf{u} \qquad (2.119)$$

şeklindeki cezalandırılmış en küçük kareler kriteri minimum yapılarak elde edilir. Eş. (2.119)'daki model, karışık doğrusal model yaklaşımı kullanılarak

$$\mathbf{y} = \boldsymbol{\Lambda}(\boldsymbol{\beta},\gamma) + \mathbf{T}\mathbf{u} + \boldsymbol{\varepsilon} \qquad (2.120)$$

şeklinde ifade edilir. Burada

$$\boldsymbol{\Lambda} = \begin{pmatrix} z_{11} & \cdots & z_{1d} & 1 & x_1 & \cdots & x_1^p \\ z_{21} & \cdots & z_{1d} & 1 & x_2 & \cdots & x_2^p \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{1d} & 1 & x_n & \cdots & x_n^p \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} (x_1-\kappa_1)_+^p & \cdots & (x_1-\kappa_K)_+^p \\ (x_2-\kappa_1)_+^p & \cdots & (x_2-\kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_n-\kappa_1)_+^p & \cdots & (x_n-\kappa_K)_+^p \end{pmatrix}$$

olup, $\mathbf{u} = (u_1,...,u_K)^T \sim (0,\sigma_u^2)$ ve $\boldsymbol{\varepsilon} = (\varepsilon_1,...,\varepsilon_n) \sim (0,\sigma_\varepsilon^2)$'dir. Eş. (2.120)'deki model için varyans-kovaryans matrisi $\mathbf{V} = \mathbf{T}\mathbf{T}^T + \lambda\mathbf{I}$ olup, $\lambda = \sigma_\varepsilon^2/\sigma_u^2$'dır. Eş. (2.119)'daki gösterime benzer olarak cezalı en küçük kareler,

$$\frac{1}{\sigma_\varepsilon^2}(\mathbf{y}-\boldsymbol{\Lambda}\boldsymbol{\theta}-\mathbf{T}\mathbf{u})^T(\mathbf{y}-\boldsymbol{\Lambda}\boldsymbol{\theta}-\mathbf{T}\mathbf{u}) + \lambda\mathbf{u}^T\mathbf{u} \qquad (2.120)$$

şeklinde yazılır. Burada $\boldsymbol{\theta} = (\boldsymbol{\beta},\gamma)$' dır. Eş. (2.120)'ün $\boldsymbol{\theta}$ ve $\mathbf{u}$'ya göre türevi alınıp sıfıra eşitlenirse,

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Lambda}^T\mathbf{V}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\mathbf{V}^{-1}\mathbf{y} \qquad (2.121)$$

$$\mathbf{u} = \mathbf{T}^T\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\Lambda}\hat{\boldsymbol{\theta}}) \qquad (2.122)$$

olarak bulunur.

Cezalandırılmış eğrisel çizgi yaklaşımında düzleştirme parametresinin (λ) ve düğüm sayılarının seçilmesi (K) gerekir. Bu iki parametre içinde düzleştirme parametresinin seçimi daha önemlidir. Parametrik olmayan regresyonda ve yarı parametrik regresyonda düzleştirme parametresinin seçimi için bazı ölçütler geliştirilmiştir. Ancak karışık doğrusal model yaklaşımında düzleştirme parametresi $\lambda = \sigma_{\varepsilon}^2 / \sigma_{u}^2$ olduğu için λ'nın seçiminde herhangi bir ölçüte gerek duyulmaz (Ruppert vd., 2003; Liang, 2006).

3. DOĞRUSAL REGRESYON MODELİNDE ETKİ ANALİZİ

Basit doğrusal regresyon modelinde, kestirilen model doğru olsa bile elde edilen sonuçlar üzerinde dikkatli bir denetim yapmadan kestirilen modeli kullanmamak gerekir. Bu denetim süreci, genellikle, modelin yeterliliğinin incelenmesi süreci olarak bilinir. Kestirilen modelin geçerliliğini araştırmak için kullanılan ölçütlerden belirtme katsayısının büyük olması ya da t ve F test istatistiklerinin anlamlı olması kestirilen modelin her zaman incelenen veriye uygun bir model olduğunu göstermez. Bu nedenle, modelin yeterliğine ilişkin bazı ölçütlerde, test istatistiklerinin ve katsayıların belirlenmesinde her bir gözlemin rolüne önem verilerek ayrıntılı bir şekilde inceleme yapılır. Çünkü elde edilen sonuçlar, sadece bir gözleme bile bağlı olabilir. Bu tür gözlemlerin incelenmesi gerekir (Alpar, 2003). Bu gözlemler, tek bir bağımsız değişkenin olduğu modelde, x değerlerine karşılık gelen y değerlerinin dağılımını gösteren saçılım grafiği çizilerek kabaca belirlenebilir. Sonuç olarak gözlemlerin grafiksel olarak incelenmesi oldukça önemlidir (Chatterjee ve Price, 1991). Verinin saçılım grafiği çizildiğinde, veri kümesindeki gözlenen değerlerin çoğu kestirilen regresyon doğrusu etrafında dağılırken, bir ya da daha fazla gözlem değeri diğer gözlem değerlerinden uzakta bulunabilir. Bu gözlemler genellikle üç grupta ele alınır: aykırı değer (outlier), büyük kaldıraç değeri (high leverage) ve etkili gözlemler (influential observations).

X uzayının merkezine yakın ve büyük artık değerine sahip olan gözlemler y ekseni yönünde aykırı değer olarak tanımlanır. Aykırı değer (outlier) ifadesi genellikle y ekseni yönündeki aykırı değerler için kullanılır. Bu değerler ölçüm yanlışlığı, verilerin yanlış girilmesi gibi nedenlerden dolayı ortaya çıkabilir. Model denkleminin kestiriminde, bu gözlemlerin varlığı, modeldeki parametre kestirimlerini, hata varyansının kestirimini, parametre kestirimlerinin varyanslarını, bunlara bağlı olarak test istatistiklerini önemli ölçüde değiştirebilir. Kısaca bu gözlemlerin varlığında kestirim denklemine güvenilmez. Bu nedenle, eğer mümkünse bu değerler düzeltilmeli, düzeltilemiyorsa veri kümesinden çıkarılmalıdır. y ekseni yönünde aykırı değerler, kutu-çizgi grafiklerinden, saçılım grafiklerinden, artık grafiklerinden yararlanılarak ortaya çıkarılabilir. İkiden daha fazla açıklayıcı değişken olması durumunda, bu noktaların basit grafiksel yöntemlerle ortaya çıkarılması zordur.

Herhangi bir gözlemin x değeri, X uzayının merkezine uzakta ise bu gözlem, x ekseni yönünde aykırı değer olarak tanımlanır. Şapka matrisi $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$'un köşegen elemanı h_{ii} değerine, kaldıraç (leverage) değeri denir. Büyük h_{ii} değerine sahip gözlemlere büyük kaldıraç değerine sahiptir (high leverage) denir. Çünkü bu gözlemler, kestirilen regresyon doğrusunun altında ise doğruyu aşağı çekerek, kestirilen regresyon doğrusunun üstünde ise doğruyu yukarı çekerek kaldıraç görevi görmektedir. X ekseni yönündeki aykırı değerler, kaldıraç değeri h_{ii}'den yararlanarak ortaya çıkarılabilir (Türkan, 2008).

Hem büyük artık değerine sahip hem de X uzayının merkezine uzak olan gözlemler hem x, hem de y yönünde aykırı değerlerdir.

Modelden çıkarıldığında, modelin kestirim değerlerinde önemli ölçüde değişiklik meydana getiren gözlemler etkili gözlem olarak tanımlanır. y ekseni yönünde aykırı gözlemler ve x ekseni yönünde aykırı gözlemler etkili gözlem olabilir de olmayabilir de (Chatterjee vd., 2000; Albayrak, 2006; Kutner vd., 2004). Bu tez çalışmasında analiz sonuçlarını etkileyen gözlemler genel olarak etkili gözlemler olarak adlandırılacaktır.

Doğrusal regresyon modelinde bu tür gözlemleri ortaya çıkarmak için ilk olarak saçılım grafiğinden yararlanılır ancak bazen yalnız bu grafiklere bakılarak bu gözlemleri ortaya çıkarmak mümkün olmayabilir. Bu nedenle bu tür gözlemleri ortaya çıkarmak için Cook(1977), Belsley vd. (1980), Cook ve Weisberg (1982), Atkinson (1985), Hadi (1992) ve Pena (2005) tarafından çeşitli ölçütler geliştirilmiştir. Bu ölçütlerin hiçbirinin en iyi ölçüt olduğu söylenemez (Freund, 1998). Bu ölçütlerden en yaygın olarak kullanılanları Cook uzaklığı, Hadi'nin ölçütü, Pena'nın ölçütü ve COVRATIO ölçütüdür. Bu ölçütler sonraki alt bölümde ayrıntılı olarak incelenecektir.

3.1. Cook Uzaklığı

Cook, 1977'de doğrusal regresyon modelinde i.gözlem değerinin veri kümesinden çıkarılmasının parametre kestirimleri üzerindeki etkisini araştırmak için, tüm veri

kümesi kullanılarak elde edilen parametre kestirimleri $\hat{\boldsymbol{\beta}}$ ile i.gözlem veri kümesinden çıkarıldıktan sonra elde edilen parametre kestirimleri $\hat{\boldsymbol{\beta}}_{-i}$ arasındaki farka dayanan ve Cook uzaklığı adı verilen ,

$$C_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^T (\mathbf{X}^T \mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{p\hat{\sigma}^2} \qquad (3.1)$$

ölçütünün kullanılmasını önermiştir (Cook, 1977). Eş. (3.1)'de **X**, doğrusal modeldeki tasarım matrisi, n örneklemdeki gözlem sayısı, p parametre sayısı, **e** artıklar vektörü olup, artık varyansı $\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p}$'dır. Ayrıca $\hat{\boldsymbol{\beta}}_{-i}$, i.gözlem veri kümesinden çıkarıldıktan sonra elde edilen parametre kestirimleri ; $\mathbf{X}_{-i}$, i.gözleme ilişkin satır, **X** matrisinden çıkarıldıktan sonra kalan tasarım matrisi; $\mathbf{y}_{-i}$, i. gözlem çıkarıldıktan sonraki gözlemler vektörü olmak üzere $\hat{\boldsymbol{\beta}}_{-i} = (\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}$ 'dır. Bu eşitlikte $(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1} = (\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\,\mathbf{x}_i^T)^{-1}$'dır, burada , $\mathbf{x}_i^T$, **X** matrisinin i. satır vektörüdür. $(\mathbf{X}_{-i}^T\mathbf{X}_{-i})$ matrisinin tersi Sherman-Morrison-Woodbury teoreminden yararlanılarak aşağıdaki gibi bulunabilir.

Sherman-Morrison-Woodbury Teoremi: G tekil olmayan bir matris, a bir sabit, t ve c kolon vektörleri olmak üzere **,**

$$(G - atc^T)^{-1} = G^{-1} + \frac{aG^{-1}tc^TG^{-1}}{(1 - ac^TG^{-1}t)}$$

dır.

Bu teoreme göre G= $\mathbf{X}^T\mathbf{X}$, a=1, t= $\mathbf{x}_i$ ve $c^T = \mathbf{x}_i^T$ olmak üzere $(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}$,

$$(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i} \qquad (3.2)$$

şeklinde yazılır. Eş. (3.2)'de paydadaki ifade, $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ matrisinin köşegen elemanlarından yararlanılarak 1-h_{ii} yazılabilir ($h_{ii}=\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$). Eş. (3.2)'deki $(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}$ ifadesi, $\hat{\boldsymbol{\beta}}_{-i}=(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}$ eşitliğinde yerine konulursa,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{-i} &= \left[(\mathbf{X}^T\mathbf{X})^{-1}+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1-h_{ii}}\right]\mathbf{X}_{-i}^T\mathbf{y}_{-i} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y}-\mathbf{x}_i y_i)+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}} \\
&= \hat{\boldsymbol{\beta}}-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}} \qquad (3.3)
\end{aligned}$$

bulunur. Eş. (3.3)'den $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{-i}$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{-i} &= \hat{\boldsymbol{\beta}}-\left[\hat{\boldsymbol{\beta}}-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}}\right] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i-\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_{-i}^T\mathbf{y}_{-i}}{1-h_{ii}} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i-\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}[\mathbf{X}^T\mathbf{y}-\mathbf{x}_i y_i]}{1-h_{ii}} \\
&= \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i(1-h_{ii})-(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}[\mathbf{X}^T\mathbf{y}-\mathbf{x}_i y_i]}{1-h_{ii}} \\
&= \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}}-\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}}-\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}}{1-h_{ii}} \\
&\quad+\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}} \\
&= \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i y_i}{1-h_{ii}}-\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T\hat{\boldsymbol{\beta}}}{1-h_{ii}}
\end{aligned}$$

$$= \frac{(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}x_i\left[y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right]}{1-h_{ii}}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i} = \frac{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_i e_i}{1-h_{ii}} \tag{3.4}$$

olarak bulunur. Eş. (3.1)'deki $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}$ ifadesi yerine Eş. (3.4)'de bulunan $\frac{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}x_i e_i}{1-h_{ii}}$ ifadesi konulup, $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ olduğu da dikkate alınarak C_i ölçütü,

$$C_i = \frac{h_{ii}}{1-h_{ii}}\frac{r_i^2}{p} \tag{3.5}$$

şeklinde ifade edilebilir. Bu ölçüt tüm veri kümesi kullanılarak kestirilen gözlem değerlerinin vektörü $\hat{\mathbf{y}}$ ile i.gözlem değeri veri kümesinden çıkarıldıktan sonra kestirilen gözlem değerlerinin vektörü $\hat{\mathbf{y}}_{-i}$ arasındaki farka dayalı olarak elde edilen Cook uzaklığı,

$$C_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i})^{\mathrm{T}}(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i})}{p\hat{\sigma}^2} \tag{3.6}$$

şeklinde de ifade edilebilir. Cook (1977), C_i uzaklıklarının yaklaşık olarak p ve n-p serbestlik dereceli F dağıldığını belirtmiştir. C_i, bir önemlilik testi olarak kullanılmamasına rağmen, Cook her bir C_i değerinin p ve n-p serbestlik F ile karşılaştırılmasını önermiştir.

3.2. Hadi'nin Ölçütü

Hadi (1992) her bir gözlemin analiz sonuçları üzerindeki etkilerini araştırmak için bir ölçüt önermiştir. Bu ölçüt, etkili gözlemleri ortaya çıkarmak için kullanılan çok sayıda ölçütün sahip olmadığı birçok özelliklere sahip yeni bir ölçüttür (Ullah ve Pasha, 2009). Hadi'nin ölçütü,

$$H_i^2 = \frac{p}{(1-h_{ii})}\frac{d_i^2}{1-d_i^2} + \frac{h_{ii}}{(1-h_{ii})} \qquad (3.7)$$

şeklinde tanımlanır (Hadi, 1992). Eş. (3.7)'de p, parametre sayısı; e_i, i.gözlem için artık değeri; $\mathbf{e}^T\mathbf{e}$, tüm gözlemler için artık kareler toplamı olmak üzere, $d_i^2 = e_i^2/\mathbf{e}^T\mathbf{e}$ i. normalleştirilmiş artığın karesidir.

Hadi'nin ölçütünü diğer ölçütlerden ayıran bazı özellikler vardır. Bu ölçüt,

- Bağımlı değişkende konum ve ölçeğe göre değişmezdir,
- Açıklayıcı değişkenlerin tekil olmayan dönüşümlerine göre değişmezdir,
- Artıkların ve şapka matrisinin köşegen elemanlarının (h_{ii}) toplamsal bir fonksiyonudur,
- Karesel artıklara ve kaldıraç değerlerine (leverage) göre monoton olarak artar (Hadi, 1992).

Hadi Eş. (3.7)'deki ölçüt için bir kesim noktası önermiştir. Buna göre H_i^2

$$H_i^2 > \text{ ortalama}\,(H_i^2) + c\,\sqrt{\text{Var}(H_i^2)} \qquad (3.8)$$

ise i. gözlem etkili gözlemdir. Burada c, 2 ya da 3 gibi seçilmiş sabit bir değerdir. Ancak uç değerler ortalama ve varyans değerlerini şişirdiği için kesim noktası daha büyük olmaktadır. Bu nedenle Hadi, Eş. (3.8)'deki kesim noktasında ortalama ve varyans yerine bu değerlerin sağlam tahmin edicileri olan medyan ve medyan mutlak sapmanın (MAD) kullanılmasını önermiştir. Buna göre H_i^2

$$H_i^2 > \text{ortanca}(H_i^2) + 4.5\,\text{MAD}(H_i^2) \qquad (3.9)$$

ise i. gözlem etkili gözlemdir. Eş.(3.9)'da, $\text{MAD}(H_i^2)$= ortanca$\{\left|H_i^2 - \text{ortanca}(H_i^2)\right|\}/0.6745$'dir (Hadi, 1992).

3. 3. Pena'nın Ölçütü

Pena (2005) etkili gözlemleri ortaya çıkarmak için yeni bir ölçüt geliştirmiştir. Pena'nın yaklaşımında, veri kümesinden çıkarılan bir gözlemin parametre kestirimlerini nasıl etkilediğini ölçmek yerine, her bir gözlemin veri kümesindeki diğer gözlemler tarafından nasıl etkilendiği üzerinde durulur. Yani her bir gözlem için, veri kümesindeki diğer gözlemler çıkarılarak tahmin değerlerinde meydana gelen değişim ölçülür. Doğrusal regresyon modelinde her bir gözlemin veri kümesindeki diğer gözlemlerden nasıl etkilendiği

$$\mathbf{s_i} = (\hat{y}_i - \hat{y}_{i(1)}, \ldots, \hat{y}_i - \hat{y}_{i(n)})^T \tag{3.10}$$

vektörü dikkate alınarak incelenebilir (Pena, 2005). Burada $\hat{y}_{i(j)}$, j. gözlem veri kümesinden çıkarıldığında i. gözlem değerinin tahminidir. Pena, standartlaştırılmış $\mathbf{s_i}$ vektörünün karesel biçimi olarak ifade edilen S_i ölçütünü önermiştir. Buna göre Pena'nın geliştirdiği ölçüt aşağıdaki gibi tanımlanabilir:

$$S_i = \frac{\mathbf{s}_i^T \mathbf{s}_i}{p\,\mathrm{var}(\hat{y}_i)} \tag{3.11}$$

Eş. (3.11)'de $\mathrm{var}(\hat{y}_i) = s^2 h_{ii}$ olup, $s^2 = \mathbf{e}^T\mathbf{e}/(n-p)$ ve $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$'dir. Eş. (3.10)'daki $\hat{y}_i - \hat{y}_{i(j)}$ ifadesi, j. gözlem veri kümesinden çıkarıldığında i. gözlemin kestirimindeki değişim miktarıdır ve,

$$\hat{y}_i - \hat{y}_{i(j)} = \mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{-j} \tag{3.12}$$

dir. Eş. (3.12)'de $\hat{\boldsymbol{\beta}}_{-j} = (\mathbf{X}_{-j}^T\mathbf{X}_{-j})^{-1}\mathbf{X}_{-j}^T\mathbf{y}_{-j}$ olup bu eşitlikteki $(\mathbf{X}_{-j}^T\mathbf{X}_{-j})^{-1}$ ifadesi Sherman-Morrison-Woodbury teoremine göre Eş. (3.2)'dekine benzer şekilde,

$$(\mathbf{X}_{-j}^T\mathbf{X}_{-j})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - \mathbf{x}_j^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_j} \tag{3.13}$$

şeklinde yazılır. Eş. (3.13)'deki $(\mathbf{X}_{-j}^{T}\mathbf{X}_{-j})^{-1}$ ifadesi $\hat{\boldsymbol{\beta}}_{-j} = (\mathbf{X}_{-j}^{T}\mathbf{X}_{-j})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}$ eşitliğinde yerine konulursa $\hat{\boldsymbol{\beta}}_{-j}$,

$$\hat{\boldsymbol{\beta}}_{-j} = \left[(\mathbf{X}^{T}\mathbf{X})^{-1} + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}\right]\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}$$

$$= (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j} + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$= (\mathbf{X}^{T}\mathbf{X})^{-1}(\mathbf{X}^{T}\mathbf{y} - \mathbf{x}_j y_j) + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$= (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y} - (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}_{-j}^{T}\mathbf{y}_{-j}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$= \hat{\boldsymbol{\beta}} - (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}[\mathbf{X}^{T}\mathbf{y} - \mathbf{x}_j\mathbf{y}_j]}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j(1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j) + (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}[\mathbf{X}^{T}\mathbf{y} - \mathbf{x}_j y_j]}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

$$\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j} + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j} + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{y}}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$
$$- \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j y_j}{1-\mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j}$$

elde edilir. Burada $h_{jj} = \mathbf{x}_j^{T}(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j$ olduğu göz önüne alınırsa $\hat{\boldsymbol{\beta}}_{-j}$,

$$\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{y}_j}{1-h_{jj}} + \frac{(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{x}_j\mathbf{x}_j^{T}\hat{\boldsymbol{\beta}}}{1-h_{jj}}$$

$$= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_j(y_j - \mathbf{x}_j^{\mathsf{T}}\hat{\boldsymbol{\beta}})}{1 - h_{jj}}$$

$$\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_j e_j}{1 - h_{jj}} \tag{3.14}$$

şeklinde bulunur. Eş. (3.14)'den yararlanarak Eş. (3.12),

$$\hat{y}_i - \hat{y}_{i(j)} = \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}} - \mathbf{x}_i^{\mathsf{T}}\left[\hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_j e_j}{1 - h_{jj}}\right]$$

$$= \frac{\mathbf{x}_i^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_j e_j}{1 - h_{jj}} \tag{3.15}$$

olarak bulunur. Eş. (3.15)'den yararlanarak Eş. (3.11)'deki Pena ölçütü aşağıdaki gibi elde edilir:

$$S_i = \frac{1}{ps^2 h_{ii}} \sum_{j=1}^{n} \frac{h_{ji}^2 e_j^2}{(1 - h_{jj})^2} \tag{3.16}$$

(Pena, 2005). Pena'nın önerdiği S_i ölçütünün Cook uzaklığı ölçütüne göre bazı avantajları vardır:

- S_i ölçütünün beklenen değeri, aykırı değerler olmadığı hipotezi altında ve tüm h_{ii} değerleri küçük olduğunda, yaklaşık olarak 1/p 'dir(Pena, 2005). Başka bir deyişle, veri kümesinde aykırı değerler ya da büyük kaldıraç değerine sahip gözlemler olmadığında, tüm gözlemler aynı beklenen duyarlılığa sahiptir. S_i ölçütünün bu özelliği, beklenen değeri h_{ii} değerlerine bağlı olan Cook uzaklığı ölçütüne göre önemli bir avantaj sağlamaktadır.

- Çok sayıda açıklayıcı değişkenin olduğu büyük örneklemler için, S_i ölçütünün dağılımı yaklaşık olarak normal dağılacaktır. Bu özelliği de karmaşık asimtotik bir dağılıma sahip olan Cook uzaklığı ölçütüne göre üstünlük sağlamaktadır. Çünkü S_i ölçütünün normal dağılması, bu ölçüt için kesim noktalarının (cut-off values) bulunmasına olanak sağlamaktadır. Buna göre S_i ölçütü,

$$\frac{S_i - E(S_i)}{s(S_i)} \qquad (3.17)$$

değerinden büyük ise i. gözlem etkili gözlemdir. Burada $s(S_i)$, S_i ölçütünün standart sapmasını göstermektedir. S_i ölçütünün ortalaması ve standart sapması etkili gözlemlerden etkilendiğinden Pena, S_i ölçütü için Eş. (3.17)'deki kesim noktasından farklı olarak yeni bir kesim noktası önermiştir. Buna göre S_i ölçütü,

$$|S_i| \geq \text{ortanca}(S_i) + 4.5\ \text{MAD}(S_i) \qquad (3.18)$$

değerinden büyük ise i. gözlem etkili gözlemdir. Burada $\text{ortanca}(S_i)$, S_i değerinin ortancası ve $\text{MAD}(S_i) = \text{ortanca}|S_i - \text{ortanca}(S_i)|$ 'dir (Nurunnabi vd., 2010).

- Veri kümesinde bir grup aykırı değer ya da büyük kaldıraç değerine sahip gözlemler varsa, S_i ölçütü bu aykırı değerleri ya da büyük kaldıraç değerine sahip noktaları diğer noktalardan ayırt edecektir.

3.4. COVRATIO Ölçütü

Belsley vd. 1980'de doğrusal regresyon modelinde i.gözlem değerinin veri kümesinden çıkarılmasının parametre kestirimlerine ilişkin varyans-kovaryans matrisi üzerindeki etkisini araştırmak için, i.gözlem veri kümesinden çıkarıldıktan sonra elde edilen parametre kestirimlerine ilişkin varyans-kovaryans matrisinin determinantının, tüm veri kümesi kullanılarak elde edilen parametre kestirimlerine ilişkin varyans-kovaryans matrisinin determinantına oranı olarak tanımlanan,CR_i,

$$CR_i = \frac{\left|\hat{\sigma}^2_{-i}(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}\right|}{\left|\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}\right|}$$

$$= \left(\frac{\hat{\sigma}^2_{-i}}{\hat{\sigma}^2}\right)^p \frac{\left|(\mathbf{X}^T\mathbf{X})\right|}{\left|(\mathbf{X}_{-i}^T\mathbf{X}_{-i})\right|} \qquad (3.19)$$

ölçütünün kullanılmasını önermişlerdir. Bu ölçüye COVRATIO ölçütü denilmektedir. Varyans-kovaryans matrisinin determinantına genelleştirilmiş varyans da denilmektedir. Eş. (3.19)'daki ölçüt, i.gözlemin, katsayıların kestiriminin etkinliği üzerindeki etkisinin bir ölçüsü olarak yorumlanabilir (Belsley vd., 1980). CR_i ölçütünün değerinin 1 olması, parametre tahminlerinin etkinliği üzerinde i. gözlemin etkili olmadığını; 1'den büyük olması, i. gözlem veri kümesinden çıkarıldıktan sonra elde edilen katsayıların genelleştirilmiş varyanslarının tahminlerinde i. gözlemin bir iyileşme (azalma) oluşturduğunu; 1'den küçük olması ise i. gözlemin veri kümesi içinde bulunmasının parametre tahminlerinin genelleştirilmiş varyanslarını artırdığını gösterir (Rawlings vd., 1998). Belsley vd. (1980), CR_i ölçütü için bir kesim noktası önermiştir. Buna göre CR_i,

$$CR_i \begin{cases} < 1 - 3p/n \\ > 1 - 3p/n \end{cases} \qquad (3.20)$$

ise i. gözlemin etkili olduğunu gösterir.

CR_i ölçütü, aşağıdaki teoremden yararlanılarak standartlaştırılmış artıklara (r_i) ve kaldıraç değerlerine (h_{ii}) bağlı olarak da ifade edilebilir.

Teorem 1: B ve C, kxm boyutlu matrisler ya da (kx1) boyutlu vektörler olsun. A matrisi (kxk) boyutlu tekil olmayan bir matris ise

$$\left| \mathbf{A} - \mathbf{B}\mathbf{C}^T \right| = \left| \mathbf{A} \right| \left| \mathbf{I} - \mathbf{C}^T \mathbf{A}^{-1} \mathbf{B} \right| \qquad (3.21)$$

dır (Belsley vd., 1980). Eş. (3.19)'daki formülde $\left| (\mathbf{X}_{-i}^T \mathbf{X}_{-i}) \right|$ ifadesi Teorem 1'e göre, $\mathbf{A} = (\mathbf{X}^T \mathbf{X})$ ve $\mathbf{B} = \mathbf{C} = \mathbf{x}_i$ olmak üzere

$$\begin{aligned} \left| \mathbf{X}_{-i}^T \mathbf{X}_{-i} \right| &= \left| \mathbf{X}^T \mathbf{X} - \mathbf{x}_i \mathbf{x}_i^T \right| \\ &= \left| \mathbf{X}^T \mathbf{X} \right| (1 - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i) \\ &= \left| \mathbf{X}^T \mathbf{X} \right| (1 - h_{ii}) \end{aligned} \qquad (3.22)$$

şeklinde yazılabilir. Ayrıca Eş. (3.19)'daki formülde $\hat{\sigma}^2_{-i}$, i.gözlem veri kümesinden çıkarıldıktan sonra elde edilen artık varyansının tahminini ,$\hat{\sigma}^2$ ise tüm gözlemlerin bulunduğu veri kümesi için artık varyansının tahminini göstermekte olup, aralarında,

$$\hat{\sigma}^2_{-i} = \hat{\sigma}^2 \frac{n-p-r_i^2}{n-p-1} \tag{3.23}$$

şeklinde bir ilişki vardır (Belsley vd., 1980). Burada, $r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, standartlaştırılmış artık olup, n gözlem sayısını, p parametre sayısını göstermektedir. Eş. (3.22) ve Eş. (3.23)'den yararlanılarak Eş. (3.19)'daki COVRATIO ölçütü, r_i ve h_{ii}'ye bağlı olarak aşağıdaki gibi yazılabilir:

$$CR_i = \left(\frac{n-p-r_i^2}{n-p-1}\right)^p \frac{1}{(1-h_{ii})} \tag{3.24}$$

Eş. (3.24)'deki ifadede r_i^2 değeri küçük ve h_{ii} değeri büyük olduğunda CR_i ölçütünün değeri 1'den büyük olur; r_i^2 değeri büyük ve h_{ii} değeri küçük olduğunda CR_i ölçütünün değeri 1'den küçük olur. Her iki durumda da i. gözlemin, katsayıların genelleştirilmiş varyansları üzerinde etkili olduğu söylenir. Hem r_i^2 değeri hem de h_{ii} değeri büyük olduğunda (ya da her ikisi de küçük olduğunda) CR_i ölçütünün değeri 1'e yaklaşacaktır. Bu durumda i. gözlem etkili gözlem değildir.

4. YARI PARAMETRİK REGRESYON MODELİNDE ETKİ ANALİZİ

Regresyon modellerinde aykırı değerleri ve etkili gözlemleri ortaya çıkarmak için geliştirilen ölçütlerle ilgili çok sayıda çalışma vardır. Ancak, parametrik olmayan regresyon modellerinde, bu gözlemleri ortaya çıkarmak için geliştirilen ölçütleri ile ilgili çalışmalar son yıllarda artmıştır. Eubank (1985), Thomas (1991) ve Kim (1996) eğrisel çizgi düzleştirme yönteminde artıklar, kaldıraç değerleri (leverage) ve Cook uzaklığı üzerine çalışmalar yapmışlardır. Kim ve Kim (1998) çekirdek yoğunluk kestirimi için Cook uzaklığını önermişlerdir. Kim vd. (2001) yerel polinom regresyonda Cook uzaklığını önermiştir. Zhongyi ve Baocheng (2001) yarı parametrik doğrusal olmayan regresyon modelinde etkili gözlem ölçütlerini, Kim vd. (2002) yarı parametrik regresyon modelinde etkili gözlem ölçütlerini, Fung vd. (2002), yarı parametrik karışık modellerde etkili gözlem ölçütlerini ve Zhang vd. (2007), kısmi değişen katsayılı modellerde (partially varying-coefficient models) doğrusal kısma ilişkin katsayılar vektörü için Cook uzaklığı ölçütünü önermişlerdir.

Yarı parametrik regresyon modelinde etkili gözlem ölçütleri parametrik regresyon ve parametrik olmayan regresyon modellerindeki ölçütlerden farklı özelliklere sahiptir. Örneğin, $\boldsymbol{\beta}$'nın tahmini üzerinde etkili olan bir gözlem, **m** 'in tahmini üzerinde etkili olmayabilir. Bu nedenle yarı parametrik regresyonda etkili gözlem ölçütleri $\boldsymbol{\beta}$ ve **m**'nin tahminleri için ayrı ayrı incelenmelidir.

Bu tez çalışmasında, yarı parametrik regresyon modelinde etkili gözlem ölçütleri parametrik olmayan f(x) fonksiyonunun yerel polinom regresyon düzleştiricisi ile tahmin edildiği durum için önerilmiştir. Bundan sonraki alt bölümde yarı parametrik regresyon modelinde önerilen ölçütler ayrıntılı olarak incelenecektir.

4.1. Yarı Parametrik Regresyonda $\hat{\boldsymbol{\beta}}$ için Cook Uzaklığı

Yarı parametrik regresyon modelinde $\hat{\boldsymbol{\beta}}$ için Cook uzaklığı, $\tilde{C}_i$, Kim vd. (2002) tarafından parametrik regresyon modelindeki Cook uzaklığına benzer olarak aşağıdaki biçimde tanımlanmıştır:

$$\tilde{C}_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})^T (\tilde{\mathbf{Z}}^T \tilde{\mathbf{Z}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{\hat{\sigma}^2 iz(\tilde{\mathbf{H}})} \tag{4.1}$$

Burada $\tilde{\mathbf{Z}}$, Alt Bölüm 2.4.4'de tanımlanan yarı parametrik modelde, düzleştirme matrisine bağlı olarak ifade edilen tasarım matrisini; $\hat{\sigma}^2$ ise tüm gözlemlerin bulunduğu veri kümesi için artık varyansının tahminini; $\tilde{\mathbf{H}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T$ ise şapka matrisini göstermektedir. Eş. (4.1)'de $\hat{\boldsymbol{\beta}}_{-i} = (\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}$ olup, bu eşitlikteki $(\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1}$ ifadesi Sherman-Morrison-Woodbury teoremine göre,

$$(\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1} = (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} + \frac{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}}{1 - \tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i} \tag{4.2}$$

şeklinde yazılır. Eş. (4.2)'deki $(\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1}$ ifadesi $\hat{\boldsymbol{\beta}}_{-i} = (\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}$ eşitliğinde yerine konulursa $\hat{\boldsymbol{\beta}}_{-i}$,

$$\hat{\boldsymbol{\beta}}_{-i} = \left[(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1} + \frac{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}}{1 - \tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}\right]\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}$$

$$= (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i} + \frac{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}}{1 - \tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

olup, bu ifadede $\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i} = \left[\tilde{\mathbf{Z}}^T\tilde{\mathbf{y}} - \tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i\right]$ olduğu göz önüne alınarak,

$$\hat{\boldsymbol{\beta}}_{-i} = (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{y}} - \tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i) + \frac{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}}{1 - \tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$= (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T\tilde{\mathbf{y}} - (\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i + \frac{(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{y}}_{-i}}{1 - \tilde{\mathbf{z}}_i^T(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$\hat{\boldsymbol{\beta}}_{-i} = \hat{\boldsymbol{\beta}} - (\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i + \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^{\top}\tilde{\mathbf{y}}_{-i}}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i} \qquad (4.3)$$

bulunur. Eş. (4.3)'deki ifadeden,

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i} = (\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}_{-i}^{\top}\tilde{\mathbf{y}}_{-i}}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i} = (\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\left[\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{y}} - \tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i\right]}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i(1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i) - (\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\left[\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{y}} - \tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i\right]}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{\mathbf{z}}_i'(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{y}}}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$+ \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{h}_{ii}} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{h}_{ii}} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{y}}}{1-\tilde{h}_{ii}}$$

$$+ \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{h}_{ii}}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{h}_{ii}} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}\hat{\boldsymbol{\beta}}}{1-\tilde{h}_{ii}}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{y}}_i}{1-\tilde{h}_{ii}} - \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{\mathbf{z}}_i^{\top}\hat{\boldsymbol{\beta}}}{1-\tilde{h}_{ii}}$$

$$= \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\left[\tilde{y}_i - \tilde{\mathbf{z}}_i^{\top}\hat{\boldsymbol{\beta}}\right]}{1-\tilde{h}_{ii}}$$

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i} = \frac{(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{e}_i}{1-\tilde{h}_{ii}} \qquad (4.4)$$

biçiminde bulunur. Burada $\tilde{e}_i = \tilde{y}_i - \tilde{\mathbf{z}}_i^{\top}\hat{\boldsymbol{\beta}}$'dır. Eş.(4.1)'de verilen Cook uzaklığı formülünde, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}$ ifadesinin Eş.(4.4)'deki değeri yerine yazılırsa, artık değerlerine ve kaldıraç değerlerine bağlı olarak elde edilen Cook uzaklığı, $\tilde{C}_i$,

$$\tilde{C}_i = \frac{\tilde{\mathbf{z}}_i^{\top}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})(\tilde{\mathbf{Z}}^{\top}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_i\tilde{e}_i}{\hat{\sigma}^2(1-\tilde{h}_{ii})\,\text{İz}(\tilde{\mathbf{H}})}$$

$$= \frac{1}{\hat{\sigma}^2\,\text{İz}(\tilde{\mathbf{H}})}\frac{\tilde{h}_{ii}\tilde{e}_i^2}{(1-\tilde{h}_{ii})^2} \qquad (4.5)$$

şeklinde elde edilir (Kim vd., 2002). Yarı parametrik regresyon modeli için bulunan Cook uzaklığının dağılımı bilinmemektedir. Bu durumda gözlem numaraları x ekseninde ve $\tilde{C}_i$ değerleri y ekseninde olmak üzere Cook uzaklarının saçılım grafiği çizilir. Saçılım grafiğinde Cook uzaklığı aykırı olan gözlemlerin etkili gözlem olduğu sonucuna varılır.

4.2. Yarı Parametrik Regresyonda $\hat{m}$ için Cook Uzaklığı

Yerel polinom regresyon modelinde Alt Bölüm 2.3.2'de verilen, tasarım matrisi, $\mathbf{X}_x$, ve çekirdek ağırlıkları $K\{h^{-1}(x_i - x)\}$ olan ağırlık matrisi, $\mathbf{W}_x$, sırasıyla aşağıdaki gibidir:

$$\mathbf{X}_x = \begin{bmatrix} 1 & (x_1 - x) & (x_1 - x)^2 & \cdots & (x_1 - x)^p \\ 1 & (x_2 - x) & (x_2 - x)^2 & \cdots & (x_2 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & (x_n - x)^2 & \cdots & (x_n - x)^p \end{bmatrix}$$

$$\mathbf{W_x} = \begin{bmatrix} K_h(x_1 - x) & 0 & \cdots & 0 \\ 0 & K_h(x_2 - x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & K_h(x_n - x) \end{bmatrix}$$

Bu matrislerden yararlanarak, Alt Bölüm 2.3.2'de x noktası komşuluğundaki parametre kestirimi $\hat{\boldsymbol{\beta}}_{\mathbf{x}}$,

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}} = (\mathbf{X_x}^{\mathrm{T}}\mathbf{W_x}\mathbf{X_x})^{-1}\mathbf{X_x}^{\mathrm{T}}\mathbf{W_x}\mathbf{y} \tag{4.6}$$

olarak verilmişti. $\mathbf{t}_x(x_i) = (1,(x_i - x),\ldots,(x_i - x)^p)$ olmak üzere $m(x_i)$ fonksiyonunun yerel polinom düzleştirici $\hat{m}(x_i) = \mathbf{t}_x(x_i)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{x}}$'dir. Bu tanımlamalardan **m** vektörünün tahmini

$$\hat{\mathbf{m}} = \left(\hat{m}_x(x_1),...,\hat{m}_x(x_n)\right)^{\mathrm{T}} = \mathbf{H_x}\mathbf{y} \tag{4.7}$$

olarak yazılabilir. Eş. (4.7)'de $\mathbf{H_x} = \mathbf{X}_x(\mathbf{X_x}^{\mathrm{T}}\mathbf{W_x}\mathbf{X_x})^{-1}\mathbf{X_x}^{\mathrm{T}}\mathbf{W}_x$ olup, yerel şapka matrisi olarak adlandırılır ve $\mathbf{H_x}$'in (j, i)' inci elemanı, $h_x(j,i)$, aşağıdaki gibidir:

$$h_x(j,i) = \mathbf{t}_x(x_j)^{\mathrm{T}}(\mathbf{X_x}^{\mathrm{T}}\mathbf{W_x}\mathbf{X_x})^{-1}\mathbf{t}_x(x_i)K_h(x_i - x) \tag{4.8}$$

(Kim vd., 2001).

Yerel polinom regresyon modelinde i. gözleme ilişkin Cook uzaklığını elde etmek için ilk olarak veri kümesinin tüm gözlemleri kullanılarak elde edilen tahminin ($\hat{m}(x_i)$'nin) ve i. gözlem çıkarıldığında elde edilen tahminin ($\hat{m}_{-i}(x_i)$'in) bulunması gerekir. Burada i. gözlem çıkarıldığında $m(x_i)$'nin tahmini $\hat{m}_{-i}(x_i) = \mathbf{t}_{x_j}(x_i)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{x_i},-i}$ dir. Bu eşitlikte i. gözlem çıkarıldıktan sonra elde edilen parametre kestirimi $\hat{\boldsymbol{\beta}}_{x,-i} = (\mathbf{X}^{\mathrm{T}}_{x,-i}\mathbf{W}_{x,[-i]}\mathbf{X}_{x,-i})^{-1}\,\mathbf{X}^{\mathrm{T}}_{x,-i}\mathbf{W}_{x,[-i]}\mathbf{y}_{-i}$ olup, aşağıdaki tanımlamalar yardımıyla daha açık olarak elde edilebilir:

A ve **B** herhangi bir matris ya da vektör ve $\mathbf{W_x}$ yerel polinom regresyondaki ağırlık matrisi olmak üzere parçalanmış matrisler,

$$\mathbf{A}=\begin{bmatrix}\mathbf{a_i^T}\\ \mathbf{A}_{-i}\end{bmatrix},\quad \mathbf{B}=\begin{bmatrix}\mathbf{b_i^T}\\ \mathbf{B}_{-i}\end{bmatrix},\quad \mathbf{W}=\begin{bmatrix}K_h(a_i-x) & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_{x,[-i]}\end{bmatrix}$$

şeklinde tanımlanabilir. $\mathbf{A^TWB}$ matris çarpımı aşağıdaki gibidir:

$$\mathbf{A^TWB}=\begin{bmatrix}\mathbf{a_i} & \mathbf{A}_{-i}\end{bmatrix}\begin{bmatrix}K_h(a_i-x) & \mathbf{0}\\ \mathbf{0} & \mathbf{W_{x,[-i]}}\end{bmatrix}\begin{bmatrix}\mathbf{b_i^T}\\ \mathbf{B}_{-i}\end{bmatrix}$$

$$=K_h(a_i-x)\mathbf{a}_i\mathbf{b_i^T}+\mathbf{A_{-i}^T}\mathbf{W}_{x,[-i]}\mathbf{B}_{-i} \qquad (4.9)$$

Eş. (4.9)'da $\mathbf{A_{-i}^T}\mathbf{W}_{x,[-i]}\mathbf{B}_{-i}$ matris çarpımı, yalnız bırakılırsa,

$$\mathbf{A_{-i}^T}\mathbf{W}_{x,[-i]}\mathbf{B}_{-i}=\mathbf{A^TWB}-K_h(a_i-x)\mathbf{a}_i\mathbf{b_i^T} \qquad (4.10)$$

elde edilir. $(\mathbf{A_{-i}^T}\mathbf{W}_{x,[-i]}\mathbf{B}_{-i})$ matrisinin tersi,

$$(\mathbf{A_{-i}^T}\mathbf{W}_{x,[-i]}\mathbf{B}_{-i})^{-1}=[\,\mathbf{A^TWB}-K_h(a_i-x)\mathbf{a}_i\mathbf{b_i^T}\,]^{-1} \qquad (4.11)$$

şeklinde yazılabilir. Eş. (4.11)'deki ifade Sherman-Morrison-Woodbury teoreminden yararlanılarak $G=\mathbf{A^TWB}$, $a=K_h(a_i-x)$, $t=\mathbf{a}_i$ ve $c^T=\mathbf{b}_i$ olmak üzere

$$(\mathbf{A_{-i}^T}\ \mathbf{W}_{x,[-i]}\mathbf{B}_{-i})^{-1}=(\mathbf{A^TWB})^{-1}+\frac{(\mathbf{A^TWB})^{-1}\mathbf{a}_i\mathbf{b_i^T}(\mathbf{A^TWB})^{-1}K_h(a_i-x)}{1-\mathbf{b_i^T}(\mathbf{A^TWB})^{-1}\mathbf{a}_iK_h(a_i-x)} \qquad (4.12)$$

şeklinde yazılabilir. Yukarıdaki tanımlamalardan yararlanarak x noktası komşuluğunda i.gözlem çıkarıldıktan sonra elde edilen parametre kestirimi $\hat{\boldsymbol{\beta}}_{x,-i}$,

$$\hat{\boldsymbol{\beta}}_{x,-i} = (\mathbf{X}_{x,-i}^{T}\mathbf{W}_{x,[-i]}\mathbf{X}_{x,-i})^{-1}\,\mathbf{X}_{x,-i}^{T}\mathbf{W}_{x,[-i]}\mathbf{y}_{-i}$$

$$= [(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1} - \mathbf{t}_x(x_i)\,\mathbf{t}_x^{T}(x_i)K_h(x_i - x)]^{-1}[\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y} - \mathbf{t}_x(x_i)y_iK_h(x_i - x)]$$

$$= \left[(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1} + \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}}{1 - t_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}\right] *$$

$$[\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y} - \mathbf{t}_x(x_i)y_iK_h(x_i - x)]$$

$$= (\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\,\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y} \; - \; (\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\,\mathbf{t}_x(x_i)y_iK_h(x_i - x)$$

$$+ \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y}}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$- \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)y_iK_h(x_i - x)}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$\hat{\boldsymbol{\beta}}_{x,-i} = \hat{\boldsymbol{\beta}}_x - (\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\,\mathbf{t}_x(x_i)y_iK_h(x_i - x)$$

$$+ \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y}}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$- \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)y_iK_h(x_i - x)}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)} \qquad (4.13)$$

şeklinde yazılabilir. Eş. (4.13)'de $\hat{\boldsymbol{\beta}}_x$ dışındaki terimlerin paydaları eşitlenirse $\hat{\boldsymbol{\beta}}_{x,-i}$,

$$\hat{\boldsymbol{\beta}}_{x,-i} = \hat{\boldsymbol{\beta}}_x - \frac{(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)y_iK_h(x_i - x)}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$+ \frac{(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)y_iK_h(x_i - x)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$+ \frac{K_h(x_i - x)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)\mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{y}}{1 - \mathbf{t}_x^{T}(x_i)(\mathbf{X}_x^{T}\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{t}_x(x_i)K_h(x_i - x)}$$

$$- \frac{K_h(x_i - x)(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{t}_x(x_i) \mathbf{t}_x^T(x_i)(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{t}_x(x_i) y_i K_h(x_i - x)}{1 - \mathbf{t}_x^T(x_i)(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{t}_x(x_i) K_h(x_i - x)}$$

$$\hat{\boldsymbol{\beta}}_{x,-i} = \hat{\boldsymbol{\beta}}_x - \frac{(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{t}_x(x_i) y_i K_h(x_i - x) \left\{ y_i - \mathbf{t}_x^T(x_i) \hat{\boldsymbol{\beta}}_x \right\}}{1 - \mathbf{t}_x^T(x_i)(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{t}_x(x_i) K_h(x_i - x)} \qquad (4.14)$$

şeklinde elde edilir. i.gözlem çıkarıldıktan sonra x_j noktası etrafındaki $m_{-i}(x_j)$ fonksiyonunun yerel polinom tahmini, $\hat{m}_{-i}(x_j) = \mathbf{t}_{x_j}{}^T(x_j) \hat{\boldsymbol{\beta}}_{\mathbf{x}_j,-i}$ olup Eş. (4.14)'den yararlanarak,

$$\hat{m}_{-i}(x_j) = \mathbf{t}_{x_j}{}^T(x_j) \hat{\boldsymbol{\beta}}_{\mathbf{x}_j,-i}$$

$$= \mathbf{t}_{x_j}{}^T(x_j) \left[\hat{\boldsymbol{\beta}}_{x_j} - \frac{(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{\mathbf{x}_j})^{-1} \mathbf{t}_{x_j}(x_i) y_i K_h(x_i - x) \left\{ y_i - \mathbf{t}_{x_j}{}^T(x_i) \hat{\boldsymbol{\beta}}_{x_j} \right\}}{1 - \mathbf{t}_{x_j}{}^T(x_i)(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{\mathbf{x}_j})^{-1} \mathbf{t}_{x_j}(x_i) K_h(x_i - x)} \right]$$

$$= \mathbf{t}_{x_j}{}^T(x_j) \hat{\boldsymbol{\beta}}_{x_j} - \frac{\mathbf{t}_{x_j}(x_j)(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{\mathbf{x}_j})^{-1} \mathbf{t}_{x_j}(x_i) y_i K_h(x_i - x) \left\{ y_i - \mathbf{t}_{x_j}{}^T(x_i) \hat{\boldsymbol{\beta}}_{x_j} \right\}}{1 - \mathbf{t}_{x_j}{}^T(x_i)(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{\mathbf{x}_j})^{-1} \mathbf{t}_{x_j}(x_i) K_h(x_i - x)}$$

$$= \hat{m}(x_j) - \frac{\mathbf{t}_{x_j}(x_j)(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{x_j})^{-1} \mathbf{t}_{x_j}(x_i) y_i K_h(x_i - x) \left\{ y_i - \mathbf{t}_{x_j}{}^T(x_i) \hat{\boldsymbol{\beta}}_{x_j} \right\}}{1 - \mathbf{t}_{x_j}{}^T(x_i)(\mathbf{X}_{x_j}{}^T \mathbf{W}_{x_j} \mathbf{X}_{x_j})^{-1} \mathbf{t}_{x_j}(x_i) K_h(x_i - x)} \qquad (4.15)$$

olarak bulunur. Yerel polinom regresyon modelinde i. gözlem için Cook uzaklığı,C_i,

$$C_i = \sum_{j=1}^{n} \left\{ \hat{m}(x_j) - \hat{m}_{-i}(x_j) \right\}^2 \qquad (4.16)$$

şeklinde tanımlanabilir. Eş. (4.16)'daki eşitliği, artıklara ve kaldıraç değerlerine (leverage) bağlı olarak ifade edebiliriz. Yerel polinom regresyon modelinde yerel artıklar aşağıdaki gibi ifade edilir:

$$e_x(i) = y_i - \hat{m}_x(x_i) \qquad (4.17)$$

Eş. (4.17)'deki yerel artıklardan ve Eş. (4.8)'deki yerel şapka matrisinin elemanlarından yararlanarak, i.gözlem çıkarıldıktan sonra x_j noktasında yerel polinom tahmininde ortaya çıkacak olan değişim miktarı aşağıdaki gibi yazılabilir:

$$\hat{m}(x_j) - \hat{m}_{-i}(x_j) = h_{x_j}(j,i)\, e_{x_j}(i) / \left\{1 - h_{x_j}(i,i)\right\} \qquad (4.18)$$

Eş. (4.18)'den yararlanarak, yerel polinom regresyon modeli kullanıldığında, i.gözlem için Cook uzaklığı, C_i,

$$C_i = \sum_{j=1}^{n} \left\{\hat{m}(x_j) - \hat{m}_{-i}(x_j)\right\}^2 = \sum_{j=1}^{n} \left[h_{x_j}(j,i)\, e_{x_j}(i) / \left\{1 - h_{x_j}(i,i)\right\} \right]^2 \qquad (4.19)$$

şeklinde ifade edilir. $\hat{\mathbf{m}}$'nın yerel özelliğinden dolayı, $\hat{m}(x_j)$'ye yakın noktalardan daha çok etkilendiği dikkate alınarak, C_i'nin basitleştirilmiş bir uyarlaması, C_i^*, x_j yerine x_i yazarak,

$$C_i^* = \left\{\hat{m}(x_i) - \hat{m}_{-i}(x_i)\right\}^2 \qquad (4.20)$$

biçiminde tanımlanabilir. C_i^*, her zaman C_i'den daha küçüktür. Kim vd.(2001), etkili gözlemleri ortaya çıkarmada C_i^*'ın C_i'den etkili gözlemleri ortaya çıkarmada daha iyi bir ölçüt olduğunu belirtmişlerdir (Kim vd., 2001). x_j yerine x_i kullanılması nedeniyle, i. gözlemin veri kümesinde bulunmamasının $\hat{\mathbf{m}}$ kestiriminde oluşturacağı değişim miktarının, olduğundan ne kadar düşük tahmin edileceği, x_i noktası cıvarındaki yoğunluğa bağlı olacaktır. x_i noktası cıvarındaki yoğunluk az ise, olduğundan düşük tahmin miktarı küçük olacaktır, x_i noktası cıvarında çok gözlem bulunuyorsa değişim miktarının olduğundan düşük tahmin miktarı büyük olacaktır.

i.gözlem çıkarıldıktan sonra x_i noktası etrafındaki $m_{-i}(x_i)$ fonksiyonunun yerel polinom düzleştiricisi $\hat{m}_{-i}(x_i) = \mathbf{t}_{x_j}(x_i)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}_i,-i}$ olup, bu eşitlikte $\hat{\boldsymbol{\beta}}_{x_i,-i}$ değeri yerine yazıldığında ve $\mathbf{t}_{ii} \equiv \mathbf{t}_{x_i}(x_i)$, $\hat{m}_{x_i}(x_i) = \hat{m}(x_i)$ ve $\hat{m}_{-i}(x_i) = \hat{m}_{x_i,-i}(x_i)$ kısaltmaları kullanıldığında $\hat{m}_{-i}(x_i)$

$$\hat{m}_{-i}(x_i) = \mathbf{t_{ii}}\hat{\boldsymbol{\beta}}_{x_i} - \frac{\mathbf{t_{ii}}(\mathbf{X}_{x_i}{}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t_{ii}}K_h(x_i - x_i)\{y_i - \hat{m}_{x_i}(x_i)\}}{1 - \mathbf{t_{ii}}(\mathbf{X}_{x_i}{}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t_{ii}}K_h(x_i - x_i)}$$

$$= \hat{m}(x_i) - \frac{\mathbf{t_{ii}}(\mathbf{X}_{x_i}{}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t_{ii}}K_h(x_i - x_i)\{y_i - \hat{m}(x_i)\}}{1 - \mathbf{t_{ii}}(\mathbf{X}_{x_i}{}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t_{ii}}K_h(x_i - x_i)}$$

$$= \hat{m}(x_i) - \frac{h_{x_i}(i,i)e_{x_i}(i)}{1 - h_{x_i}(i,i)} \qquad (4.21)$$

olarak elde edilir. Eş. (4.21)'den yararlanarak yerel polinom modelinde i.gözlem için Cook uzaklığı yerel artıklara ve kaldıraç değerlerine göre aşağıdaki gibi ifade edilir:

$$C_i^* = \{\hat{m}(x_i) - \hat{m}_{-i}(x_i)\}^2 = \left(\frac{h_{x_i}(i,i)e_{x_i}(i)}{1 - h_{x_i}(i,i)}\right)^2 \qquad (4.22)$$

(Kim vd., 2001). Eş. (4.22)'de $h_{x_i}(i,i) = (\mathbf{X}_{x_i}{}^{T}\mathbf{W}_{x_i}\mathbf{X}_{\mathbf{x_i}})^{-1}K_h(0)$'dır. Eş. (4.22)'deki ifade, regresyon modelindeki Cook uzaklığına benzemektedir, ancak $p\hat{\sigma}^2$ sabiti yoktur. p serbestlik derecesini gösterdiği için, yerel polinom modelinde p için şapka matrisinin izi kullanılabilir, ancak burada sadece yerel şapka matrisi vardır. Şapka matrisinin yerine, $\sum_{i=1}^{n} h_{ii}$ ifadesi de kullanılabilir. Buna göre $\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 / \{n - \sum_{i=1}^{n} h_{ii}\}$ olarak elde edilebilir.

Yarı parametrik regresyon modelinde **β** ve **m**'nin yerel polinom tahminleri, Alt Bölüm 2.4.4.'de verildiği gibi sırasıyla, aşağıdaki gibidir:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{y}} \quad \text{ve} \quad \hat{\mathbf{m}}(\mathbf{x}) = \mathbf{S}_{\mathbf{x}}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \qquad (4.23)$$

Yarı parametrik regresyon modelinde gözlem değerleri vektörü $\hat{\mathbf{y}}$, Eş. (2.83)'den aşağıdaki gibi yazılabilir:

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} + \hat{\mathbf{m}}(\mathbf{x}) \qquad (4.24)$$

Burada $\mathbf{Z}\hat{\boldsymbol{\beta}}$, yarı parametrik regresyonda parametrik kısma karşılık gelmektedir ve $\mathbf{Z}\hat{\boldsymbol{\beta}}=\mathbf{Z}(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{y}}$'dır. Bu ifadede $\mathbf{Z}$, tasarım matrisi; $\tilde{\mathbf{Z}}$, düzleştirme matrisine bağlı olarak ifade edilen tasarım matrisi; $\tilde{\mathbf{y}}$, düzleştirme matrisine bağlı olarak ifade edilen gözlemler vektörünün tahmini olmak üzere, Alt bölüm 2.4.4'de verilen $\tilde{\mathbf{Z}}=(\mathbf{I}-\mathbf{S}_x)\mathbf{Z}$ ve $\tilde{\mathbf{y}}=(\mathbf{I}-\mathbf{S}_x)\mathbf{y}$ gösterimlerinden yararlanarak, $\mathbf{Z}\hat{\boldsymbol{\beta}}$,

$$\mathbf{Z}\hat{\boldsymbol{\beta}}=(\mathbf{I}-\mathbf{S}_x)^{-1}\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{T}(\mathbf{I}-\mathbf{S}_x)\mathbf{y}=\hat{\mathbf{H}}\mathbf{y} \tag{4.25}$$

olarak yazılabilir. Burada $\hat{\mathbf{H}}=(\mathbf{I}-\mathbf{S}_x)^{-1}\tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^{T}(\mathbf{I}-\mathbf{S}_x)$'dır. Eş. (4.24)'deki parametrik olmayan kısma karşılık gelen, $\hat{\mathbf{m}}(\mathbf{x})$, Eş. (4.23)'den yararlanarak aşağıdaki gibi ifade edilebilir:

$$\hat{\mathbf{m}}(\mathbf{x})=\mathbf{S}_x(\mathbf{y}-\mathbf{Z}\hat{\boldsymbol{\beta}})=\mathbf{S}_x(\mathbf{y}-\hat{\mathbf{H}}\mathbf{y})=\mathbf{S}_x(\mathbf{I}-\hat{\mathbf{H}})\mathbf{y}=\mathbf{H}^{*}\mathbf{y} \tag{4.26}$$

Eş. (4.26)'da $\mathbf{H}^{*}=\mathbf{S}_x(\mathbf{I}-\hat{\mathbf{H}})$'dır. Kim vd. (2002), Eş. (4.26)'dan yararlanarak yarı parametrik regresyonda $\hat{\mathbf{m}}$ için Cook uzaklığını aşağıdaki gibi ifade etmişlerdir:

$$C_i^{*}=\frac{\left\{\hat{m}(x_i)-\hat{m}_{-i}(x_i)\right\}^2}{\hat{\sigma}^2 \text{İz}(\mathbf{H}^{*})} \tag{4.27}$$

Eş.(4.22)'dekine benzer olarak, Eş. (4.27)'deki $\left\{\hat{m}(x_i)-\hat{m}_{-i}(x_i)\right\}^2$ ifadesi artıklara ve kaldıraç değerlerine göre $\left\{\hat{m}(x_i)-\hat{m}_{-i}(x_i)\right\}^2=\frac{(h_{ii}^{*}e_i^{*})^2}{(1-h_{ii}^{*})^2}$ şeklinde yazılırsa, C_i^{*},

$$C_i^{*}=\frac{(h_{ii}^{*}e_i^{*})^2}{(1-h_{ii}^{*})^2\hat{\sigma}^2 \text{İz}(\mathbf{H}^{*})} \tag{4.28}$$

olarak elde edilir. Burada $\mathbf{e}^{*}=(\mathbf{I}-\mathbf{H}^{*})\mathbf{y}$'dir (Kim vd., 2002). C_i^{*}'nın dağılımı bilinmemektedir. Bu durumda gözlem numaraları x ekseninde ve C_i^{*} değerleri y

ekseninde olmak üzere Cook uzaklarının saçılım grafiği çizilir. Saçılım grafiğinde Cook uzaklığı aykırı olan gözlemlerin etkili gözlem olduğu sonucuna varılır.

4.3. Yarı Parametrik Regresyonda $\hat{\mathbf{y}}$ için Cook Uzaklığı

Yarı parametrik regresyon modelinde gözlem değerleri vektörüne ilişkin Cook uzaklığını bulmak için, gözlem değerleri vektörünün tahmini olan $\hat{\mathbf{y}}$'yı $\mathbf{y}$'ye bağlı olarak elde etmek gerekmektedir.

Yarı parametrik regresyon modelinde gözlemler vektörü $\hat{\mathbf{y}}$, $\mathbf{y}$'ye bağlı olarak Eş. (4.25)'den ve Eş. (4.26)'dan yararlanarak aşağıdaki gibi elde edilir:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{Z}\hat{\boldsymbol{\beta}} + \hat{\mathbf{m}}(\mathbf{x}) \\ &= \widehat{\mathbf{H}}\mathbf{y} + \mathbf{H}^{*}\mathbf{y} \\ &= \breve{\mathbf{H}}\mathbf{y}\end{aligned} \tag{4.29}$$

Burada $\breve{\mathbf{H}}$ doğrusal regresyon modelindeki şapka matrisine benzemektedir ve $\breve{\mathbf{H}} = \widehat{\mathbf{H}} + \mathbf{H}^{*}$ dır.

Yarı parametrik regresyon modelinde $\hat{\mathbf{y}}$ için Cook uzaklığını artık değerlerine ve kaldıraç değerlerine bağlı olarak Eş. (3.6)'dakine benzer biçimde aşağıdaki gibi ifade edilir:

$$\begin{aligned}\breve{C}_i &= \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i})^{T}(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{-i})}{\hat{\sigma}^{2}\text{İz}(\breve{\mathbf{H}})} \\ &= \frac{\breve{h}_{ii}\breve{e}_i^{2}}{(1-\breve{h}_{ii})^{2}\hat{\sigma}^{2}\text{İz}(\breve{\mathbf{H}})}\end{aligned} \tag{4.30}$$

(Kim vd., 2002). Burada $\breve{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \breve{\mathbf{H}})\mathbf{y}$ ve $\breve{h}_{ii}$, $\breve{\mathbf{H}}$ matrisinin köşegen elemanıdır. $\breve{C}_i$'nın dağılımı bilinmemektedir. Bu durumda gözlem numaraları x ekseninde ve $\breve{C}_i$ değerleri y ekseninde olmak üzere Cook uzaklarının saçılım

grafiği çizilir. Saçılım grafiğinde Cook uzaklığı aykırı olan gözlemlerin, etkili gözlem olduğu sonucuna varılır.

4.4. Yarı Parametrik Regresyon Modeli için Hadi'nin Ölçütü

Bu bölümde Hadi'nin doğrusal regresyon modeli için önerdiği ölçüt, yarı parametrik regresyon modeli için yeniden tanımlanmıştır. Yarı parametrik regresyon modeli için Hadi'nin ölçütü, regresyon modelindekine benzer olarak Alt Bölüm 4.3'de verilen artık değerlerinden ve kaldıraç değerlerinden yararlanarak,

$$\tilde{H}_i^2 = \frac{p}{(1-\breve{h}_{ii})}\frac{\breve{d}_i^2}{1-\breve{d}_i^2} + \frac{\breve{h}_{ii}}{(1-\breve{h}_{ii})} \qquad (4.31)$$

şeklinde tanımlanabilir. Eş. (4.31)'de $\breve{d}_i^2 = \breve{e}_i^2/\breve{\mathbf{e}}^T\breve{\mathbf{e}}$, i. normalleştirilmiş artığın karesidir ve $\breve{h}_{ii}$, $\breve{\mathbf{H}}$ şapka matrisinin köşegen elemanıdır. Hadi (1992)'nin doğrusal regresyon modeli için önerdiği kesim noktası, yarı parametrik regresyon modeli için de kullanılabilir. Bu durumda c, 2 ya da 3 gibi uygun olarak seçilen sabit bir değer olmak üzere, $\tilde{H}_i^2$,

$$\tilde{H}_i^2 > \text{ ortalama } (\tilde{H}_i^2) + c\sqrt{\text{Var}(\tilde{H}_i^2)} \qquad (4.32)$$

ya da

$$\tilde{H}_i^2 > \text{ortanca}(\tilde{H}_i^2) + 4.5\text{MAD}(\tilde{H}_i^2) \qquad (4.33)$$

ise, i. gözlem etkili gözlemdir. Burada ortanca($\tilde{H}_i^2$), $\tilde{H}_i^2$ ölçütünün ortancasıdır . Eş.(4.33)'de MAD($\tilde{H}_i^2$), $\tilde{H}_i^2$ ölçütünün ortancasının mutlak sapması olup ,

$$\text{MAD}(\tilde{H}_i^2) = \text{ortanca}\left|\tilde{H}_i^2 - \text{ortanca}(\tilde{H}_i^2)\right|$$

olarak yazılır.

4.5. Yarı Parametrik Regresyon Modeli için Pena Ölçütü

Bu bölümde Pena (2005)'nın etkili gözlemleri ortaya çıkarmak için önerdiği ölçüt, yarı parametrik regresyon modeli için geliştirilmiştir. **m** vektörünü tahmin etmek için yerel polinom modeli kullanıldığında elde edilen kestirim değerleri vektörü, $\hat{\mathbf{y}}$ aşağıdaki gibi yazılabilir:

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{Z}\hat{\boldsymbol{\beta}} + \hat{\mathbf{m}}(\mathbf{x}) \\ &= \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{S}_x(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{S}_x\mathbf{y} - \mathbf{S}_x\mathbf{Z}\hat{\boldsymbol{\beta}} \\ &= (\mathbf{I} - \mathbf{S}_x)\mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{S}_x\mathbf{y} \\ \hat{\mathbf{y}} &= \tilde{\mathbf{Z}}\hat{\boldsymbol{\beta}} + \mathbf{S}_x\mathbf{y}\end{aligned} \tag{4.34}$$

Burada $\mathbf{S}_x=\mathbf{X}_x(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x$ dır. i. gözlemin kestirim değeri, $\hat{y}_i$'yı elde etmek için $\mathbf{S}_x$ matrisinin i. satır vektörü $\mathbf{S}_{x_i} = \mathbf{t}^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x$, **y** vektörü ile çarpılıp, $\tilde{\mathbf{Z}}\hat{\boldsymbol{\beta}}$'nın i. satır elemanı $\tilde{\mathbf{z}}_i^T\hat{\boldsymbol{\beta}}$ ile toplanır (Hamilton ve Truong, 1997). Buna göre , i. gözlemin kestirim değeri,

$$\hat{y}_i = \tilde{\mathbf{z}}_i^T\hat{\boldsymbol{\beta}} + \mathbf{t}_{x_i}(x_i)\hat{\boldsymbol{\beta}}_{x_i} \tag{4.35}$$

şeklinde yazılabilir. $\mathbf{t}_{ii} \equiv \mathbf{t}_{x_i}(x_i)$ tanımından yararlanarak Eş. (4.35),

$$\hat{y}_i = \tilde{\mathbf{z}}_i^T\hat{\boldsymbol{\beta}} + \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} \tag{4.36}$$

olarak ifade edilir. j. gözlem çıkarıldıktan sonra elde edilen kestirim değeri, $\hat{y}_{i,-j}$,

$$\hat{y}_{i,-j} = \tilde{\mathbf{z}}_i^T\hat{\boldsymbol{\beta}}_{-j} + \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i,-j} \tag{4.37}$$

biçiminde ifade edilir. Eş. (4.37)'de j. gözlem çıkarıldıktan sonra elde edilen parametrik kısma karşılık gelen $\hat{\boldsymbol{\beta}}_{-j}$ değeri, Eş. (4.4)'dekine benzer olarak,

$$\hat{\boldsymbol{\beta}}_{-j} = \hat{\boldsymbol{\beta}} - \frac{(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_j\tilde{e}_j}{1-\tilde{h}_{jj}} \tag{4.38}$$

şeklinde elde edilir. Eş. (4.38)'de j. gözlem çıkarıldıktan sonra elde edilen parametrik olmayan kısma karşılık gelen $\mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i,-j}$ ifadesi, Eş.(4.14)'den yararlanarak,

$$\begin{aligned}
\mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i,-j} &= \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} - \frac{\mathbf{t}_{ii}(\mathbf{X}_{x_i}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t}_{ij}K_h(x_j - x_i)\{y_j - \hat{m}_{x_i}(x_j)\}}{1-\mathbf{t}_{ij}^{T}(\mathbf{X}_{x_i}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t}_{ij}K_h(x_j - x_i)} \\
&= \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} - \frac{\mathbf{t}_{ii}(\mathbf{X}_{x_i}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t}_{ij}K_h(x_j - x_i)\{y_j - \hat{m}_{x_i}(x_j)\}}{1-\mathbf{t}_{ij}^{T}(\mathbf{X}_{x_i}^{T}\mathbf{W}_{x_i}\mathbf{X}_{x_i})^{-1}\mathbf{t}_{ij}K_h(x_j - x_i)} \\
&= \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} - \frac{h_{\mathbf{x}_i}(i,j)\,e_{x_i}(j)}{1-h_{\mathbf{x}_i}(j,j)}
\end{aligned} \tag{4.39}$$

şeklinde bulunur. Eş. (4.38) ve Eş. (4.39)'dan yararlanarak $\hat{y}_i - \hat{y}_{i,-j}$,

$$\hat{y}_i - \hat{y}_{i,-j} = \tilde{\mathbf{z}}_i^{T}\hat{\boldsymbol{\beta}} + \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} - \tilde{\mathbf{z}}_i^{T}\hat{\boldsymbol{\beta}}_{-j} - \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i,-j}$$

$$\begin{aligned}
&= \tilde{\mathbf{z}}_i^{T}\hat{\boldsymbol{\beta}} + \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} - \tilde{\mathbf{z}}_i^{T}\hat{\boldsymbol{\beta}} + \frac{\tilde{\mathbf{z}}_i^{T}(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_j\tilde{e}_j}{1-\tilde{h}_{jj}} - \mathbf{t}_{ii}\hat{\boldsymbol{\beta}}_{x_i} + \frac{h_{\mathbf{x}_i}(i,j)\,e_{x_i}(j)}{1-h_{\mathbf{x}_i}(j,j)} \\
&= \frac{\tilde{h}_{ij}\tilde{e}_j}{1-\tilde{h}_{jj}} + \frac{h_{\mathbf{x}_i}(i,j)\,e_{x_i}(j)}{1-h_{\mathbf{x}_i}(j,j)}
\end{aligned} \tag{4.40}$$

olarak elde edilir. Eş. (4.40)'da $\tilde{h}_{ij} = \tilde{\mathbf{z}}_i^{T}(\tilde{\mathbf{Z}}^{T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{z}}_j$'dır. Buna göre yarı parametrik regresyon modeli için $\mathbf{s}_i = (\hat{y}_i - \hat{y}_{i(1)},\ldots,\hat{y}_i - \hat{y}_{i(n)})^{T}$ olmak üzere, Pena ölçütü, $\tilde{S}_i$,

$$\tilde{S}_i = \frac{\mathbf{s}_i^T \mathbf{s}_i}{p\,\mathrm{var}(\hat{y}_i)}$$

$$= \frac{1}{p\ \mathrm{var}(\hat{y}_i)} \sum_{j=1}^{n} \left(\frac{\tilde{h}_{ij}\tilde{e}_j}{1-\tilde{h}_{jj}} + \frac{h_{\mathbf{x}_i}(i,j)\, e_{x_i}(j)}{1-h_{\mathbf{x}_i}(j,j)} \right)^2 \qquad (4.41)$$

şeklinde elde edilir.

Pena'nın doğrusal regresyon modeli için önerdiği kesim noktası yarı parametrik regresyon modeli için de kullanılabilir. Buna göre $\tilde{S}_i$ ölçütü,

$$\left|\tilde{S}_i\right| \geq \mathrm{ortanca}(\tilde{S}_i) + 4.5\ \mathrm{MAD}\,(\tilde{S}_i) \qquad (4.42)$$

ise i. gözlem etkili gözlemdir. Burada ortanca($\tilde{S}_i$), $\tilde{S}_i$ değerinin ortancasıdır ve $\tilde{S}_i$ ölçütünün ortancasının mutlak sapması aşağıdaki gibidir:

$$\mathrm{MAD}\,(\tilde{S}_i) = \mathrm{ortanca}\left|\tilde{S}_i - \mathrm{ortanca}(\tilde{S}_i)\right|$$

4.6. Yarı Parametrik Regresyon Modeli için COVRATIO Ölçütü

Bu bölümde, parametrik regresyon modelinde etkili gözlemleri ortaya çıkarmak için kullanılan COVRATIO ölçütü, yarı parametrik regresyon modeli için geliştirilmiştir. Yarı parametrik regresyon modelinde **m** vektörünü tahmin etmek için yerel polinom düzleştiricisi kullanıldığında elde edilen artık terimleri vektörü, Eş. (4.34)'den,

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \tilde{\tilde{\mathbf{Z}}}\hat{\boldsymbol{\beta}} - \mathbf{S}_{\mathbf{x}}\mathbf{y}$$

$$= (\mathbf{I} - \mathbf{S}_{\mathbf{x}})\mathbf{y} - \tilde{\tilde{\mathbf{Z}}}\hat{\boldsymbol{\beta}}$$

$$\mathbf{e} = \tilde{\mathbf{y}} - \tilde{\tilde{\mathbf{Z}}}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{e}} \qquad (4.43)$$

şeklinde elde edilir. Eş. (4.43)'den görüldüğü gibi yarı parametrik regresyon modelinde artık vektörü, modelin sadece doğrusal kısmına karşılık gelen artık

vektörüne bağlı olarak da ifade edilebilir. Yarı parametrik regresyon modelinde $\hat{\sigma}^2_{-i}$ ile $\hat{\sigma}^2$ arasındaki ilişki, Eş. (3.23)'dekine benzer bir ifade

$$\hat{\sigma}^2_{-i} = \hat{\sigma}^2 \frac{n-p-\tilde{r}_i^2}{n-p-1} \tag{4.44}$$

şeklinde yazılabilir. Burada $\hat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-p} = \frac{\tilde{\mathbf{e}}^T\tilde{\mathbf{e}}}{n-p}$ ve $\tilde{r}_i = \frac{\tilde{e}_i}{\hat{\sigma}\sqrt{1-\tilde{h}_{ii}}}$ olup, p parametre sayısını göstermektedir.

Parametrik regresyon modelindeki COVRATIO ölçütüne benzer olarak yarı parametrik regresyon modelinde COVRATIO ölçütü,

$$\begin{aligned} \widetilde{CR}_i &= \frac{\left|\hat{\sigma}^2_{-i}(\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})^{-1}\right|}{\left|\hat{\sigma}^2(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\right|} \\ &= \left(\frac{\hat{\sigma}^2_{-i}}{\hat{\sigma}^2}\right)^p \frac{\left|(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})\right|}{\left|(\tilde{\mathbf{Z}}_{-i}^T\tilde{\mathbf{Z}}_{-i})\right|} \end{aligned} \tag{4.45}$$

şeklinde elde edilir. Eş. (3.22)'den yararlanarak COVRATIO ölçütü standartlaştırılmış artıklara ve kaldıraç değerlerine bağlı olarak aşağıdaki gibi elde edilir:

$$\widetilde{CR}_i = \left(\frac{n-p-\tilde{r}_i^2}{n-p-1}\right)^p \frac{1}{(1-\tilde{h}_{ii})} \tag{4.46}$$

Yarı parametrik regresyon modelinde de COVRATIO için kesim noktası, doğrusal regresyonda olduğu gibi 1 alınabilir.

5. UYGULAMA

Bu bölümde yarı parametrik regresyon modelinde etkili gözlemleri ortaya çıkarmak için önerilen ölçütlerin bu gözlemleri ortaya çıkarmadaki başarılarını araştırmak amacıyla ilk olarak Kim vd. (2002)'in kullandığı diyabet hastalarına ilişkin veri kümesi incelenmiştir. Kim vd. (2002)'in çalışmasında Cook uzaklığı ölçütü kullanılarak belirlenen etkili gözlemlerin, bu tez çalışmasında önerilen ölçütler tarafından da etkili olarak bulunup bulunmadığı araştırılmıştır. Daha sonra regresyon modelinde büyük örneklemlerde hem büyük kaldıraç değeri, hem de aykırı değer olan gözlemler bulunduğunda, bu gözlemleri belirlemede diğer ölçütlere göre daha başarılı olan Pena ölçütünün yarı parametrik regresyon modelinde de bu gözlemleri belirlemede başarılı olup olmadığını araştırmak için yapay bir veri kümesi türetilmiş, ölçüt değerleri elde edilmiş ve elde edilen sonuçlar tartışılmıştır. Son olarak yarı parametrik regresyon modelinde hangi durumlarda hangi ölçülerin etkili gözlemleri belirlemede daha başarılı olduklarını araştırmak amacıyla simülasyon çalışması yapılmış ve elde edilen sonuçlar tartışılmıştır.

Bu tez çalışmasında veri kümesinden her bir gözlem tek tek çıkartıldığında meydana gelen değişim incelenmiştir. Gizleme(masking) ve sürükleme(sweeping) etkisi dikkate alınmamıştır.

5.1. Gerçek Bir Veri Kümesi Üzerinde Uygulama

Bu tez çalışmasında önerilen ölçütlerin etkili gözlemleri ortaya çıkarmadaki başarılarını incelemek için Kim vd. (2002)'in çalışmalarında kullandığı diyabet hastalarına ilişkin veri kümesinden yararlanılmıştır. Bu veri kümesi için

$$y_i = \mathbf{z}_i^T\boldsymbol{\beta} + m(x_i) + \varepsilon_i, \quad (1 \le i \le n) \tag{5.1}$$

şeklindeki yarı parametrik regresyon modeli kullanılmıştır. Bu modelde bağımlı değişken y=C-peptid hormonu konsantrasyonu, doğrusal kısma ilişkin açıklayıcı değişken z = kord kanı baz açığı (base deficit) ve doğrusal olmayan kısma ilişkin açıklayıcı değişken x=yaş olarak alınmıştır. Parametre tahminlerini elde etmek için

yerel doğrusal düzleştirici kullanılmış ve bant genişliği (h) çapraz geçerlilik ölçütüne göre h= 5.6 olarak bulunmuştur.

Kim vd. (2002) bu veri kümesini kullanarak yarı parametrik regresyon modelinde $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{m}}$ ve $\hat{\mathbf{y}}$ üzerinde etkili olan gözlemleri ortaya çıkarmak için önerdikleri Cook uzaklığı ölçütlerini ($\tilde{C}_i, C_i^*$ ve $\breve{C}_i$) incelemişlerdir. $\tilde{C}_i$ ölçüt değerlerine göre, $\hat{\boldsymbol{\beta}}$ üzerinde etkili olan gözlemler, 6. ve 34. gözlemler; C_i^* ölçüt değerlerine göre, $\hat{\mathbf{m}}$ üzerinde etkili olan gözlemler, 22. ve 13. gözlemler; $\breve{C}_i$ ölçüt değerlerine göre, $\hat{\mathbf{y}}$ üzerinde etkili olan gözlem, 34. gözlem olarak bulunmuştur.

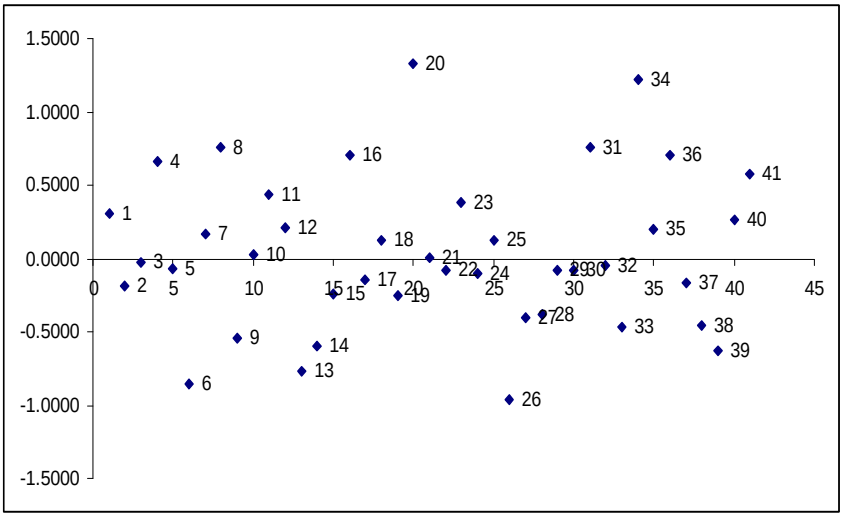
Bu tez çalışmasında aynı veri kümesine uygulanan yarı parametrik regresyon modelinde etkili gözlemleri ortaya çıkarmak için önerilen Hadi'nin ölçütü ($\tilde{H}_i^2$), Pena'nın ölçütü ($\tilde{S}_i$) ve COVRATIO ölçütü ($\tilde{CR}_i$) değerleri elde edilmiştir. Eş. (5.1)'deki model için yerel doğrusal düzleştirici kullanılarak tahmin edilen artık değerleri ($\tilde{e}_i$, $\breve{e}_i$ ve $e_{x_i}(i)$) ve kaldıraç değerleri ($\tilde{h}_{ii}$, $\breve{h}_{ii}$ ve $h_{\mathbf{x}_i}(i,i)$) Çizelge 5.1'de verilmiştir:

Çizelge 5.1. Diyabet verisine ilişkin artık değerleri ve kaldıraç değerleri

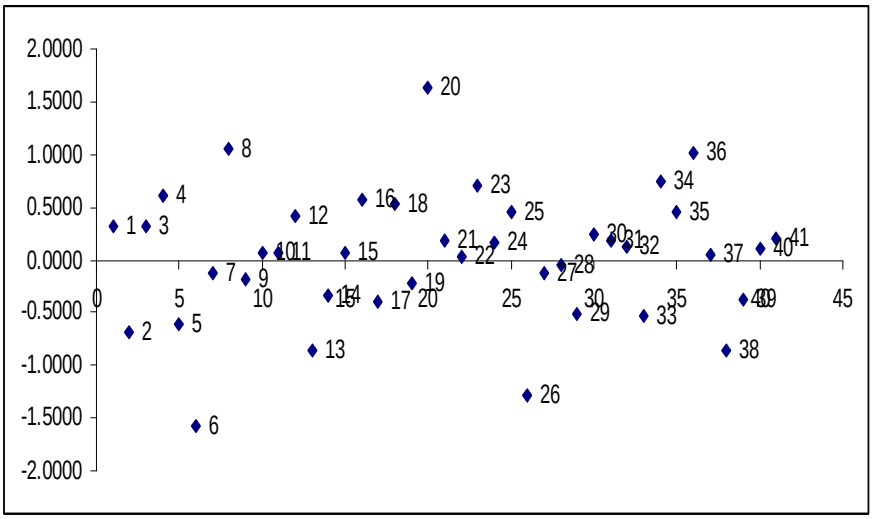
	y	x	z	$\tilde{e}_i$	$\breve{e}_i$	$e_{x_i}(i)$	$\tilde{h}_{ii}$	$\breve{h}_{ii}$	$h_{\mathbf{x}_i}(i,i)$
1	4.8	5.2	-8.1	0.3114	0.3114	0.3194	0.0000	0.0575	0.0575
2	4.1	8.8	-16	-0.1891	-0.1891	-0.6953	0.0600	0.0920	0.0311
3	5.2	10.5	-0.9	-0.0278	-0.0278	0.3116	0.0270	0.0577	0.0310
4	5.5	10.6	-7.8	0.6619	0.6619	0.6071	0.0007	0.0321	0.0313
5	3.4	1.8	-19	-0.0696	-0.0696	-0.6094	0.0683	0.2113	0.1441
6	3.4	12.7	-19	-0.8553	-0.8553	-1.5668	0.1186	0.1684	0.0494
7	4.9	15.6	-11	0.1629	0.1629	-0.1174	0.0184	0.1767	0.1577
8	5.6	5.8	-2.8	0.7564	0.7564	1.0547	0.0208	0.0704	0.0503
9	3.9	2.2	-3.1	-0.5404	-0.5404	-0.1745	0.0314	0.1592	0.1274
10	4.5	4.8	-7.8	0.0320	0.0320	0.0655	0.0003	0.0633	0.0631
11	4.8	7.9	-14	0.4361	0.4361	0.0675	0.0318	0.0670	0.0344
12	4.9	5.2	-4.5	0.2063	0.2063	0.4194	0.0106	0.0676	0.0575
13	3	0.9	-12	-0.7737	-0.7737	-0.8527	0.0015	0.1942	0.1930
14	4.6	11.8	-2.1	-0.5961	-0.5961	-0.3398	0.0154	0.0536	0.0384
15	4.8	7.9	-2	-0.2419	-0.2419	0.0675	0.0224	0.0561	0.0344
16	5.5	11.5	-9	0.7041	0.7041	0.5707	0.0042	0.0402	0.0359
17	4.5	10.6	-11	-0.1444	-0.1444	-0.3929	0.0145	0.0460	0.0313
18	5.3	8.5	-0.2	0.1209	0.1209	0.5246	0.0382	0.0694	0.0320
19	4.7	11.1	-6.1	-0.2504	-0.2504	-0.2141	0.0003	0.0337	0.0334
20	6.6	12.8	-1	1.3235	1.3235	1.6306	0.0221	0.0730	0.0511

21	5.1	11.3	-3.6	0.0016	0.0016	0.1781	0.0073	0.0417	0.0346
22	3.9	1	-8.2	-0.0824	-0.0824	0.0292	0.0029	0.1899	0.1866
23	5.7	14.5	-0.5	0.3836	0.3836	0.6964	0.0229	0.1202	0.0977
24	5.1	11.9	-2	-0.1039	-0.1039	0.1569	0.0159	0.0551	0.0393
25	5.2	8.1	-1.6	0.1233	0.1233	0.4527	0.0254	0.0582	0.0335
26	3.7	13.8	-12	-0.9649	-0.9649	-1.2917	0.0250	0.0988	0.0734
27	4.9	15.5	-0.7	-0.4019	-0.4019	-0.1163	0.0191	0.1693	0.1508
28	4.8	9.8	-1.2	-0.3843	-0.3843	-0.0539	0.0256	0.0551	0.0300
29	4.4	11	-14	-0.0803	-0.0803	-0.5101	0.0433	0.0765	0.0329
30	5.2	12.4	-0.8	-0.0818	-0.0818	0.2415	0.0245	0.0694	0.0450
31	5.1	11.1	-17	0.7592	0.7592	0.1859	0.0770	0.1108	0.0334
32	4.6	5.1	-5.1	-0.0503	-0.0503	0.1307	0.0077	0.0661	0.0588
33	3.9	4.8	-9.5	-0.4711	-0.4711	-0.5345	0.0009	0.0642	0.0631
34	5.1	4.2	-17	1.2166	1.2166	0.7391	0.0534	0.1273	0.0732
35	5.1	6.9	-3.3	0.1987	0.1987	0.4493	0.0147	0.0545	0.0404
36	6	13.2	-0.7	0.7017	0.7017	1.0209	0.0239	0.0822	0.0586
37	4.9	9.9	-3.3	-0.1687	-0.1687	0.0409	0.0103	0.0401	0.0300
38	4.1	12.5	-14	-0.4542	-0.4542	-0.8613	0.0388	0.0855	0.0464
39	4.6	13.2	-1.9	-0.6300	-0.6300	-0.3791	0.0147	0.0732	0.0586
40	4.9	8.9	-10	0.2582	0.2582	0.0984	0.0060	0.0371	0.0309
41	5.1	10.8	-14	0.5802	0.5802	0.1983	0.0342	0.0665	0.0320

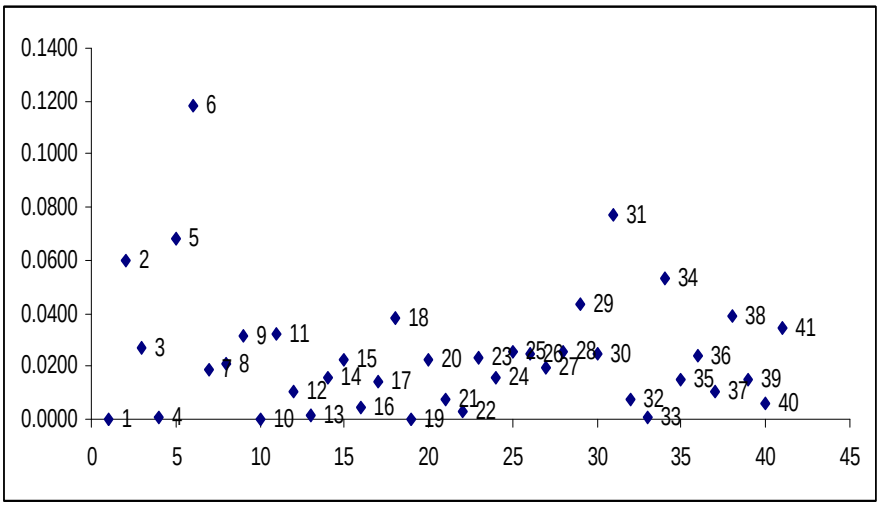
Çizelge 5.1'deki artık değerlerine ve kaldıraç değerlerine ilişkin saçılım grafikleri Şekil 5.1- Şekil 5.5'deki gibi elde edilmiştir:



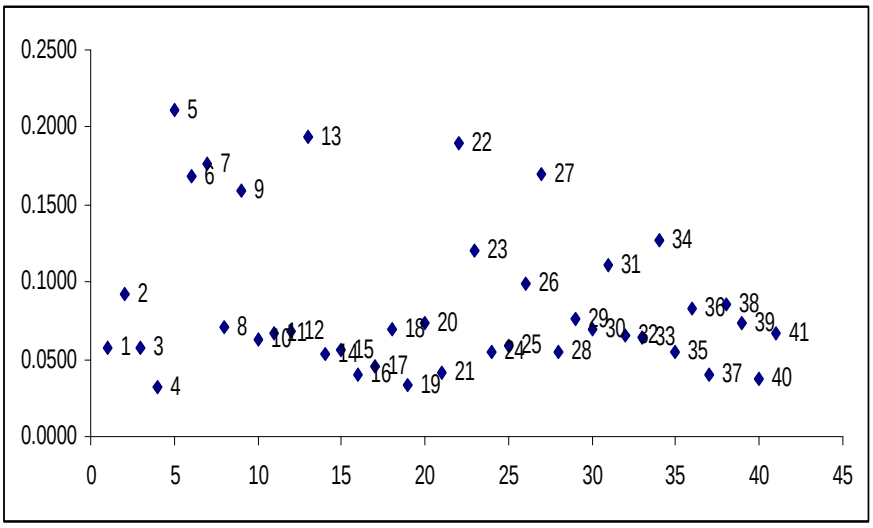
Şekil 5.1. $\tilde{e}_i$ ve $\breve{e}_i$ 'nin saçılım grafiği



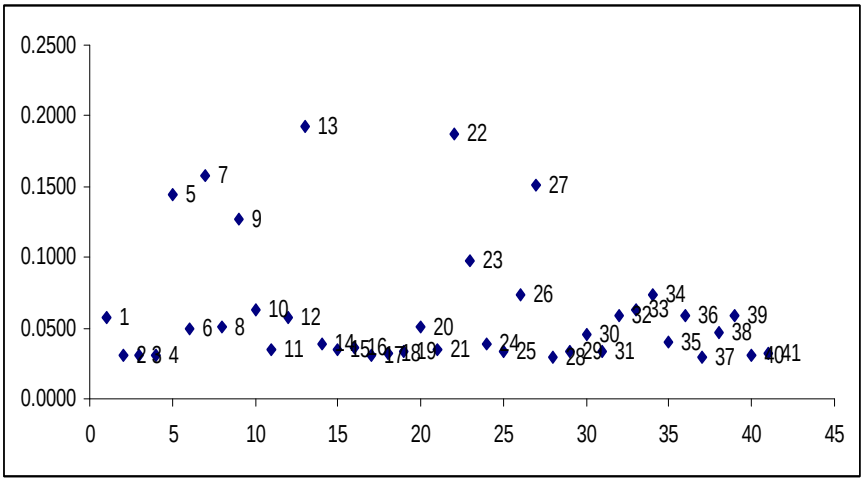
Şekil 5.2. $e_{x_i}(i)$ 'nin saçılım grafiği



Şekil 5.3. $\tilde{h}_{ii}$'nin saçılım grafiği



Şekil 5.4. $\breve{h}_{ii}$'nin saçılım grafiği



Şekil 5.5. $h_{x_i}(i,i)$'nin saçılım grafiği

Şekil 5.1 ve Şekil 5.2 incelendiğinde artık değerlerine göre 6., 20., 26. ve 34. gözlemlerin diğer gözlemlere göre daha büyük artık değerlerine sahip olduklarını yani diğer gözlemlere göre aykırı değer olduklarını söyleyebiliriz. Şekil 5.3- Şekil 5.5 incelendiğinde 5., 6., 7., 9. ve 22. gözlemlerin diğer gözlemlere göre kaldıraç değerlerinin daha büyük olduğunu söyleyebiliriz. Ancak bu gözlemler için bulunan kaldıraç değerleri büyük değildir. h_{ii} değeri 1'e yakın gözlemlere, büyük kaldıraç değerine sahiptir denilmektedir. Oysa bu gözlemlere ilişkin Çizelge 5.1'de elde edilen h_{ii} değerleri incelendiğinde ($\tilde{h}_{ii}, \breve{h}_{ii}$ ve $h_{x_i}(i,i)$) bu değerlerin çok büyük değerler olmadığı görülmektedir. Artık değerlerine ve kaldıraç değerlerine göre, incelenen veri kümesinde hem aykırı değer hem de büyük kaldıraç değerine sahip gözlem olmadığı söylenebilir.

Çizelge 5.1'de elde edilen artık değerlerinden ve kaldıraç değerlerinden yararlanarak yarı parametrik regresyon modelinde etkili gözlemleri ortaya çıkarmak için önerilen ölçüt ($\tilde{C}_i, C_i^*, \breve{C}_i, \tilde{S}_i, \tilde{CR}_i$ ve $\tilde{H}_i^2$) değerleri Çizelge 5.2'deki gibi elde edilmiştir:

Çizelge 5.2. Diyabet verisine ilişkin etkili gözlem ölçüt değerleri

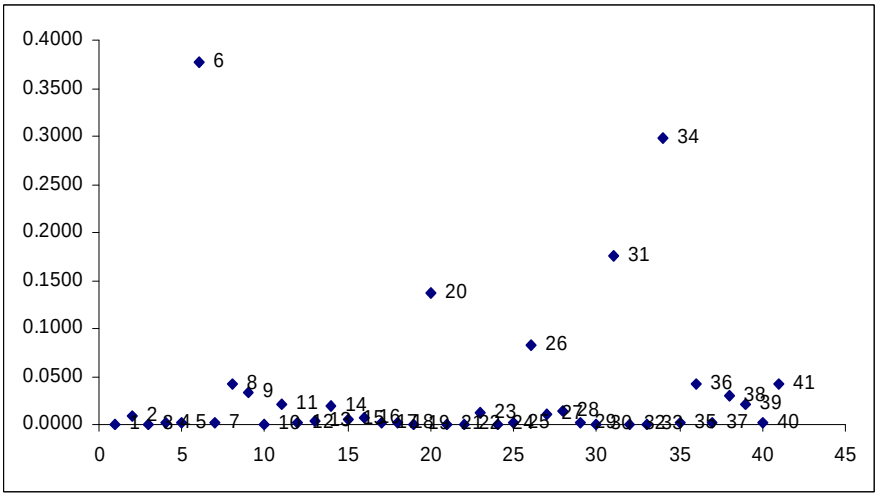
	y	x	z	$\tilde{\tilde{C}}_i$	C_i^*	$\breve{C}_i$	$\tilde{\tilde{S}}_i$	$C\tilde{\tilde{R}}_i$	$\tilde{H}_i^2$
1	4.8	5.2	-8.1	0.0000	0.0001	0.0060	11.1510	1.0168	0.0088
2	4.1	8.8	-16.1	0.0082	0.0005	0.0038	16.0521	1.0870	0.0673
3	5.2	10.5	-0.9	0.0001	0.0000	0.0000	13.5566	1.0534	0.0278
4	5.5	10.6	-7.8	0.0010	0.0000	0.0144	18.8341	0.9886	0.0419
5	3.4	1.8	-19.2	0.0013	0.0119	0.0016	14.3334	1.0996	0.0737
6	3.4	12.7	-18.9	0.3779	0.0007	0.1708	19.2155	1.0832	0.2148
7	4.9	15.6	-10.6	0.0017	0.0064	0.0066	22.6835	1.0419	0.0212
8	5.6	5.8	-2.8	0.0421	0.0037	0.0447	10.2300	0.9963	0.0769
9	3.9	2.2	-3.1	0.0330	0.0308	0.0630	11.6529	1.0318	0.0603
10	4.5	4.8	-7.8	0.0000	0.0011	0.0001	11.1819	1.0252	0.0004
11	4.8	7.9	-13.9	0.0219	0.0000	0.0140	15.1180	1.0415	0.0509
12	4.9	5.2	-4.5	0.0016	0.0000	0.0032	10.3782	1.0323	0.0147
13	3	0.9	-11.6	0.0030	0.1402	0.1716	12.8737	0.9757	0.0587
14	4.6	11.8	-2.1	0.0191	0.0029	0.0204	17.8466	1.0100	0.0493
15	4.8	7.9	-2	0.0046	0.0008	0.0035	10.0913	1.0433	0.0284
16	5.5	11.5	-9	0.0071	0.0000	0.0207	22.1797	0.9870	0.0513
17	4.5	10.6	-11.2	0.0011	0.0001	0.0010	19.2633	1.0382	0.0166
18	5.3	8.5	-0.2	0.0020	0.0001	0.0011	10.8473	1.0643	0.0411
19	4.7	11.1	-6.1	0.0001	0.0007	0.0022	19.5299	1.0200	0.0060
20	6.6	12.8	-1	0.1370	0.0149	0.1425	19.2925	0.8932	0.2148
21	5.1	11.3	-3.6	0.0000	0.0002	0.0000	17.7140	1.0325	0.0074
22	3.9	1	-8.2	0.0001	0.0264	0.0019	12.3405	1.0274	0.0035
23	5.7	14.5	-0.5	0.0120	0.0033	0.0219	20.7277	1.0360	0.0372
24	5.1	11.9	-2	0.0006	0.0003	0.0006	18.0407	1.0407	0.0172
25	5.2	8.1	-1.6	0.0014	0.0000	0.0010	10.2748	1.0504	0.0275
26	3.7	13.8	-11.9	0.0829	0.0085	0.1085	23.8819	0.9684	0.1198
27	4.9	15.5	-0.7	0.0108	0.0112	0.0380	20.3512	1.0308	0.0346
28	4.8	9.8	-1.2	0.0135	0.0011	0.0087	12.1895	1.0387	0.0401
29	4.4	11	-14.3	0.0010	0.0000	0.0006	19.0181	1.0708	0.0458
30	5.2	12.4	-0.8	0.0006	0.0001	0.0005	18.2220	1.0501	0.0257
31	5.1	11.1	-16.8	0.1763	0.0000	0.0774	18.2580	1.0533	0.1429
32	4.6	5.1	-5.1	0.0001	0.0011	0.0002	10.4507	1.0327	0.0080
33	3.9	4.8	-9.5	0.0007	0.0047	0.0156	11.9106	1.0072	0.0214
34	5.1	4.2	-17	0.2986	0.0000	0.2370	14.5896	0.9431	0.2194
35	5.1	6.9	-3.3	0.0020	0.0000	0.0023	9.7417	1.0369	0.0186
36	6	13.2	-0.7	0.0418	0.0051	0.0461	19.7461	1.0064	0.0721
37	4.9	9.9	-3.3	0.0010	0.0005	0.0012	12.8508	1.0332	0.0130
38	4.1	12.5	-13.6	0.0293	0.0003	0.0202	21.7346	1.0475	0.0602
39	4.6	13.2	-1.9	0.0204	0.0054	0.0324	20.9487	1.0057	0.0527
40	4.9	8.9	-10	0.0014	0.0000	0.0026	14.7360	1.0255	0.0121
41	5.1	10.8	-13.5	0.0417	0.0000	0.0246	18.9809	1.0307	0.0678

$\tilde{\tilde{C}}_i$: $\hat{\boldsymbol{\beta}}$ için Cook uzaklığı; C_i^* : $\hat{\mathbf{m}}$ için Cook uzaklığı; $\breve{C}_i$: $\hat{\mathbf{y}}$ için Cook uzaklığı

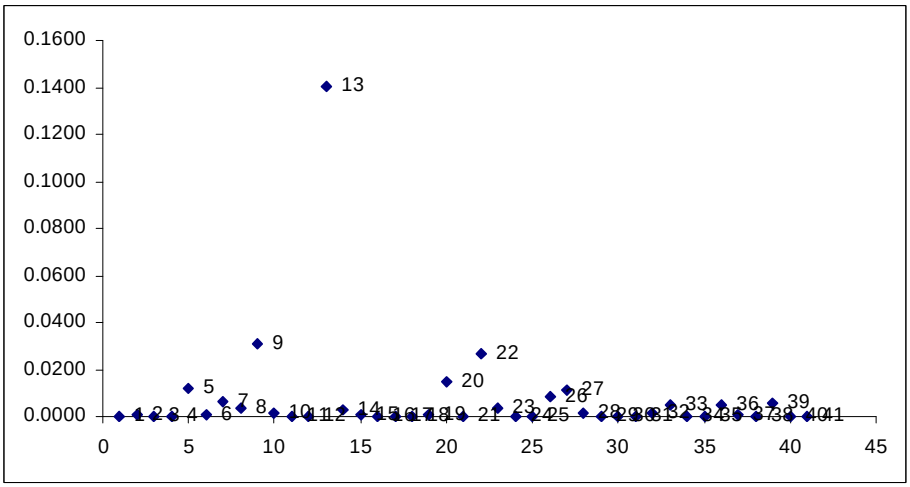
$\tilde{\tilde{S}}_i$: Önerilen Pena ölçütü; $C\tilde{\tilde{R}}_i$: Önerilen COVRATIO ölçütü; $\tilde{H}_i^2$: Önerilen Hadi'nin ölçütü

Parametre tahminleri üzerinde etkili olan gözlemleri daha net görebilmek amacıyla Çizelge 5.2'deki etkili gözlem ölçüt değerlerinin saçılım grafikleri çizdirilmiştir. Bu

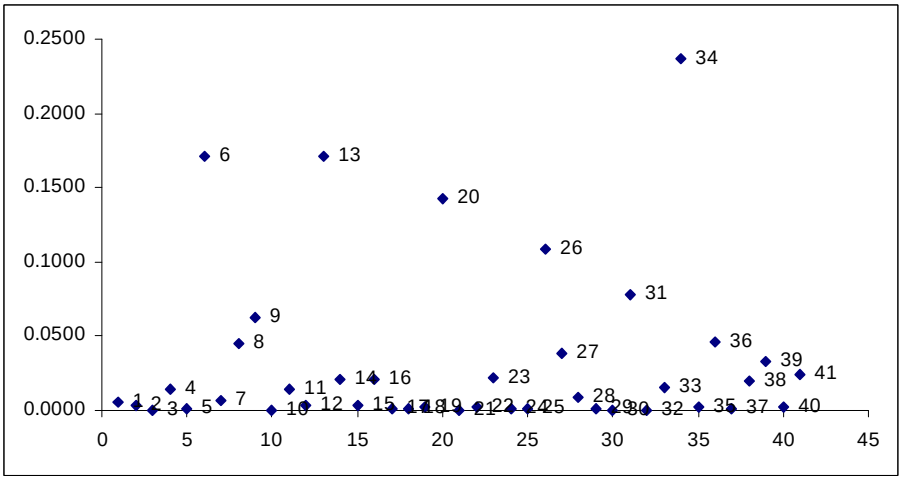
ölçüt değerlerine ilişkin saçılım grafikleri Şekil 5.6 - Şekil 5.11'deki gibi elde edilmiştir:



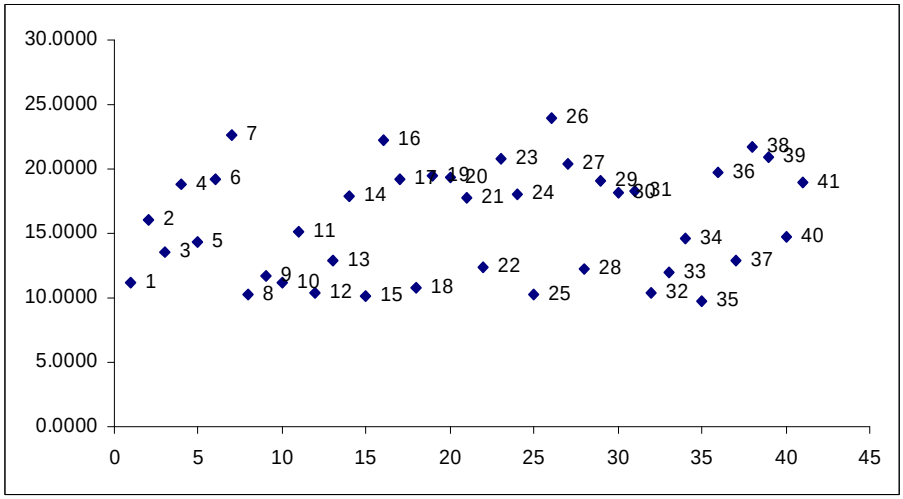
Şekil 5.6. $\tilde{C}_i$'nin saçılım grafiği



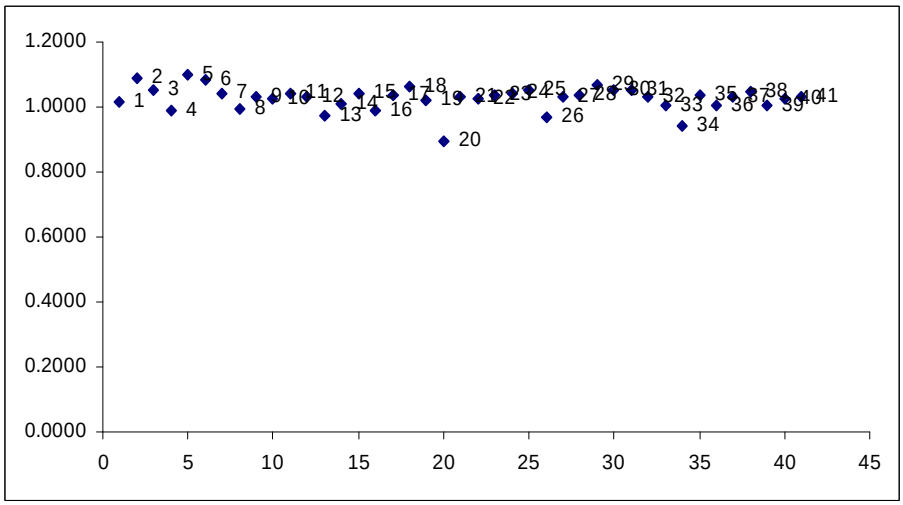
Şekil 5.7. C_i^*'nin saçılım grafiği



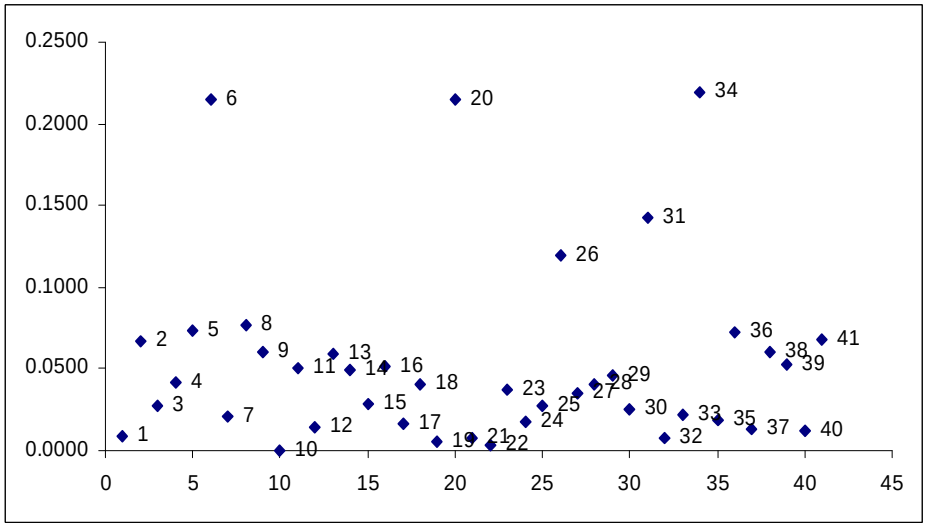
Şekil 5.8. $\breve{C}_i$'nin saçılım grafiği



Şekil 5.9. $\tilde{S}_i$'nin saçılım grafiği



Şekil 5.10. $C\tilde{R}_i$'nin saçılım grafiği



Şekil 5.11. $\tilde{H}_i^2$'nin saçılım grafiği

Saçılım grafikleri incelendiğinde, önerilen $\tilde{S}_i$ ölçütüne göre kestirim değerleri üzerinde etkili gözlem bulunmamıştır. Çünkü $\tilde{S}_i$ ölçütü özellikle büyük örneklemlerde büyük kaldıraç değerine sahip aykırı değerleri (high leverage outliers) ortaya çıkarmada daha başarılıdır. Daha öncede belirtildiği gibi aykırı değerlere ve kaldıraç değerlerine göre veri kümesinde hem büyük kaldıraç değeri hem de aykırı değer olan gözlem bulunmadığında $\tilde{S}_i$ ölçütünün değeri diğer gözlemlerin değerlerinden çok büyük olmayacaktır. Bu nedenle bu veri kümesi için önerilen $\tilde{S}_i$ ölçütü hiçbir gözlemi etkili gözlem olarak belirlememiştir. Önerilen $C\tilde{R}_i$ ölçütünün saçılım grafiği incelendiğinde $C\tilde{R}_i$ ölçütüne göre 20. ve 34. gözlemlerin etkili gözlemler olduğu görülmektedir. Çünkü $C\tilde{R}_i$ ölçütü artık değeri büyük olup kaldıraç değeri küçük olan gözlemleri genel olarak etkili gözlemler olarak belirlemektedir. 20. ve 34. gözlemlerin artık değerleri büyük olup kaldıraç değerleri küçük olduğundan $C\tilde{R}_i$ ölçütü tarafından etkili gözlem olarak belirlenmiştir. Aynı şekilde önerilen $\tilde{H}_i^2$ ölçütünün saçılım grafiği incelendiğinde 6., 20., 26., 31. ve 34. gözlemlerin etkili gözlemler olduğu görülmektedir. Kim vd. (2002) 'nin çalışmalarında Cook uzaklığı ölçütüne göre etkili gözlem olarak belirlenen gözlemler bu çalışmada önerilen $C\tilde{R}_i$ ve $\tilde{H}_i^2$ ölçütlerine göre de etkili gözlem olarak belirlenmiştir.

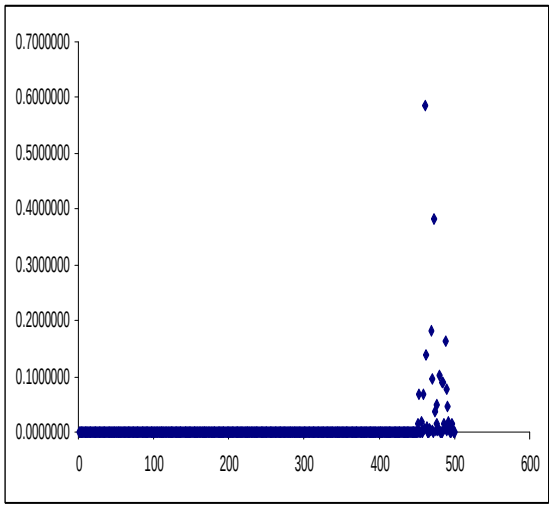
5.2. Yapay Bir Veri Kümesi Üzerinde Uygulama

Bu çalışmada $\tilde{S}_i$ ölçütünün büyük örneklemlerde hem büyük kaldıraç değerine sahip olduğu, hem de aykırı değer olan gözlemleri ortaya çıkarmada diğer ölçütlere göre daha başarılı olup olmadığını göstermek için aşağıdaki yarı parametrik regresyon modeli kullanılarak yapay bir veri kümesi türetilmiştir:

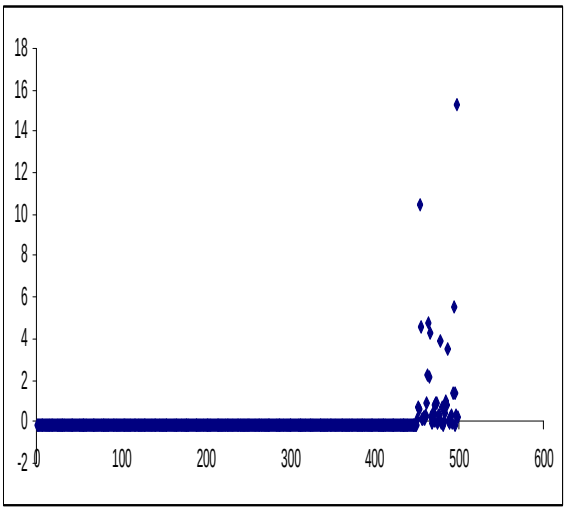
$$y_i = 0.5\,\mathbf{z}_i + (x_i - 0.5)^2 + \varepsilon_i, \quad i = 1,...,500 \tag{5.2}$$

Eş. (5.2)'deki model kullanılarak 500 gözlemli bir veri kümesi türetilmiş ve son 50 gözlem hem büyük kaldıraç değeri hem de aykırı değer olacak şekilde türetilmiştir. Eş. (5.2)'deki model kullanılarak türetilen veri kümesinde ilk 450 veri için hata

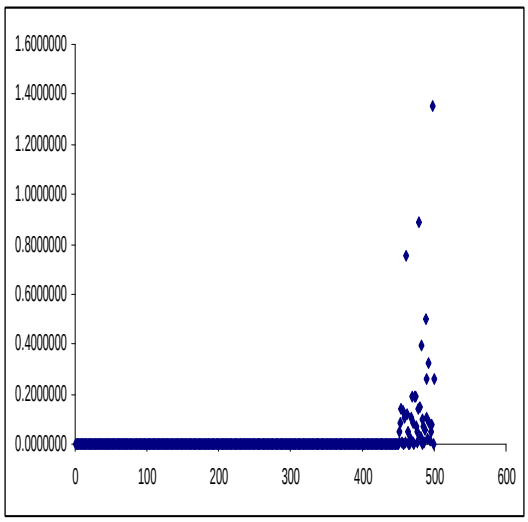
terimleri $\varepsilon_i \sim N(0,0.02)$ dağılımından ve $x_i \sim U(0,1)$ dağılımından türetilmiş ve $z_i = i / 450$ olarak alınmıştır. Hem büyük kaldıraç değeri hem de aykırı değer olan gözlemler oluşturulurken son 50 gözlem için hata terimleri $\varepsilon_i \sim N(5,2)$ dağılımından ve $x_i \sim U(5,10)$ dağılımından türetilmiş ve $z_i = i / 50$ olarak alınmıştır. Türetilen veri kümesi için $\tilde{C}_i, C_i^* \breve{C}_i$, $\tilde{S}_i$, $C\tilde{R}_i$ ve $\tilde{H}_i^2$ ölçüt değerleri hesaplanmış ve bu ölçüt değerlerine ilişkin saçılım grafikleri, Şekil 5.12- Şekil 5.17'deki gibi elde edilmiştir:



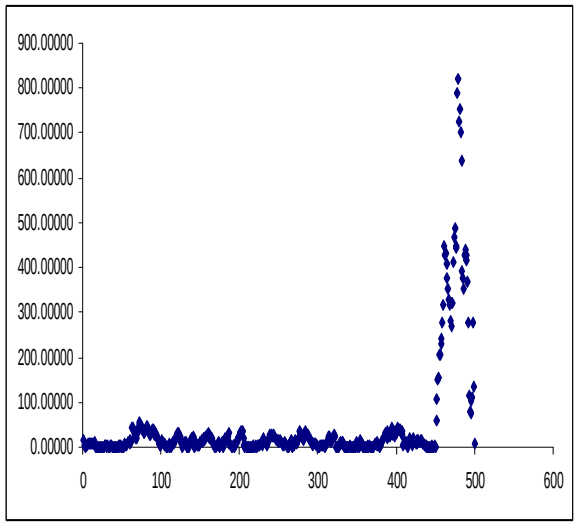
Şekil 5.12. $\tilde{C}_i$'nin saçılım grafiği



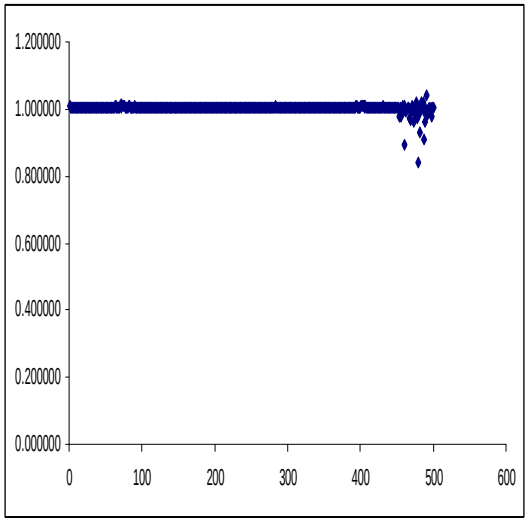
Şekil 5.13. C_i^*'nin saçılım grafiği



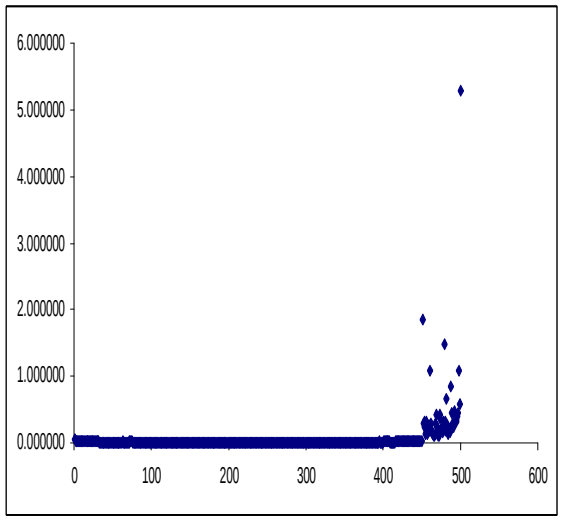
Şekil 5.14. $\breve{C}_i$'nin saçılım grafiği



Şekil 5.15. $\tilde{S}_i$'nin saçılım grafiği



Şekil 5.16. $C\tilde{R}_i$'nin saçılım grafiği



Şekil 5.17. $\tilde{H}_i^2$'nin saçılım grafiği

Saçılım grafikleri incelendiğinde, $\tilde{S}_i$ ölçütünün türetilen veri kümesindeki hem büyük kaldıraç değeri hem de aykırı değer olan gözlemlerin neredeyse hepsini ortaya çıkardığı, diğer ölçütlerin bu gözlemleri ortaya çıkarmada $\tilde{S}_i$ ölçütü kadar başarılı olmadığı görülmüştür. Ayrıca $\tilde{S}_i$ ölçütünün saçılım grafiği incelendiğinde verideki heterojen yapının diğer ölçütlerin saçılım grafiğinden daha net ortaya çıktığı görülmüştür. Bu gözlemleri ortaya çıkarmada $\tilde{S}_i$ ölçütünden sonra $\tilde{C}_i, C_i^*$, ve $\tilde{H}_i^2$ ölçütlerinin daha başarılı olduğunu söyleyebiliriz. Çünkü veri kümesinde hem büyük kaldıraç değeri hem de aykırı değer olduğunda , $C\tilde{R}_i$ ölçütü değeri 1'e yakın bir değer olur ve bu da veri kümesinde etkili gözlem olmadığını gösterir. Buna göre, $C\tilde{R}_i$ ölçütü veri kümesinde hem büyük kaldıraç değeri hem de aykırı değer olduğunda bu gözlemleri ortaya çıkarmada başarılı değildir. Bu nedenle türetilen bu veri kümesi için $C\tilde{R}_i$ ölçütü hem büyük kaldıraç değeri hem de aykırı değer olan son 50 gözlemi etkili gözlem olarak belirleyememiştir.

5.3. Simülasyon Çalışması

Bu tez çalışmasında yarı parametrik regresyon modeli için önerilen ölçütlerin veri kümesindeki etkili gözlemleri ortaya çıkarmadaki başarılarını incelemek amacıyla bir simülasyon çalışması yapılmıştır. Simülasyon çalışmasında Eş. (5.2)'deki yarı

parametrik regresyon modeli kullanılarak n=50, 100 ve 250 olmak üzere 3 farklı örneklem büyüklüğü alınarak veriler türetilmiştir. Veriler türetilirken veri kümesinin son %10 ve %20 'i gözlemi

a) Aykırı değer ve büyük kaldıraç değeri (high leverage)
b) Aykırı değer ve kaldıraç değeri (low leverage)
c) Sadece büyük kaldıraç değeri

olacak şekilde türetilmiştir. Başka bir deyişle, n=50, 100 ve 250 örneklem büyüklükleri için son %10 gözlemi yani her bir örneklem büyüklüğü için sırasıyla 5, 10 ve 25 gözlem ve son %20 gözlemi yani her bir örneklem büyüklüğü için sırasıyla 10, 20 ve 50 gözlem hem aykırı değer hem de büyük kaldıraç değeri, hem aykırı değer hem de kaldıraç değeri ve sadece kaldıraç değeri olacak şekilde türetilmiştir. Buna göre Eş. (5.2)'deki model kullanılarak n=50 için,

- Son %10 gözlem yani son 5 gözlem aykırı değer ve büyük kaldıraç değeri olarak türetilirken; veri kümesinde ilk 45 veri için hata terimleri $\varepsilon_i \sim N(0,0.02)$ dağılımından ve $x_i \sim U(0,1)$ dağılımından türetilmiş ve z_i = i / 45 olarak alınmış ve son 5 gözlem için hata terimleri $\varepsilon_i \sim N(5,2)$ dağılımından ve $x_i \sim U(5,10)$ dağılımından türetilmiş ve z_i = i / 5 olarak alınmıştır.

- Son %10 gözlem yani son 5 gözlem aykırı değer ve kaldıraç değeri olarak türetilirken veri kümesinde ilk 45 veri için hata terimleri $\varepsilon_i \sim N(0,0.02)$ dağılımından ve $x_i \sim U(0,1)$ dağılımından türetilmiş ve z_i = i / 45 olarak alınmış ve son 5 gözlem için hata terimleri $\varepsilon_i \sim N(5,2)$ dağılımından ve $x_i \sim U(5,10)$ dağılımından türetilmiş ve z_i = i / 50 olarak alınmıştır.

- Son %10 gözlem yani son 5 gözlem büyük kaldıraç değeri olarak türetilirken veri kümesinde ilk 45 veri için hata terimleri $\varepsilon_i \sim N(0,0.02)$ dağılımından ve $x_i \sim U(0,1)$ dağılımından türetilmiş ve z_i = i / 45 olarak alınmış ve son 5

gözlem için hata terimleri $\varepsilon_i \sim N(0,0.02)$ dağılımından ve $x_i \sim U(5,10)$ dağılımından türetilmiş ve $z_i = i / 5$ olarak alınmıştır.

n=100 ve 250 için de yukarıda açıklandığı gibi son %10 ve %20 'i gözlem aykırı değer ve büyük kaldıraç değeri, aykırı değer ve kaldıraç değeri ve sadece büyük kaldıraç değeri olarak türetilmiş ve türetilen veri kümeleri için $\tilde{C}_i, \breve{C}_i$, $\tilde{S}_i$, $C\tilde{R}_i$ ve $\tilde{H}_i^2$ ölçüt değerleri hesaplanmıştır. Bu işlemler 100 kez tekrar edilmiştir. 100 tekrar sonucunda her bir örneklem büyüklüğü için önerilen ölçütlerin veri kümelerinde türetilen etkili gözlemleri belirleme yüzdeleri hesaplanmıştır ve Çizelge 5.3-Çizelge 5.5'deki gibi elde edilmiştir:

Çizelge 5.3. Çeşitli örneklem büyüklükleri için önerilen ölçütlerin aykırı değer olmayan ancak büyük kaldıraç değeri olan gözlemleri belirleme yüzdeleri

Örneklem büyüklüğü	Aykırı değer yüzdesi	Aykırı değer olmayan ancak büyük kaldıraç değeri olan gözlemlerin varlığında					
		$\tilde{C}_i$	C_i^*	$\breve{C}_i$	$\tilde{S}_i$	$C\tilde{R}_i$	$\tilde{H}_i^2$
n=50	10%	33	60	60	68	18	43
	20%	16	19	39	36	24	25
n=100	10%	23	11	39	45	31	37
	20%	17	14	38	35	34	30
n=250	10%	49	50	69	72	37	71
	20%	43	17	75	76	27	69

$\tilde{C}_i$: $\hat{\boldsymbol{\beta}}$ için Cook uzaklığı; C_i^* : $\hat{\mathbf{m}}$ için Cook uzaklığı; $\breve{C}_i$: $\hat{\mathbf{y}}$ için Cook uzaklığı
$\tilde{S}_i$: Önerilen Pena ölçütü; $C\tilde{R}_i$: Önerilen COVRATIO ölçütü; $\tilde{H}_i^2$: Önerilen Hadi'nin ölçütü

Çizelge 5.4. Çeşitli örneklem büyüklükleri için önerilen ölçütlerin hem aykırı değer hem de büyük kaldıraç değeri olan gözlemleri belirleme yüzdeleri

Örneklem büyüklüğü	Aykırı değer yüzdesi	Hem aykırı değer hem de büyük kaldıraç değeri olan gözlemlerin varlığında					
		$\tilde{C}_i$	C_i^*	$\breve{C}_i$	$\tilde{S}_i$	$C\tilde{R}_i$	$\tilde{H}_i^2$
n=50	10%	51	70	72	80	26	64
	20%	46	44	68	84	28	65
n=100	10%	49	66	75	91	28	73
	20%	45	23	65	92	29	68
n=250	10%	52	52	71	98	30	78
	20%	44	19	62	98	30	69

$\tilde{C}_i$: $\hat{\boldsymbol{\beta}}$ için Cook uzaklığı; C_i^* : $\hat{\mathbf{m}}$ için Cook uzaklığı; $\breve{C}_i$: $\hat{\mathbf{y}}$ için Cook uzaklığı
$\tilde{S}_i$: Önerilen Pena ölçütü; $C\tilde{R}_i$: Önerilen COVRATIO ölçütü; $\tilde{H}_i^2$: Önerilen Hadi'nin ölçütü

Çizelge 5.5. Çeşitli örneklem büyüklükleri için önerilen ölçütlerin hem aykırı değer hem de kaldıraç değeri olan gözlemleri belirleme yüzdeleri

Örneklem büyüklüğü	Aykırı değer yüzdesi	Hem aykırı değer hem de kaldıraç değeri olan gözlemlerin varlığında					
		$\tilde{C}_i$	C_i^*	$\breve{C}_i$	$\tilde{S}_i$	$C\tilde{R}_i$	$\tilde{H}_i^2$
n=50	10%	40	48	81	82	4	56
	20%	32	34	70	86	4	53
n=100	10%	32	39	77	86	4	60
	20%	23	27	66	89	7	54
n=250	10%	20	31	73	94	8	65
	20%	14	17	63	96	10	56

$\tilde{C}_i$: $\hat{\boldsymbol{\beta}}$ için Cook uzaklığı; C_i^*: $\hat{\mathbf{m}}$ için Cook uzaklığı; $\breve{C}_i$: $\hat{\mathbf{y}}$ için Cook uzaklığı
$\tilde{S}_i$: Önerilen Pena ölçütü; $C\tilde{R}_i$: Önerilen COVRATIO ölçütü; $\tilde{H}_i^2$: Önerilen Hadi'nin ölçütü

Çizelge 5.3 incelendiğinde veri kümesinde aykırı değer olmayan ancak büyük kaldıraç değeri olan gözlemlerin varlığında özellikle küçük örneklemlerde tüm ölçütler bu gözlemleri belirlemede aynı ölçüde başarılı olmuşlardır. Örneklem büyüklüğü arttıkça $\breve{C}_i$, $\tilde{S}_i$ ve $\tilde{H}_i^2$ ölçütleri bu gözlemleri ortaya çıkarmada diğer ölçütlere göre daha başarılı olmuşlardır.

Çizelge 5.4 incelendiğinde $\tilde{S}_i$ ölçütü tüm örneklem büyüklüklerinde diğer ölçütlere göre hem aykırı değer hem de büyük kaldıraç değeri olan gözlemleri ortaya çıkarmada daha başarılı olmuştur. Özellikle örneklem büyüklüğü arttıkça $\tilde{S}_i$ ölçütünün veri kümesinde hem aykırı değer hem de büyük kaldıraç değeri olan gözlemleri ortaya çıkarmada diğer ölçütlere göre önemli ölçüde daha başarılı olduğu söylenebilir.

Çizelge 5.5 incelendiğinde veri kümesinde aykırı değer ve kaldıraç değeri olduğunda bu gözlemleri ortaya çıkarmada en başarılı ölçütün $\tilde{S}_i$ ölçütü olduğu görülmektedir. Ayrıca $\breve{C}_i$ ölçütü ve $\tilde{H}_i^2$ ölçütü de hem aykırı değer ve hem de kaldıraç değeri olan gözlemleri ortaya çıkarmada diğer ölçütlere göre daha başarılı olmuştur.

6. SONUÇ VE TARTIŞMA

Bu çalışmada doğrusal regresyon modelinde etkili gözlemleri belirlemek için kullanılan ölçütlerden Hadi'nin ölçütünün, Pena'nın ölçütünün ve COVRATIO ölçütünün yarı parametrik regresyon modeli için uyarlanması amaçlanmıştır.

Çalışmanın ikinci bölümünde, regresyon modelleri hakkında genel bir bilgi verilmiştir. Daha sonra düzleştirme yöntemleri tanıtılmıştır. Düzleştirme yöntemlerinde kullanılan düzleştirme parametresinin seçimi için yaygın olarak kullanılan seçim yöntemleri hakkında bilgi verilmiştir. Daha sonra yarı parametrik regresyon modellerinde kestirim yöntemleri tanıtılmıştır. Üçüncü bölümde doğrusal regresyon modelinde etkili gözlemleri ortaya çıkarmak için yaygın olarak kullanılan ölçütler ayrıntılı olarak incelenmiştir. Dördüncü bölümde doğrusal regresyon modelinde etkili gözlemleri ortaya çıkarmak için kullanılan ölçütlerden Hadi'nin ölçütü, Pena'nın ölçütü ve COVRATIO ölçütü yarı parametrik regresyon modeli için uyarlanmıştır.

Beşinci bölümde ise, yarı parametrik regresyon modelinde etkili gözlemleri ortaya çıkarmak için geliştirilen ölçütlerin bu gözlemleri ortaya çıkarmadaki başarıları gerçek bir veri kümesi kullanılarak araştırılmıştır. Bu veri kümesi için Kim vd.(2002)'nin yaptıkları çalışmada etkili gözlem olarak belirlenen gözlemleri ortaya çıkarmada bu tez çalışmasında uyarlanan Hadi'nin ölçütü ve COVRATIO ölçütü başarılı olmalarına rağmen, Pena'nın ölçütü çok başarılı olmamıştır. Ancak uyarlanan Pena ölçütü büyük örneklemlerde veri kümesinde hem aykırı değer hem de büyük kaldıraç değeri olan gözlemleri belirlemede diğerlerine göre daha başarılıdır. Kim vd.(2002)'nin inceledikleri veri kümesinde bu özellikte gözlemler olmadığı için Pena'nın ölçütü, etkili gözlemleri belirleyememiştir. Pena'nın ölçütünün doğrusal regresyon modelinde olduğu gibi yarı parametrik regresyon modelinde de büyük örneklemlerde hem aykırı değer hem de büyük kaldıraç değeri olan gözlemleri belirlemede daha başarılı olduğunu göstermek için yapay bir veri kümesi türetilmiştir. Türetilen veri kümesi incelendiğinde, Pena'nın önerdiği ölçütün yarı parametrik regresyon modeli için de büyük örneklemlerde hem aykırı değer hem de kaldıraç değeri olan gözlemleri ortaya çıkarmada diğer ölçütlere göre daha başarılı olduğu söylenebilir.

Yarı parametrik regresyon modeli için uyarlanan ölçütlerin veri kümesindeki etkili gözlemleri belirlemedeki başarılarını daha ayrıntılı incelemek için bir simülasyon çalışması yapılmıştır. Simülasyon çalışmasında veriler farklı örneklem büyüklüklerinde ve veri kümesinde sadece aykırı değerlerin bulunduğu, hem aykırı değer hem de büyük kaldıraç değerine sahip gözlemlerin bulunduğu ve sadece büyük kaldıraç değerine sahip gözlemlerin bulunduğu durumlar göz önüne alınarak türetilmiştir. Simülasyon çalışması sonucunda elde edilen sonuçlar incelendiğinde gerçek veri kümesi ve yapay veri kümesi kullanılarak elde edilen sonuçlar ile simülasyon sonuçları birbirini desteklemektedir. Genel olarak büyük örneklemlerde etkili gözlemleri belirlemede uyarlanan ölçütlerden Pena'nın ölçütünün ve Hadi'nin ölçütünün bu gözlemleri belirlemede diğer ölçütlere göre daha başarılı olduğu, özellikle veri kümesinde hem aykırı hem de büyük kaldıraç değerine sahip gözlemlerin bulunduğu büyük örneklemlerde Pena'nın ölçütünün diğer ölçütlere göre önemli ölçüde başarılı olduğu görülmüştür. Küçük örneklem büyüklükleri için uyarlanan ölçütlerin etkili gözlemleri belirlemede aynı oranda başarılı oldukları görülmüştür.

KAYNAKLAR

Akaike, H., 1973, Information Theory and an Extension of the Maximum Likelihood Principle, in Petrov and Csaki, eds., Proceedings of the Second International Symposium on Information Theory, 267–281.

Akdeniz, F., Akdeniz D. E., 2009, Liu-type Estimator in Semiparametric Regression Model. Journal of Statistical Computation Simulation.

Akdeniz, F., Tabakan, G., 2009, Restricted Ridge Estimators of the Parameters in Semiparametric Regression Model, Communications in Statistics-Theory and Methods, 38: 1852-1869.

Alpar, R., 2003, Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş : I, Nobel, Ankara.

Atkinson, A.C. 1985, Plots, Transformations and Regression, Oxford: Clarendon Press.

Aydın, D., 2005, Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı ile Tahmin ve Çıkarsamalar, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Doktora Tezi, Eskişehir.

Belsley, D.A., Kuh, E., Roy, E.W., 1980, Regression Diagnostics : Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York.

Chatterjee, S., A. S. Hadi., 1986, Influential Observations, High Leverage Points, and Outliers in Linear Regression, Statistical Science , 1: 379–416.

Chatterjee, S., Price, B., 1991, Regression Analysis By Example, John Wiley & Sons, New York.

Chatterjee, S., Hadi, A.S., Price, B., 2000, Regression Analysis By Example, John Wiley & Sons, New York.

Chen, H., 1988, Convergence Rates for Parametric Components in a Partly Linear Model, Annals Statistics, 16:136-146.

Cook, R.D., 1977, Detection of Influential Observations in Linear Regression, Technometrics, 19, 15-18.

Cook, R.D., Weisberg, S., 1982, Residuals and Influence in Regression, Chapman and Hall, New York.

Cula, S., 1998, Çok Değişkenli Olasılık Yoğunluk Fonksiyonunun Çekirdek Fonksiyonlarıyla Kestirimi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Doktora Tezi, Ankara.

Demir, S., 2005, Regresyon Fonksiyonlarının Uyarlanabilir Nadaraya-Watson Çekirdek Kestirimleri, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Doktora Tezi, Ankara.

Dillane, D.M., 2005, Deletion Diagnostics for The Linear Mixed Model, Ph.D. Thesis, 141p.

Duran, E.A., Akdeniz, F., Hu, H., 2011, Efficiency of a Liu-type Estimator in Semiparametric Regression Models. J. Computational Applied Mathematics 235(5): 1418-1428.

Engle, R. F., Granger, C. W. J., Rice, J., Weiss, A., 1985, Semiparametric Estimates of The Relation Between Weather and Electricity Sales, J. Amer. Statist. Assoc. 81: 310-320.

Eubank, R.L., 1985, Diagnostics for Smoothing Splines. J. Roy. Statist. Soc. Ser. B 47, 332–341.

Eubank, R.L., 1988, Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York.

Fan, J., 1993, Local Linear Smoothers and Their Minimax Effiencies., Annals of Statistics, 21:196-216.

Fox, J., 2002, An R and S-Plus Companion to Applied Regression, Sage Publications.

Freund, R.J., 1998, Regression Analysis : Statistical modeling of a response variable, Academic Pres, San Diego.

Fung, W.K., Zhu, Z.Y., Wei, B.C., He, X., 2002. Infuence Diagnostics and Outlier Tests for Semiparametric Mixed Models, J. Roy. Statist. Soc. Ser. B, 64: 565-579.

Gökmen, D., 2002, Bant Genişliği Seçiminde Kullanılan Yöntemlerin Simetrik ve Simetrik Olmayan Dağılımlarda Karşılaştırılması, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Bilim Uzmanlığı Tezi, Ankara.

Green, P., Yandell, B.S., 1985, Semiparametric Generalized Linear Models, Proceedings of The International Conference on Generalized Linear Models, (R.Gilchrist ed.), Springer-Verlag.

Green, P., Jennison, C., Seheult, A., 1985, Analysis of Field Experiments by Least Squares Smoothing, Journal of Royal Statistical Society B, 47: 299-315.

Green, P. J., Silverman, B. W., 1994, Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, Chapman and Hall, London.

Green, P.J., Silverman, B.W., 2000, Nonparametric Regression and Generalized Linear Models, Chapman & Hall, New York.

Hadi, A.S., 1992, A New Measure of Overall Potential Influencein Linear Regression, Computational Statistics and Data Analysis, 14, 1-27.

Hamilton, S. A., Truong, Y. K., 1997, Local Linear Estimation in Partly Linear Models, Journal of Multivariate Analysis, 60: 1-19.

Härdle, W., 1994, Applied Nonparametric Regression, Cambridge University Press, Cambridge.

Härdle, W., Müler, M., Sperlich, S., Werwatz, A., 2004, Nonparametric and Semiparametric Models, Springer, New York.

Hastie, T., Tibshirani, R.J., 1990, Generalized Additive Models, Chapman & Hall, London.

Heckman, N. , 1986, Spline Smoothing in Partial Linear Model, Journal of Royal Statistical Society B, 244-248.

Hong, S.Y., 1998, Automatic Bandwidth Selection and Data-Driven Estimators in a Semiparametric Regression Model, A Dissertation Degree of Doctor of Philosophy, Northwestern University.

Hurvich, C.M., Simonoff, J.S., 1998, Smoothing Parameter Selection in Nonparametric Regression Using An Improved Akaike Information Criterion, J.R. Statist. Soc. B, 60, 271-293.

Keele, L., 2008, Semiparametric Regression for The Social Sciences, John Wiley & Sons, England.

Kim, C. ,1996, Cook's Distance in Spline Smoothing. Statistics and Probability Letters, 31:139-144.

Kim, C., Kim, W., 1998, Some Diagnostics Results in Nonparametric Density Estimation. Comm. Statist. Theory Methods 27, 291–303.

Kim, C., Park, B.U., Kim, W., 2001, Cook’s Distance in Local Polynomial Regression, Statistics & Probability Letters, 54, 33-40.

Kim, C., Park, B.U., Kim, W., 2002, Influential Diagnostics in Semiparametric Regression Models, Statistics & Probability Letters, 60, 49-58.

Kutner, M.H., Nachtsheim, C., Neter, J., 2004, Applied Linear Regression Models, McGraw-Hill/Irwin, Boston.

Lee, C.M., 2003, Smoothing Parameter Selection for Smoothing Splines: A Simulation Study, Comput. Statistics & Data Analysis, 42, 139-148.

Liang, H., 2006, Estimation in Partially Linear Models and Numerical Comparisons, Computational Statistics and Data Analysis 50(3): 675-687.

Mallows, C.L., 1973, Some Comments on C_P, Technometrics 15 (4): 661–675.

McCulloch, C.E., Searle, S.R., 2001, Generalized, Linear, and Mixed Models, John Wiley & Sons, New York.

Omay, R.E., 2007, Regresyonda Pürüzlülük Ceza Yaklaşımı, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Doktora Tezi, Eskişehir.

Pena, D., 2005, A New Statistic for Influence in Linear Regression, Technometrics, 47, 1.

Rawlings, J.O., Pantula, S.G., Dickey, D.A., 1998, Applied Regression Analysis : A research tool, Springer, New York.

Robinson., P.M., 1988, Root-n-consistent Semiparametric Regression, Econometrica, 56: 931-954.

Rosenblatt, M., 1956, Remarks on Some Nonparametric Estimates of a Density Function, Annals of Mathematical Statistics, 27: 832–837.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003, Semiparametric Regression, Cambrige University Pres.

Sheather, S.J., 2009, A Modern Approach to Regression with R, Springer Texts in Statistics.

Shi, X., 2009, Applications of Nonparametric and Semiparametric Methods in Economics and Finance, Dissertation Degree of Doctor of Philosophy, Economics in the Graduate School of Binghamton University, New York.

Speckman, P., 1988, Kernel Smoothing in Partial Linear Models, Journal of Royal Statistical Society B, 50, No.3, 413-436.

Stone, C. J., 1977, Consistent Nonparametric Regression, The Annals ofStatistics, 5: 595-620.

Tabakan, G., 2009, Yarı Parametrik Regresyonda Tahmin Metodları, Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Doktora Tezi, Adana.

Thomas, W., 1991. Infuence Diagnostics for The Cross-validated Smoothing Parameter in Spline Smoothing. J. Amer. Statist. Assoc. 86, 693–698.

Türkan, S., 2008, Karışık Doğrusal Modellerde Artık analizi ve Etki analizi, Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 137s.

Ullah, M.A., Pasha, G.R., 2009, The Orjin and Developments of Influence Measures in Regression. Pakistan Journal of Statistic, 25 (3): 295-307.

Yao, F., Lee, T.C.M., 2008, An Improved Knot Placement Scheme for Penalized Spline Regression, Journal of the Korean Statistical Society, 37: 259-267.

Wahba, G., 1990, Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

Zhang, C., Mei, C., Zhang, J., 2007, Influence Diagnostics in Partially Varying Coefficient Models, Acta. Math. Appl. Sinica, 23 (4): 619-628

Zhongyi, Z., Baocheng, W., 2001, Influence Analysis in Semiparametric Nonlinear Regression Models. Acta. Math. Appl. Sinica vol. 24 (4): 568-581.

ÖZGEÇMİŞ

Adı Soyadı : Semra Türkan

Doğum Yeri : Ankara

Doğum Yılı : 1983

Medeni Hali : Bekar

Eğitim ve Akademik Durumu:

Lise 1997.-2001…Eryaman Süper Lisesi

Lisans 2001.-2005…H.Ü. İstatistik Bölümü

Yabancı Dil : İnglizce

İş Tecrübesi :

2005.-… H.Ü. İstatistik Bölümü