REGRESYONDA PÜRÜZLÜLÜK CEZA YAKLAŞIMI

Rabia Ece OMAY

Doktora Tezi

İstatistik Anabilim Dalı

Temmuz-2007

JÜRİ VE ENSTİTÜ ONAYI

Rabia Ece OMAY'ın "**Regresyonda Pürüzlülük Ceza Yaklaşımı"** başlıklı **İstatistik** Anabilim Dalındaki, Doktora Tezi 29.06.2007 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü EğitimÖğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

		Adı-Soyadı	İmza
Üye (Tez Danışmanı)	**:**	**Prof.Dr. Memmedağa MEMMEDLİ**	**……**
Üye	**:**	**Prof.Dr. Aydın ERAR**	**……**
Üye	**:**	**Prof.Dr. Embiya AĞAOĞLU**	**……**
Üye	**:**	**Prof.Dr. Hasan DURUCASU**	**……**
Üye	**:**	**Prof.Dr. Ali Fuat YÜZER**	**……**

Anadolu Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …… tarih ve …… sayılı kararıyla onaylanmıştır.

Enstitü Müdürü

ÖZET

Doktora Tezi

REGRESYONDA PÜRÜZLÜLÜK CEZA YAKLAŞIMI

Rabia Ece OMAY

Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışman: Prof. Dr. Memmedağa MEMMEDLİ
2007, 121 sayfa

Tez çalışmasında splayn düzeltme ve regresyon splaynı ile regresyonda pürüzlülük ceza yaklaşımı ele alınmıştır. Toplamsal ve genelleştirilmiş toplamsal regresyon modellerinin tahmini için farklı algoritmalar ve onların yakınsamaları incelenmiştir. Bazı açıklayıcı değişkenlerin yanıt değişkenini doğrusal etkilemediği karmaşık regresyon problemlerinde, splayn düzeltme ve regresyon splaynı yöntemlerinin çok önemli olduğu, uygulamalarla ortaya konmuştur. Regresyonda pürüzlülülük ceza yaklaşımında iyi bir modelin elde edilmesi, optimum düzeltme parametresi veya uygun serbestlik derecesinin seçimi ile gerçekleştirilir. Bu nedenle, düzeltme parametresinin seçimi ile ilgili bazı kriterler ve bazı serbestlik derecesi tanımları incelenmiştir. Büyük veri setine sahip ve çok boyutlu problemlerde, regresyon splaynı ile pürüzlülük ceza yaklaşımının daha kullanışlı olduğu gösterilmiştir. Diğer taraftan, ince tabakalı splaynları içeren genelleştirilmiş toplamsal modellerin çok yararlı olabileceği gözlenmiştir.

Anahtar Kelimeler: Pürüzlülük Ceza Yaklaşımı, Splayn Düzeltme Yaklaşımı, Regresyon Splayn Yaklaşımı, Toplamsal Model, Genelleştirilmiş Toplamsal Model, İnce Tabakalı Splayn

ABSTRACT

PhD Dissertation

ROUGHNESS PENALTY APPROACH IN REGRESSION

Rabia Ece OMAY

Anadolu University
Graduate School of Sciences
Statistics Program

Supervisor: Prof. Dr. Memmedağa MEMMEDLİ
2007, 121 pages

In this thesis study, roughness penalty approach is examined by using smoothing spline and regression spline. Different algorithyms and convergence of those are studied for estimation of additive and generalized additive regression models. It is proved in the applications that the smoothing spline and regression spline methods are very important in complicated problems considering some explanotory variables have not linear effect on response variables. Obtaining a good model in roughness penalty approach in regression can be done by choosing the optimum smoothing parameter or appropriate degrees of freedom. Therefore some criteria about choosing the smoothing parameter and some defines of degrees of freedom are examined. It is shown that, when regression spline is used, roughness penalty method is more convenient in problems with large data input and multiple dimension. At the same time, it is observed that generalized additive models with thin plate spline can be beneficial.

Keywords: Roughness penalty approach, Smoothing spline approach, Regression spline approach, Additive Model, Generalized additive model, Thin plate spline

TEŞEKKÜR

Tez çalışmamda değerli bilgileri ve önerileri ile bana büyük destek olan ve
bu tezin ortaya çıkmasında çok büyük rol oynayan
kıymetli ve saygı değer hocam
Prof. Dr. Memmedağa MEMMEDLİ'ye,

çalışmalarımda benden yardımlarını esirgemeyen değerli hocalarım
Prof. Dr. Embiya AĞAOĞLU
Prof. Dr. Hasan DURUCASU
Prof. Dr. Ali Fuat YÜZER
ve diğer bölüm hocalarıma,

büyük özveri ile desteğini esirgemeyen ve attığım her adımda
benimle birlikte olan çok sevdiğim eşim Barış'a
anneme ve babama
ve
hayatımıza girdiği andan itibaren varlığıyla yuvamıza
daha fazla mutluluk ve uğur getiren
hayatımın anlamı, minik meleğim
oğlum Halil Ege'ye

teşekkür ederim…

Rabia Ece OMAY
Temmuz-2007

İÇİNDEKİLER

Sayfa

ÖZET……i

ABSTRACT ii

TEŞEKKÜR……..iii

İÇİNDEKİLER……iv

ŞEKİLLER DİZİNİ……..vi

TABLOLAR DİZİNİ……..vii

SİMGELER VE KISALTMALAR DİZİNİ……..…..viii

1. GİRİŞ 1

2. ÖN BİLGİLER 5

2.1. Pürüzlülük Cezası, Cezalı En Küçük Kareler Regresyonu…….5

2.2. Kübik Splayn Enterpolasyonu…….7

2.3. Genelleştirilmiş Doğrusal Modeller (GLM)…….10

2.3.1. Dağılımların üstel ailesi……10

2.3.2. Üstel aile dağılımlarının özellikleri……...11

2.3.3. GLM için maksimum olabilirlik tahmini……..14

2.3.4. Log-olabilirlik oran istatistiği……...17

2.3.5. Sapma (deviance) için örnekleme dağılımı……..18

3. SPLAYN DÜZELTME İLE REGRESYON 20

3.1. Nonparametrik Regresyon Modeli……20

3.2. Semiparametrik (Kısmiparametrik) Regresyon……21

3.3. Splayn Düzeltme ile Toplamsal Regresyon Modelleri…….26

3.3.1. Toplamsal modeller için tahmin denklemleri……...27

3.3.2. İki düzeltici ile backfitting algoritması…….29

3.3.3. Çözümün varlığı ve tekliği, Backfitting algoritmasının yakınsaması…….32

3.4. Düzeltme Parametresinin Seçimi, Serbestlik Derecesi…….37

3.4.1. Düzeltme parametresinin seçimi……...38

3.4.2. Serbestlik derecesi…….40

4. KÜBİK REGRESYON SPLAYNI İLE PÜRÜZLÜLÜK CEZA YAKLAŞIMI **42**

4.1. Cezalı Regresyon Splaynı ile Nonparametrik Regresyon……...42

4.1.1. Taban Fonksiyonlar Yardımıyla Modelleme……43

4.1.2. Kübik Splayn Taban……..44

4.2. Regresyon Splayn ile Toplamsal ve Genelleştirilmiş Toplamsal Regresyon Modelleri……...46

4.2.1. Toplamsal Modeller……..46

4.2.2. Genelleştirilmiş Toplamsal Modeller (GAM)……...47

4.3. Serbestlik Derecesi ve Düzeltme Parametresinin Seçimi……...49

4.4. İnce Tabakalı Splayn (Thin Plate Spline-TPS)……...51

4.4.1. İki Boyutlu İnce Tabakalı Splayn…….52

4.4.2. d Boyutlu TPS…….54

4.4.3. İnce Tabakalı Regresyon Splayn (TPRS)…….56

4.4.4. Tenzor Çarpım Splaynı…….57

5. SPLAYN DÜZELTME VE REGRESYON SPLAYNI İLE UYGULAMA **60**

5.1. Evlerin Özelliklerinin Satış Fiyatları ve Kira Fiyatları Üzerindeki Etkisinin Splayn Düzeltme ile İncelenmesi……60

5.2. Hava Kirliliğinin Ölüm Oranı Üzerine Etkisinin Regresyon Splayn ile İncelenmesi……..79

6. SONUÇ VE ÖNERİLER **112**

KAYNAKLAR **116**

ŞEKİLLER DİZİNİ

5.1. Evlerin satış fiyatlarının (a) komşu gelirlerine (b) brüt kullanım alanına göre değişimi ve güven aralıkları……..……...66
5.2. Evlerin kira fiyatlarının (a) depozitolarına (b) yaşlarına göre değişimi ve %95 güven aralıkları……...…….71
5.3. Evlerin fiyatlarının (a) yaşlarına (b) kullanım alanlarına göre değişimi ve %95 güven aralıkları……75
5.4. Doğrusal model için kontrol grafikleri……...87
5.5. Semiparametrik toplamsal model için kontrol grafikleri……...92
5.6. Semiparametrik toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri…….93
5.7. Genelleştirilmiş toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri…….95
5.8. İnce tabakalı splayn içeren son model için kontrol grafikleri……97
5.9. İnce tabakalı splayn içeren son modelden elde edilen pürüzsüz fonksiyon tahminleri……97
5.10.Doğrusal model için kontrol grafikleri……100
5.11.Tam toplamsal model için kontrol grafikleri……...104
5.12.Tam toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri…….104
5.13.Genelleştirilmiş toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri……...108
5.14.İnce tabakalı splayn içeren son model için kontrol grafikleri……..110
5.15.İnce tabakalı splayn içeren son modelden elde edilen pürüzsüz fonksiyon tahminleri……...111

TABLOLAR DİZİNİ

2.1. Bazı üstel aile dağılımlarının karakteristikleri……12
5.1. Doğrusal regresyon modeline ait sonuçlar……..62
5.2. Semiparametrik toplamsal regresyon modeline ait sonuçlar……..63
5.3.Toplamsal regresyon modeline ait sonuçlar……64
5.4. Uygun semiparametrik toplamsal regresyon modeline ait sonuçlar……...65
5.5. Modellerin belirlilik katsayıları ve sapmaları…….67
5.6. Uygun semiparametrik toplamsal regresyon modeli sonuçları……...69
5.7. Modellerin belirlilik katsayıları ve sapmaları…….71
5.8. Uygun semiparametrik toplamsal regresyon modeli sonuçları……..74
5.9. Doğrusal regresyon modeli sonuçları…….....76
5.10. Semiparametrik$_{YAŞ}$ regresyon modeli sonuçları……..77
5.11. Semiparametrik$_{ALAN}$ regresyon modeli sonuçları……78
5.12. Semiparametrik toplamsal model ve diğer modeller için bazı sonuçlar…..78
5.13. Doğrusal model için özet istatistikler……...86
5.14. Genelleştirilmiş doğrusal model için özet istatistikler……..89
5.15. Kurulan modellerden çıkarılan sapan gözlem adayları…….90
5.16. Sapan gözlem adayları çıkarılarak kurulan modellere ait sonuçlar……..90
5.17. Semiparametrik toplamsal model için özet istatistikler……91
5.18. Genelleştirilmiş toplamsal model için özet istatistikler……94
5.19. Semiparametrik toplamsal ve genelleştirilmiş toplamsal model için hipotez testi……...95
5.20. İnce tabakalı splayn modeli için özet istatistikler…….96
5.21. Doğrusal model için özet istatistikler……..100
5.22. Genelleştirilmiş doğrusal model için özet istatistikler……102
5.23. Tam toplamsal model için özet istatistikler……103
5.24. Semiparametrik toplamsal model için özet istatistikler……..105
5.25. Genelleştirilmiş toplamsal model için özet istatistikler……..107
5.26. Yeni ince tabakalı splayn modeli için özet istatistikler……...109

KISALTMALAR DİZİNİ

AIC : Akaike bilgi kriteri

AIC_C : Düzeltilmiş Akaike bilgi kriteri

ASR : Ortalama artık kareler

CV : Çapraz geçerlilik

EKK : En küçük kareler

GAM : Genelleştirilmiş toplamsal modeller

GCV : Genelleştirilmiş çapraz geçerlilik

GLM : Genelleştirilmiş doğrusal model

IRLS : İteratif olarak yeniden ağırlıklandırılmış en küçük kareler

MSE : Hata kareler ortalaması

NCS : Doğal kübik splayn

PIRLS : Cezalı iteratif olarak yeniden ağırlıklandırılmış en küçük kareler

PPR : İzdüşüm takip regresyonu

RSS : Artık kareler toplamı

TPS : İnce tabakalı splayn

TPRS : İnce tabakalı regresyon splayn

1. GİRİŞ

Regresyon problemlerinde, verilen veya elde edilen veri kümesine dayanarak bağımsız (açıklayıcı) değişkenlerin bağımlı (yanıt) değişkene fonksiyonel etkisi istatistiksel olarak analiz edilmektedir. Doğrusal regresyon modeli (Draper ve Smith, 1998; Myers, 1990;Montgomery ve ark., 2001) en basit regresyon modeli biçimi olarak, bazı koşullar sağlandığında, açıklayıcı değişkenlerin yanıt değişkenini doğrusal olarak etkilemesi ve yanıt değişkeninin normal dağılıma sahip olması varsayımları üzerine incelenmektedir. Doğrusal regresyon problemi için geniş teorik ve pratik çalışmalar yapılmıştır ve elde edilen sonuçlar daha karmaşık regresyon modellerinin incelenmesi için teorik ve pratik taban oluşturmaktadır. Bir takım koşullar sağlandığında, doğrusal regresyon yaklaşımı bazı pratik tahmin problemlerinin çözümünde makul sonuçlar elde edilmesini sağlamaktadır. Fakat maalesef, çoğu pratik tahmin probleminde açıklayıcı değişkenlerin bir kısmı yanıt değişkenini doğrusal olarak etkilememektedir. Bu nedenle tam doğrusal olmayan, daha karmaşık bağıntı içeren regresyon modellerinin incelenmesi ihtiyacı ciddi bir şekilde ortaya çıkmaktadır.

Genelleştirilmiş regresyon modelleri (GLM) (McCullagh ve Nelder,1989; Myers ve ark., 2002; Dobson, 2002; Nelder ve Wedderburn, 1972) doğrusal regresyon modellerinin geliştirilmiş bir şeklidir. GLM' de yanıt dağılımı üstel aileye ait olan rassal değişkendir ve onun beklenen değeri açıklayıcı değişkenlerin doğrusal ifadesinin (doğrusal tahmincinin) bir monoton fonksiyonudur (link fonksiyonu). Link fonksiyonu doğrusal olmadığında GLM'nin doğrusal olmadığı da açıktır. Doğrusal regresyon, link fonksiyonu beklenen değerin özdeşlik fonksiyonu olan bir GLM' dir. GLM ve doğrusal regresyon parametrik regresyon problemleri sınıfına aittir. Bu problemlerde regresyon fonksiyonlarının biçimi bilinmektedir ve bu fonksiyonların tahmini için sadece sonlu sayıda parametreyi tahmin etmek gerekmektedir. Bu nedenle hata veya log-olabilirlik fonksiyonuna optimum değer veren parametrelerin bulunması problemi ortaya çıkmaktadır. Sonuç olarak, parametre tahmini normal denklemler veya skor denklemleri sisteminin çözümü ile belirlenmektedir. Maksimum log-olabilirlik probleminin çözümünde ağırlıklandırılmış tekrarlı en küçük kareler (IRLS - Iteratively

Reweighted Least Squares) algoritması uygulanabilmektedir (McCullagh ve Nelder, 1989; Nelder ve Wedderburn, 1972).

Uygulamada regresyon denklemlerinde, açıklayıcı değişkenlerin yanıt değişkenine etkisinin sonlu parametre içeren fonksiyonel bir ilişkiden daha da karmaşık bir ilişkiye sahip olduğu birçok problemle karşılaşılmaktadır. Bu durumda tahmin edilen ilişkilendirici fonksiyon belirli sınıf sonsuz boyutlu fonksiyonlar uzayından seçilir. Bu durumda, sonlu sayıda parametre tahmini problemi sonsuz boyutlu uzaydan bir eleman tahmini problemine, diğer bir deyişle parametrik olmayan (nonparametrik) regresyon problemine dönüşmüş olur.

Nonparametrik regresyon modellerinden en basiti bir açıklayıcı değişken içeren, yani bir düzeltici fonksiyonu olan modeldir (Green ve Silverman, 1994; Wahba, 1990; Eubank, 1999). Bu modelin gelişmiş biçimleri olarak aşağıdaki modelleri sıralamak mümkündür: Tek bir açıklayıcı değişkeni nonparametrik kısma, kalanlarını ise doğrusal kısma alan kısmi parametrik (semiparametrik) model; bileşenlerinin her biri tek bir açıklayıcı değişkenin fonksiyonu olan toplamasal (additive) model; GLM' nin genel biçimi olan genelleştirilmiş toplamsal model (generalized additive model-GAM). Bunun yanı sıra, bazı bileşenleri birden fazla açıklayıcı değişkenin fonksiyonu olan ince tabakalı splayn (thin plate splline-TPS) modelleri daha karmaşık ve gelişmiş modellerdir (Hastie ve Tibshirani, 1999; Buja ve ark. 1989; Wood, 2003).

Nonparametrik regresyon modellerinin analizinde yer alan temel yaklaşım pürüzlülük ceza yaklaşımıdır (Green ve Silverman, 1994; Hastie ve Tibshirani, 1999). Bu yaklaşımda, modeldeki bilinmeyen fonksiyonları tahmin etmek için, hata kareler toplamına veya log-olabilirlik fonksiyonuna bir ceza terimi eklenir. Ceza teriminin eklenmesi mevcut problemi, geleneksel enterpolasyon probleminden gerçek regresyon problemine dönüştürmüş olur. Ceza terimi farklı şekillerde tanımlanabilir (Wahba, 1990; Gu, 2002). Geleneksel olarak ceza terimi $\int_a^b \{f''(x)\}^2 dx$ integralinin bir λ düzeltme parametresine çarpımı olarak kullanılır ve $\hat{f}$ tahmin fonksiyonu ikinci mertebeden sürekli türeve sahip $f(.)$ fonksiyonlar uzayından seçilir. Schonberg'in 1964 yılında yayınladığı polinomial splaynların

optimumluk özelliğine ilişkin çalışması (bkz. Teorem 2.2), bahsedilen sonsuz boyutlu problemi sonlu boyutlu parametrik bir probleme dönüştürmek için imkan sağlamaktadır. Diğer bir ifadeyle, cezalı hataya optimum değer veren fonksiyonun bir kübik splayn olması bulgusu nonparametrik regresyonda tahmin problemini parametrik hale getirir. Dolayısıyla, nonparametrik regresyon probleminin polinomial splayn fonksiyonlarıyla incelenmesi, modelin analizini çok kolaylaştırmaktadır. Pürüzlülük ceza yaklaşımında splayn fonksiyonları farklı iki yöntemle kullanılabilir. Bunlardan biri splayn düzeltme, diğeri ise regresyon splaynı olarak adlandırılır (Hastie ve Tibshirani, 1999; Wood, 2002).

Splayn düzeltmede her bir nonparametrik açıklayıcı değişkenin tüm gözlem değerlerine uygun değerleri, düğüm noktaları olarak kullanılır. Bu durum, özellikle büyük veri setine sahip olan veya çok boyutlu toplamsal modeller için hesaplama zorluklarını ortaya çıkarmaktadır. Splayn düzeltme ile toplamsal modeller ve bu modellerde parametre tahmini için geriye uyum (backfitting) algoritması Buja ve ark. (1989) ve Hastie ve Tibshirani, (1999) tarafından ayrıntılı şekilde incelenmiştir.

Regresyon splaynı ile pürüzlülük ceza yaklaşımında splayn fonksiyonu belirli taban splayn fonksiyonlarının toplamı şeklinde ifade edilir ve burada bilinmeyenler, kullanılan taban fonksiyonların katsayılarıdır. Bu katsayılar cezalı hatanın minimizasyonu probleminden bulunur. Regresyon splaynında düğüm noktalarının sayısı (konumu) veya taban fonksiyonlarının sayısı özel olarak seçilerek, tüm gözlemlerin sayısından çok daha küçük alınabilir. Bu işlem, splayn düzeltme yöntemindeki hesaplama karmaşıklığını epey hafifletmiş olur. Eiler ve Marx ceza teriminin kesikli yaklaşımını ve taban fonksiyonları olarak B-splayınları kullanarak regresyon splaynı uygulamışlardır. Bu çalışmalar literatürde P-splayn yaklaşımı olarak ifade edilmektedir (Eilers ve Marx, 1996; Marx ve Eilers, 1998). Wood taban fonksiyonları olarak kübik splayn taban fonksiyonlarını (Gu, 2002) ve sürekli ceza terimini kullanarak regresyon splaynını uygulamıştır (Wood, 2000; Wood ve ark., 2002, Wood, 2003; Wood, 2004). Wood' un yaklaşımı daha genel olup Eiler ve Marx'ın P-splayınları için de uygulanabilir (Wood, 2002). Bu nedenle tez çalışmasında regresyon splaynı yönteminin incelendiği bölümlerde Wood'un yaklaşımı temel alınmıştır.

Splayn fonksiyonu ile nonparametrik regresyon analizinde düzelticiler doğrusaldır ve onların her biri bir düzeltici matrisle belirlenir. Doğrusal düzelticiler, verilmiş düzeltme parametreleri için yanıt değişkeninin sonlu doğrusal bir dönüşümünü yaparak, problemi enterpolasyondan regresyon problemi haline dönüştürmektedir (Green ve Siverman, 1994; Hastie ve Tibshirani, 1999). Nonparametrik regresyonda düzeltme parametresinin seçimi ve onunla ilişkili olan serbestlik derecesinin hesaplanması önemli problemlerden biridir ve bu konuda sürekli olarak çalışmalar yapılmaktadır (bkz. Aydın, 2005).

Bu tez çalışmasında genel yaklaşımın pürüzlülük ceza yaklaşımı olması nedeniyle, splayn düzeltme ve regresyon splaynı ile farklı nonparametrik regresyon modelleri ve tahmin algoritmaları incelenmiştir. Bu amaçla önce splayn enterpolasyonu ve GLM hakkında ön bilgiler verilmiş (Bölüm 2), daha sonra ise splayn düzeltme (Bölüm 3) ve regresyon splaynı (Bölüm 4) ile ilgili incelemeler verilmiştir. Son bölümde (Bölüm 5) ise iki sınıf uygulama problemli için en iyi sonuç veren uygun modeller belirlenmiş, genelleştirilmiş toplamsal regresyon modellerinin yüksek düzeydeki önemi sergilenmiştir. Uygulamada R ve S-Plus paket programlarından yararlanılmıştır.

2. ÖN BİLGİLER

Bu bölümde, tezde açıklanan temel konuların kolay anlaşılması için yardımcı kavram ve bilgiler verilmiştir. İlk olarak, regresyonda pürüzlülük cezası kavramı, cezalı en küçük kareler yaklaşımı özet olarak açıklanmıştır. Bir sonraki alt bölümde ise pürüzlülük ceza yaklaşımında önemli rol oynayan splayn fonksiyonları, kübik splayn enterpolasyonu, doğal kübik splayn enterpolantının optimumluk özelliği hakkında gerekli bilgiler verilmiştir. Son alt bölümde ise genelleştirilmiş toplamsal modeller için alt yapı oluşturan genelleştirilmiş doğrusal modeller (GLM) ele alınmıştır. Bu konuyla ilgili üstel aile dağılımları, karakteristikleri ve özellikleri, GLM için maksimum olabilirlik tahmini, skor denklemleri ve bu denklemlerin çözümü için IRLS (Iteratively Reweighted Least Squares) algoritması, log-olabilirlik oran istatistiği ve sapma (deviance) hakkında bilgiler verilmiştir.

2.1. Pürüzlülük Cezası, Cezalı En Küçük Kareler Regresyonu

Birçok tahmin problemi için doğrusal modeller, verilen veri kümesine uymaz ve bu gibi durumlarda, doğrusal olmayan uyum eğrilerinin oluşturulması zorunluluğu ortaya çıkar. Bir t açıklayıcı değişkenine sahip doğrusal olmayan regresyon modeli aşağıda (2.1) denkleminde verilmiştir.

$$y = f(t) + hata \tag{2.1}$$

Burada $f(.)$ belirli niteliklere sahip bilinmeyen bir fonksiyondur. Verilen (t_i, y_i) , $i = 1,...,n$ veri kümesi için, $f(.)$ fonksiyonu en küçük kareler (EKK) yöntemiyle belirli bir sınıftan belirlendiğinde, klasik enterpolasyon problemi ortaya çıkar. Örneğin, $f(.)$ k dereceden bir polinom olarak arandığında, katsayıları EKK yöntemiyle polinomial enterpolasyon probleminden belirlenir. $f(.)$ ikinci mertebeden sürekli türeve sahip eğri olarak arandığında, EKK metodu uygun uzayda $f(t_i) = y_i, i = 1,...,n$ koşullu bir enterpolasyon problemi halini almış olur. Bu durumda oluşturulan $\hat{f}(.)$ fonksiyonunun eğimi hızlı şekilde değişebilir ve veri kümesinde hata kareler toplamının sıfır olmasına karşın, ortaya çıkan "gürültülü" dış veriler için toplam hata çok büyük olabilir. Diğer bir

ifadeyle (2.1) regresyon probleminin enterpolasyon problemine dönüştürülmesi tahmin için istenen bir durum değildir.

Pürüzlülük ceza yaklaşımının temel fikri, eğrinin hızlı eğim değişimini ölçerek, tahmin problemleri için verilen verilere uyum ve dalgalanma arasında uygun bir uzlaşma sağlamaktır.

$[a,b]$ aralığında tanımlanan f eğrisinin "pürüzlülüğü" farklı yollarla ölçülebilir. İkinci mertebeden sürekli türeve sahip f eğrisi için popüler pürüzlülük ölçüsü olarak

$$\int_a^b \{f''(t)\}^2 dt \tag{2.2}$$

integralinin değeri dikkate alınabilir. (2.2) integralinin minimum değerinin bir doğal kübik splaynda gerçekleştiği, Schonberg'in (1964a, 1964b) çalışmalarının sonuçlarının özel bir durumu olarak, bir sonraki alt bölümde Teorem 2.2' de verilmiştir. Bu önemli teorem, (2.2) integralinin minimum değerinin, bir kare formun değeri olarak hesaplanmasını sağlar (bak. Teorem2.1, formül (2.10)).

(2.1) regresyon problemi için pürüzlülük ceza yaklaşımı ile eğri tahmini problemini ele alınsın. $[a,b]$ aralığında tanımlanan ve ikinci mertebeden türeve sahip f fonksiyonu ve λ düzeltme parametresi verildiğinde cezalı en küçük kareler toplamı, aşağıdaki gibi tanımlanır.

$$S(f) = \sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int_a^b \{f''(x)\}^2 dx \tag{2.3}$$

Cezalı en küçük kareler tahmincisi, $S(f)$ fonksiyoneline ikinci mertebeden türeve sahip f fonksiyonlar uzayında minimum değer veren $\hat{f}$ fonksiyonu olur. (2.3) ifadesinde $\lambda \int_a^b \{f''(x)\}^2 dx$ pürüzlülük ceza terimi, $\sum_{i=1}^{n} \{y_i - f(t_i)\}^2$ ise hata kareler toplamıdır. $S(f)$ fonksiyonelinin tanımladığı eğri hata kareler toplamının etkisiyle verilere uyum sağlamakla beraber pürüzlülük ölçüsüyle de değerlendirilir. Verilmiş λ için $S(f)$'nin minimizasyonu pürüzsüzlük ve verilere uyum arasında en iyi biçimde uzlaşma sağlar. λ düzeltme parametresi artık hatalar ile yerel değişim (dalgalanma) arasındaki "değişme oranını" ifade eder. λ yeteri kadar küçük olduğunda $S(f)$ ifadesinde hata kareler toplamının ağırlığı

büyük olur ve $\hat{f}$ eğri tahmincisi verileri daha yakından izler. λ'nın büyük değerlerinde ise $S(f)$ ifadesinde pürüzlülük ceza teriminin ağırlığı büyük olur ve bu durumda $\hat{f}$ tahmincisi çok küçük eğrilik sergiler.

2.2. Kübik Splayn Enterpolasyonu

Tez çalışmasında splayn fonksiyonlarının önemli rol oynaması nedeniyle bu alt bölümde kübik splayn fonksiyonlarının tanımı, yapısı ve bazı özellikleri hakkında bilgi verilmektedir.

$t_1, t_2, ..., t_n$ noktaları, $[a,b]$ parçasının $a < t_1 < t_2 < ... < t_n < b$ koşulunu sağlayan noktaları olsun. t_i, $i = 1,2,...,n$ *düğüm noktaları* olarak adlandırılır.

Tanım 2.1. $[a,b]$ parçasında tanımlanmış $f(t)$ fonksiyonuna, aşağıdaki iki koşulu sağladığında $a < t_1 < t_2 < ... < t_n < b$ düğüm noktalarıyla *kübik splayn* denir.

a. Her bir $[t_i, t_{i+1}]$, $i = 0,1,2,...,n$ aralığında $f(t)$ kübik polinomdur, yani

$$t_i \le t \le t_{i+1} \text{ için, } f(t) = a_{0i} + a_{1i}(t - t_i) + a_{2i}(t - t_i)^2 + a_{3i}(t - t_i)^3 \qquad (2.4)$$

b. $[a,b]$ parçasında (t_i düğüm noktaları dahil olarak) $f(t), f'(t)$ ve $f''(t)$ fonksiyonları süreklidir.

Tanımdan anlaşıldığı gibi kübik splayn, düğüm noktalarında pürüzsüz olarak kübik polinom parçalarının birleşimidir.

Tanım 2.2. $[a,b]$ parçasında verilmiş kübik splaynın *a* ve *b* uç noktalarında ikinci ve üçüncü mertebeden türevleri sıfır olduğunda, bu splayna *doğal kübik splayn* (*Natural Cubic Splayn*-NCS) denir.

$$f''(a) = f''(b) = 0, \; f'''(a) = f'''(b) = 0 \qquad (2.5)$$

koşullarına da *doğal sınır koşulları* denir.

(2.5) koşullarından, $a_{30} = a_{20} = a_{3n} = a_{2n} = 0$ sonucuna varılır. Bu sonuç, $f(t)$ fonksiyonunun $[a,t_1]$ ve $[t_n,b]$ sınır parçalarında doğrusal fonksiyon olduğunu ifade etmektedir.

$f(.)$ fonksiyonu $t_1 < ... < t_n$ düğüm noktalarıyla bir doğal kübik splayn olsun ve

$$f_i = f(t_i) \text{ ve } \gamma_i = f''(t_i),\ i = 1,2,...,n \tag{2.6}$$

olarak tanımlansın. Doğal kübik splayn (NCS) için $\gamma_1 = \gamma_n = 0$ olur. $\mathbf{f} = (f_1,...,f_n)^T$, $\boldsymbol{\gamma} = (\gamma_2,...,\gamma_{n-1})^T$ ve $h_i = t_{i+1} - t_i,\ i = 1,2,...,n-1$ olsun. $t_1, t_2, ..., t_n$ düğüm noktalarının yardımıyla **Q** ve **R** *bant* matrisleri tanımlanabilir.

$n \times (n-2)$ boyutlu $\mathbf{Q}(q_{ij})$ matrisinin elemanları aşağıdaki gibi hesaplanır:

$$\begin{aligned} &q_{ij} = 0, \text{ eğer } |i-j| \geq 2 \text{ ise };\ i = 1,...,n;\ j = 2,...,n-1 \\ &q_{j-1,j} = h_{j-1}^{-1},\ q_{jj} = -h_{j-1}^{-1} - h_j^{-1},\ q_{j+1,j} = h_j^{-1},\ j = 2,...,n-1 \end{aligned} \tag{2.7}$$

Q matrisinin ilk sütunu $j = 2$ ile işaretlenmiştir. $(n-2) \times (n-2)$ boyutlu simetrik $\mathbf{R}(r_{ij})$ matrisinin r_{ij} elemanları ise aşağıdaki gibi hesaplanır.

$$\begin{aligned} &r_{ij} = 0 \text{ eğer } |i-j| \geq 2 \text{ ise} \\ &r_{ii} = \frac{1}{3}(h_{i-1} + h_i),\ i = 2,...,n-1 \\ &r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i,\ i = 2,...,n-2 \end{aligned} \tag{2.8}$$

R matrisi simetriktir ve kesin köşegen dominant matristir, diğer bir ifadeyle $|r_{ii}| > \sum_{i \neq j} |r_{ij}|,\ i = 2,...,n-1$'dir. Bu nedenle de R simetrik *pozitif tanımlı* matristir (Golub ve Loan, 1996).

Doğal kübik splaynlar için aşağıdaki iki özellik anahtar rol oynamaktadır (Green ve Silverman, 1994).

Teorem 2.1. (2.6) ile tanımlanan **f** ve **γ** vektörleri ancak ve ancak

$$\mathbf{Q}^T\mathbf{f} = \mathbf{R}\boldsymbol{\gamma} \tag{2.9}$$

koşulu sağlandığında bir $f(.)$ doğal kübik splaynını belirtir. Bu durumda pürüzlülük ceza terimi şöyle hesaplanabilir:

$$\int_a^b f''(t)^2\,dt = \boldsymbol{\gamma}^T\mathbf{R}\boldsymbol{\gamma} = \mathbf{f}^T\mathbf{K}\mathbf{f} \tag{2.10}$$

Burada **K**,

$$\mathbf{K} = \mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T. \tag{2.11}$$

olarak ifade edilen *ceza matrisi*dir.

Enterpolasyon Splaynı

(t_i, y_i), $i = 1, 2, ..., n$ değerlerinin verildiği ve $a < t_1 < ... < t_n < b$ koşulunun sağlandığı varsayılsın. $f(t_i) = y_i, i = 1, ..., n$ koşullu splayn enterpolasyonu problemi ele alınsın. $n \geq 2$ $t_1 < t_2 < ... < t_{n-1} < t_n$ olduğunda verilen herhangi $y_1, y_2, ..., y_n$ değerleri için $f(t_i) = y_i, i = 1, ..., n$ koşulunu sağlayan bir tek doğal kübik splayn vardır. Çünkü Teorem 2.1'de $\mathbf{f}$ vektörü yerine $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ vektörü dikkate alındığında (2.9) eşitliği $\mathbf{Q}^T\mathbf{y} = \mathbf{R}\boldsymbol{\gamma}$ şekline dönüşür. $\mathbf{R}$ matrisi (kesin) pozitif tanımlı bir matris olduğundan, bu matrisin tersi vardır ve $\boldsymbol{\gamma} = \mathbf{R}^{-1}\mathbf{Q}\mathbf{y}$ ile elde edilen $\boldsymbol{\gamma}$ tektir.

Bu aşamada, $\mathbf{R}$ ve $\mathbf{Q}$ matrislerinin yapısına bağlı olarak bazı hesaplamalar açıklanacaktır. $\mathbf{R}$ matrisi üçköşegen (tridiagonal) matristir, çünkü $|i - j| \geq 2$ için $r_{ij} = 0$'dır. Bu nedenle $\mathbf{R}\boldsymbol{\gamma} = \mathbf{z}$ denklemi, $\mathbf{R}^{-1}$ bulunmadan , $O(n)$ oranda işlem sonucunda çözülebilir (Golub ve Loan, 1996). $\mathbf{Q}$ matrisinin de üçköşegen yapısı, $\mathbf{Q}^T\mathbf{f}$ çarpımının $\mathbf{f}$'den $O(n)$ oranda işlem sonucu elde edilebileceği sonucunu doğurur. $\mathbf{Q}^T\mathbf{f}$ 'nin kolay hesabı aşağıdaki gibi yapılır.

$$(\mathbf{Q}^T\mathbf{f})_i = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}} \tag{2.12}$$

Burada, $h_i = t_{i+1} - t_i$, $i = 1, 2, ..., n-1$'dir. Sonuç olarak, n sayıda (t_i, y_i) noktası için doğal kübik splayn enterpolantının $O(n)$ oranda işlem sonucu elde edilebilirliği söylenebilir.

Algoritma: NCS Enterpolasyonu

1. $f_i = y_i, i = 1, ..., n$ kabul et.
2. (2.12) formülünü kullanarak $\mathbf{z} = \mathbf{Q}^T\mathbf{f}$ vektörünü bul.
3. $\mathbf{R}\boldsymbol{\gamma} = \mathbf{z}$ denkleminden $\boldsymbol{\gamma}$'yı bul.

Optimumluk Özelliği

Doğal kübik splayn enterpolantı aşağıdaki optimumluk özelliğine sahiptir (Schoenberg 1964a, 1964b).

Teorem 2.2. $n \geq 2$ ve $a < t_1 < ... < t_n < b$ olduğunda, $f(.)$ fonksiyonu $y_1, y_2, ..., y_n$ değerleri için $t_1, t_2, ..., t_n$ noktalarında NCS enterpolantı ise, herhangi bir $\tilde{f}(.) \in C^2[a,b]$ ve $\tilde{f}(t_i) = y_i$, $i = 1, ..., n$ koşullu $\tilde{f}(.)$ fonksiyonu için

$$\int_a^b [f''(t)]^2 dt \leq \int_a^b [\tilde{f}''(t)]^2 dt \tag{2.13}$$

olur ve eşitlik $\tilde{f}(.) = f(.)$ durumunda sağlanır.

Regresyon modellerinde NCS enterpolantının bu önemli özelliği, pürüzsüz fonksiyonlar uzayında pürüzlülük ceza ifadesinin minimum değerinin bir doğal kübik splaynda gerçekleştiğini gösterir. Bu durum, sonsuz boyutlu optimizasyon probleminin sonlu boyutlu bir probleme dönüştürülmesine katkı sağlar.

2.3. Genelleştirilmiş Doğrusal Modeller (GLM)

Bir genelleştirilmiş doğrusal modelin (GLM) temel düşüncesi, yanıt değişkeninin beklenen değerinin uygun bir (link) fonksiyonu için doğrusal bir model geliştirmektir. Bir GLM aşağıda ifade edilen temel yapıya sahiptir.

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2.14}$$

Burada $\mu_i \equiv E(Y_i)$; g, pürüzsüz (smooth) bir *link fonksiyonu*; $\mathbf{x}_i^T$, $\mathbf{X}$ model matrisinin *i*. satırı ve $\boldsymbol{\beta}$, bilinmeyen parametre vektörüdür. Burada yer alan η_i ise, *doğrusal kestirici* (linear predictor) olarak adlandırılır. Ek olarak, bir GLM iki varsayımı dikkate alır: Y_i'ler bağımsızdır ve Y_i, *üstel aile*den gelen bir dağılıma sahiptir. Dağılımların *üstel ailesi*, Poisson, Binom, Gamma ve Normal dağılım gibi birçok yaygın dağılımları içerir.

2.3.1. Dağılımların üstel ailesi

Bir GLM'deki yanıt değişkeni üstel aileden herhangi bir dağılıma sahip olabilir. Eğer bir dağılımın olasılık yoğunluk fonksiyonu aşağıdaki biçimde yazılabiliyorsa, o dağılım üstel aileden gelmektedir (Nelder ve Wedderburn, 1972).

$$f_\theta(y) = \exp\{[y\theta - b(\theta)]/a(\phi) + c(y,\phi)\} \tag{2.15}$$

Burada $b(\cdot)$, $a(\cdot)$ ve $c(\cdot)$ keyfi fonksiyonlar; ϕ, keyfi bir *ölçek parametresi* (scale parameter) ve θ, dağılımın *kanonik parametresi* olarak bilinir. Örneğin, normal dağılımın üstel ailenin bir dağılımı olduğu kolaylıkla görülebilir.

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2} \right] = \exp\left[\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2} - \ln\left(\sigma\sqrt{2\pi}\right) \right]$$
$$= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \ln\left(\sigma\sqrt{2\pi}\right) \right].$$

$\theta = \mu$, $b(\theta) = \theta^2/2 \equiv \mu^2/2$, $a(\phi) = \phi = \sigma^2$ ve $c(\phi, y) = -y^2/(2\phi) - \ln\left(\sqrt{\phi 2\pi}\right) \equiv y^2/\left(2\sigma^2\right) - \ln\left(\sigma\sqrt{2\pi}\right)$ olarak ifade edildiğinde normal dağılımın, (2.15)'de verilen üstel aile biçimine uygun olduğu görülebilmektedir.

Poisson dağılımının da üstel aileden olduğu hemen görülebilir:

$$f_\mu(y) = \frac{\mu^y \exp(-\mu)}{y!} = \exp[y \ln \mu - \mu - \ln(y!)].$$

Buna göre de Poisson dağılımı için $\theta = \ln(\mu)$, $\mu = \exp(\theta)$, $b(\theta) = \mu$, $a(\phi) = \phi = 1$ ve $c(\phi, y) = -\ln(y!)$ elde edilir. Tablo2.1'de üstel ailedeki bazı dağılımların karakteristikleri yer almaktadır.

2.3.2. Üstel aile dağılımlarının özellikleri

Olasılık yoğunluk fonksiyonunun tanımından

$$\int f_\theta(y) dy = 1 \tag{2.16}$$

olduğu bilinmektedir. (2.16)'nın her iki tarafında da θ'ya göre türev alınırsa,

$$\frac{d}{d\theta} \int f_\theta(y) dy = 0 \tag{2.17}$$

olur. (2.17)'de türev alma işlemi integral altında yapılabilir:

$$\int \frac{df_\theta(y)}{d\theta} dy = 0 \tag{2.18}$$

Benzer şekilde

$$\int \frac{d^2 f_\theta(y)}{d\theta^2} dy = 0 \tag{2.19}$$

eşitliği elde edilir. Bu sonuçlar herhangi bir üstel aile rassal değişkeninin beklenen değerinin ve varyansının elde edilmesi için kullanılabilir. Bu amaçla (2.15)'den,

Tablo 2.1. Bazı üstel aile dağılımlarının karakteristikleri

	Normal	Poisson	Binomial	Gamma	Ters Gamma
Notasyon	$N(\mu,\sigma^2)$	$P(\mu)$	$B(m,\pi)/m$	$G(m,\nu)$	$IG(\mu,\sigma^2)$
$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$	$\frac{\mu^y\exp(-\mu)}{y!}$	$\binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}$	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1}\exp\left(-\frac{\nu y}{\mu}\right)$	$\sqrt{\frac{\gamma}{2\pi y^3}}\exp\left[\frac{-\gamma(y-\mu)^2}{2\mu^2 y}\right]$
Aralık	$-\infty < y < \infty$	$y = 0,1,2,...$	$y = 0,1,...,n$	$y > 0$	$y > 0$
θ	μ	$\log(\mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$
ϕ	σ^2	1	1	$\frac{1}{\nu}$	$\frac{1}{\gamma}$
$a(\phi)$	$\phi(=\sigma^2)$	$\phi(=1)$	$\phi(=1)$	$\phi\left(=\frac{1}{\nu}\right)$	$\phi\left(=\frac{1}{\gamma}\right)$
$b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$n\log\left(1+e^\theta\right)$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
$c(y,\phi)$	$-\frac{1}{2}\left[\frac{y^2}{\theta}+\log(2\pi\phi)\right]$	$-\log(y!)$	$\log\binom{n}{y}$	$\nu\log(\nu y)-\log\{y\Gamma(\nu)\}$	$-\frac{1}{2}\left[\log(2\pi y^3\phi)+\frac{1}{\phi y}\right]$
$V(\mu)$	1	μ	$\mu\left(1-\mu/n\right)$	μ^2	μ^3
$g_c(\mu)$	μ	$\log(\mu)$	$\frac{\mu}{n-\mu}$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$

$$\frac{df_\theta(y)}{d\theta} = \left\{\left[y - b'(\theta)\right]/a(\phi)\right\} f_\theta(y)$$

eşitliği, (2.18)'den ise

$$\int \left\{\left[y - b'(\theta)\right]/a(\phi)\right\} f_\theta(y)dy = 0 \text{ ve } \frac{1}{a(\phi)}\int y f_\theta(y)dy - \frac{b'(\theta)}{a(\phi)}\int f_\theta(y)dy = 0$$

eşitliği elde edilir. Beklenen değerin tanımına göre $E[Y] = \int y f_\theta(y)dy$ olduğu için

$$\frac{1}{a(\phi)}E[Y] - \frac{b'(\theta)}{a(\phi)} = 0 \tag{2.20}$$

eşitliği bulunur. Buradan

$$E[Y] = b'(\theta) \tag{2.21}$$

olarak elde edilir.

Benzer bir düzenleme, üstel aile rassal değişkeninin varyansının bulunmasında da kullanılabilir. Bu amaçla,

$$\frac{d^2 f_\theta(y)}{d\theta^2} = \left[-b''(\theta)/a(\phi)\right] f_\theta(y) + \left\{\left[y - b'(\theta)\right]/a(\phi)\right\}^2 f_\theta(y) \tag{2.22}$$

Denklemi elde edilir. (2.21) kullanılarak, (2.22) eşitliği aşağıdaki gibi yazılır.

$$\frac{d^2 f_\theta(y)}{d\theta^2} = \left[-b''(\theta)/a(\phi)\right] f_\theta(y) + \left[1/a(\phi)\right]^2 \left\{y - E[Y]\right\}^2 f_\theta(y).$$

(2.19)'dan,

$$\int \frac{d^2 f_\theta(y)}{d\theta^2}dy = \left[-b''(\theta)/a(\phi)\right]\int f_\theta(y)dy + \left[1/a(\phi)\right]^2 \int \left\{y - E[Y]\right\}^2 f_\theta(y)dy = 0$$

ifadesi elde edilir. (2.16)'dan ve varyansın tanımından ($Var(Y) = \int \left\{y - E[Y]\right\}^2 f_\theta(y)dy$),

$$-\frac{b''(\theta)}{a(\phi)} + \frac{1}{\left[a(\phi)\right]^2} Var(Y) = 0 \tag{2.23}$$

elde edilir. (2.23) tekrar düzenlenirse,

$$Var(Y) = a(\phi)b''(\theta) \tag{2.24}$$

eşitliği bulunur. (2.21) ve (2.24) ifadeleri sırasıyla, $E(Y) = b'(\theta) = \frac{db(\theta)}{d\theta}$ ve $Var(Y) = b''(\theta)a(\phi) = \frac{d^2 b(\theta)}{d\theta^2}.a(\phi) = \frac{d\mu}{d\theta}a(\phi)$ olarak da ifade edilebilir. Burada Var_μ, $a(\phi)$ dışında yanıtın varyansı olsun; Var_μ, yanıtın varyansının

ortalamasıyla bağımlılığını gösterir. Bu nedenle $Var_\mu = \frac{Var(Y)}{a(\phi)} = \frac{d\mu}{d\theta}$ ya da $\frac{d\theta}{d\mu} = \frac{1}{Var_\mu}$ olarak ifade edilebilir.

2.3.3. GLM için maksimum olabilirlik tahmini

Bir genelleştirilmiş doğrusal modelin özelliklerini taşıyan $Y_1,...,Y_N$ bağımsız rassal değişkenleri göz önünde bulundurulsun. Söz konusu genelleştirilmiş doğrusal model,

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad (E(Y_i) = \mu_i)$$

olarak ifade edilir. Bu modeldeki $\boldsymbol{\beta}$ parametrelerini tahmin etmek gerekir.

Her bir Y_i için log-olabilirlik fonksiyonu,

$$l_i = \left[y_i\theta_i - b_i(\theta_i)\right]/a(\phi) + c_i(y_i,\phi) \tag{2.25}$$

olarak ifade edilir. Burada, (2.21), (2.24) ve (2.14)'e göre

$$E(Y_i) = \mu_i = b'(\theta_i),\ Var(Y_i) = a(\phi)b''(\theta_i),\ g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$$

olur, $\mathbf{x}_i$, $x_{ij}, j = 1,..., p$ elemanlarına sahip bir vektördür.

Tüm Y_i'ler için log-olabilirlik fonksiyonu,

$$l = \sum_{i=1}^{N} l_i = \sum_{i=1}^{N} \left\{\left[y_i\theta_i - b_i(\theta_i)\right]/a(\phi) + c_i(y_i,\phi)\right\}$$

olarak yazılabilir. β_j parametresinin maksimum olabilirlik tahminini elde etmek için zincir kuralı kullanılarak türev alınırsa,

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right] \tag{2.26}$$

$$\frac{\partial l_i}{\partial \theta_i} = \left[y_i - b_i'(\theta_i)\right]\Big/a(\phi) = \left[y_i - \mu_i\right]/a(\phi)$$

$$\frac{\partial \theta_i}{\partial \mu_i} = 1\Big/\left(\frac{\partial \mu_i}{\partial \theta_i}\right) = 1/b_i''(\theta_i),\ \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\mathbf{x}_i$$

elde edilir. Bu durumda (2.26) denklemi aşağıdaki hali alır.

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N}\left[\frac{(y_i-\mu_i)}{a(\phi)}\frac{1}{b_i''(\theta_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\mathbf{x}_i\right]$$

İncelenen genelleştirilmiş doğrusal model için dikkate alınan link fonksiyonunun *kanonik link* olması nedeniyle, $\eta_i = \theta_i$ ve dolayısıyla $\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta_i)$'dir. Bu durumda, $U_j = \sum_{i=1}^{N}\left[\frac{(y_i-\mu_i)}{a(\phi)}\frac{1}{b_i''(\theta_i)}b_i''(\theta_i)\mathbf{x}_i\right]$ ve

$$U_j = \frac{1}{a(\phi)}\sum_{i=1}^{N}(y_i-\mu_i)\mathbf{x}_i \tag{2.27}$$

elde edilir. Böylece, **β** için (2.28) denklem sistemi çözülerek parametrelerin maksimum olabilirlik tahminleri bulunabilir.

$$\frac{1}{a(\phi)}\sum_{i=1}^{N}(y_i-\mu_i)\mathbf{x}_i = \mathbf{0} \tag{2.28}$$

Birçok durumda $a(\phi)$'nin bir sabit olması nedeniyle denklem sistemi,

$$\sum_{i=1}^{N}(y_i-\mu_i)\mathbf{x}_i = \mathbf{0}$$

halini alır. Bu ifade aslında her bir model parametresi için $p = k+1$ denklemden oluşan bir sistemdir. Bu denklemler matris formunda aşağıdaki gibi yazlabilir.

$$\mathbf{U} = \mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0} \tag{2.29}$$

Burada $\boldsymbol{\mu}^T = [\mu_1, \mu_2, ..., \mu_p]$'dir. (2.29) denklem sistemi (**U**) *skor denklemleri* olarak adlandırılır (Nelder ve Wedderburn, 1972). Bu denklemlerin çözümü *IRLS (Iteratively Reweighted Least Squares)* algoritması kullanılarak çözülebilir. İlk olarak η_i^* çözümünün komşuluğunda birinci mertebeden Taylor serisi açılımı yapılır:

$$y_i - \mu_i \approx \frac{d\mu_i}{d\eta_i}(\eta_i^* - \eta_i).$$

Bir kanonik link için $\eta_i = \theta_i$ ve buna göre de

$$y_i - \mu_i \approx \frac{d\mu_i}{d\theta_i}(\eta_i^* - \eta_i)$$

olacaktır. Buradan,

$$\eta_i^* - \eta_i \approx (y_i - \mu_i)\frac{d\theta_i}{d\mu_i} \tag{2.30}$$

olur. Burada $Var_\mu = \frac{d\mu_i}{d\theta_i}$'dir. Bu nedenle (2.30) aşağıdaki gibi yazılabilir:

$$\eta_i^* - \eta_i \approx \frac{(y_i - \mu_i)}{Var_\mu} \text{ veya } \qquad y_i - \mu_i \approx Var_\mu(\eta_i^* - \eta_i) \tag{2.31}$$

(2.31) ifadesi (2.28)'de yerine koyulursa,

$$\frac{1}{a(\phi)}\sum_{i=1}^{N} Var_\mu(\eta_i^* - \eta_i)\mathbf{x}_i = \mathbf{0} \tag{2.32}$$

elde edilir. $\mathbf{V} = diag\{Var_\mu\}$ olsun. (2.32) matris formunda aşağıdaki gibi ifade edilecektir.

$$\mathbf{y} - \boldsymbol{\mu} \approx \frac{1}{a(\phi)}\mathbf{V}(\boldsymbol{\eta}^* - \boldsymbol{\eta})$$

Eğer $a(\phi)$ sabit ise skor denklemi aşağıdaki biçimlerde yazılabilir:

$$\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0} \ , \ \mathbf{X}^T\mathbf{V}(\boldsymbol{\eta}^* - \boldsymbol{\eta}) = \mathbf{0} \text{ ve ya } \mathbf{X}^T\mathbf{V}(\boldsymbol{\eta}^* - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

Bu durumda $\boldsymbol{\beta}$'nın maksimum olabilirlik tahmini,

$$\mathbf{b} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\boldsymbol{\eta}^*$$

olarak elde edilir. Burada $\boldsymbol{\eta}^*$ bilinmemektedir, bu nedenle

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$$

ifadesine dayalı iteratif bir metot izlenir. Newton-Raphson metoduna dayanan IRLS kullanıldığında,

$\mathbf{b} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{z}$' den

çözüm elde edilir.

Newton-Raphson metoduna dayanan IRLS aşağıdaki adımlarla verilebilir (Montgomery ve ark., 2001).

1. β'nın bir başlangıç ($\mathbf{b}_0$) tahminini EKK ile elde et;
2. $\mathbf{V}$ ve μ'yü tahmin etmek için $\mathbf{b}_0$'ı kullan.
3. $\boldsymbol{\eta}_0 = \mathbf{X}\mathbf{b}_0$ olarak al.
4. $\boldsymbol{\eta}_0$'a dayalı olarak $\mathbf{z}_1$'i al.

5. Yeni bir $\mathbf{b}_1$ tahmini elde et ve uygun yakınsama kriteri sağlanana kadar iterasyona devam et.

$\boldsymbol{\beta}$'nın tahmini olarak **b**, IRLS algoritması ile elde edilen son değer olsun. Eğer model varsayımları (örneğin, link fonksiyonunun seçimi) doğru ise, asimptotik olarak,

$$E(\mathbf{b}) = \boldsymbol{\beta}$$

olduğu gösterilebilir, çünkü **b**, (2.29) skor denklemlerinin çözümüdür. Tahmincilerin *enformasyon matrisi* $\mathbf{I}(\mathbf{b})$, skor denklemlerinin varyansı ile aşağıdaki eşitlilkle verilir.

$$\mathbf{I}(\mathbf{b}) = Var\left\{\frac{1}{a(\phi)}\left[\mathbf{X}^T(\mathbf{y}-\boldsymbol{\mu})\right]\right\} = \frac{\mathbf{X}^T\mathbf{V}\mathbf{X}}{[a(\phi)]^2} \tag{2.33}$$

Burada $\mathbf{V} = diag\{\sigma_i^{\,2}\}$'dir ve $\sigma_i^{\,2}$, μ_i'nin bir fonksiyonudur ve dikkate alınan dağılıma bağımlıdır. Bu nedenle **b**'nin asimptotik varyans-kovaryans matrisi,

$$Var(\mathbf{b}) = \mathbf{I}^{-1}(\mathbf{b}) = \left[\mathbf{X}^T\mathbf{V}\mathbf{X}\right]^{-1}[a(\phi)]^2 \tag{2.34}$$

olarak elde edilir.

2.3.4. Log-olabilirlik oran istatistiği

İlgilenilen modelin uyum iyiliğini (goodness-of-fit) test etmenin ve modelleri karşılaştırmanın bir yolu, tahmin edilebilecek maksimum parametreyi içeren ve *doymuş* (*saturated*) *model* olarak adlandırılan daha genel bir modelle, ilgilenilen modeli karşılaştırmaktadır. Doymuş model, ilgilenilen modelle aynı dağılıma ve link fonksiyonuna sahiptir ve aynı zamanda *maksimal* ya da *full model* olarak da adlandırılır.

Genel olarak *m*, tahminlenebilecek maksimum parametre sayısı olsun. $\boldsymbol{\beta}_{\max}$, doymuş model için parametre vektörünü ve $\mathbf{b}_{\max}$, $\boldsymbol{\beta}_{\max}$'ın maksimum olabilirlik tahmincisini göstersin. $L(\mathbf{b}_{\max}, \mathbf{y})$ olabilirlik fonksiyonu, gözlemler için aynı dağılım ve link fonksiyonuna sahip diğer olabilirlik fonksiyonlarından daha büyük olacaktır. Çünkü bu model verilerin en tam tanımını sağlar. $L(\mathbf{b}, \mathbf{y})$, ilgilenilen model için olabilirlik fonksiyonunun maksimum değerini göstersin. Bu durumda olabilirlik oranı,

$$\gamma = \frac{L(\mathbf{b}_{\max}; \mathbf{y})}{L(\mathbf{b}; \mathbf{y})} \tag{2.35}$$

olarak ifade edilir. Pratikte olabilirlik oranının logaritması kullanılır ve bu ifade aşağıda verilmiştir.

$$\log(\gamma) = l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y}) \tag{2.36}$$

$\log(\gamma)$'nın büyük değerleri ilgilenilen modelin, verileri doymuş modele göre zayıf tanımladığı anlamına gelir.

$2\log(\gamma)$, bir χ^2 dağılımı gösterir. Bu nedenle istatistikte daha yaygın olarak, $\log(\gamma)$ yerine $2\log(\gamma)$ kullanılır. Bu değer Nelder ve Wedderburn (1972) tarafından *sapma (deviance)* olarak adlandırılmıştır. Genelleştirilmiş modeller için sapma, *artık kareler toplamı* (RSS) rolünü üstlenir ve uyum iyiliği testi (goodness-of-fit) için ve modelleri karşılaştırmak için kullanılabilir.

Nonparametrik ve toplamsal modeller için sapma (deviance), modelleri ve bu modellerin farklarını değerlendirmek için kullanılır. Dağılım teorisi geliştirilmemiş olmasına karşın, χ^2 dağılımı, modelleri karşılaştırmak için referans dağılım olarak kullanılır (Hastie ve Tibshirani, 1999).

2.3.5. Sapma (deviance) için örnekleme dağılımı

Sapma (deviance), log-olabilirlik oran istatistiğinde tanımlandığı gibi,

$$\mathbf{D} = 2\left[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})\right] \tag{2.37}$$

dir. Eğer $\mathbf{b}$, $\boldsymbol{\beta}$ parametresinin en çok olabilirlik tahmincisi ise (dolayısıyla $\mathbf{U}(\mathbf{b}) = \mathbf{0}$), yaklaşık olarak,

$$l(\boldsymbol{\beta}) - l(\mathbf{b}) = -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{b})^T \mathbf{I}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b}) \tag{2.38}$$

olur. Bu durumda,

$$2\left[l(\mathbf{b}; \mathbf{y}) - l(\boldsymbol{\beta}; \mathbf{y})\right] = (\boldsymbol{\beta} - \mathbf{b})^T \mathbf{I}(\mathbf{b})(\boldsymbol{\beta} - \mathbf{b}) \tag{2.39}$$

istatistiği $\chi^2(p)$ dağılımına sahiptir. Burada *p*, parametre sayısını göstermektedir.

Sapma için örneklem dağılımının bu sonucundan,

$$\begin{aligned} \mathbf{D} &= 2\left[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})\right] \\ &= 2\left[l(\mathbf{b}_{\max}; \mathbf{y}) - l(\boldsymbol{\beta}_{\max}; \mathbf{y})\right] - 2\left[l(\mathbf{b}; \mathbf{y}) - l(\boldsymbol{\beta}; \mathbf{y})\right] + 2\left[l(\boldsymbol{\beta}_{\max}; \mathbf{y}) - l(\boldsymbol{\beta}; \mathbf{y})\right] \end{aligned} \tag{2.40}$$

türetilebilir. (2.40) ifadesindeki köşeli parantezler içindeki ilk terim bir $\chi^2(m)$ dağılımına sahiptir. Burada *m*, doymuş modeldeki parametre sayısıdır. İkinci terim bir $\chi^2(p)$ dağılımına sahiptir ve *p*, ilgilenilen modeldeki parametre sayısıdır. Üçüncü terim $\nu = 2\left[l(\boldsymbol{\beta}_{\max};\mathbf{y}) - l(\boldsymbol{\beta};\mathbf{y})\right]$ ise, pozitif bir sabittir. Eğer ilgilenilen model verilere neredeyse doymuş model kadar iyi uyum yapıyorsa bu değer sıfıra yakın olacaktır. Bu durumda sapma (deviance) için örnekleme dağılımı yaklaşık olarak,

$$\mathbf{D} \approx \chi^2(m-p,\nu) \tag{2.41}$$

olacaktır. Burada ν, merkezsel olmayan parametredir (Dobson, 2002).

3. SPLAYN DÜZELTME İLE REGRESYON

Bu bölümde nonparametrik ve toplamsal regresyon modelleri için splayn düzeltme ile pürüzlülük ceza yaklaşımı ele alınmıştır. Splayn düzeltmede açıklayıcı değişkenin verilen tüm değerleri oluşturulan splayn fonksiyonu için düğüm noktaları olarak göz önüne alınır. Nonparametrik ve toplamsal regresyon modellerinde cezalı hata kareler toplamına minimum değer veren fonksiyonun uygun bir splayn olması bilgisinden yola çıkarak, fonksiyonlar uzayında ele alınan optimizasyon problemi sonlu bir parametrik optimizasyon problemine dönüştürülebilir.

3.1. Nonparametrik Regresyon Modeli

$$y_i = f(t_i) + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}(0,\sigma^2),\ i = 1,2,...,n \qquad (3.1a)$$

regresyon modeli göz önüne alınsın. Burada $(y_i, t_i), i = 1,2,...,n$, gözlem verileri; $f(t)$, kendisi, birinci ve ikinci mertebeden türevleri $[a,b]$ parçasında sürekli olan pürüzsüz bilinmeyen bir fonksiyon; ε_i, i. hata terimidir.

$t_1, t_2, ..., t_n$ noktalarının $[a,b]$ parçasının içinde olup, $a < t_1 < t_2 < ... < t_n < b$ koşulunu sağladığı varsayılır. $y_1, y_2, ..., y_n$ ise bu düğüm noktalarına uygun gözlem değerleridir ve $n \geq 3$'dür.

Cezalı en küçük kareler problemi ele alınsın. İkinci mertebeden sürekli türeve sahip fonksiyonlar uzayında

$$\sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 \, dt \qquad (3.2a)$$

cezalı hata kareler toplamını minimum yapan fonksiyon bulunur. Burada $\lambda > 0$ sayısına *düzeltme parametresi* denir. $\lambda = 0$ olduğunda (3.2a) minimizasyon problemi $f(t_i) = y_i$ koşullu enterpolasyon problemine dönüşür. $\lambda \to \infty$ olduğunda (3.2a) minimizasyon problemi, doğrusal regresyon uyumunu verir ($f''(t) = 0,\ \ t \in [a,b]$).

$\mathbf{y} = (y_1, y_2, ..., y_n)^T$, $\mathbf{f} = (f(t_1), f(t_2), ..., f(t_n))^T$ ve $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ vektörleri tanımlansın. Bu durumda (3.1a) regresyon denklemi aşağıdaki gibi yazılabilir:

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2\mathbf{I}) \tag{3.1b}$$

(3.2a) cezalı hata kareler toplamında ceza teriminin yerine (2.10) ifadesi yazılırsa

$$\begin{aligned} S(f) &= (\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f}) + \lambda \mathbf{f}^T\mathbf{K}\mathbf{f} \\ &= \mathbf{f}^T(\mathbf{I}+\lambda\mathbf{K})\mathbf{f} - 2\mathbf{y}^T\mathbf{f} + \mathbf{y}^T\mathbf{y} \end{aligned} \tag{3.2b}$$

kare formu elde edilir. $\lambda > 0$ sayısı için $\lambda\mathbf{K}$ yarı pozitif tanımlı matris olduğundan, $(\mathbf{I}+\lambda\mathbf{K})$ kesin pozitif tanımlı matristir ve (3.2b) kare formunun bir tek minimumu vardır ve bu minimum (3.3) eşitliğinde verilmiştir.

$$\mathbf{f} = (\mathbf{I}+\lambda\mathbf{K})^{-1}\mathbf{y} \tag{3.3}$$

Burada $\mathbf{S}_\lambda = (\mathbf{I}+\lambda\mathbf{K})^{-1}$ matrisine *düzeltme matrisi* denir. $\mathbf{S}_\lambda$, (3.1a) regresyon denklemi için aynı zamanda *şapka (hat) matristir* ve $\hat{\boldsymbol{\mu}} = E(\mathbf{y}) = \mathbf{f} = \mathbf{S}_\lambda\mathbf{y}$.

(3.3) formülü ile hesaplanan $\mathbf{f} = (f_1, f_2, ..., f_n)^T$ vektörü ile $f(t_i) = f_i,\ i = 1,2,...,n$ enterpolasyon koşulları tanımlanır.

Teorem 3.1. (Reinsch, 1967) $n \geq 3$ olsun ve $t_1, t_2, ..., t_n$ noktalarının $a < t_1 < t_2 < ... < t_n < b$ koşullarını sağladığını varsayalım. $y_1, y_2, ..., y_n$ gözlem değerleri ve $\lambda > 0$ sayısı verildiğinde, $\hat{f}$, $\mathbf{f} = (\mathbf{I}+\lambda\mathbf{K})^{-1}\mathbf{y}$ değer-vektörü ve $t_1, t_2, ..., t_n$ düğüm noktalarına göre, $\tilde{f}(t_i) = f_i$ koşullu bir doğal kübik splayn olsun. Bu durumda $\forall f \in C^2[a,b]$ için,

$$S(\hat{f}) \leq S(f) \tag{3.4}$$

olur. (3.4)'deki eşitlik yalnız ve yalnız $\hat{f}$ ve f aynı olduğunda sağlanır.

Bu önemli teorem, (3.2) cezalı hata kareler toplamının minimumunun bir doğal kübik splayn olduğunu gösterir. Bu doğal kübik splayn, $t_1, t_2, ..., t_n$ düğüm noktaları ve (3.3) değer-vektörüyle tanımlanan $f(t_i) = f_i,\ i = 1,2,...,n$ koşullu enterpolasyon splaynıdır.

3.2. Semiparametrik (Kısmiparametrik) Regresyon

Her bir $y_i, i = 1,2,...,n$ gözlem değerine $p+1$ açıklayıcı değişkenin (p-boyutlu bir $\mathbf{x}_i$ vektörü ve bir t_i skaleri) uygun olduğu varsayılsın.

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + f(t_i) + \varepsilon_i,\ \varepsilon_i \sim \mathrm{N}(0, \sigma^2) \tag{3.5}$$

şeklinde ifade edilen regresyon modeline *semiparametrik regresyon modeli* denir. $\boldsymbol{\beta}$, regresyon katsayılarının p boyutlu vektörü; $f(\cdot) \in C^2[a,b]$, ikinci mertebeden sürekli türeve sahip tahmin edilecek bir pürüzsüz fonksiyon; ε_i, hata terimidir. (3.5)'de, $\mathbf{x}_i^T \boldsymbol{\beta}$ modelin parametrik kısmını, $f(t_i)$ ise nonparametrik kısmını ifade eder. (3.5) modeli için cezalı hata kareler toplamı,

$$S(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left\{ y_i - \mathbf{x}_i^T \boldsymbol{\beta} - f(t_i) \right\}^2 + \lambda \int_a^b \left\{ f''(t) \right\}^2 dt \tag{3.6a}$$

şeklindedir. Semiparametrik regresyonda gözlem noktaları $(y_i, \mathbf{x}_i, t_i)$, $i = 1, 2, ..., n$ şeklinde ifade edilen $p+2$ boyutlu noktalardır. Burada t_i, $i = 1, 2, ..., n$ düğümleri için $a < t_1 < ... < t_n < b$ koşulu sağlanmayabilir. Diğer bir ifadeyle, üst üste düşen (tekrarlanan) düğüm noktaları olabilir. Bu yetersizliği ortadan kaldırmak için **N** *benzerlik matrisi* kullanılır. $t_1, t_2, ..., t_n$ noktalarının artan değerleri sırası $s_1, s_2, ..., s_q$ olsun ($q \le n$). **N** matrisi şöyle tanımlanır:

$$N_{ij} = \begin{cases} 1 & \text{eğer } t_i = s_j \\ 0 & \text{eğer } t_i \neq s_j \end{cases} \quad i = 1, 2, ..., n;\ j = 1, 2, ..., q$$

N matrisi, $n \times q$ boyutlu matristir. Bu matrisinin her satırının sadece bir elemanı 1, kalan elemanları ise sıfırdır. *j*. sütunda s_j elemanına eşit olan t_i'ler için 1, kalanları ise 0'dır. Benzerlik matrisi kullanılarak, (3.6a) cezalı hata kareler toplamı aşağıdaki gibi yazılabilir.

$$S(\boldsymbol{\beta}, f) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f} \right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f} \right) + \lambda \int_a^b \left\{ f''(t) \right\}^2 dt$$

$\int_a^b \left\{ f''(t) \right\}^2 dt = \mathbf{f}^T \mathbf{K} \mathbf{f}$ olduğu bilinerek, (3.6a) ifadesi (3.2b) kare formuna benzer olarak,

$$S(\boldsymbol{\beta}, f) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f} \right)^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f} \right) + \lambda \mathbf{g}^T \mathbf{K} \mathbf{g} \tag{3.6b}$$

şeklinde yazılabilir. (3.6b) kare formunun minimizasyon problemi, aşağıdaki blok matrisli denklemin $\boldsymbol{\beta}$ ve **f** çözümünün bulunması problemine getirilir.

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{N} \\ \mathbf{N}^T\mathbf{X} & \mathbf{N}^T\mathbf{N}+\lambda\mathbf{K} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{f} \end{bmatrix}=\begin{bmatrix} \mathbf{X}^T \\ \mathbf{N}^T \end{bmatrix}\mathbf{y} \tag{3.7}$$

$\mathbf{X}$ matrisi tam sütun ranklı ise ve $i=1,2,...,n$ için $\delta_1+\delta_2 t_i$ doğrusal formuna eşit olan bir $\mathbf{x}_i^T\boldsymbol{\beta}$ doğrusal kombinasyonu yok ise (3.7) denklemler sisteminin bir tek çözümü vardır ve

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{N} \\ \mathbf{N}^T\mathbf{X} & \mathbf{N}^T\mathbf{N}+\lambda\mathbf{K} \end{bmatrix}$$

matrisi herhangi bir $\lambda>0$ için pozitif tanımlı matristir (Green ve Silverman, 1994). Bu koşullar, $(\mathbf{x}_i,1,t_i)$ açıklayıcı değişkenlerine sahip tam parametrik doğrusal model için en küçük kareler tahmincisinin tek olması koşullarıdır. (3.7) ifadesi, $p+q$ denklemler sistemidir ve bu sistemi doğrudan çözmek büyük boyut nedeniyle elverişli olmayabilir. Bu nedenle, farklı nümerik yöntemlerin kullanımı faydalıdır.

Backfitting (Geri Uyum) Algoritması

(3.7) denklemler sistemi,

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}=\mathbf{X}^T(\mathbf{y}-\mathbf{N}\mathbf{f}) \tag{3.8}$$

$$(\mathbf{N}^T\mathbf{N}+\lambda\mathbf{K})\mathbf{f}=\mathbf{N}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{3.9}$$

matris denklemler çifti şeklinde yazılabilir. (3.8)'de $\mathbf{f}$ bilindiğinde, y_i'den $f(t_i)=(\mathbf{N}\mathbf{f})_i$ çıkartılarak $\boldsymbol{\beta}$ en küçük kareler yöntemiyle bulunabilir. Tersine, $\boldsymbol{\beta}$ bilindiğinde, (3.9) denklemi $y_i-\mathbf{x}_i^T\boldsymbol{\beta}$ formlarına göre kübik splayn düzeltmeyi gerçekleştirmeye imkan sağlar.

Backfitting olarak bilinen iterasyon prosedürü Gauss-Seidel metoduna benzemektedir (Friedman ve Stuetzle, 1981; Buja ve ark., 1989). (3.8) ve (3.9) denklemleri kullanılarak, (3.7) için backfitting iterasyonu aşağıdaki gibi yazılabilir.

$$\mathbf{f}^{(n)}=(\mathbf{N}^T\mathbf{N}+\lambda\mathbf{K})^{-1}\mathbf{N}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(n-1)}) \tag{3.10}$$

$$\boldsymbol{\beta}^{(n)}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{N}\mathbf{f}^{(n)}) \tag{3.11}$$

(3.10), (3.11) iteratif süreci herhangi bir $f^{(0)}$ vektöründe başlatıldığında, (3.7) denkleminin tek çözümüne yakınsar (Green ve Silverman, 1994). (3.10) ve (3.11)

sürecinin her bir döngüsü, denklemleri hesaplama açısından kolaydır ve geleneksel metotlarla çözülebilir. (3.11), basit en küçük kareler denklemlerini, (3.10) ise λ parametreli, t_i noktalarında gözlem değerleri $y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ olan splayn düzeltme problemini içerir. (3.10) problemi örneğin, Reinsch algoritması (Reinsch, 1967) kullanılarak, $O(n)$ işlem oranında yerine getirilebilir.

Direkt (Dolaysız) Metot (Green ve ark., 1985)

(3.9) denklemlerinden,

$$\mathbf{Nf} = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

olduğu bulunur. $\mathbf{S}_\lambda$,

$$\mathbf{S}_\lambda = \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T \tag{3.12}$$

olarak tanımlandığında,

$$\mathbf{Nf} = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.13}$$

yazılabilir. Burada (3.12) ile tanımlanan $\mathbf{S}_\lambda$ düzeltme matrisinin *şapka (hat) matris* olduğu kanaatine varılır. (3.13) ifadesini (3.8) denkleminde göz önüne alarak, **β**'nın bulunması için, (köşegen olmayan $(\mathbf{I} - \mathbf{S})$ ağırlık matrisli) *genelleştirilmiş en küçük karelerin normal denklemleri* alınmış olur:

$$\mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T(\mathbf{I} - \mathbf{S})\mathbf{y} \tag{3.14}$$

(3.14) probleminin çözümü, **S** şapka (hat) matrisinin özel bant yapıya sahip olmasından dolayı, *Reinsch algoritmasının* yardımıyla kolaylaştırılabilir. (3.14) $p \times p$ doğrusal sistemi $O(p^2)$ işlem sonucunda standart metotlarla bulunabilir. Son olarak (3.13)'den **Nf** hızlı ve verimli olarak bulunabilir.

Direkt çözüm metodunun dezavantajı, onun ortogonal ayrışım metodlarıyla birleştirilmesinin mümkün olmamasıdır. Bu durum, hata yuvarlanmasının sonucu olarak büyük riske neden olur.

Speckman Algoritması

Kısmi splayn regresyonunda alternatif bir metot, Speckman (1988) tarafından verilmiştir . (3.5) semiparametrik modeli ele alınsın.

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + f(t_i) + \varepsilon_i \tag{3.15}$$

$\mathbf{x}_i$ açıklayıcı değişkenlerinin (3.16) denklemi ile ifade edilen t_i'lerle bir *regresyon bağımlılığı* varsayılsın.

$$\mathbf{x}_i = \boldsymbol{\xi}(t_i) + \boldsymbol{\eta}_i \tag{3.16}$$

Burada $\boldsymbol{\xi}$, t değişkeninin vektör-pürüzsüz fonksiyonudur; $\boldsymbol{\eta}_i$, hataların vektörüdür. f_0 fonksiyonu aşağıdaki gibi tanımlansın.

$$f_0(t_i) = \boldsymbol{\xi}^T(t_i)\boldsymbol{\beta} + f(t_i) \tag{3.17}$$

Bu durumda,

$$y_i = f_0(t_i) + \text{hata}$$

şeklinde yazılabilir. (3.15)'den (3.17) denklemi çıkartılırsa,

$$y_i - f_0(t_i) = \left\{\mathbf{x}_i - \boldsymbol{\xi}(t_i)\right\}^T \boldsymbol{\beta} + \text{hata} \tag{3.18}$$

denklemi elde edilir.

Verilmiş λ düzeltme parametresi için splayn düzeltmede $\mathbf{S}_\lambda$, düzeltme matrisi olsun. $\mathbf{X}$, satırları $\mathbf{x}_i^T$'ler ($i = 1, 2, ..., n$), $\boldsymbol{\Sigma}$ ise satırları $\boldsymbol{\xi}^T(t_i)$'ler olan matrislerdir. (3.18) ifadesi aşağıdaki sürecin (algoritmanın) ileri sürülmesine imkan sağlamaktadır:

a) $\left\{f_0(t_i)\right\}$ ve $\boldsymbol{\Sigma}$ splayn düzeltmelerinin tahmini için uygun olarak sırasıyla, $\mathbf{f}_0 = \mathbf{S}_\lambda \mathbf{y}$ ve $\boldsymbol{\xi} = \mathbf{S}_\lambda \mathbf{X}$'i hesapla.

b) $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$ ve $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}$ artıklarını hesapla.

c) (3.18)'e uygun olan $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \text{hata}$ doğrusal regresyon denkleminin $\boldsymbol{\beta}$ tahminini bul:

$$\hat{\boldsymbol{\beta}} = \left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{y}} \tag{3.19}$$

d) (3.19)'da elde edilen $\hat{\boldsymbol{\beta}}$ ifadesini (3.15)'de yerine koyarak, $y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ değerleri için splayn düzeltme ile $\hat{f}$ tahmin fonksiyonunu bul.

Bu algoritma, (3.16) regresyon bağlılığının makul olup olmamasına bakılmaksızın uygulanabilir. (a) adımında $\mathbf{y}$ ve $\mathbf{X}$ için aynı $\mathbf{S}_\lambda$ düzeltme matrisi uygulanmaktadır.

Speckman sürecinin şapka matrisini elde etmek için tahminlenen değerler vektörü kullanılsın. Bu durumda,

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{S}_\lambda\mathbf{y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\hat{\boldsymbol{\beta}}$$
$$= \mathbf{S}_\lambda\mathbf{y} + (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$$

$$\hat{\boldsymbol{\mu}} = \left[\mathbf{S}_\lambda + \tilde{\mathbf{X}}\left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{S}_\lambda)\right]\mathbf{y} \tag{3.20}$$

olacaktır. Burada $\mathbf{H} = \mathbf{S}_\lambda + \tilde{\mathbf{X}}\left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{S}_\lambda)$ *şapka matristir*.

Not: t_i, $i = 1, 2, ..., n$ düğümleri için $a < t_1 < ... < t_n < b$ koşulu sağlandığında (3.5) regresyon denkleminin *direkt* ve S*peckman* metotlarıyla tahmincileri uygun olarak aşağıdaki formüllerle hesaplanır.

a) Direkt metot için

$$\boldsymbol{\beta}_d = \left\{\tilde{\mathbf{X}}^T\mathbf{X}\right\}^{-1}\tilde{\mathbf{X}}^T\mathbf{y}$$
$$\mathbf{f}_d = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_d) \tag{3.21}$$
$$\boldsymbol{\mu}_d = \mathbf{X}\boldsymbol{\beta}_d + \mathbf{f}_d = \mathbf{H_d}\mathbf{y}$$
$$\mathbf{H_d} = \left[\mathbf{S}_\lambda + \tilde{\mathbf{X}}\left\{\tilde{\mathbf{X}}^T\mathbf{X}\right\}^{-1}\tilde{\mathbf{X}}^T\right]$$

b) Speckman metodu için

$$\boldsymbol{\beta}_s = \left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y}$$
$$\mathbf{f}_s = \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_s) \tag{3.22}$$
$$\boldsymbol{\mu}_s = \mathbf{X}\boldsymbol{\beta}_s + \mathbf{f}_s = \mathbf{H_s}\mathbf{y}$$
$$\mathbf{H_s} = \left[\mathbf{S}_\lambda + \tilde{\mathbf{X}}\left\{\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right\}^{-1}\tilde{\mathbf{X}}^T(\mathbf{I} - \mathbf{S}_\lambda)\right]$$

(3.21) ve (3.22) formüllerinde $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}$'dir. Burada $\mathbf{H_d}$ ve $\mathbf{H_s}$ matrisleri uygun metotların şapka (hat) matrisleridir.

3.3. Splayn Düzeltme ile Toplamsal Regresyon Modelleri

Toplamsal (additive) model,

$$y = \sum_{j=1}^{p} f_j(t_j) + \varepsilon \tag{3.23}$$

şeklinde tanımlanır. Burada $t_1, t_2, ..., t_p$ açıklayıcı değişkenler, y ise bağımlı değişkendir. ε hatası t_j'lerden bağımsız olduğunda $E(\varepsilon) = 0,\ Var(\varepsilon) = \sigma^2$ koşullarını sağlamaktadır. f_j uygun t_j değişkenlerinin pürüzsüz bir fonksiyonudur.

Gözlem değerleri $(y_i, t_{i1}, ..., t_{ip}), i = 1, 2, ..., n$ olduğunda, (3.23) modeli aşağıdaki yazılabilir.

$$y_i = \sum_{j=1}^{p} f_j(t_{ij}) + \varepsilon_i, \quad i = 1, 2, ..., n \tag{3.24a}$$

veya vektör şeklinde,

$$\mathbf{y} = \sum_{j=1}^{p} \mathbf{f}_j + \boldsymbol{\varepsilon}. \tag{3.24b}$$

(3.24b) denkleminde $\mathbf{f}_j = \left(f_j(t_{1j}), ..., f_j(t_{nj}) \right)^T$ n-boyutlu bilinmeyen kestirici vektördür. $\mathbf{y} = \left(y_1, y_2, ..., y_n \right)^T$ ve $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n \right)^T$'dir.

Toplamsal modelin önemli özelliklerinden biri, bir açıklayıcı değişkenin yanıt yüzeyine etkisinin diğer değişkenlerden bağımsız olmasıdır. Pratikte bu p sayıda ayrılmış fonksiyonların yardımıyla, kestirici değişkenlerin yanıta etkisi incelenebilir.

Toplamsal modeller Friedman ve Stuetzle (1981) tarafından ileri sürülen *izdüşüm takip regresyonu* (PPR-*projection pursuit regression*) modelinin özel bir halidir.

Toplamsal modeldeki tahmin edilen fonksiyonlar doğrusal regresyondaki β katsayılarına benzemektedir. Doğrusal modellerde karşıya çıkabilen tüm zorluklarla (tuzaklarla) ilgili yorumlar toplamsal modellere de uygulanabilir.

3.3.1. Toplamsal modeller için tahmin denklemleri

(3.23) toplamsal regresyon modelinin tahmini için splayn düzeltme ile pürüzlülük ceza yaklaşımı uygulandığında, ikinci mertebeden sürekli türeve sahip fonksiyonlar uzayında, verilmiş $\lambda_j > 0$ düzeltme parametreleri ile

$$\sum_{i=1}^{n}\left\{y_j-\sum_{j=1}^{p}f_j(t_{ij})\right\}^2+\sum_{i=1}^{p}\lambda_j\int_a^b\left\{f_j''(t)\right\}^2dt \tag{3.25}$$

genelleştirilmiş cezalı hata kareler toplamını minimum yapan $\hat{f}_j, j=1,2,...,p$ kestirici fonksiyonlarını bulmak gerekir.

Teorem3.1'e göre, (3.25) ifadesine minimum değeri veren f_j fonksiyonlarının kübik splayn olduğu kanaatine varılır. Bu durumda (3.25) cezalı hata kareler toplamı aşağıdaki kare form şeklinde yazılabilir:

$$\left(\mathbf{y}-\sum_{j=1}^{p}\mathbf{f}_j\right)^T\left(\mathbf{y}-\sum_{j=1}^{p}\mathbf{f}_j\right)+\sum_{j=1}^{p}\lambda_j\mathbf{f}_j{}^T\mathbf{K}_j\mathbf{f}_j \tag{3.26}$$

Burada $\mathbf{K}_j$, $\hat{f}_j$ kestiricisine uygun *ceza matrisidir* ve tek bir kestirici durumundaki **K** matrisine benzer olarak tanımlanır (Hastie ve Tibshirani, 1999). (3.26) ifadesi $\mathbf{f}_j,\ j=1,2,...,p$ vektörlerine göre bir kare formdur. Bu ifadenin $\mathbf{f}_j$'lere göre türevini sıfıra eşitleyerek,

$$\lambda_j\mathbf{K}_j\mathbf{f}_j-\left(\mathbf{y}-\sum_j\mathbf{f}_j\right)=\mathbf{0},\quad j=1,2,...,p \tag{3.27}$$

denklemler sistemi elde edilir. (3.27) denklemler sistemi, *tahmin denklemleri* olarak adlandırılır ve aşağıdaki şekilde yazılabilir:

$$\begin{pmatrix}\mathbf{I}&\mathbf{S}_1&\mathbf{S}_1&\cdots&\mathbf{S}_1\\\mathbf{S}_2&\mathbf{I}&\mathbf{S}_2&\cdots&\mathbf{S}_2\\\vdots&\vdots&\vdots&\ddots&\vdots\\\mathbf{S}_p&\mathbf{S}_p&\mathbf{S}_p&\cdots&\mathbf{S}_p\end{pmatrix}\begin{pmatrix}\mathbf{f}_1\\\mathbf{f}_2\\\vdots\\\mathbf{f}_p\end{pmatrix}=\begin{pmatrix}\mathbf{S}_1\mathbf{y}\\\mathbf{S}_2\mathbf{y}\\\vdots\\\mathbf{S}_p\mathbf{y}\end{pmatrix} \tag{3.28}$$

Burada $\mathbf{S}_j=\mathbf{S}_{\lambda_j}=\left(\mathbf{I}+\lambda_j\mathbf{K}_j\right)^{-1}$, $j=1,2,...,p$ uygun *j. kestiricinin düzeltme matrisidir.* (3.28) denklemler sistemi doğrusal regresyonda *normal denklemeler sisteminin* benzeridir, burada $\boldsymbol{\beta}$ katsayıları yerine (3.28)'i sağlayan $\mathbf{f}_j, j=1,2,...,p$ vektörleri aranır. (3.28) sistemi kısa şekilde,

$$\hat{\mathbf{P}}\mathbf{f}=\hat{\mathbf{Q}}\mathbf{y} \tag{3.29}$$

olarak yazılabilir. (3.28) ve (3.29), $(np\times np)$-boyutlu doğrusal denklemler sistemidir. Büyük veri seti ve çok sayıda açıklayıcı değişkenin bulunması

durumunda, (3.29) denklemler sisteminin geleneksel direkt metotlarla çözümü zorlaşabildiğinden, daha elverişli iteratif çözüm yöntemleri önerilir.

(3.27) eşitliği kullanılarak, (3.28) denklemler sistemi

$$\hat{\mathbf{f}}_k = \mathbf{S}_k \left(\mathbf{y} - \sum_{j \neq k} \hat{\mathbf{f}}_j \right), \quad k = 1, 2, ..., p \tag{3.30}$$

şeklinde yazılabilir. (3.30) ifadesi doğrusal denklemler sisteminin çözümü için kullanılan Gauss-Seidel sürecinin $\mathbf{f}_1, \mathbf{f}_2, ... \mathbf{f}_p$ vektörlerinin tahmini amacıyla, (3.28) blok sistemine uygulanmasına imkan sağlar.

***Backfitting veya Gauss-Seidel algoritması*:**

Tanımla: $\mathbf{f}_i = \mathbf{f}_i^0$, $i = 1, 2, ..., p$

Döngü:

$$\begin{aligned} &j = 1, 2, ..., p \\ &\mathbf{f}_j = \mathbf{S}_j \left(\mathbf{y} - \sum_{k \neq j} \mathbf{f}_k \right) \end{aligned} \tag{3.31}$$

Koşul: Bireysel fonksiyonlar (vektörler) değişmeyinceye kadar döngüye devam et.

(3.31) backfitting algoritmasının yakınsaklığı önemli problemlerden biridir (Buja ve ark., 1989). Bu konuya açıklık getirmeden önce, çalışmanın bu aşamasında, iki kestiriciye sahip toplamsal modeller ele alınmıştır.

3.3.2. İki düzeltici ile backfitting algoritması

Bir çok uygulamada iki nonparametrik açıklayıcı değişken içeren toplamsal modellerin incelenmesi gerekir. Diğer taraftan, iki düzeltici içeren regresyon modeli için backfitting algoritması ile ilgili incelemelerin sonuçları genelde p sayıda düzeltici için de genişletilebilir.

(3.23) regresyon denklemlerinde $p = 2$ olduğu durumda (3.30) tahmin denklemleri,

$$\begin{aligned} \mathbf{f}_1 &= \mathbf{S}_1 \left(\mathbf{y} - \mathbf{f}_2 \right) \\ \mathbf{f}_2 &= \mathbf{S}_2 \left(\mathbf{y} - \mathbf{f}_1 \right) \end{aligned} \tag{3.32}$$

şekline dönüşür. Bu denklemler,

$$(\mathbf{I}-\mathbf{S}_1\mathbf{S}_2)\mathbf{f}_1=\mathbf{S}_1(\mathbf{I}-\mathbf{S}_2)\mathbf{y}$$
$$(\mathbf{I}-\mathbf{S}_2\mathbf{S}_1)\mathbf{f}_2=\mathbf{S}_2(\mathbf{I}-\mathbf{S}_1)\mathbf{y}$$

şeklinde yazılarak, uygun ters matrislerin varlığı durumunda,

$$\begin{aligned}\mathbf{f}_1&=(\mathbf{I}-\mathbf{S}_1\mathbf{S}_2)^{-1}\mathbf{S}_1(\mathbf{I}-\mathbf{S}_2)\mathbf{y}\\ \mathbf{f}_2&=(\mathbf{I}-\mathbf{S}_2\mathbf{S}_1)^{-1}\mathbf{S}_2(\mathbf{I}-\mathbf{S}_1)\mathbf{y}\end{aligned} \tag{3.33a}$$

gibi tek çözüme sahip olurlar. Ters matrislerin varlığı için $\|\mathbf{S}_1\mathbf{S}_2\|<1$ koşulu ortak bir koşuldur (Buja ve ark., 1989). (3.33a) ifadelerinin,

$$\begin{aligned}\mathbf{f}_1&=\left\{\mathbf{I}-(\mathbf{I}-\mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I}-\mathbf{S}_1)\right\}\mathbf{y}\\ \mathbf{f}_2&=\left\{\mathbf{I}-(\mathbf{I}-\mathbf{S}_2\mathbf{S}_1)^{-1}(\mathbf{I}-\mathbf{S}_2)\right\}\mathbf{y}\end{aligned} \tag{3.33b}$$

ifadelerine denk olduğu doğrudan kanıtlanabilir.

$\mathbf{f}_1^0,\mathbf{f}_2^0$ başlangıç vektörler olduğunda, (3.32) sistemine uygun backfitting aşağıdaki süreçle gerçekleştirilir:

$$\begin{aligned}\mathbf{f}_1^{(m)}&=\mathbf{S}_1\left(\mathbf{y}-\mathbf{f}_2^{(m-1)}\right)\\ \mathbf{f}_2^{(m)}&=\mathbf{S}_2\left(\mathbf{y}-\mathbf{f}_1^{(m)}\right)\end{aligned} \tag{3.34a}$$

Burada $\mathbf{f}_1^{(m)}$ ve $\mathbf{f}_2^{(m)}$, algoritmanın m. adımındaki tahmin vektörleridir. Tümevarım metoduyla $m\geq 1$ durumunda (3.34a) formüllerinden,

$$\begin{aligned}\mathbf{f}_1^{(m)}&=\mathbf{y}-\sum_{j=0}^{m-1}(\mathbf{S}_1\mathbf{S}_2)^j(\mathbf{I}-\mathbf{S}_1)\mathbf{y}-(\mathbf{S}_1\mathbf{S}_2)^{m-1}\mathbf{S}_1\mathbf{f}_2^{(0)}\\ \mathbf{f}_2^{(m)}&=\mathbf{S}_2\sum_{j=0}^{m-1}(\mathbf{S}_1\mathbf{S}_2)^j(\mathbf{I}-\mathbf{S}_1)\mathbf{y}+\mathbf{S}_2(\mathbf{S}_1\mathbf{S}_2)^{m-1}\mathbf{S}_1\mathbf{f}_2^{(0)}\end{aligned} \tag{3.34b}$$

olduğu görülebilir. $\|\mathbf{S}_1\mathbf{S}_2\|<1$ koşulu sağlandığında $\mathbf{f}_1^{(m)}$, $\mathbf{f}_2^{(m)}$ yakınsak olur ve bu durumda (3.34b) formüllerinde $m\to\infty$ koşuluyla limite geçerek aşağıdaki eşitlikler bulunur:

$$\begin{aligned}\mathbf{f}_1^{(\infty)}&=\left\{\mathbf{I}-(\mathbf{I}-\mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I}-\mathbf{S}_1)\right\}\mathbf{y}\\ \mathbf{f}_2^{(\infty)}&=\mathbf{S}_2(\mathbf{I}-\mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I}-\mathbf{S}_1)\mathbf{y}\\ &=\left\{\mathbf{I}-(\mathbf{I}-\mathbf{S}_2\mathbf{S}_1)^{-1}(\mathbf{I}-\mathbf{S}_2)\right\}\mathbf{y}\end{aligned} \tag{3.35}$$

(3.33b) ve (3.35) ifadelerinin aynı olduğu görülmektedir. (3.35) ifadesini kullanarak, $p=2$ durumunda (3.23) regresyon denklemi için,

$$E(\mathbf{y}) = \hat{\mathbf{y}} = \mathbf{f}_1^{(\infty)} + \mathbf{f}_2^{(\infty)} = \left\{\mathbf{I} - (\mathbf{I} - \mathbf{S}_2)(\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1)\right\}\mathbf{y} \tag{3.36}$$

olduğu sonucuna varılır. Burada

$$\mathbf{H} = \mathbf{I} - (\mathbf{I} - \mathbf{S}_2)(\mathbf{I} - \mathbf{S}_1\mathbf{S}_2)^{-1}(\mathbf{I} - \mathbf{S}_1),$$

(3.23) regresyon denklemine uygun *şapka (hat) matristir* ve bu matris $\mathbf{S}_1$ ve $\mathbf{S}_2$'ye göre simetriktir.

Böylece eğer $\|\mathbf{S}_1\mathbf{S}_2\| < 1$ koşulu sağlanıyorsa, (3.32) tahmin denklemleri sistemi *tutarlıdır*, *çözüm tektir* ve (3.34) *backfitting algoritması çözüme yakınsamaktadır.*

İki bağımlı $\mathbf{t}_1$ ve $\mathbf{t}_2$ gözlem vektörlü splayn düzeltme ile bir backfitting algoritması göz önüne alınsın. Eğer veriler katı kollineerlik sergiliyorsa ($\mathbf{t}_2 = c\mathbf{t}_1$ şeklinde ise) $\|\mathbf{S}_1\mathbf{S}_2\| = 1$ olur. Genel *p*-bağımlı $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p$ veri vektörleri için $\|\mathbf{S}_1\mathbf{S}_2\| = 1$ koşulu doğrusal bağımlılık örneğidir. $\mathbf{S}_1$ ve $\mathbf{S}_2$ matrisleri simetrikse ve özdeğerleri $(-1,1]$ aralığında yer alıyorsa, (3.32) tahmin denklemleri *tutarlıdır* ve (3.34) *backfitting algoritması yakınsaktır.* Bu durumda $\mathbf{f}_1^{(\infty)} + \mathbf{f}_2^{(\infty)}$ toplamı başlangıç $\mathbf{f}_2^{(0)}$ noktasından bağımsız olarak tek değer alır, ancak $\mathbf{f}_1^{(\infty)}$ ve $\mathbf{f}_2^{(\infty)}$ ayrı ayrı farklı değerler alabilir (Buja ve ark., 1989). Böylece eğer $\mathbf{S}_1$ ve $\mathbf{S}_2$ matrisleri, özdeğerleri $(-1,1]$ aralığında yer alan simetrik matrisler ise, (3.32) tahmin denklemleri en az bir çözüme sahiptir ve (3.34a) backfitting algoritması bu çözümlerden birine yakınsar. Bu çözüm başlangıç $\mathbf{f}_2^{(0)}$ vektörüne bağımlıdır.

Not: Kübik splayn düzelticiler simetriktir ve özdeğerleri $(0,1]$ aralığında yer alır. Bu nedenle de kübik splayn düzeltmede tahmin denklemleri tutarlıdır ve backfitting algoritması yakınsaktır. Eğer $\|\mathbf{S}_1\mathbf{S}_2\| < 1$ ise çözüm tektir ve $\mathbf{f}_2^{(0)}$ başlangıç noktasından bağımsız olarak backfitting algoritması bu tek çözüme yakınsar. Eğer $\|\mathbf{S}_1\mathbf{S}_2\| = 1$ ise çözüm vardır ve backfitting algoritması $\mathbf{f}_2^{(0)}$ başlangıç noktasına bağımlı olarak bir çözüme yakınsar.

Not: Doğrusal regresyonda $\mathbf{X}$ matrisi singular olduğunda (tam sütun ranklı olmadığında), $\hat{\boldsymbol{\beta}}$ tahmin vektörü tek olmasa bile $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ için tahmin tektir.

$\hat{\mathbf{y}} = \mathbf{f}_1^{(\infty)} + \mathbf{f}_2^{(\infty)}$ tahmininin tek olması durumu doğrusal regresyonun bu özelliğine benzerdir.

3.3.3. Çözümün varlığı ve tekliği, Backfitting algoritmasının yakınsaması

Bu aşamada, genel *p*-düzeltici durumuna uygun (3.27) veya (3.28) normal tahmin denklemler sisteminin çözümünün varlığı ve tekliği problemi incelenmiştir.

$\Re(\mathbf{S}_j)$ ile $\mathbf{S}_j$ matrisinin aldığı değer-vektörleri uzayı ve $\mathrm{M}_\alpha(\mathbf{S}_j)$ ile $\mathbf{S}_j$ matrisinin α özdeğerlerine uygun özvektörlerinin oluşturduğu uzay işaret edilsin.

(3.28) normal denklemler sisteminin tutarlı olduğunu göstermek için $\forall y \in R^n$ için (3.29)'u sağlayan $\mathbf{f}$ vektörünün varlığı kanıtlanmalıdır. Burada $\mathbf{f} = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_p)^T$, *np*-boyutlu vektördür.

Teorem 3.2 (Buja ve ark., 1989). Eğer $\mathbf{S}_j, j = 1, 2, ..., p$ düzeltici matrisleri, özdeğerleri $[0,1]$ parçasında yerleşen simetrik matrisler ise, (3.28) normal denklemleri $\forall y \in R^n$ gözlemi için tutarlıdır.

İspat: (3.28) sisteminin $\mathbf{f}$ çözümü, (3.26) cezalı hata kareler toplamının minimum noktasıdır. Teoremde verilen koşullar sağlandığında herhangi bir $\mathbf{f}_j \in \Re(\mathbf{S}_j)$ için $\mathbf{f}_j^T \mathbf{K}_j \mathbf{f}_j \geq 0$ olur. Buna göre de (3.26) kare formu, $\mathbf{f}_j \in \Re(\mathbf{S}_j)$ olduğunda herhangi bir $\mathbf{f}$ vektörü için negatif olmayan değerler alır. Kare form aşağıdan sınırlı olduğundan, onun en az bir minimum noktası ve buna göre de (3.28) sisteminin en az bir çözümü vardır.■

Not: Simetrik $\mathbf{S}_j$ matrislerinin özdeğerleri $[0,1)$ aralığında yerleşirse, (3.28) normal denklemlerinin çözümleri aşağıdaki gibi yazılabilir (Breiman ve Friedman, 1985).

$$\mathbf{f}_j = \left(\mathbf{I} - \mathbf{S}_j\right)^{-1} \mathbf{S}_j \left(I + \sum_j \left(\mathbf{I} - \mathbf{S}_j\right)^{-1} \mathbf{S}_j \right)^{-1} \mathbf{y} \qquad (3.37)$$

Teorem 3.2'nin koşulları sağlandığında (3.28) tahmin denklemler sisteminin en az bir çözümü vardır. Bu durumda, çözümün hangi koşullarda tek veya birden çok olduğu sorusu ortaya çıkar.

(3.29) sistemi göz önüne alınsın. $\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}, \mathbf{g} \neq \mathbf{0}$ koşulunu sağlayan $\mathbf{g} \in R^{n \times p}$ vektörünün var olduğu varsayılsın. Bu durumda $\hat{\mathbf{P}}\mathbf{f} = \hat{\mathbf{Q}}\mathbf{y}$ şeklinde olan (3.29) denklemlerinin sonsuz sayıda çözümü olacaktır, çünkü $\left\{\mathbf{f}_j^{(\infty)} : j = 1,..., p\right\}$ (3.29) için çözüm ise $\left\{\mathbf{f}_j^{(\infty)} + c\mathbf{g}_j : j = 1,..., p\right\}$ de herhangi bir "c" için çözüm olacaktır.

$\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}$ koşulunu sağlayan $\mathbf{g}$ vektörlerinin oluşturduğu uzaya (3.28) sisteminin *mutabıklık uzayı* (concurvity space) denir.

İki düzeltici (iki nonparametrik değişken) durumunda $\|\mathbf{S}_1 \mathbf{S}_2\| < 1$ ve $\|\mathbf{S}_2 \mathbf{S}_1\| < 1$ koşulları ancak ve ancak mutabıklık uzayı boş olduğunda sağlanır. Böylece mutabıklık, backfitting algoritmasının davranışında önemli rol oynar.

$\mathbf{S}_j$, $j = 1, 2,..., p$ matrisleri, özdeğerleri $[0,1]$ aralığında yerleşen simetrik matrisler olsun. $M_1(\mathbf{S}_j)$, $\mathbf{S}_j$, $j = 1, 2,..., p$ matrisinin 1 özdeğerine uygun özvektörlerin oluşturduğu uzay olsun. Bu durumda ancak ve ancak $\mathbf{g}_j \in M_1(\mathbf{S}_j)$, $j = 1, 2,..., p$ ve $\mathbf{g}_+ = \mathbf{g}_1 + ... + \mathbf{g}_p = \mathbf{0}$ koşulları sağlandığında $\hat{\mathbf{P}}\mathbf{g} = \mathbf{0}$ olur (Buja ve ark., 1989). Diğer bir ifadeyle, mutabıklık (concurvity) ancak ve ancak $M_1(\mathbf{S}_j)$, $j = 1, 2,..., p$ uzayları doğrusal bağımlı olduğunda ortaya çıkabilir. Diğer bir deyişle, hiç olmazsa biri sıfırdan farklı öyle $\mathbf{g}_j \in M_1(\mathbf{S}_j)$, $j = 1, 2,..., p$ vardır ki, $\mathbf{g}_+ = \mathbf{g}_1 + ... + \mathbf{g}_p = \mathbf{0}$'dır. Verilmiş böyle bir doğrusal yozlaşma, (3.29) sisteminin herhangi $\mathbf{f}_1,...,\mathbf{f}_p$ çözümü için $\mathbf{f}_1 + c\mathbf{g}_1,...,\mathbf{f}_p + c\mathbf{g}_p$ şeklinde ek çözümlerinin ortaya çıkmasını sağlar.

Böylece, mutabıklık (concurvity) ancak 1 özdeğerine uygun özvektörler uzayının fonksiyonlarını içerir. Kübik splayn düzelticileri için bu özuzaylar her bir kestiricinin doğrusal fonksiyonuna (vektöre) uygundur ve bu durumda kesin mutabıklık (concurvity) yalnız kestiriciler kesin kollineer olduğunda ortaya çıkar.

p-düzeltici durumuna uygun (3.31) backfitting algoritmasının yakınsaması problemi ele alınsın. Concurvity tanımı bu soruna cevap vermekte yardımcı olur. Özdeğerleri $[0,1]$ aralığında yerleşen simetrik $\mathbf{S}_j,\ j=1,2,...,p$ düzelticileri ile (3.31) backfitting algoritması incelensin.

Eğer $\mathbf{S}_j$ düzelticileri için concurvity uzayı boş ise bu durumda (3.31) algoritması (3.28) denkleminin bir tek çözümüne yakınsar. Eğer concurvity durumu var ise backfitting (3.28) denkleminin çözümlerinden birine (başlangıç noktaya bağlı olarak) yakınsar.

3.3.2 ve **3.3.3** alt bölümlerinden sonuç olarak şunlar özetlenebilir:

I. İki $\mathbf{S}_1$ ve $\mathbf{S}_2$ düzelticisi için

a) Eğer $\|\mathbf{S}_1\mathbf{S}_2\|<1$ ise bu durumda (3.32) tahmin denklemleri sistemi bir tek çözüme sahiptir ve (3.34a) backfitting algoritması bu tek çözüme yakınsar (herhangi bir başlangıç nokta için).

b) Eğer $\mathbf{S}_1$ ve $\mathbf{S}_2$ özdeğerleri $(-1,1]$ aralığında yerleşen simetrik matrisler ise (3.32) denklemlerinin en az bir çözümü vardır ve backfitting algoritması bu çözümlerden birine yakınsar. Çözüm başlangıç $\mathbf{f}_2^{(0)}$ noktasına bağımlıdır.

II. $\mathbf{S}_1,...,\mathbf{S}_p$ simetrik ve özdeğerleri $[0,1]$ parçasında yerleşen düzelticiler için,

a) Herhangi bir $\mathbf{y}\in R^n$ için (3.28) tahmin denklemleri en az bir çözüme sahiptir.

b) Aşağıdaki her bir koşul ayrılıkta tahmin denklemlerinin concurvity uzayını verir:

b$_1$) $\hat{\mathbf{P}}\mathbf{g}=\mathbf{0},\ \mathbf{g}\neq\mathbf{0}$

b$_2$) $\exists\mathbf{g}_j\in M_1(\mathbf{S}_j),\ j=1,2,...,p,\ \mathbf{g}_+=\mathbf{g}_1+...+\mathbf{g}_p=\mathbf{0}$

c) Eğer concurvity uzayı boş ise (3.28) tahmin denklemlerinin bir tek çözümü vardır ve backfitting bu çözüme yakınsar.

d) Eğer concurvity uzayı boş değilse, (3.28) tahmin denklemlerinin en az bir çözümü vardır ve backfitting bu çözümden birine yakınsar.

Semiparametrik Toplamsal Regresyon Modeli

(3.24) toplamsal regresyon modelinde terimlerden biri doğrusal olduğunda, bu modele *semiparametrik toplamsal model* denir (Green ve Silverman, 1994):

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \sum_{j=2}^{p} \mathbf{f}_j(t_{ij}) + \varepsilon_i, \quad i = 1,2,...,n \tag{3.38a}$$

veya

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{j=2}^{p} \mathbf{f}_j + \boldsymbol{\varepsilon} \tag{3.38b}$$

Bu durumda, (3.28) normal denkleminde yer alan $\mathbf{S}_1$ düzeltme matrisi $\mathbf{f}_1 = \mathbf{X}\boldsymbol{\beta}$ fonksiyonuna uygun olur ve

$$\mathbf{S}_1 = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{3.39}$$

şeklinde hesaplanır. (3.30) formülüne göre, $k = 1$ olduğunda,

$$\hat{\mathbf{f}}_1 = \mathbf{S}_1\left(\mathbf{y} - \sum_{j=2}^{p} \hat{\mathbf{f}}_j\right)$$

yazılabilir. Böylece (3.31) backfitting algoritması, (3.38) denklemleri için de aynı derecede uygulanabilir, sadece $\mathbf{S}_1$ hesaplandığında (3.39) formülünü dikkate almak gerekir.

Düzeltilmiş Backfitting Algoritması

Çoğu düzelticiler, izdüşüm (projection) ve çekme (shrinking) gibi iki kısma sahiptirler. Örneğin kübik splayn düzeltme kestiricileri, sabit ve doğrusal fonksiyonların uygun birim özdeğerlerine sahiptir (izdüşüm kısmı), diğer özvektörler ise birden küçük özdeğerlere uygundur (shrinking). Temel fikir, tüm kestirimler için izdüşüm işlemlerini bir büyük izdüşüm işleminde birleştirmek ve iteratif backfitting tipli işlemlerde düzelticinin izdüşüm olmayan kısmını kullanmaktır.

Bu düzeltmenin birkaç avantajı vardır. Splayn düzeltme uygulandığında, bazı kestiriciler arasında korelasyon olabilir. Tüm izdüşümleri bir izdüşüm işleminde birleştirmekle tüm fonksiyon eğrileri benzer olarak tahmin edilebilir.

$\mathbf{G}_j$ matrisi $M_1(\mathbf{S}_j)$ üzerine ortogonal izdüşüm yapan matris olsun, burada $M_1(\mathbf{S}_j)$, *j*. $\mathbf{S}_j$ düzelticinin 1 özdeğerine uygun özvektörleri uzayıdır. Düzeltilmiş $\mathbf{S}_j$ matrisi,

$$\tilde{\mathbf{S}}_j = \left(\mathbf{I} - \mathbf{G}_j\right)$$

şeklinde tanımlanır. $\tilde{\mathbf{S}}_j$ matrisi $M_1(\mathbf{S}_j)$ uzayı üzerindeki düzeltme değeri bileşenini çıkarma etkisine sahiptir.

Algoritma: (*Düzeltilmiş Backfitting Algoritması*)

1. $\tilde{\mathbf{f}}_1,...,\tilde{\mathbf{f}}_p$ başlangıç kestiricilerini tanımla ve $\tilde{\mathbf{f}}_+ = \tilde{\mathbf{f}}_1 + ... + \tilde{\mathbf{f}}_p$ olarak hesapla.
2. $\mathbf{G}$, R^n'de $M_1(\mathbf{S}_1) + ... + M_1(\mathbf{S}_p)$ üzerine ortogonal (dik) izdüşüm olduğunda, $\mathbf{g} = \mathbf{G}\left(\mathbf{y} - \tilde{\mathbf{f}}_+\right)$ vektörünü hesapla.
3. $\tilde{\mathbf{S}}_i$ düzelticilerini kullanarak $(\mathbf{y} - \mathbf{g})$ için backfitting tekrarlamalarını (devirlerini) yap; Bu adımda toplamsal $\tilde{\mathbf{f}}_+ = \tilde{\mathbf{f}}_1 + ... + \tilde{\mathbf{f}}_p$ güncellenir.
4. **2.** ve **3.** adımları yakınsama olana kadar tekrar et. Toplam uyum için son tahmin $\mathbf{f}_+ = \mathbf{g} + \tilde{\mathbf{f}}_+$ olarak alınır.

Düzeltilmiş backfitting algoritmasının yakınsaması ile ilgili aşağıdaki koşullar sağlanır (Hastie ve Tibshirani, 1999):

a. $\mathbf{S}_j$, $j = 1,..., p$ düzelticileri simetrik ise ve özdeğerleri $[0,1]$ aralığında yer alıyorsa, $\mathbf{g}$ ve $\tilde{\mathbf{f}}_1,...,\tilde{\mathbf{f}}_p$ vektörlerinin yakınsaması anlamında, düzeltilmiş backfitting algoritması yakınsaktır.

b. $\tilde{\mathbf{S}}_i$ düzelticileri ile düzeltilmiş backfitting algoritmasının $\tilde{\mathbf{f}}_j$ ve $\mathbf{g}_j \in M_1(\mathbf{S}_j)$ fonksiyonları bakımından yakınsak olduğu varsayılsın. Bu durumda $\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$ bileşenleri $\mathbf{S}_j^* = \mathbf{G}_j + \left(\mathbf{I} - \mathbf{G}_j\right)\mathbf{S}_j$ düzelticilerine uygun tahmin denklemlerinin çözümüdür.

c. Eğer $\mathbf{S}_j$ düzelticileri simetrik ise $\mathbf{S}_j^* = \mathbf{S}_j$ olur ve düzeltilmiş algoritmanın çözümü $\mathbf{S}_j$ düzelticilerine uygun tahmin denklemlerinin çözümü olur.

d. Eğer $\mathbf{S}_j$ simetrik ve özdeğerleri $[0,1]$ parçasında ise, bu durumda $\tilde{\mathbf{S}}_j = \mathbf{S}_j - \mathbf{G}_j$ ve $\left\|\tilde{\mathbf{S}}_j\right\| < 1$'dir. Splayn düzelticiler bu özelliğe sahiptirler ve splayn düzeltme durumunda düzeltilmiş backfitting algoritması her zaman yakınsaktır.

e. Tüm açıklayıcı değişkenler için kübik splayn kullanıldığında **G** matrisi, $\left(\mathbf{1}, \mathbf{x}_1, ..., \mathbf{x}_p\right)$ üzerine en küçük kareler regresyonunun şapka (hat) matrisidir.

f. Yakınsama sonucunda $\mathbf{g} = \sum_{j=1}^{p} \mathbf{g}_j$ ayrışımı yapılırsa, $\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$ final bileşenleri bulunabilir. $\mathbf{S}_j$ kübik splayn düzeltici ise ve **y** başlangıçta verilmişse, bu durumda $\mathbf{g}_j = \hat{\beta}_j \mathbf{x}_j$ olur. Burada $\hat{\beta}_1, ..., \hat{\beta}_p$, $\mathbf{x}_1, ..., \mathbf{x}_p$ üzerinde $\mathbf{y} - \tilde{\mathbf{f}}_+$'nin çoklu doğrusal regresyonundan elde edilen katsayılardır.

g. $\tilde{\mathbf{A}}_j = \left(\mathbf{I} - \tilde{\mathbf{S}}_j\right)^{-1} \tilde{\mathbf{S}}_j$, $\tilde{\mathbf{A}} = \sum_{j=1}^{p} \tilde{\mathbf{A}}_j$ ve $\mathbf{B} = \left(\mathbf{I} + \tilde{\mathbf{A}}\right)^{-1} \tilde{\mathbf{A}}$ olduğunda (3.28) tahmin denklemlerinin çözümü aşağıdaki gibi yazılabilir.

$$\tilde{\mathbf{f}}_+ = \left(\mathbf{I} - \mathbf{BG}\right)^{-1} \mathbf{B}\left(\mathbf{I} - \mathbf{G}\right)\mathbf{y} \text{ ve } \mathbf{g} = \mathbf{G}\left(\mathbf{y} - \tilde{\mathbf{f}}_+\right)$$

Bu formüller aşağıdaki gibi birleştirilebilir:

$$\mathbf{f}_+ = \left\{\mathbf{G} + \left(\mathbf{I} - \mathbf{G}\right)\left(\mathbf{I} - \mathbf{BG}\right)^{-1} \mathbf{B}\left(\mathbf{I} - \mathbf{G}\right)\right\}\mathbf{y}$$

$$\tilde{\mathbf{f}}_j = \left(\mathbf{I} - \tilde{\mathbf{S}}_j\right)^{-1}\left(\mathbf{y} - \mathbf{g} - \tilde{\mathbf{f}}_+\right) \qquad (3.40)$$

$$\mathbf{f}_j = \mathbf{g}_j + \tilde{\mathbf{f}}_j$$

Burada $\mathbf{g}_j \in M_1(\mathbf{S}_j)$ ve $\sum_{j=1}^{p} \mathbf{g}_j = \mathbf{g}$'dir.

3.4. Düzeltme Parametresinin Seçimi, Serbestlik Derecesi

Yukarıda ele alınan, pürüzlülük ceza yaklaşımı uygulanan regresyon modellerinde, iyi (optimum) düzeltme parametresinin seçimi önemli konulardan biridir. Eğer λ çok büyük seçilirse veriler gerekenin üstünde düzeltilmiş olur, buna

karşın λ çok küçük seçilirse düzeltme gerçek durumun altında olur. λ parametresinin optimum seçimi $\hat{f}$ tahmin fonksiyonunun, gerçek *f* fonksiyonuna mümkün olduğu kadar yakın olmasını sağlar.

Bu alt bölümde, nonparametrik regresyonda düzeltme parametresinin seçim kriterleri ve farklı serbestlik derecesi kavramları verilmiştir.

3.4.1. Düzeltme parametresinin seçimi

(3.1) nonparametrik regresyon modeli için düzeltme parametresinin seçim kriterlerinden biri olan çapraz geçerlilik (Cross Validation–CV) fonksiyonu,

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}_\lambda^{-i}(t_i)\right\}^2 \tag{3.41a}$$

olarak tanımlanır. Burada $\hat{f}_\lambda^{-i}$, *i.* y_i gözlemi çıkarıldığında (3.1) modeli ile elde edilen tahmin fonksiyonudur. (3.41a) formülü ile tanımlanan çapraz geçerlilik skoru şöyle hesaplanabilir:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(t_i)}{1 - S_{ii}(\lambda)}\right\}^2 \tag{3.41b}$$

Burada $\hat{f}_\lambda(.)$, tüm $(t_i, y_i), i = 1,2,...,n$ gözlemleri ve λ parametresine uygun splayn düzeltici, $S_{ii}(\lambda)$ ise (3.3) ifadesindeki $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$ düzeltme (şapka) matrisinin *i.* köşegen elemanıdır. Uygun λ düzeltme parametresi, (3.41b) fonksiyonunun minimumu probleminden belirlenir.

Genelleştirilmiş çapraz geçerlilik (Generalized Cross Validation–GCV) ise çapraz geçerlilik fonksiyonunun düzeltilmiş halidir ve düzeltme parametresinin seçiminde çok yaygın olarak kullanılan bir metottur. İlk olara Craven ve Wahba'nın 1979 yılında yayımlanmış olan makalelerinde yer almıştır. GCV skoru yapılanmasında temel fikir, CV skorundaki S_{ii} yerine, onların ortalaması olan $\frac{1}{n}tr(\mathbf{S}_\lambda)$ ifadesinin yazılmasıdır. Böylece GCV(λ) fonksiyonu şöyle tanımlanır:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(t_i)}{1 - \frac{1}{n}tr(\mathbf{S}_\lambda)}\right\}^2 \tag{3.42}$$

(3.42)'de ifade edilen GCV fonksiyonuna minimum değer veren λ, uygun bir parametre olarak seçilir.

Eğer tüm $S_{ii}(\lambda)$ köşegen elemanları eşit ise (örneğin t_i noktaları eşit aralıklarla yerleşmiş ise) bu durumda GCV ve CV aynı olur. Genelde ise bu iki metot farklı sonuçlar verir ve GCV skoru, düzeltme parametresinin seçiminde daha makul görünür (Aydın, 2005).

(3.42) şeklinde ifade edilen GCV skor fonksiyonu aşağıdaki şekilde getirilebilir (Green ve Silverman, 1994):

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-S_{ii}(\lambda)}{1-\frac{1}{n}tr(\mathbf{S}_\lambda)}\right)^2\left\{y_i - \hat{f}^{-i}(t_i)\right\}^2$$

Regresyon kaynaklarında, şapka matrisin köşegen elemanlarına ($S_{ii}(\lambda), i = 1,...,n$) kuvvet (leverage) değerleri denir. S_{ii}, t_i noktasındaki y_i gözlem değerinin $\hat{f}(t_i)$ tahmin değerine olan etki miktarını belirler. Son formül GCV'nin CV skorunun büyük kuvvet (leverage) değerine sahip noktalardan çıkarılan artıkların biraz düşük ağırlıklı hali olduğunu gösterir.

Mallows'un C_p kriteri de düzeltme parametresinin seçimi için kullanılabilir ve aşağıdaki gibi hesaplanır:

$$C_p(\lambda) = \frac{1}{n}\left\|\left(\mathbf{S}_\lambda - \mathbf{I}\right)\mathbf{y}\right\|^2 + 2\sigma^2 tr(\mathbf{S}_\lambda) + \sigma^2 \qquad (3.43)$$

(3.43) formülünde σ^2 bilinmediğinde, onun farklı $\hat{\sigma}^2$ tahminleri kullanılabilir (Gasser ve ark., 1986). Örneğin,

$$\hat{\sigma}^2_{\lambda_p} = \frac{\left\|\left(\mathbf{S}_{\lambda_p} - \mathbf{I}\right)\mathbf{y}\right\|^2}{tr\left(\mathbf{1} - \mathbf{S}_{\lambda_p}\right)}$$

Burada pilot (yönetici) λ_p değeri önceden CV veya diğer bir kriterden belirlenebilir.

(Hurvich ve ark.,1998) makalesinde klasik AIC (Akaike Information Criteri) kriterinin düzeltilmiş versiyonu olan düzeltilmiş Akaike bilgi kriteri AIC_C, λ düzeltme parametresinin seçimi için tasarlanmıştır. Bu kriter aşağıdaki gibi hesaplanır:

$$AIC_C(\lambda) = \log \frac{\left\|(\mathbf{S}_\lambda - \mathbf{I})\mathbf{y}\right\|^2}{n} + 1 + \frac{2\{tg(\mathbf{S}_\lambda)+1\}}{n - tr(\mathbf{S}_\lambda) - 2} \tag{3.44}$$

Yukarıda olduğu gibi, λ parametresi (3.44) fonksiyonunun minimizasyonu probleminden belirlenir.

3.4.2. Serbestlik derecesi

Klasik regresyonda serbestlik derecesi, modeldeki etkili parametre sayısı ile belirlenir. Nonparametrik regresyonda tanımlanan serbestlik derecesi kavramları doğrusal regresyona benzer olarak verilebilir. Bu açıdan üç farklı serbestlik derecesi tanımı verilebilir (Buja ve ark., 1989).

a. Doğrusal regresyon için $\sum_i \text{var}(\hat{y}_i) = p\sigma^2$'dir ve p-parametre sayısı serbestlik derecesidir. Splayn düzeltmede ise $\text{cov}\,\hat{\mathbf{y}} = \mathbf{S}\mathbf{S}^T\sigma^2$ olduğu için, serbestlik derecesi benzer olarak aşağıdaki gibi tanımlanabilir:

Tanım3.1. Düzelticisi $\mathbf{S}_\lambda$ olan nonparametrik regresyon modelinin serbestlik derecesi,

$$df_1(\lambda) = tr\left(\mathbf{S}_\lambda \mathbf{S}_\lambda^{\,T}\right) \tag{3.45}$$

sayısına denir.

b. Nonparametrik model için $\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}$ olduğunda artık kareler toplamının beklenen değeri,

$$E\left[(\mathbf{y} - \mathbf{S}_\lambda \mathbf{y})^T (\mathbf{y} - \mathbf{S}_\lambda \mathbf{y})\right] = \left[n - tr\left(2\mathbf{S}_\lambda - \mathbf{S}_\lambda^{\,T}\mathbf{S}_\lambda\right)\right]\sigma^2 - \mathbf{f}^T(\mathbf{I} - \mathbf{S}_\lambda)^T(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{f}$$

olur. Burada son terim sapmayı gösterir. Doğrusal regresyon durumunda bu ifadedeki birinci terim $(n - p)$'ye uygundur. Buna göre de serbestlik derecesi şöyle tanımlanabilir:

Tanım3.2. Düzelticisi $\mathbf{S}_\lambda$ olan nonparametrik regresyon modelinin serbestlik derecesi,

$$df_2(\lambda) = tr\left(2\mathbf{S}_\lambda - \mathbf{S}_\lambda \mathbf{S}_\lambda^{\,T}\right) \tag{3.46}$$

sayısına denir.

c. $tr(\mathbf{S}_\lambda)$, ortalama artık kareler (average squared error-ASR) için C_p düzeltmesi olarak motive edilebilir (kullanılabilir). Örneğin C_p istatistiği ASR'ye $\frac{\hat{\sigma}^2}{n} 2tr(\mathbf{S}_\lambda)$ eklenerek düzeltilebilir. Burada $\hat{\sigma}^2$, σ^2'nin tahminidir. Bu, doğrusal regresyonda hata kareler ortalamasına (MSE) $2p\hat{\sigma}^2$ eklenmesine benzerdir. Bu bakımdan $tr(\mathbf{S}_\lambda)$ doğrusal regresyonda *p*-parametre sayısına karşılık gelir.

Tanım3.3. Düzelticisi $\mathbf{S}_\lambda$ olan nonparametrik regresyon modelinin serbestlik derecesi,

$$df_3(\lambda) = tr\left(\mathbf{S}_\lambda\right) \tag{3.47}$$

sayısına denir.

Splayn düzeltmede serbestlik derecesinin (3.47) tanımı çok popülerdir (Green ve Yandell, 1985; O'Sullivan ve ark., 1986; Silverman, 1985).

Not:

1. **S** simetrik izdüşüm matrisi ise, $tr(\mathbf{S}_\lambda)$, $tr\left(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^T\right)$ ve $tr\left(\mathbf{S}_\lambda\mathbf{S}_\lambda^T\right)$ serbestlik dereceleri üst üste düşer. Bu durum, doğrusal ve polinomial regresyon düzeltmede ve regresyon splaynda ortaya çıkar.
2. **S** splayn düzeltici ise,
   - **a.** $tr\left(\mathbf{S}_\lambda\mathbf{S}_\lambda^T\right) \le tr(\mathbf{S}_\lambda) \le tr\left(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda^T\right)$
   - **b.** (3.45), (3.46) ve (3.47) fonksiyonlarının her üçü de λ parametresinin artan fonksiyonlarıdır.

Yukarıda tanımlanan serbestlik derecelerinden herhangi birisi düzeltme parametresinin değerinin belirlenmesinde uygulanabilir. Bu, otomatik parametre seçimi yapılabilir olmadığında, düzelticiler sınıfında ayarlanmış düzeltme parametreleri arasından seçim için makul bir metot sağlar. Hesaplama açısından, $\mathbf{S}_\lambda$ matrisinin köşegen elemanlarının toplamı olan $df_3(\lambda) = tr\left(\mathbf{S}_\lambda\right)$ avantaja sahiptir (Buja ve ark., 1989).

4. KÜBİK REGRESYON SPLAYNI İLE PÜRÜZLÜLÜK CEZA YAKLAŞIMI

Splayn düzeltmede regresyon modelinin yapılanması için elde olan gözlem verilerindeki tüm $t_i,\ i=1,2,...,n$ noktaları, aranan splayn fonksiyonu için düğüm noktaları olarak kullanılır. Diğer bir ifadeyle, splayn fonksiyonu taban fonksiyonlarla ifade edildiğinde bu taban fonksiyonların sayısı n'den az olmamalıdır. İncelenen problemde çok fazla veri sayısı istenmediği durumlarda, splayn düzeltme ile regresyonda aşırı zorluklar ortaya çıkmaz. Fakat veri sayısı çok olan nonparametrik modellerde veya çok değişken içeren toplamsal modellerde bir dizi sayısal zorluk ortaya çıkabilir.

Bu bölümde *regresyonda pürüzlülük ceza yaklaşımı* için, taban fonksiyonları kullanılarak *regresyon splaynı* modelleri ele alınmıştır. Bu amaçla önce, genel olarak, regresyonda tahmin fonksiyonunun *taban fonksiyonların* sonlu doğrusal kombinasyonu şeklinde aranması tekniği incelenerek, sonsuz boyutlu minimizasyon probleminin sonlu boyutlu probleme (uygun bir kare formun minimizasyonu problemine) dönüşümü açıklanmıştır. Sonra ise kübik splayn taban fonksiyonları kullanılarak, *kübik regresyon splaynı* modelleri incelenmiştir. Parametrik olmayan regresyon için kullanılan bu teknik bir sonraki aşamada *toplamsal* ve *genelleştirilmiş toplamsal* regresyon modellerine uygulanmıştır. Son olarak, bir veya birkaç değişkenin pürüzsüz fonksiyonu olarak tahmin edilen *ince tabakalı splayn* regresyon modeli incelenmiştir.

4.1. Cezalı Regresyon Splaynı ile Nonparametrik Regresyon

Cezalı regresyon splaynı fikrinin ortaya koyulduğu ilk makaleler olarak Wahba, (1980) ve Parker ve Rice, (1985)'ın çalışmaları örnek gösterilebilir. Splayn fonksiyonu taban fonksiyonların yardımıyla ifade edildiğinde, açıklayıcı değişkenler için gerçek veri seti yerine, splaynda çok az düğüm noktasının kullanımı, hesaplamaların önemli ölçüde kolaylaşmasına ve modelin tahmin açısından daha da esnek olmasına neden olur (Wood, 2002).

(3.1) nonparametrik regresyon modeli ele alınsın:

$$y_i = f(t_i) + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2),\ i = 1,2,...,n$$
$$\text{veya}\ \mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0,\sigma^2\mathbf{I}) \tag{4.1}$$

$t_1, t_2, ..., t_n$ noktalarının $a < t_1 < t_2 < ... < t_n < b$ koşulunu sağladığı varsayılır. $y_1, y_2, ..., y_n$ bu düğüm noktalarındaki gözlem değerleridir ($n \geq 3$).

Bölüm 3'de, (3.1) regresyon modeli için (3.2) cezalı en küçük kareler toplamının minimum değerini bir *doğal kübik splaynda* aldığı ifade edilmişti. Bu bölümde taban fonksiyonları kullanılarak doğal kübik splayn için alternatif bir gösterim incelenmektedir.

4.1.1. Taban Fonksiyonlar Yardımıyla Modelleme

Bir değişkenli $f(t)$ fonksiyonu aşağıdaki gibi aransın:

$$f(t) = \sum_{j=1}^{m} \alpha_j b_j(t) \tag{4.2}$$

Burada $\{b_j(t) : j = 1,2,...,m\}$ uygun özelliklere sahip *taban fonksiyonları* kümesidir. (4.1) modelinde $y_i \sim N(f(t_i), \sigma^2)$'dir ve n-sayıda ($i = 1,2,...,n$) gözlem değerleri mevcuttur.

(4.1) regresyon modeli için (3.2) cezalı kareler toplamı dikkate alınsın:

$$\sum_{i=1}^{n} \{y_i - f(t_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2\, dt \tag{4.3}$$

Burada λ, düzeltme parametresidir. $\hat{f}(t)$ tahmin fonksiyonu, *ikinci mertebeden sürekli türeve sahip fonksiyonlar uzayında* (4.3) ifadesine minimum değerini veren bir fonksiyondur ve bu fonksiyonun *kübik splayn* olduğu bilinmektedir (Reinsch, 1967). (4.3) ifadesi aşağıdaki gibi de yazılabilir:

$$\|\mathbf{X}\boldsymbol{\alpha} - \mathbf{y}\|^2 + \lambda \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} \tag{4.4}$$

Burada, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)^T$; $\mathbf{X} = \begin{bmatrix} \mathbf{b}(t_1)^T \\ \vdots \\ \mathbf{b}(t_n)^T \end{bmatrix}$; $\mathbf{b}(t_i) = [b_1(t_i), ..., b_n(t_i)]^T$ ve

$\mathbf{y} = (y_1, ..., y_n)^T$'dir. (4.4) ifadesindeki $\mathbf{H}$ matrisi, $\mathbf{J}(f) = \int [f''(t)]^2 dt$ ceza terimi hesaplanarak açıklanabilir. $f''(t)$ ve $[f''(t)]^2$ ifadeleri açık yazılsın:

$$f''(t) = \sum_{j=1}^{m} \alpha_j b_j''(t) = \mathbf{b}''(t)^T \boldsymbol{\alpha}$$

$$\left[f''(t)\right]^2 = \left(\mathbf{b}''(t)^T \boldsymbol{\alpha}\right)^T \mathbf{b}''(t)^T \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{b}''(t)^T \mathbf{b}''(t) \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{S}(t) \boldsymbol{\alpha}$$

Burada $s_{ij}(t) = b_i''(t) b_j''(t),\ i, j = 1, 2, ..., m$'dir ve $\mathbf{S}(t) = (s_{ij}(t))$, $m \times m$ boyutlu bir matristir. Bu durumda ceza terimi aşağıdaki gibi ifade edilebilir:

$$\mathbf{J}(f) = \boldsymbol{\alpha}^T \left(\int \mathbf{S}(t) dt\right) \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha} \tag{4.5}$$

Burada $\mathbf{H} = \int \mathbf{S}(t) dt$, $m \times m$ boyutlu bir *ceza matrisidir.*

Böylece (4.3) ifadesinin minimizasyonu problemi, (4.4) kare formunun minimizasyonu haline gelir. (4.4) ifadesinin minimum noktası

$$\hat{\boldsymbol{\alpha}} = \left[\mathbf{X}^T \mathbf{X} + \lambda \mathbf{H}\right]^{-1} \mathbf{X}^T \mathbf{y} \tag{4.6}$$

olarak bulunur. (4.2) ifadesine dayanarak, $\hat{\mathbf{f}}(t) = \hat{\boldsymbol{\alpha}}^T \mathbf{b}(t)$ şeklinde tahmin edilir.

4.1.2. Kübik Splayn Taban

En popüler taban fonksiyonlarından biri *kübik splayn taban*dır. Bu durumda da farklı alternatif kübik splayn tabanları vardır (Gu, 2002). Konunun iyi anlaşılması için, burada kübik splayn tabanlardan en basit olanı kullanılmıştır. $\left\{x_j^* : j = 1, 2, ..., m\right\}$, $\hat{f}(t)$ splayn fonksiyonunun düğüm noktaları kümesi olsun. $b_j(t) = \left|t - x_j^*\right|^3$, $j = 1, 2, ..., m$; $b_{m+1}(t) = 1$, $b_{m+2}(t) = t$ şeklinde ifade edilen $(m+2)$ sayıda taban fonksiyonu için,

$$f(t) = \sum_{j=1}^{m+2} \alpha_j b_j(t) \tag{4.7}$$

bir *kübik splayn fonksiyonunu* oluşturur. $f(t)$'in *doğal kübik splayn* olması için,

$$\sum_{j=1}^{m} \alpha_j = 0, \sum_{j=1}^{m} \alpha_j x_j^* = 0 \tag{4.8a}$$

koşullarının sağlanması gerekir. (4.8a) koşulu, $\mathbf{C}\boldsymbol{\alpha} = \mathbf{0}$ şeklinde de ifade edilebilen bir koşuldur ve bu koşul açık olarak (4.8b)'de verilmiştir.

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ x_1^* & x_2^* & x_3^* & \cdots & x_m^* & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{m+2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{4.8b}$$

Burada, $\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ x_1^* & x_2^* & x_3^* & \cdots & x_m^* & 0 & 0 \end{bmatrix}$, $(2 \times m+2)$ boyutlu bir matristir. Bu durumda, *doğal kübik splayn* durumunda, (4.4) denklemine (4.8) koşulları eklenmiş olur.

Genel olarak, (4.4) denklemine eklenen (4.8) koşulundaki **C** matrisinin $(q \times m)$ $(q < m)$ boyutlu bir matris olduğu varsayılsın:

$$\mathbf{C}\boldsymbol{\alpha} = \mathbf{0}. \tag{4.9}$$

(4.8b) ifadesinde $q = 2$'dir, m ise m+2'dir. Şimdi **C** matrisi için **QR** ayrışımı yapılsın ($\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$) (Golub ve ark., 1996):

$$\mathbf{C}\mathbf{Q} = \begin{bmatrix} \mathbf{0}_{q,m-q} & \mathbf{T}_{q,q} \end{bmatrix} \text{ veya } \mathbf{Q}^T\mathbf{C}^T = \begin{bmatrix} \mathbf{0} \\ \mathbf{T} \end{bmatrix}. \tag{4.10}$$

Burada **T** matrisi $(q \times q)$ boyutlu üst üçgen bir matristir. **Q** matrisi, **Z** matrisi $m \times (m-q)$ boyutlu, **Y** ise $(m \times q)$ boyutlu matrisler olduğunda, $\mathbf{Q} = [\mathbf{Z} \quad \mathbf{Y}]$ şeklinde yazılabilir. (4.9) ve (4.10)'a göre,

$$\mathbf{C}\mathbf{Q} = \mathbf{C}[\mathbf{Z} \quad \mathbf{Y}] = [\mathbf{C}\mathbf{Z} \quad \mathbf{C}\mathbf{Y}] = [\mathbf{0} \quad \mathbf{T}]$$

ve buradan,

$$\mathbf{C}\mathbf{Z} = \mathbf{0},\ \mathbf{C}\mathbf{Y} = \mathbf{T} \tag{4.11}$$

olarak alınır. $(m-q)$ boyutlu $\boldsymbol{\alpha}_Z$ vektörü, $\boldsymbol{\alpha} = \mathbf{Z}\boldsymbol{\alpha}_Z$ eşitliğini sağlayan herhangi bir vektör ise,

$$\mathbf{C}\boldsymbol{\alpha} = (\mathbf{C}\mathbf{Z})\boldsymbol{\alpha}_Z = \mathbf{0} \tag{4.12}$$

olur. Bu durumda (4.9) koşullu (4.4) minimizasyon problemi, koşulsuz problem olarak aşağıdaki şekilde yazılabilir.

$$\min_{\boldsymbol{\alpha}_Z} \left\| \mathbf{X}\mathbf{Z}\boldsymbol{\alpha}_Z - \mathbf{y} \right\|^2 + \lambda \boldsymbol{\alpha}_Z^T \mathbf{Z}^T \mathbf{H} \mathbf{Z} \boldsymbol{\alpha}_Z \tag{4.13}$$

(4.6) formülünde **X** matrisi **XZ** ile ve **H** matrisi $\mathbf{Z}^T\mathbf{H}\mathbf{Z}$ ile değiştirilerek,

$$\hat{\boldsymbol{\alpha}}_Z = \left[\mathbf{Z}^T\mathbf{X}^T\mathbf{X}\mathbf{Z} + \lambda\mathbf{Z}^T\mathbf{H}\mathbf{Z}\right]^{-1}\mathbf{Z}^T\mathbf{X}^T\mathbf{y} \tag{4.14}$$

şeklinde bulunabilir ve bu durumda, $\hat{\boldsymbol{\alpha}} = \mathbf{Z}\boldsymbol{\alpha}_Z$ olarak belirlenir (Wood ve Augustin, 2002).

4.2. Regresyon Splayn ile Toplamsal ve Genelleştirilmiş Toplamsal Regresyon Modelleri

Bu aşamada, *cezalı regresyon splaynı* uygulanarak toplamsal ve genelleştirilmiş toplamsal modellerin yapılanması incelenmiştir.

4.2.1. Toplamsal Modeller

Kolaylık için önce *iki düzeltici terime* sahip olan toplamsal model ele alınsın:

$$y_i = \beta_0 + f_1(t_{1i}) + f_2(t_{2i}) + \varepsilon_i,\ \ \varepsilon_i \sim N(0,\sigma^2) \tag{4.15a}$$

Burada $(y_i, t_{1i}, t_{2i}), i = 1,2,...,n$ gözlem değerleridir. (4.15a) modeli,

$$E(y_i) = \mu_i = \beta_0 + f_1(t_{1i}) + f_2(t_{2i}),\ \ y_i \sim N(\mu_i,\sigma^2) \tag{4.15b}$$

olarak ifade edilebilir. Bu denklemde f_1 ve f_2, cezalı kübik regresyon splaynı ile aranan düzeltme fonksiyonlarıdır ve (4.16) denklemi ile verilmiştir.

$$f_1(t) = \sum_{j=1}^{q_1}\beta_j b_{1j}(t),\quad f_2(t) = \sum_{j=1}^{q_2}\beta_{j+q_1} b_{2j}(t) \tag{4.16}$$

(4.16) ifadesindeki $b_{1j}(\cdot)$ ve $b_{2j}(\cdot)$ taban fonksiyonlarının, sırasıyla f_1 ve f_2 için *kübik splayn taban fonksiyonları* olduğu varsayılır. (4.15) modeli için cezalı hata kareler toplamı aşağıdaki gibi ifade edilir:

$$\sum_{i=1}^{n}(y_i - \beta_0 - f_1(t_{1i}) - f_2(t_{2i}))^2 + \lambda_1\int[f_1''(t_1)]dt_1 + \lambda_2\int[f_2''(t_2)]dt_2\,. \tag{4.17}$$

Bu aşamada problem, (4.17) ifadesinin minimize edilmesi ve doğal kübik splayn olarak f_1 ve f_2 düzeltici fonksiyonlarının bulunmasıdır. Bölüm 4.1'de ifade edilen tasarımı göz önünde bulundurarak, (4.17) problemi aşağıdaki gibi ifade edilebilir.

$$\min_{\substack{\boldsymbol{\beta}\\ \mathbf{C}\boldsymbol{\beta}=\mathbf{0}}} \left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right\|^2 + \lambda_1\boldsymbol{\beta}^T\mathbf{H}_1\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}^T\mathbf{H}_2\boldsymbol{\beta}\,. \tag{4.18}$$

Burada $\mathbf{X} = \begin{bmatrix} \mathbf{b}_1(t_1)^T & \mathbf{b}_2(t_1)^T \\ \vdots & \vdots \\ \mathbf{b}_1(t_n)^T & \mathbf{b}_2(t_n)^T \end{bmatrix}$, $n \times (q_1 + q_2 + 1)$ boyutlu bir matris olduğunda,

$\mathbf{b}_k(t_i) = [b_{k1}(t_i), ..., b_{kn}(t_i)]^T$, $k = 1, 2$; $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_{q_1}, ..., \beta_{q_1+q_2}]^T$; $\mathbf{H}_i$, $i = 1, 2$ matrisleri $(q_1 + q_2 + 1) \times (q_1 + q_2 + 1)$ boyutlu kare matrislerdir ve $\mathbf{C}$ matrisi $6 \times (q_1 + q_2 + 1)$ boyutlu bir matristir (Wood ve Augustin, 2002). $\mathbf{C}$ matrisinin ilk dört satırı *doğal kübik splayn* koşulları, son iki satırı ise *modelin kesin tanımlanması* koşullarıdır:

$$\sum_{i=1}^{n} f_1(x_{1i}) = 0, \quad \sum_{i=1}^{n} f_2(x_{2i}) = 0 \tag{4.19}$$

λ_1 ve λ_2 düzeltme parametreleri, her iki terim için serbestlik derecesinin etkisini belirlemede, özel yöntemlerle bulunur (Wood, 2004). Bu konuya sonraki bölümlerde değinilecektir.

Verilmiş λ_1 ve λ_2 düzeltme parametreleri için (4.18) ifadesindeki kare form $\mathbf{S} = \lambda_1 \boldsymbol{\beta}^T \mathbf{H}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{H}_2 \boldsymbol{\beta}$ olarak tanımlanarak, aşağıdaki şekilde yazılabilir:

$$\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 + \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta} = \left\| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \boldsymbol{\beta} \right\|^2$$

Burada $\mathbf{B}$ matrisi $\mathbf{B}^T \mathbf{B} = \mathbf{S}$ eşitliliğini sağlayan *kök* matristir. Tek düzeltici durumuna benzer olarak, son ifadenin sağ kısmının $\boldsymbol{\beta}$'ya göre minimizasyonu sıradan en küçük kareler problemidir ve standart doğrusal regresyon modeli gibi çözülebilir. Sonuç tahmin $\hat{\boldsymbol{\beta}} = \mathbf{B}\mathbf{y}$'dir. Burada $\mathbf{B}$,

$$\mathbf{B} = \left(\mathbf{X}^T \mathbf{X} + \lambda_1 \boldsymbol{\beta}^T \mathbf{H}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{H}_2 \boldsymbol{\beta} \right)^{-1} \mathbf{X}^T$$

şeklinde ifade edilebilir. Bu durumda, $\boldsymbol{\mu} = \mathbf{A}\mathbf{y}$ ve $\mathbf{A}$ *şapka* (*hat) matrisi*,

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^T \mathbf{X} + \lambda_1 \boldsymbol{\beta}^T \mathbf{H}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{H}_2 \boldsymbol{\beta} \right)^{-1} \mathbf{X}^T .$$

4.2.2. Genelleştirilmiş Toplamsal Modeller (GAM)

(4.15) toplamsal modeli *genelleştirilmiş toplamsal modele* şöyle genişletilebilir:

$$g(\mu) = \mathbf{X}\boldsymbol{\beta}, \; Y_i \sim \text{üstel aile} \tag{4.20}$$

Burada **X**, Bölüm 4.2.1' de yapılanan toplamsal modelin *tasarım matrisi*, g ise pürüzsüz monoton *"link"* fonksiyonudur. (4.15), (4.20) *genelleştirilmiş toplamsal modelini* oluşturur. Düzeltici f_i fonksiyonlarının kestirimi için *cezalı log-olabilirlik* (log-likelihood) yöntemi kullanılabilir:

$$l_p(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \frac{1}{2}\left(\lambda_1 \boldsymbol{\beta}^T \mathbf{H}_1 \boldsymbol{\beta} + \lambda_2 \boldsymbol{\beta}^T \mathbf{H}_2 \boldsymbol{\beta}\right) \to \min_{\boldsymbol{\beta}} \tag{4.21a}$$

(4.21a) formülünde $l(\boldsymbol{\beta})$ log-olabilirlik fonksiyonu $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)]$ olarak tanımlanır, burada $f_{\theta_i}(y_i)$ üstel aileye uygun bir olasılık yoğunluk fonksiyonudur (Bkz., Bölüm 2.3).

(4.21a) ifadesi kısa olarak aşağıdaki gibi yazılabilir.

$$l_p(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T \mathbf{H} \boldsymbol{\beta} \to \min_{\boldsymbol{\beta}} \tag{4.21b}$$

Burada $\mathbf{H} = \sum_j \lambda_j \mathbf{H}_j$ ve λ_j düzeltme parametrelerinin verildiği varsayılır. $l_p(\boldsymbol{\beta})$'ın minimizasyonu için β_j'lere uygun türevler sıfıra eşitlenir:

$$\frac{\partial l_p}{\partial \beta_j} = -\frac{\partial l}{\partial \beta_j} + \left[\mathbf{H}\boldsymbol{\beta}\right]_j = -\frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} + \left[\mathbf{H}\boldsymbol{\beta}\right]_j = 0 .$$

Bölüm2.3.'de anlatıldığı gibi, alınan denklem, $\mathrm{var}(y_i)$ teriminin hesaplandığı varsayılarak, aşağıdaki doğrusal olmayan cezalı en küçük kareler problemi olarak çözülebilir.

$$S_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mathrm{var}(y_i)} + \boldsymbol{\beta}^T \mathbf{H} \boldsymbol{\beta} .$$

Negatif cezalı log-olabilirliğin minimizasyonu problemi, ***IRLS*** *(Iteratively Re-weighted Least Squares)* algoritması yardımı ile gerçekleştirilebilir (Nelder ve Wedderburn, 1972). λ_i'ler verildiğinde, pratikte modelin tahmini için aşağıdaki cezalı IRLS (PIRLS) şeması kullanılabilir:

1. y_i'nin μ_i ortalamasına bağlı olan varyansı $V(\mu_i)$ olsun. Her bir k. adımda,

$$\mathbf{z}^{[k]} = \mathbf{X}\boldsymbol{\beta}^{[k]} + \Gamma^{[k]}\left(\mathbf{y} - \boldsymbol{\mu}\right)^{[k]} \tag{4.22}$$

yapay veriyi (pseudodata) hesaplayınız. Burada $\Gamma^{[k]}$, $\mathbf{\Gamma}_{ii}^{[k]} = g'(\mu_i^{[k]})$ olarak ifade edilen bir köşegen matristir.

2. Aşağıda tanımlanan köşegen **W** ağırlık matrisini hesaplayınız:

$$\mathbf{W}_{ii} = \left[\Gamma_{ii}^{[k]} \sqrt{V(\mu_i^{[k]})} \right]^{-1}.$$

3. Cezalı maksimum olabilirlik kestiriminin iteratif çözümünü ($\boldsymbol{\beta}^{[k+1]}$)

$$\left\| \mathbf{W}^{[k]} \left(\mathbf{X}\boldsymbol{\beta} - \mathbf{z}^{[k]} \right) \right\|^2 + \boldsymbol{\beta}^T \mathbf{H} \boldsymbol{\beta} \rightarrow \min_{\substack{\boldsymbol{\beta} \\ \mathbf{C}\boldsymbol{\beta}=\mathbf{0}}} \tag{4.23}$$

minimizasyon probleminden bulunuz (O'Sullivan ve ark., 1986; Wood ve Augustin, 2002).

Yaygın olarak g linki log fonksiyonudur, bu durumda $g'(\mu) = \mu^{-1}$ olur. Gamma dağılımı için ise $V(\mu_i) = \mu_i^2$' dir.

4.3. Serbestlik Derecesi ve Düzeltme Parametresinin Seçimi

Serbestlik derecesi ve artık varyans tahmini.

Düzeltme parametrelerinin ölçeklenmesi konusunda *serbestlik derecesi* önemli rol oynamaktadır. Genelleştirilmiş toplamsal modellerde (GAM) serbestlik derecesinin belirlenmesi için farklı fikirler geliştirilebilir. Düzeltme parametreleri sıfıra eşitlendiğinde serbestlik derecesinin, $\boldsymbol{\beta}$'nın boyutuna eşit olduğu açıktır. Diğer ekstrem durumda, düzeltme parametreleri çok büyük olduğunda, model aşırı esnektir ve bu nedenle de farklı serbestlik derecesine sahiptir. Kurulmuş esnek modelin serbestlik derecesinin ölçülmesi için bir yol, *etkin serbestlik derecesinin* $tr(\mathbf{A})$ olarak hesaplanmasıdır, burada **A** şapka matristir. Genelleştirilmiş toplamsal model için *şapka matrisi*,

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \sum_i \lambda_i \mathbf{H}_i \right)^{-1} \mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \mathbf{H} \right)^{-1} \mathbf{X}^T \tag{4.25}$$

olarak hesaplanır.

Serbestlik derecesinin belirlenmesi için bir diğer yaklaşım $\boldsymbol{\beta}$ vektörünün her bir bileşeni için cezayı göz önüne almaktır, diğer bir ifadeyle ayrı ayrı her bir düzeltme fonksiyonu için serbestlik derecesini kullanmaktır. Bu amaçla,

$\mathbf{P} = \left(\mathbf{X}^T\mathbf{X} + \mathbf{H}\right)^{-1}\mathbf{X}^T$ matrisi tanımlansın. Öyle ki, $\boldsymbol{\beta} = \mathbf{P}\mathbf{y}$ ve $tr(\mathbf{A}) = tr(\mathbf{XP})$'dir. $\mathbf{P}$'nin *i.* satırı hariç tutulup kalan satırları sıfırlanarak elde edilen matris $\mathbf{P}_{\mathrm{i}}^0$ olsun. Bu durumda $\mathbf{P}_{\mathrm{i}}^0\mathbf{y}$ vektörünün *i.* bileşeni β_i, kalan bileşenleri ise sıfır olur ve $tr(\mathbf{A}) = \sum_{i=1}^{p} tr(\mathbf{X}\mathbf{P}_{\mathrm{i}}^0)$'dır. Bu durumda, $tr(\mathbf{X}\mathbf{P}_{\mathrm{i}}^0)$, *i.* parametreye bağlı etkili serbestlik derecesi olarak düşünülebilir. Diğer taraftan, $tr(\mathbf{X}\mathbf{P}_{\mathrm{i}}^0) = (\mathbf{PX})_{\mathrm{ii}}$ olduğu için, model parametreleri için etkili serbestlik derecesi vektörü

$$\mathbf{F} = \mathbf{PX} = \left(\mathbf{X}^T\mathbf{X} + \mathbf{H}\right)^{-1}\mathbf{X}^T\mathbf{X}$$

matrisinin esas köşegenidir. Burada $tr(\mathbf{A}) = tr(\mathbf{F})$ olduğu da not edilmelidir. Ceza terimi olmadığında parametre tahmini $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ formülü ile verilir. Ceza teriminin varlığı durumunda ise tahmin vektörü $\hat{\boldsymbol{\beta}} = \mathbf{F}\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \mathbf{H}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ olarak verilebilir. Bu durumda $\mathbf{F}$ matrisi, cezasız parametre tahmincisini cezalı tahminciye dönüştüren matristir. $F_{ii} = \partial\hat{\beta}_i \big/ \partial\tilde{\beta}_i$ ifadesi, F_{ii} cezasız $\tilde{\beta}_t$ parametresinin birim değişimine uygun cezalı $\hat{\beta}_t$' nin değişim ölçüsünü göstermektedir. Bu durumda, F_{ii}'nin niçin $i.$ceza parametresi için etkili serbestlik derecesi ölçütü olduğu anlaşılır: Cezasız parametre 1 serbestlik derecesine sahiptir, ama cezalar F_{ii} faktörü yardımıyla serbestliği etkili olarak küçültür.

Özdeş link ve normal hatalar durumunda doğrusal regresyona benzer olarak, σ^2 aşağıdaki formülle tahmin edilebilir.

$$\hat{\sigma}^2 = \frac{\left\|\mathbf{y} - \mathbf{A}\mathbf{y}\right\|^2}{n - tr(\mathbf{A})}, \tag{4.26}$$

Burada $tr(\mathbf{A}) = tr(\mathbf{F})$'dir. (4.26) sapmasız tahmin değildir, çünkü

$$E\left(\left\|\mathbf{y} - \mathbf{A}\mathbf{y}\right\|^2\right) = \sigma^2\left[n - 2tr(\mathbf{A}) + tr(\mathbf{A}^T\mathbf{A})\right] + \mathbf{b}^T\mathbf{b}, \tag{4.27}$$

burada $\mathbf{b} = \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\mu}$ sapmayı tasvir etmektedir.

Genelleştirilmiş toplamsal model durumunda, ölçek parametresi (artık varyansı) çoğunlukla Pearson-benzerlik ölçek tamincisi yardımıyla tahmin edilir:

$$\hat{\phi} = \frac{\sum_i V(\hat{\mu}_i)^{-1}(y_i - \mu_i)^2}{n - tr(\mathbf{A})}.$$

Düzeltme parametresinin seçimi

GCV kullanılarak, λ düzeltme parametresinin seçimi için etkili bir yaklaşım verilebilir (Wood ve Augustin, 2002; Wood, 2004). GCV değeri,

$$V = \frac{n\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2}{\left[n - tr(\mathbf{A})\right]^2} \tag{4.28}$$

olarak tanımlanır. Burada **A** şapka (hat) matristir ($\hat{\boldsymbol{\mu}} = \mathbf{A}\mathbf{y}$ dönüşümünü yapan matristir). (4.28) denklemindeki $tr(\mathbf{A})$ iz değeri *modelin serbestlik derecesini* tahmin eder. Doğrusal regresyonda şapka matrisi $\mathbf{A} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ olarak tanımlanır ve $tr(\mathbf{A}) = p$ değeri doğrusal modelin serbestlik derecesini (modeldeki parametre sayısını) ifade eder. GCV değeri verilerin varyans tahmincisinin artıkların serbestlik derecesine orantısıdır. V, düzeltme parametrelerinin (λ'ların) bir fonksiyonudur ve en iyi $\boldsymbol{\lambda}^0 = \left(\lambda_1^0, ..., \lambda_k^0\right)$ vektörü, $V(\lambda)$ fonksiyonunun minimizasyonun probleminden bulunur. Wood (2004) makalesinde $\mathbf{X} = \mathbf{QR}$ ayrışımı tekniğini kullanarak, genelleştirilmiş toplamsal modelin düzeltme parametrelerinin bulunmasını kolaylaştıracak bir yaklaşım vermiştir.

4.4. İnce Tabakalı Splayn (Thin Plate Spline-TPS)

İnce tabakalı splayn (TPS) çok değişkenli düzeltme fonksiyonunun tahmini problemi için, gürültülü gözlemler verildiğinde, çok zarif ve genel bir çözüm yaklaşımıdır. TPS için teorik temeller Duchon (1975, 1976, 1977) ve Meinguet (1979)'in makalelerinde verilmiştir. Bir sonraki bulgu ve uygulamalar Wahba ve Wendelberger (1980), Hutchinson ve Bishof (1983) ve Seaman ve Hutchinson (1985)'un makaleleriyle devam etmiştir.

x d-boyutlu bir vektör olsun. n-sayıda $(y_i, \mathbf{x}_i), i = 1, 2, ..., n$ gözlemleri kullanılarak, $f(\mathbf{x})$ düzeltme fonksiyonunun tahmin edilmesi problemi ele alınsın.

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \tag{4.29}$$

ε_i, rassal hata terimidir.

4.4.1. İki Boyutlu İnce Tabakalı Splayn

Kolay anlaşılması bakımından, ilk olarak $d = 2$ kabul edilsin ($\mathbf{x} \in \Re^2$). Bir ince tabakalı splayn tanımlamadan önce, bazı temel kavramlara ihtiyaç vardır. Bu nedenle $r \in \Re$ için aşağıdaki gibi bir $\eta(r)$ fonksiyonu tanımlansın (Green ve Silverman, 1994):

$$\eta(r) = \begin{cases} \dfrac{1}{16\pi} r^2 \log r^2, & r > 0 \text{ise} \\ 0 \qquad\qquad , & \text{d.d.} \end{cases} \tag{4.30}$$

$\mathbf{x} = (t, z) \in \Re^2$ için aşağıdaki üç ϕ_j fonksiyon tanımlanır:

$$\begin{aligned} \phi_1(t, z) &= 1 \\ \phi_2(t, z) &= t \\ \phi_3(t, z) &= z \end{aligned} \tag{4.31}$$

Elemanları $T_{jk} = \phi_j(\mathbf{t}_k)$ olan $(3 \times n)$ boyutlu **T** matrisi aşağıdaki gibi tanımlansın:

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix} \tag{4.32}$$

Tanım 4.1. Uygun δ_i ve a_j sabitleri için $f(\mathbf{x})$, ancak ve ancak

$$f(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta\left(\left\|\mathbf{x} - \mathbf{x}_i\right\|\right) + \sum_{j=1}^{3} a_j \phi_j(\mathbf{x}) \tag{4.33}$$

formunda ise, $f(\mathbf{x})$ fonksiyonuna $\mathbf{x}_1, ..., \mathbf{x}_n$ noktalarında bir *ince tabakalı splayn (TPS)* denir. Ek olarak, $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ vektörü için,

$$\mathbf{T}\boldsymbol{\delta} = 0 \tag{4.34a}$$

koşulu sağlandığında *f*'ye *doğal ince tabakalı splayn* denir.

Not: (4.32)'e göre, (4.34a) aşağıdaki koşula denktir.

$$\sum_{i=1}^{n} \delta_i = 0, \quad \sum_{i=1}^{n} \delta_i \mathbf{x}_i = 0 \tag{4.34b}$$

$(n \times n)$ boyutlu bir **E** matrisi tanımlansın:

$$E_{ij} = \eta\left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right) = \frac{1}{16\pi}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \log\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \tag{4.35}$$

η'nın (4.30) tanımına göre $E_{ii} = 0, \quad i = 1, 2, ..., n$.

Bu aşamada, doğal ince tabakalı splaynın iki önemli özelliğini verilsin (Green ve Silverman, 1994).

Teorem 4.1 Aşağıda ifade edilen ceza terimi,

$$J(f)=\iint_{\Re^2}\left(\frac{\partial^2 f}{\partial t_1^2}+2\frac{\partial^2 f}{\partial t_1 \partial t_2}+\frac{\partial^2 f}{\partial t_2^2}\right)dt_1 dt_2 \quad \text{ceza terimi} \tag{4.36}$$

i. ancak ve ancak f doğal ince tabakalı splayn olduğunda sonlu olur,

ii. f doğal ince tabakalı splayn olduğunda

$$J(f)=\boldsymbol{\delta}^T\mathbf{E}\boldsymbol{\delta} \tag{4.37}$$

kare formuna dönüşür.

Teorem 4.2. $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n$, $\Re^2$ uzayında kollineer olmayan farklı noktalar ve $z_1,z_2,...,z_n$ verilen sayılar olduğunda,

$$f(\mathbf{x}_i)=z_i,\ i=1,2,...,n \tag{4.38}$$

koşullarını sağlayan ince tabakalı splayn vardır ve tektir.

Not: Teorem 4.2'nin ispatı sürecinde (Green ve Silverman, 1994) uygun TPS'nin,

$$\begin{bmatrix}\mathbf{E} & \mathbf{T}^T\\ \mathbf{T} & \mathbf{0}\end{bmatrix}\begin{pmatrix}\boldsymbol{\delta}\\ \mathbf{a}\end{pmatrix}=\begin{pmatrix}\mathbf{z}\\ \mathbf{0}\end{pmatrix} \tag{4.39}$$

denkleminin çözümünden belirlendiği görülmektedir. (4.39) eşitliğinin sol tarafındaki blok matris tam ranka sahiptir ve bu durumda (4.39) ifadesinin bir tek $\begin{pmatrix}\boldsymbol{\delta}_0\\ \mathbf{a}_0\end{pmatrix}$ çözümü vardır.

Şimdi verilmiş $(y_i,\mathbf{x}_i), i=1,2,...,n$ gözlem değerleri için cezalı en küçük karelerden, uygun TPS'nin bulunuşu açıklansın:

$$S(f)=\sum_i\{y_i-f(\mathbf{t}_i)\}^2+\lambda J(f)\to\min \tag{4.40}$$

(4.40) ifadesi, **T** ve **E** matrisleri ile ve (4.37) dikkate alınarak aşağıdaki gibi ifade edilebilir.

$$\min_{\substack{\boldsymbol{\delta},\mathbf{a}\\ \mathbf{T}\boldsymbol{\delta}=\mathbf{0}}} S(f)=\left(\mathbf{Y}-\mathbf{E}\boldsymbol{\delta}-\mathbf{T}^T\mathbf{a}\right)^T\left(\mathbf{Y}-\mathbf{E}\boldsymbol{\delta}-\mathbf{T}^T\mathbf{a}\right)+\lambda\boldsymbol{\delta}^T\mathbf{E}\boldsymbol{\delta} \tag{4.41}$$

(4.41) probleminin çözümü olan TPS'nin katsayıları,

$$\begin{bmatrix} \mathbf{E}+\lambda\mathbf{I} & \mathbf{T}^T \\ \mathbf{T} & \mathbf{0} \end{bmatrix}\begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{a}} \end{pmatrix}=\begin{pmatrix} \mathbf{Y} \\ \mathbf{0} \end{pmatrix} \tag{4.42}$$

denklemlerinden bulunur. (4.42) denklemleri tek bir çözüme sahiptir (sol kısımdaki blok matris tam ranklı olduğu için). (4.42) denklemleri

$$\begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{T} & -\lambda\mathbf{I} \end{bmatrix}$$

matrisiyle çarpılarak

$$\begin{bmatrix} \mathbf{E}^2+\lambda\mathbf{E} & \mathbf{E}\mathbf{T}^T \\ \mathbf{T}\mathbf{E} & \mathbf{T}\mathbf{T}^T \end{bmatrix}\begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{a}} \end{pmatrix}=\begin{pmatrix} \mathbf{E} \\ \mathbf{T} \end{pmatrix}\mathbf{Y}$$

denklemleri alınır, bu da $\hat{\boldsymbol{\delta}}$ ve $\hat{\mathbf{a}}$'nın (4.41) probleminin çözümü olduğunu gösterir. (4.33)'le tanımlanan $\hat{f}$ uygun *doğal ince tabakalı splayn* olur.

4.4.2. d Boyutlu TPS

x vektörü d-boyutlu olduğunda (4.29) regresyon problemi ele alınsın. Bu durumda f fonksiyonunun TPS tahmini problemi

$$S(f)=\|\mathbf{y}-\mathbf{f}\|+\lambda J_{md}(f) \tag{4.43}$$

ifadesine minimum değer veren $\hat{f}$ fonksiyonu olur, burada **y** bileşenleri y_i verileri olan vektör, $\mathbf{f}=\left[f(\mathrm{x}_1),...,f(\mathrm{x}_\mathrm{n})\right]^\mathrm{T}$'dir, $J_{md}(f)$ ceza fonksiyonu, λ ise düzeltme parametresidir. Ceza fonksiyonu aşağıdaki gibi tanımlanır:

$$J_{md}(f)=\int\!\!...\!\int_{R^d}\sum_{\nu_1+...+\nu_d}\frac{m!}{\nu_1!...\nu_d!}\left(\frac{\partial^m f}{\partial x_1^{\nu_1}...\partial x_d^{\nu_d}}\right)^2 dx_1...dx_d \tag{4.44}$$

(4.44)'de $2m>d$ koşulunun sağlandığı varsayılır (Wahba, 1990). $m=2$ ve $d=2$ durumunda (4.44) ifadesi (4.36) şekline dönüşür. (4.43) cezalı hata kareler toplamını minimize eden $\hat{f}$ fonksiyonu, $m=2$ ve $d=2$ durumundaki (4.33) fonksiyonuna benzer olarak,

$$\hat{f}(\mathbf{x})=\sum_{i=1}^{n}\delta_i\eta_{md}\left(\|\mathbf{x}-\mathbf{x}_i\|\right)+\sum_{j=1}^{M}a_j\phi_j(\mathbf{x}) \tag{4.45}$$

biçiminde verilir. (4.45) ifadesindeki bilinmeyen ve tahmin edilen kat sayılar vektörleri uygun olarak $\boldsymbol{\delta}$ ve $\boldsymbol{\alpha}$'dır. $\boldsymbol{\delta}$ için $\mathbf{T}\boldsymbol{\delta}=\mathbf{0}$ koşulu doğal kübik splayn

yaklaşımını belirtir, burada $T_{ij} = \phi_j(\mathbf{x}_i)$, $j = 1,...,M$; $M = \begin{pmatrix} m+d-1 \\ d \end{pmatrix}$ olur. ϕ_j, $j = 1,...,M$ fonksiyonları R^d uzayında derecesi m'yi aşmayan polinomları oluşturan lineer bağımsız (taban) polinomlardır. ϕ_j'lerin oluşturduğu polinomlarda, (4.44)'le tanımlanan J_{md} fonksiyonu sıfır değerini alır. Diğer bir ifadeyle, ϕ_j'lerin oluşturduğu alt uzay J_{md} fonksiyonelinin "sıfır uzayı" olur. $m = 2$ ve $d = 2$ durumunda $M = 3$ ve $\phi_1(\mathbf{x}) = 1$, $\phi_2(\mathbf{x}) = x_1$, $\phi_3(\mathbf{x}) = x_2$. (4.45)' de kalan taban fonksiyonlar aşağıdaki gibi tanımlanır.

$$\eta_{md}(r) = \begin{cases} \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & d \text{ tek ise} \\ \dfrac{(-1)m+1+d/2}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r) & d \text{ çift ise} \end{cases} \tag{4.46}$$

$m = 2$ ve $d = 2$ durumunda (4.46) fonksiyonu (4.30) fonksiyonuna dönüşür. (4.35)'e benzer olarak, elemanları $E_{ij} = \eta_{md}\left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right)$ olan $\mathbf{E}$ matrisi tanımlansın. Bu durumda (4.41)'e benzer olarak, ince tabakalı splayn tahmincisi problemi aşağıdaki gibi yazılabilir.

$$\min_{\substack{\boldsymbol{\delta},\boldsymbol{\alpha} \\ \mathbf{T}\boldsymbol{\delta}=\mathbf{0}}} S(f) = \left(\mathbf{Y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}^T\boldsymbol{\alpha}\right)^T \left(\mathbf{Y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}^T\boldsymbol{\alpha}\right) + \lambda\boldsymbol{\delta}^T\mathbf{E}\boldsymbol{\delta}\,. \tag{4.47}$$

TPS *bir taraftan ideal bir düzelticidir*: O öyle tasarlanır ki, verilere uyum ile pürüzlülük arasındaki ağırlık tam olarak bağdaşmış olur ve *bulunan fonksiyon* düzeltme hedefini en iyi biçimde sağlamış olur. TPS' nin tasarımında düğüm veya taban fonksiyon sayısının seçimi gibi zor problemler ortaya çıkmaz, parametreler doğal verilere göre belirlenir. Diğer taraftan TPS uygulandığında *hesaplama maliyeti* büyük olabilir, veri çok olduğunda düzelticilerin parametre sayısı da büyür ve model tahmini için işlem sayısı parametre sayısının küpüyle orantılı olur. Bu durumda yüksek maliyete neden olur. Bu nedenle, modelde çok az parametre sayısı kullanan regresyon splaynı tercih edilmektedir (özellikle büyük hacimli veri kümesi için).

4.4.3. İnce Tabakalı Regresyon Splayn (TPRS)

İnce tabakalı regresyon splaynda temel fikir, "sıfır eğimli" ($\boldsymbol{\alpha}$ katsayıları ile bağlı) bileşenler değişmeyinceye kadar, TPS' deki "eğim" ($\boldsymbol{\delta}$ katsayıları ile bağlı) bileşenleri (wiggly components) uzayının kısa kesilmesidir (Wood, 2003). Burada esas önem (dikkat), $\boldsymbol{\delta}$ parametreleri uzayı tabanına verilmektedir. İdeal tabanda verilmiş herhangi $\boldsymbol{\delta}$ için uyum ve ceza terimlerinin değişimi minimum olur.

$\mathbf{E}$ matrisinin öz-ayrışımı $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ olsun, burada $\mathbf{D}$ matrisi $\mathbf{E}$'nin sıralanmış özdeğerlerinden oluşan köşegen matris, $\mathbf{U}$ matrisi ise sütunları uygun özvektörler olan ortogonal matristir. $\mathbf{U}_k$, $\mathbf{U}$ matrisinin ilk k sütunundan oluşan matris, $\mathbf{D}_k$ ise $\mathbf{D}$ matrisinin $k \times k$ sol üst, alt matrisi olsun.

$\boldsymbol{\delta} = \mathbf{U}_k \boldsymbol{\delta}_k$ yazılarak $\boldsymbol{\delta}$, $\mathbf{U}_k$ matrisinin sütun uzayı ile kısıtlansın. Sonuç olarak (4.47) problemi aşağıdaki gibi yazılabilir.

$$\min_{\substack{\boldsymbol{\delta}_k, \boldsymbol{\alpha} \\ \mathbf{T}^T \mathbf{U}_k \boldsymbol{\delta}_k = \mathbf{0}}} S(f) = \left\| \mathbf{y} - \mathbf{U}_k \mathbf{D}_k \boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha} \right\|^2 + \lambda {\boldsymbol{\delta}_k}^T \mathbf{D}_k \boldsymbol{\delta}_k \tag{4.48}$$

Kısıttan kurtulmak için, 4.1.2 alt bölümündeki (4.9)-(4.14) işlemlerini göz önüne alınarak, önce $\mathbf{C} = \mathbf{T}^{\mathrm{T}} \mathbf{U}_k$ matrisi için $\mathbf{C}\mathbf{Z}_k = \mathbf{T}^T \mathbf{U}_k \mathbf{Z}_k = \mathbf{0}$ koşulunu sağlayan $\mathbf{Z}_k$ ortogonal matrisi bulunur. Bu matris $\mathbf{C}^{\mathrm{T}} = \mathbf{U}_k^{\mathrm{T}} \mathbf{T}$ matrisi için $\mathbf{QR}$ ayrışımından bulunabilir, ortogonal $\mathbf{Q}$ matrisinin son $\mathbf{M}$ sütunu $\mathbf{Z}_k$'yı oluşturur. $\boldsymbol{\delta}_k = \mathbf{Z}_k \tilde{\boldsymbol{\delta}}$ şeklinde tanımlanarak $\boldsymbol{\delta}_k$, bu M sütun tabanlı uzaya sınırlanır ve (4.47) koşullu minimizasyon problemi aşağıdaki koşulsuz probleme dönüşür:

$$\min_{\tilde{\boldsymbol{\delta}}, \boldsymbol{\alpha}} S(f) = \left\| \mathbf{y} - \mathbf{U}_k \mathbf{D}_k \mathbf{Z}_k \tilde{\boldsymbol{\delta}} - \mathbf{T}\boldsymbol{\alpha} \right\|^2 + \lambda \tilde{\boldsymbol{\delta}}^T {\mathbf{Z}_k}^T \mathbf{D}_k \mathbf{Z}_k \tilde{\boldsymbol{\delta}} \tag{4.49}$$

Bu problem için hesaplama karmaşıklığı $O(k^3)$ oranındadır. (4.49) probleminden $\tilde{\boldsymbol{\delta}}$ belirlenerek, $\boldsymbol{\delta} = \mathbf{U}_k \mathbf{Z}_k \tilde{\boldsymbol{\delta}}$ ve (4.45) ifadesini kullanarak gerekli olan ince tabakalı splayn modeli oluşturulabilir. $\mathbf{E}$'nin tam öz-ayrışımı $O(n^3)$ işlem gerektirir, bu da TPRS yaklaşımının yararını oldukça sınırlamaktadır. Neyse ki Lanczos süreci ile $\mathbf{U}_k$ ve $\mathbf{D}_k$'nın bulunması için $O(n^2 k)$ işlem gerekmektedir.

4.4.4. Tenzor Çarpım Splaynı

Yüksek boyutlu uzayda düzeltme fonksiyonunu yapılandırma yöntemlerinden biri de bir değişkenli düzeltme fonksiyonları ailesini kullanan *tenzor çarpım splaynı* yaklaşımıdır (Green ve Silverman, 1994; Wood, 2006).

(4.29) regresyon problemi göz önüne alınsın:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, ..., n \tag{4.50}$$

Burada $\mathbf{x}_i$ d-boyutlu bir vektördür, diğer bir ifadeyle f, d-değişkenli bir fonksiyondur. Bu değişkenler arasında istatiksel bir ilişkinin olduğu ve f fonksiyonunun değerini birlikte (bağımlı olarak) etkilediği varsayılır.

Kolaylık için $d = 3$ ve $\mathbf{x}_i = (x, z, v)$ biçimde olduğu varsayılsın. x, z, v değişkenlerine uygun ve kübik taban fonksiyonlarla belirlenen bir değişkenli f_1, f_2, f_3 fonksiyonları aşağıdaki gibi tanımlansın.

$$f_1(x) = \sum_{i=1}^{q_1} \alpha_i a_i(x), \; f_2(z) = \sum_{j=1}^{q_2} \delta_j d_j(z), \; f_3(v) = \sum_{k=1}^{q_3} \beta_k b_k(v) \tag{4.51}$$

(4.51)'de α_i, δ_j, β_k bilinmeyen parametreler, $a_i(x)$, $d_j(z)$, $b_k(v)$ bilinen (kübik) taban fonksiyonlarıdır. f fonksiyonunun x, z, v değişkenleri arasındaki ilişki, (4.51)'de ifade edilen bir değişkenli fonksiyonların kat sayılarını etkiler. x değişkeninin fonksiyonu olan f_1'in x, z değişkeninin fonksiyonuna dönüştürülmesi için, onun α_i parametresi z değişkeninin fonksiyonu olarak $d_j(z)$ tabanı ile ifade edilebilir: $\alpha_i(z) = \sum_{j=1}^{q_2} \delta_{ij} d_j(z)$. Bu durum, (4.51)'deki birinci fonksiyonu kullanarak x, z değişkenlerinin

$$f_{1,2}(x, z) = \sum_{i=1}^{q_1} \sum_{j=1}^{q_2} \delta_{ij} d_j(z) a_i(x) \tag{4.52}$$

fonksiyonunu türetmesine imkanını sağlar. Aynı yaklaşımla, δ_{ij} parametresi v değişkeninin fonksiyonu olarak $b_k(v)$ tabanı ile ifade edilebilir: $\delta_{ij}(v) = \sum_{j=1}^{q_3} \delta_{ijk} b_k(v)$ ve bu durumda x, z, v değişkenlerinin aşağıdaki fonksiyonu tanımlanmış olur.

$$f_{1,2,3}(x,z,v)=\sum_{i=1}^{q_1}\sum_{j=1}^{q_2}\sum_{k=1}^{q_3}\delta_{ijk}b_k(v)d_j(z)\alpha_i(x) \tag{4.53}$$

x,z,v değişkenlerinin (4.53) ile tanımlanan fonksiyonu bir değişkenli farklı üç taban fonksiyonun çarpımı ile ifade edilmiştir. Bu (4.50) regresyon denklemindeki f fonksiyonunun tahmini için kullanılan *splayn fonksiyonu*dur. (4.53) ifadesinde $q_1\times q_2\times q_3$ sayıda bilinmeyen δ_{ijk} parametreleri cezalı en küçük kareler toplamından bulunurlar.

Tanım 4.2: $a:X\to R$, $d:Z\to R$ fonksiyonlarının *tenzor çarpımı* olan $a\otimes d:X\times Z\to R$ fonksiyonu, $a\otimes d(x,z)=a(x)d(z)$ biçiminde tanımlanır.

Taban fonksiyonların doğrusal kombinasyonu olarak tanımlanan

$$f_1(x)=\sum_{i=1}^{q_1}\alpha_i a_i(x),\quad f_2(z)=\sum_{j=1}^{q_2}\delta_j d_j(z)$$

fonksiyonlarının tenzor çarpımı ise aşağıdaki gibi tanımlanır.

$$f_1\otimes f_2(x,z)=\sum_{i=1}^{q_1}\sum_{j=1}^{q_2}\delta_{ij}a_i\otimes d_j(x,z)=\sum_{i=1}^{q_1}\sum_{j=1}^{q_2}\delta_{ij}d_j(z)a_i(x)\,. \tag{4.54}$$

Bu durumda (4.53) formülü ile ifade edilen $f_{1,2,3}$ fonksiyonu, (4.51)' de verilen $f_1,\ f_2,\ f_3$ fonksiyonlarının tenzor çarpımıdır ve aşağıdaki gibi ifade edilir.

$$\begin{aligned} f_1\otimes f_2\otimes f_3(x,z,v) &= \sum_{i=1}^{q_1}\sum_{j=1}^{q_2}\delta_{ijk}a_i\otimes d_j\otimes b_k(x,z,v) \\ &= \sum_{i=1}^{q_1}\sum_{j=1}^{q_2}\sum_{k=1}^{q_3}\delta_{ijk}b_k(v)d_j(z)a_i(x) \end{aligned} \tag{4.55}$$

(4.54) ve (4.55) formüllerinde $a_i\otimes d_j$ ve $a_i\otimes d_j\otimes b_k$ tenzor çarpımları uygun $f_1\otimes f_2(x,z)$ ve $f_1\otimes f_2\otimes f_3(x,z,v)$ fonksiyonları için $q_1\times q_2$ sayıda iki boyutlu ve $q_1\times q_2\times q_3$ sayıda üç boyutlu taban fonksiyonlarını oluşturur.

Tenzor çarpımı tabanını kullanarak (4.36) ceza terimi değerlendirilebilir. $d=3$ ve $\mathbf{x}_i=(x,z,v)$ olduğunu varsayarak, bu amaçla önce (4.55) ifadesindeki marjinal $f_1,\ f_2\ ,\ f_3$ düzeltme fonksiyonları için ceza teriminin değerlendirilmesi göz önüne alınsın. $f_1,\ f_2,\ f_3$ düzeltme fonksiyonları, onların değişkenlerine

uygun olarak f_x , f_z ve f_v olarak işaretlensin. Bir değişkenli bu fonksiyonlar için uygun marjinal ceza terimleri

$$J_x(f_x) = \boldsymbol{\alpha}^T \mathbf{S}_x \boldsymbol{\alpha},\ J_z(f_z) = \boldsymbol{\delta}^T \mathbf{S}_z \boldsymbol{\delta},\ J_v(f_v) = \boldsymbol{\beta}^T \mathbf{S}_v \boldsymbol{\beta} \tag{4.56}$$

formunda olur. (4.56)'da $\mathbf{S}$' ler uygun ceza matrisleri, $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ ve $\boldsymbol{\beta}$ uygun marjinal düzelticilerin kat sayı vektörleridir.

Kübik splayn durumunda, örneğin $J_x(f_x)$ ceza terimi $J_x(f_x) = \int_x \left(\partial^2 f_x \big/ \partial x^2 \right)^2 dx$ olarak tanımlanır. Şimdi $f_{x,z,v}(x,z,v)$ fonksiyonunda z ve v'yi sabit tutularak, x değişkeninin oluşturulan fonksiyonu $f_{x|z,v}(x)$ ile işaretlensin. Benzer olarak $f_{z|x,v}(z)$, $f_{v|x,z}(v)$ bir değişkenli fonksiyonları tanımlanır. Bu durumda $f_{x,z,v}(x,z,v)$ fonksiyonuna uygun genel ceza terimi doğal olarak aşağıdaki gibi tanımlanabilir:

$$J(f_{x,z,v}) = \lambda_x \int_{z,v} J_x(f_{x|z,v}) dzdv + \lambda_z \int_{x,v} J_z(f_{z|x,v}) dxdv + \lambda_v \int_{x,z} J_v(f_{v|x,z}) dxdz \tag{4.57}$$

Burada λ'lar düzeltme parametreleridir. Kübik splayn yaklaşımı örneğinde marjinal ceza terimleri kullanılarak (4.57) ifadesi şöyle yazılabilir:

$$J(f) = \int_{x,z,v} \lambda_x \left(\partial^2 f \big/ \partial x^2 \right)^2 + \lambda_z \left(\partial^2 f \big/ \partial z^2 \right)^2 + \lambda_v \left(\partial^2 f \big/ \partial v^2 \right)^2 dxdzdv \tag{4.58}$$

(4.58) integralinin sayısal değerlendirilmesi doğrudan yapılabilir. Örneğin, $f_{x|z,v}(x)$ fonksiyonu $f_{x|z,v}(x) = \sum_{i=1}^{q_1} \alpha_i(z,v) a_i(x)$ biçiminde yazılabilir ve $\boldsymbol{\alpha}(z,v) = \mathbf{M}_{zv} \boldsymbol{\beta}$ eşitliliğini sağlayan $\mathbf{M}_{zv}$ matrisi bulunabilir, burada $\boldsymbol{\beta}$ belirli sıralanmış β_{ijk} katsayılarının vektörüdür. Bu nedenle

$$J_x(f_{x|z,v}) = \boldsymbol{\alpha}(z,v)^T \mathbf{S}_x \boldsymbol{\alpha}(z,v) = \boldsymbol{\beta}^T \mathbf{M}_{zv}^T \mathbf{S}_x \mathbf{M}_{zv} \boldsymbol{\beta}$$

ve buradan

$$\int_{z,v} J_x(f_{x|z,v}) dzdv = \boldsymbol{\beta}^T \left(\int_{z,v} \mathbf{M}_{zv}^T \mathbf{S}_x \mathbf{M}_{zv} dzdv \right) \boldsymbol{\beta}. \tag{4.59}$$

Benzer olarak, (4.57) ifadesindeki ikinci ve üçüncü integral terimleri (4.59) kare formu biçiminde yazılabilir. Sonuç olarak (4.57) genel ceza terimi $\boldsymbol{\beta}$ katsayılar vektörünün bir kare formuna dönüşür (Wood, 2006b).

5. SPLAYN DÜZELTME VE REGRESYON SPLAYNI İLE UYGULAMA

Bu bölümde, evlerin bazı özelliklerinin satış ve kira fiyatlarına etkisi ve hava kirliliğinin ölüm üzerindeki etkisi gibi iki farklı sınıf problemde regresyon pürüzlülük ceza yaklaşımı uygulanmıştır. Evlerin satış ve kira fiyatları ile ilgili problemde splayn düzeltme yaklaşımı ile toplamsal modeller, ikinci problemde ise regresyon splaynı ile genelleştirilmiş toplamsal modeller (GAM) ele alınmıştır. Uygun "en iyi " modeli belirlemek için basit doğrusal regresyon modellerinden başlayarak, her aşamada daha da karmaşık olan farklı nonparametrik regresyon modelleri incelenmiştir. Yapılan uygulamalarda `R` ve `S-Plus` paket programları kullanılmıştır.

5.1. Evlerin Özelliklerinin Satış Fiyatları ve Kira Fiyatları Üzerindeki Etkisinin Splayn Düzeltme ile İncelenmesi

Evlerin satış ve kira fiyatlarının belirlenmesinde, evlerin bazı özellikleri önemli bir rol oynar. Bu özelliklerin evlerin satış ve kira fiyatlarını nasıl etkilediğinin bilinmesi, konut yatırımcılarına ve kiracılara rehber olması açısından önemli bir etkendir. Bu amaçla çalışmada üç uygulamaya yer verilmiştir: *i*) Kanada'daki evlerin özelliklerinin *satış fiyatları* üzerindeki etkisinin splayn düzeltme ile incelenmesi, *ii*) Eskişehir'deki evlerin özelliklerinin *satış fiyatları* üzerindeki etkisinin splayn düzeltme ile incelenmesi, *iii*) Eskişehir'deki evlerin özelliklerinin *kira fiyatları* üzerindeki etkisinin splayn düzeltme ile incelenmesi. Bu uygulamaların ayrıntıları aşağıda açıklanmaktadır.

Uygulama1: Kanada'daki Evlerin Satış Fiyatları ve Özellikleri (Omay ve ark., 2006a)

Çalışmada kullanılan veriler, Kanada'nın başkenti Ottawa'da 1987 yılında satılan 92 müstakil evin satış fiyatları ve evlerin karakteristiklerini gösteren değişkenlere ilişkin gözlem değerlerinden oluşmaktadır. Söz konusu veriler http://www.chass.utoronto.ca/~yatchew/ den alınmış olup, çalışmada yer verilen değişkenler aşağıdaki gibi tanımlanmıştır:

SF (*Saleprice*): Evin satış fiyatı (dolar)

ŞD (*Fireplac*) : Şömine için yapay (dummy) değişken

GD (*Garage*) : Garaj için yapay değişken

BD (*Luxbath*) : Banyo için yapay değişken

NKA (*Usespace*): Evin net kullanım alanı (square feet)

AU (*Disthwy*) : Evin ana yola uzaklığı

KG (*Avginc*) : İlgilenilen semtte yaşayan insanların ortalama geliri (dolar)

BKA (*Lotarea*): Evin bahçe dahil brüt kullanım alanı

Uygulamada kullanılan modeli değerlendirmek için, elde edilen uygun model (incelenen modeller içinde en iyisi) ile yapılan tahmin sonuçları, değişkenlerin tamamının doğrusal olarak yer aldığı çok değişkenli parametrik regresyon modeli, düzeltme gerektirmeyen yapay (dummy) değişkenler dışında kalan tüm değişkenlerin modelde yer aldığı toplamsal regresyon modeli ve hem yapay değişkenlerin hem de nonparametrik değişkenlerin yer aldığı semiparametrik toplamsal regresyon modeli sonuçları ile karşılaştırılmıştır.

Doğrusal Model

Tüm açıklayıcı değişkenlerin, evlerin satış fiyatları üzerinde doğrusal etkiye sahip olduğunu varsayarak, aşağıda ifade edilen doğrusal model kurulmuştur.

$$SF_i = \beta_0 + \text{ŞD}_i\beta_1 + GD_i\beta_2 + BD_i\beta_3 + KG_i\beta_4 + AU_i\beta_5 + BKA_i\beta_6 + NKA_i\beta_7 + \varepsilon_i \quad (5.1)$$

Bu modele ilişkin özet istatistikler Tablo5.1'de verilmiştir.

Tablo5.1'de görüldüğü gibi, ŞD ve BKA değişkenlerinin evlerin satış fiyatları üzerinde istatistiksel olarak anlamlı bir etkisi bulunmamaktadır. AU değişkeninin ise satış fiyatları üzerinde ters yönde bir etkisi gözlenmektedir. Evlerin ana yola olan uzaklıklarındaki bir birimlik artış evlerin fiyatında 11.3268 birimlik bir azalışa neden olmaktadır. Buna karşılık diğer değişkenlerin tümü ile evlerin satış fiyatları arasında aynı yönlü bir ilişki mevcuttur. Örneğin, evlerin net kullanım alanlarındaki bir birimlik bir artış, evlerin satış fiyatlarında 30.8762 birimlik bir artışa neden olmaktadır. Ancak Tablo5.1'deki t-test istatistikleri ve bu

istatistiklere karşılık gelen olasılık değerleri incelendiğinde, ŞD ve BKA değişkenleri %10 anlamlılık düzeyinde bile istatistiksel olarak anlamsızdırlar. Diğer taraftan, bu doğrusal model evlerin satış fiyatlarındaki değişmelerin sadece %49.98'ini açıklayabilmektedir. Bunun yanı sıra hata kareler toplamı (HKT=46959) ve eşdeğer olarak, sapma (deviance) değeri (46959.46) oldukça yüksektir. Bu durumda, doğrusal parametrik modelin evlerin satış fiyatları üzerinde etkili olan değişken etkilerini belirlemek için yeterli olmadığı sonucuna varılabilir. Bu sonuçtan yola çıkarak bir sonraki adımda, mevcut değişkenler için semiparametrik toplamsal model oluşturulmuştur.

Tablo5.1. Doğrusal regresyon modeline ait sonuçlar

	Katsayılar	Standart Hata	t-ist	Pr(>|t|)
(Sabit terim)	68.0650	17.3367	3.926	0.000176
ŞD	6.7690	6.4285	1.053	0.295375
GD	12.9218	5.3483	2.416	0.017856
BD	67.6431	11.0640	6.114	2.95e-08
KG	0.5886	0.2367	2.486	0.014888
AU	-11.3268	5.6897	-1.991	0.049760
BKA	1.2333	2.2176	0.556	0.579601
NKA	30.8762	10.1304	3.048	0.003081
$R^2 = 0.5383$, $R_A^2 = 0.4998$ $F = 13.99$ $df = 7$ ve 84), $p = 6.804e-12$				
HKT=46959, Sapma (Deviance) = 46959.46				

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Semiparametrik Toplamsal Model

Çalışmada yer verilen yapay değişkenler fonksiyondaki eğriliği etkilemediklerinden diğer bir ifadeyle, düzeltme gerektirmeyen değişkenler olduklarından modele parametrik kısım olarak dahil edilmiştir. Diğer taraftan, bağımlı değişkenle ilişkisinin türü kesin olarak bilinmeyen diğer değişkenler ise, modelin nonparametrik kısmını oluşturmuştur. Uygun semiparametrik modeli belirleyebilmek için hem yapay değişkenlerin hem de nonparametrik değişkenlerin birlikte yer aldığı aşağıda ifade edilen regresyon modeli kurulmuştur.

$$SF_i = \beta_0 + \text{ŞD}_i\beta_1 + GD_i\beta_2 + BD_i\beta_3 + f_1(KG_i) + f_2(AU_i) + f_3(BKA_i) + f_4(NKA_i) + \varepsilon_i \quad (5.2)$$

Bu modele ilişkin özet istatistikler Tablo5.2'de yer almaktadır.

Tablo5.2'de görüldüğü gibi, nonparametrik kısımda yer alan NKA değişkeni istatistiksel olarak anlamlı iken diğer nonparametrik değişkenler istatistiksel olarak anlamlı değildir. Dolayısıyla bu değişkenlerin modele anlamlı bir katkısı yoktur. Tablo5.2'nin parametrik kısmı incelendiğinde, parametrik katsayıların anlamlı oldukları görülmektedir. Nonparametrik değişkenlerin her birini tek bir katsayı ile temsil etmek mümkün olmadığı için bu modelde yer alan nonparametrik değişkenler, ancak grafiksel olarak görüntülenebilmektedir.

Tablo5.2. Semiparametrik toplamsal regresyon modeline ait sonuçlar

	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
(Sbt.Ter.)				
ŞD	8.1572673	6.273291	1.30031	1.934946e-01
GD	13.1886327	5.296560	2.49002	1.27735e-02
BD	66.7286007	11.093612	6.01506	1.79835e-09
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(KG)	1	3	1.7385	0.16675
s(AU)	1	3	1.6516	0.18509
s(BKA)	1	3	1.3040	0.27977
s(NKA)	1	3	2.7247	0.05044
	R^2 = 0.8061276		Sapma (Deviance) = 35704.24	

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Nonparametrik değişkenlerin modele istatistiksel olarak anlamlı katkılarının olmaması ve modelin yüksek bir sapmaya (35704.24) sahip olması nedeniyle, söz konusu semiparametrik toplamsal modelin, evlerin satış fiyatları üzerinde etkili olan bağımsız değişkenlerin belirlenmesinde uygun olmadığı sonucuna varılmıştır. Bir sonraki aşamada, yapay değişkenlerin yer almadığı ve diğer değişkenlerin ise modelde nonparametrik olarak yer aldığı toplamsal model oluşturulmuştur.

Toplamsal Model

Yapay değişkenlerin yer almadığı ve diğer tüm değişkenlerin nonparametrik kestirici değişken olarak işlem gördüğü toplamsal regresyon modeli aşağıda ifade edilmiştir. Bu tür bir modelle yapılan tahmin sonuçları özetle Tablo5.3'de yer almaktadır.

$$SF_i = f_1(KG_i) + f_2(AU_i) + f_3(BKA_i) + f_4(NKA_i) + \varepsilon_i \qquad (5.3)$$

Tablo5.3 incelendiğinde, modeldeki dört nonparametrik bileşenin gösterdikleri etkilerin tamamının istatistiksel olarak anlamlı olmadığı görülmekte ve yukarıda bahsedilen modellere göre çok daha yüksek bir sapma (58491.08) içerdiği gözlenmektedir. Ancak bu sapmaların yapay değişkenlerin etkilerinin modelde dikkate alınmamış olmasından ileri geldiği söylenebilir. Açık olarak görülmektedir ki, kurulan toplamsal model evlerin özelliklerinin satış fiyatını nasıl etkilediğini açıklamakta yeterli (uygun) bir model değildir.

Tablo5.3. Toplamsal regresyon modeline ait sonuçlar

	Df Npar	Df Npar	F	Pr(F)
(Sabit terim)	1	-	-	-
s(KG)	1	3	1.0140	0.39137
s(AU)	1	3	2.3712	0.07715
s(BKA)	1	3	1.1695	0.32711
s(NKA)	1	3	1.4540	0.23395
R^2 = 0.6530746		Sapma (Deviance) = 58491.08		

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Uygun Semiparametrik Toplamsal Model

Şu ana kadar incelen modellerin ele alınan problem için iyi sonuçlar vermediği görülmüştür. Bu durum modele dahil edilen bazı değişkenlerin katkılarının anlamlı olmaması veya doğrusal olarak alınan değişkenlerin gerçekte doğrusal bir etkiye sahip olmamasından kaynaklanabilir. Bu nedenle çeşitli modeller oluşturarak ve bu modeller arasında karşılaştırmalar yaparak probleme en iyi çözümü üreten model elde edilmeye çalışılmıştır. Bu amaçla modele bazı değişkenler eklenmiş bazıları ise çıkartılmıştır. Ayrıca, doğrusal olarak işlem gören değişkenlerin bazıları nonparametrik olarak, bazıları ise parametrik olarak

ele alınarak çok sayıda model oluşturulmuştur. Elde edilen bu modeller arasından seçilen en iyi model aşağıda ifade edilmiştir.

$$SF_i = \beta_0 + \text{ŞD}_i\beta_1 + GD_i\beta_2 + BD_i\beta_3 + AU_i\beta_4 + NKA_i\beta_5 + f_1(KG_i) + f_2(BKA_i) + \varepsilon_i \quad (5.4)$$

Bu modelle yapılan tahmin sonuçları Tablo5.4'de verilmiştir.

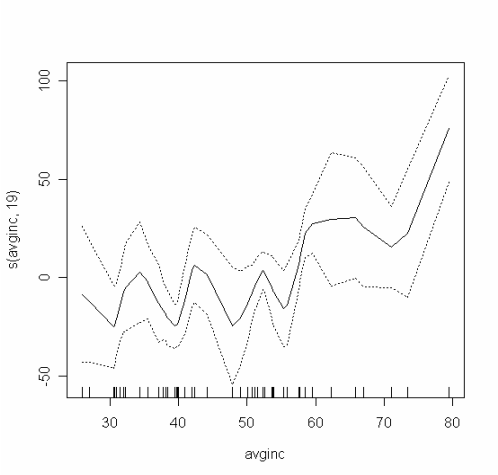
Elde edilen uygun semiparametrik toplamsal regresyon modelinin nonparametrik kısmında iki nonparametrik bileşen bulunmaktadır. Tablo5.4 incelendiğinde, bu değişkenlerin istatistiksel olarak anlamlı oldukları görülmektedir. Bu modelin parametrik kısmı incelendiğinde ise parametrik değişkenlerin tamamının istatistiksel olarak anlamlı oldukları görülmektedir. Bununla birlikte, doğrusal parametrik modelde olduğu gibi bu modelde de evlerin anayola uzaklıklarının satış fiyatları üzerinde negatif yönde bir etkiye sahip olduğu dikkati çekmektedir. Diğer bir ifadeyle, evlerin anayola uzaklıklarındaki bir birimlik bir artış, evlerin satış fiyatında 8.8009994 birimlik bir azalmaya sebep olmaktadır. Buna karşılık diğer değişkenlerin tümü ile evlerin satış fiyatları arasında aynı yönlü ilişki mevcuttur. Kurmuş olduğumuz uygun modelin belirlilik katsayısı incelendiğinde, bu modelin evlerin satış fiyatlarındaki değişmelerin %94.10'unu açıklayabildiği gözlemlenmiştir.

Tablo5.4. Uygun semiparametrik toplamsal regresyon modeline ait sonuçlar

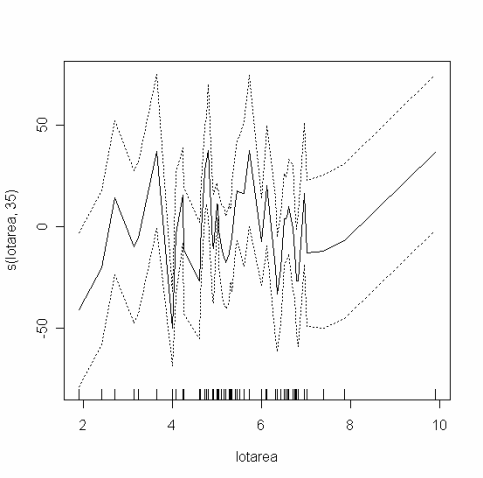
	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
(Sbt.Ter.)				
ŞD	3.547097	7.256323	0.4888352	6.24983e-01
GD	4.298754	6.130840	0.7011486	4.83103e-01
BD	58.221515	11.814254	4.9280749	8.30437e-07
AU	-8.800994	6.844451	-1.28568	1.80159e+00
NKA	35.462532	11.505871	3.0821239	2.05593e-03
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(KG)	1	18	2.2165	0.02403
s(BKA)	1	34	2.0751	0.02033
	R^2 = 0.9410437 Sapma (Deviance) = 11708.88			

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

(3.30) formülü ile elde edilen nonparametrik kısım çok sayıda katsayı içerdiğinden diğer bir deyişle vektör olarak elde edildiğinden, onu parametrik olarak ifade edebilmek mümkün değildir ve bu nedenle nonparametrik bileşenler ancak grafiksel olarak gösterilebilir. Dolayısıyla, evlerin satış fiyatları üzerindeki etkilerini gösteren eğriler Şekil5.1'de gösterilmiştir.



(a)



(b)

Şekil5.1. Evlerin satış fiyatlarının (a) komşu gelirlerine (b) brüt kullanım alanına göre değişimi ve güven aralıkları

Elde edilen tüm modelleri incelemek amacıyla, söz konusu modeller ile yapılan tahmin sonuçlarından elde edilen bazı performans göstergeleri Tablo5.5'de verilmiştir.

Tablo5.5. Modellerin belirlilik katsayıları ve sapmaları

Modeller	**R^2**	**Sapma**
Parametrik Model	0.4998	46959.46
Semipar. Toplamsal Model	0.8061276	35704.24
Toplamsal Model	0.6530746	58491.08
Uygun S. Toplamsal Model	**0.9410427**	**11708.88**

Parametrik regresyon modeli evlerin satış fiyatlarındaki değişmelerin %49.98'i açıklarken, uygun semiparametrik toplamsal modele göre büyük ölçüde sapma içermektedir. Semiparametrik toplamsal model fiyatlardaki değişimlerin %80.63 gibi önemli bir kısmını açıklamasına rağmen, uygun modelle kıyaslandığında, içerdiği sapmanın oldukça yüksek olduğu görülmektedir. Toplamsal modelin performans göstergeleri incelendiğinde, model fiyatlardaki değişmelerin %65,31'ni açıklar, ancak yapay değişkenleri içermemesi ve sahip olduğu yüksek sapma nedeniyle diğer tüm modellerden daha kötü performans sergilemektedir. Oluşturulan uygun semiparametrik toplamsal model fiyatlardaki değişmelerin %94'ünü açıklarken, bu modele ait sapma diğer modellerin sapmalarından daha düşüktür. Bu durumda, uygun semiparametrik toplamsal regresyon modelinin en iyi performans göstergeleriyle diğer modellerden daha iyi olduğu söylenebilir.

Toplamsal ve semiparametrik toplamsal regresyon modellerinin elde edilmesinde kullanılan splayn düzeltme yöntemi, $\lambda > 0$ düzeltme parametresi ve S_λ düzeltme matrisini bulundurması nedeniyle, sıradan en küçük kareler regresyon modelinden daha iyi sonuçlar vermektedir. Yapılan uygulamada, 1987 yılında Kanada'nın başkenti Ottawa'da satılan 92 müstakil evin karakteristiklerini gösteren değişkenlerin evlerin satış fiyatlarını nasıl etkilediği araştırılmış ve parametrik regresyon modeli dışında kalan tüm regresyon modelleri için splayn düzeltme yöntemi kullanılmıştır. Söz konusu evlerin satış fiyatları ile evlerin özellikleri arasındaki ilişkiler, parametrik regresyon modeli, semiparametrik toplamsal regresyon modeli ve toplamsal regresyon modeli kullanılarak analiz edilmiştir. Analizler sonucunda, hem parametrik doğrusal bileşenleri hem de iki nonparametrik bileşeni bulunduran uygun bir semiparametrik toplamsal regresyon

modeli elde edilmiştir. Bu regresyon modeline ait özet istatistikler, evlerin özelliklerinin satış fiyatları üzerindeki etkisini açıklayan en iyi modelin (elde edilen modeller içinden en iyisi) söz konusu semiparametrik toplamsal regresyon modeli olduğunu ortaya koymuştur.

Uygulama2: Eskişehir'deki Evlerin Kira Fiyatları ve Özellikleri (Omay ve ark., 2007b)

Uygulamada, Eskişehir merkezde bulunan evlerin kira fiyatlarını etkileyen ve bağımsız değişkenler olarak dikkate alınan evlerin özelliklerinin kira fiyatları üzerindeki etkileri incelenmiştir. Değişkenlerden bazılarının fiyat ile ilişkisinin şekli (doğrusal veya doğrusal olmayan) önceden bilinmeyebilir. Çalışmanın bu bölümünde, böyle bir ilişki şekli incelenerek ve diğer regresyon modelleri ile karşılaştırılarak, splayn düzeltme ile *uygun bir semiparametrik toplamsal regresyon modeli* belirlenmiştir. Yapılan istatistiksel analizlerle belirlenen modelin anlamlılığı değerlendirilmiştir.

Çalışmada kullanılan veriler, Eskişehir merkezde ikiden çok katlı binalarda yer alan kiralık evlerden şahsen elde edilmiştir. Ele alınan veriler, 2006 Mayıs ayı içersinde 108 kiralık evin kira fiyatları ve karakteristiklerini gösteren değişkenlere ilişkin gözlem değerlerinden oluşmaktadır. Söz konusu değişkenler aşağıdaki gibi tanımlanır:

Fiyat : Evlerin kira fiyatları (YTL)
Odas : Evlerde bulanan oda sayıları
Dakat : Bina içerisinde evlerin kaçıncı katta yer aldığı
Katsay: Evlerin bulunduğu binadaki kat sayısı
Kombi: Evlerde kombi sistemi olup olmadığını gösteren yapay değişken
Depozito: Evlerin kiralanması durumunda kiracıdan alınan depozito (YTL)
Yas : Evlerinin yaşı

İfade edilen bu değişkenlerden, Fiyat, Depozito ve Yas değişkenleri sürekli değişkenler, Odas, Dakat ve Katsay değişkenleri kesikli değişkenlerdir. Kombi değişkeni ise evlerde kombi sistemi olup olmadığını gösteren yapay değişkendir.

Uygulamada ele alınan modeli değerlendirmek için, oluşturulan *uygun semiparametrik toplamsal modelle* (elde edilen en iyi modelle) yapılan tahmin sonuçları, değişkenlerin tamamının doğrusal olarak yer aldığı çok değişkenli *doğrusal regresyon modeli* ve hem parametrik (düzeltme gerektirmeyen yapay değişken parametrik kısımda yer alır (Omay ve ark., 2007a)) hem de nonparametrik değişkenleri içeren *semiparametrik regresyon modeli* ile yapılan tahmin sonuçlarıyla karşılaştırılmıştır.

Uygun Semiparametrik Toplamsal Regresyon Modeli

Uygun (elde edilen en iyi) semiparametrik toplamsal modeli belirlemek amacıyla, yapay ve kesikli değişkenlerin dışında kalan mevcut bağımsız değişkenlerden bazıları nonparametrik, bazıları ise parametrik olarak ele alınarak, çeşitli regresyon modelleri oluşturulmuştur. Elde edilen bu modeller arasından seçilen en iyi (uygun) modelin formu aşağıda verilmiştir ve bu modelle yapılan tahmin sonuçları Tablo5.6'da yer almaktadır.

$$Fiyat_i = Odas_i\beta_1 + Dakat_i\beta_2 + Katsay_i\beta_3 + Kombi_i\beta_4 + f_1(Depozito_i) + f_2(Yas_i) + \varepsilon_i \quad (5.5)$$

Tablo5.6. Uygun semiparametrik toplamsal regresyon modeli sonuçları

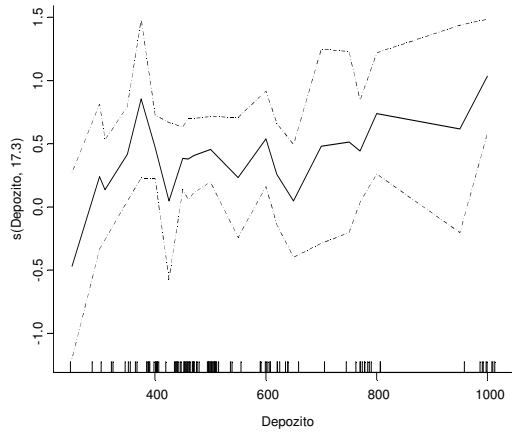
	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
Odas	0.2602	0.0291	8.948067	1.26e-12
Dakat	-0.0288	0.0166	-1.730918	5.86e-02
Katsay	0.0772	0.0162	4.742307	1.35e-05
Kombi	0.1032	0.0505	2.046119	4.52e-02
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(Depozito)	1	17.3	327.12	2.2e-16
s(Yas)	1	34	216.51	2.2e-16
	R^2 = 0.7718312 Sapma (Deviance) = 4.0313			

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

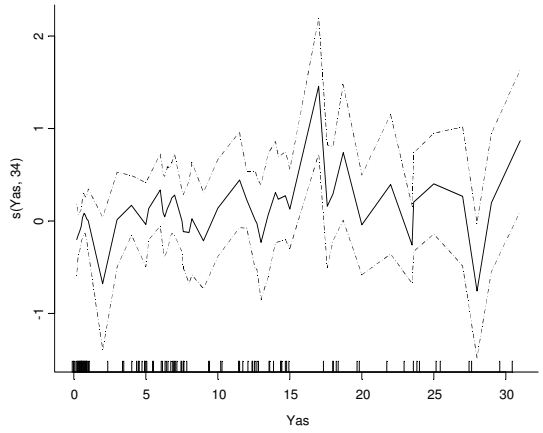
Uygun modelde iki nonparametrik bileşen bulunmaktadır: s(Depozito) ve s(Yas). Tablo5.6 incelendiğinde, hem parametrik hem de nonparametrik değişkenlerin istatistiksel olarak anlamlı oldukları görülmektedir. Bu modelde evlerin bulunduğu katlardaki bir birimlik bir artış, evlerin kira fiyatlarında 0.0288 birimlik bir azalmaya sebep olmaktadır. Diğer bir ifadeyle, binaların üst katlarında kira fiyatlarının azda olsa düştüğü söylenebilir. Buna karşılık diğer değişkenlerin tümü ile evlerin kira fiyatları arasında aynı yönlü ilişki mevcuttur. Yani, Odas, Katsay ve Kombi değişkenlerindeki bir birimlik artış kira fiyatlarına artmasına yol açmaktadır. Ancak Katsay değişkenin kira fiyatlarının üzerinde etkisinin çok düşük olduğu görülmektedir (bak. Tablo5.6). Ayrıca, uygun model tarafından evlerin kira fiyatlarındaki değişmelerin %77'sini açıklanabildiği gözlemlenmiştir.

Tablo5.6'da nonparametrik kısımda yer alan değişkenlere ilişkin katsayılar, parametrik olarak ifade edilemediğinden onlar ancak grafiksel olarak görüntülenmiştir. Söz konusu eğriler (3.30) formülü kullanılarak, sırasıyla Depozito ve Yas değişkenlerinin gözlem değerlerinden oluşan düğüm noktalarındaki kira fiyatlarını veren tahmin vektörlerinden elde edilmişlerdir. Adı geçen bu nonparametrik değişkenlerin evlerin kira fiyatları üzerindeki etkileri Şekil5.2'de görülen eğriler şeklinde ortaya çıkmıştır. Tablo5.6'da görüldüğü gibi eğrilerin fiyatlar üzerinde etkileri istatistiksel açıdan anlamlıdır.

Tablo5.7'de uygun semiparametrik toplamsal model dışında, elde edilen diğer modellere ait bazı sonuçlara da yer verilmiştir. Buna göre doğrusal regresyon modeli, evlerin kira fiyatlarındaki değişmelerin %69'unu açıklarken, *uygun semiparametrik toplamsal modele* göre büyük ölçüde sapma içermektedir. Semiparametrik model fiyatlardaki değişimlerin %62 gibi önemli bir kısmını açıklamasına rağmen, uygun semiparametrik toplamsal modelle kıyaslandığında, içerdiği sapmanın oldukça yüksek olduğu görülmektedir. Oluşturulan uygun semiparametrik toplamsal model fiyatlardaki değişmelerin %77'sini açıklarken, içerdiği sapma diğer modellerle kıyaslanmayacak ölçüde düşüktür. Modeller için hesaplanan AIC değerleri incelendiğinde en küçük AIC değerine sahip olan model uygun semiparametrik toplamsal modeldir. Bu durumda uygun modelin, en iyi performans göstergeleriyle elde edilen diğer modellerden daha iyi olduğu söylenebilir.



(a)



(b)

Şekil5.2. Evlerin kira fiyatlarının (a) depozitolarına (b) yaşlarına göre değişimi ve %95 güven aralıkları

Tablo5.7. Modellerin belirlilik katsayıları ve sapmaları

Modeller	R^2	Sapma	AIC
Parametrik Doğrusal Model	0.6921826	29.0474	178.6649
Semiparametrik Model	0.6207759	17.3936	174.3281
Semiparametrik Toplamsal Model	**0.7718312**	**4.0313**	**49.98315**

Eskişehir merkezde bulunan 105 evin kira fiyatları ile evlerin özellikleri arasındaki ilişkiler, parametrik, semiparametrik ve semiparametrik toplamsal

regresyon modelleri ile analiz edilmiştir. Bilinen geleneksel doğrusal regresyondan farklı olarak, bazı değişkenlerin kira fiyatını doğrusal etkilemediği gözlenmiştir. Bu durum, örneğin, splayn düzeltme yöntemine dayalı olan tahminlerin geleneksel (parametrik regresyon) yöntemlerden daha iyi olduğunun bir göstergesidir. Yapılan analizde, hem parametrik doğrusal bileşenleri hem de nonparametrik bileşenleri bulunduran uygun bir semiparametrik toplamsal regresyon modeli ile evlerin kira fiyatlarına ilişkin yapılan tahmin sonuçlarının elde edilen diğer modellerden daha üstün olduğu görülmüştür.

Uygulama3. Eskişehir'deki Evlerin Satış Fiyatları ve Özellikleri (Omay ve Aydın, 2007a)

Yapılan uygulamada Eskişehir merkezde bulunan satılık evlerin fiyatlarını etkileyen bağımsız değişkenlerin, ev fiyatları üzerindeki etkileri incelenmiştir. Bu değişkenlerden bazılarının fiyat üzerindeki etkisinin şekli (doğrusal olup olmadıkları) önceden bilinmeyebilir. Çalışmanın bu bölümünde böyle bir ilişki incelenmiş ve satılık ev fiyatları ile ilgili uygun (en iyi) bir *semiparametrik toplamsal regresyon modeli* elde edilmiştir. Yapılan istatistiksel analizlerle bu modelin anlamlılığı değerlendirilmiştir.

Çalışmada kullanılan veriler, Eskişehir merkezde yer alan çok katlı binalardaki satılık dairelere ilişkin bilgilerden, şahsen elde edilmiştir. İncelenen veriler, 2006 yılı Mayıs ve Haziran ayları içerisinde 105 satılık evin fiyatları ve karakteristiklerini gösteren değişkenlere ilişkin gözlem değerlerinden oluşmaktadır. Söz konusu değişkenler aşağıdaki gibi tanımlanır:

Fiyat : Evlerin satış fiyatları (YTL)
Yaş : Evlerin yaşı
Alan : Evlerin kullanım alanı (m^2)
Kgaraj : Apartmanda kapalı garaj olup olmadığını gösteren yapay değişken
Asansör: Apartmanda asansör olup olmadığını gösteren yapay değişken
Oparkı: Apartmanda oyun parkı olup olmadığını gösteren yapay değişken
Dogalgaz: Evlerde doğal gaz olup olmadığını gösteren yapay değişken
Kapıcı : Apartmanda kapıcı olup olmadığını gösteren yapay değişken
Opark: Apartmanda otopark olup olmadığını gösteren yapay değişken

Bulunan *uygun semiparametrik toplamsal regresyon modeli* ile elde edilen tahmin sonuçları, tüm açıklayıcı değişkenlerin doğrusal olarak yer aldığı *parametrik regresyon modeli* ve sadece bir tane nonparametrik değişkene sahip olan semiparametrik regresyon modelleri tahmin sonuçlarıyla karşılaştırılmıştır.

Uygun Semiparametrik Toplamsal Regresyon Modeli

Uygun semiparametrik toplamsal regresyon modelinin belirlenmesinde, düzeltme parametrelerinin ya da eşdeğer olarak serbestlik derecesinin seçimi en önemli problemlerden biridir. Bu çalışmada, nonparametrik değişkenlerin farklı serbestlik derecelerine sahip olduğu birçok model oluşturulmuştur. Elde edilen bu modeller arasından seçilen en iyi modelle diğer bir ifadeyle uygun model aşağıda verilmiştir ve bu modelle yapılan tahmin sonuçları Tablo5.8'de yer almaktadır.

$$Fiyat_i = \beta_0 + Kgaraj_i\beta_1 + Asansör_i\beta_2 + Oparkı_i\beta_3 + Doga\lg az_i\beta_4 + Kapıcı_i\beta_5 + Opark_i\beta_6 + f_1(Yaş_i) + f_2(Alan_i) + \varepsilon_i \quad (5.6)$$

Tablo5.8 göz önünde bulundurulduğunda, hem parametrik (Kgaraj, Asansör, Oparkı, Dogalgaz, Kapıcı, Opark) hem de nonparametrik (Yaş, Alan) değişkenlerin istatistiksel olarak anlamlı oldukları görülmektedir. Uygun semiparametrik toplamsal model ile evlerin satış fiyatlarındaki değişmelerin %97'sinin açıklanabildiği gözlemlenmiştir.

Tablo5.8'de nonparametrik kısımda yer alan değişkenlere ilişkin katsayılar, parametrik olarak ifade edilemediğinden onlar ancak grafiksel olarak görüntülenmiştir. Bu nonparametrik değişkenlerin evlerin fiyatları üzerindeki etkileri Şekil5.3'de gösterilmiştir.

Bu çalışmada, elde edilen uygun semiparametrik toplamsal regresyon modelinin yanı sıra kullanılan bütün değişkenlerin parametrik olarak göz önünde bulundurulduğu(bütün açıklayıcı değişkenlerin yanıt üzerindeki etkisinin doğrusal olduğu düşüncesi göz önünde bulundurularak) doğrusal regresyon modeli de elde edilmiştir ve bu iki model, sapma ve R^2 değerlerine göre karşılaştırılmıştır. Semiparametrik toplamsal regresyonda yanıt üzerinde istatistiksel olarak anlamlı etkilere sahip olduğu gözlenen birçok açıklayıcı değişkenin (Yas, Kgaraj, Asansör, Opark) parametrik modelde istatistiksel olarak anlamsız olduğu gözlenmiştir. Özellikle dikkati çeken durum, nonparametrik

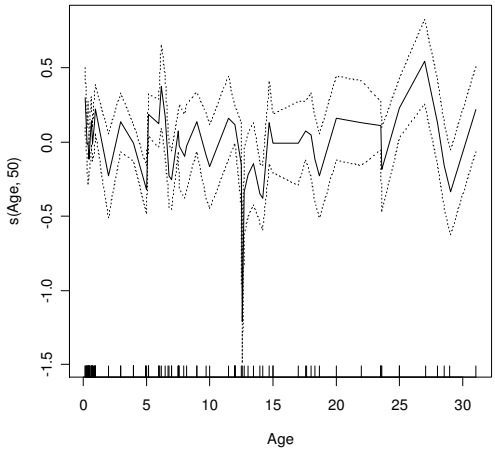
olarak değerlendirildiğinde yanıt üzerinde anlamlı etkiye sahip olan "Yas" değişkenin, doğrusal modelde anlamsızmış gibi görünmesidir. Bahsettiğimiz bu anlamsız değişkenler nedeniyle, tüm değişkenlerin modelde istatistiksel olarak anlamlı olduğu bir model kurmak için, model seçiminde kullanılan bir yaklaşım olan geriye doğru model seçim (backward model selection) yöntemi uygulanabilir. Bu yöntemde en yüksek *p*-değerine sahip olan ve belirli bir anlamlılık düzeyini aşan (örn. 0.05) terim modelden çıkarılarak model yeniden kurulur ve bu işlem modelde yer alan tüm değişkenler istatistiksel olarak anlamlı olana kadar devam eder (Wood, 2006a). Geriye doğru model seçimi (bckward elimination) yöntemi ile elde edilen model aşağıda verilmiştir ve bu modele ilişkin sonuçlar Tablo5.9'da yer almaktadır.

$$Fiyat_i = \beta_0 + Alan_i\beta_1 + Oparkı_i\beta_2 + Doga\lg az_i\beta_3 + Kapıcı_i\beta_4 + \varepsilon_i \qquad (5.7)$$

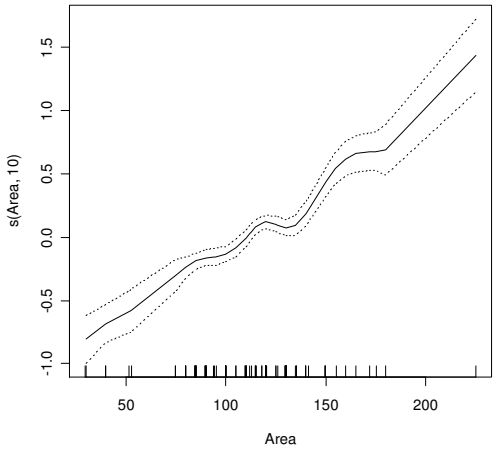
Tablo5.8. Uygun semiparametrik toplamsal regresyon modeli sonuçları

	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
(Sbt. Trm.)	10.1987	0.05999	170.01211	3.88587e-55
Kgaraj	0.0796	0.03771	2.11203	4.14965e-02
Asansör	-0.1056	0.03772	-2.79872	8.10099e-03
Oparkı	-0.1471	0.04394	-3.34829	1.87885e-03
Dogalgaz	0.0827	0.03105	2.66274	1.14053e-02
Kapıcı	0.3488	0.03353	10.40233	1.54878e-12
Opark	-0.1548	0.03094	-5.00234	1.40050e-05
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(Yaş)	1	49	2.8593	0.0006184***
s(Alan)	1	9	3.6703	0.0023109**
	R^2=0.97097 Sapma (Deviance) = 0.7494			

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1



(a)



(b)

Şekil5.3: Evlerin fiyatlarının **(a)** yaşlarına **(b)** kullanım alanlarına göre değişimi ve %95 güven aralıkları

Tablo5.9 incelendiğinde, doğrusal model için "Yas", "Kgaraj", "Asansör", "Opark" değişkenlerinin yanıt üzerinde istatistiksel olarak anlamlı etkilere sahip olmadıkları gözlenmiştir, halbuki biz kurduğumuz semiparametrik toplamsal regresyon modelinde, bahsedilen bu değişkenlerin yanıt üzerinde anlamlı etkilere sahip olduğunu görmüştük. Bu durumda, semiparametrik toplamsal regresyon modeli yerine doğrusal regresyon modeli kullandığımızda bazı değişkenleri göz ardı etmiş oluruz.

Tablo5.9. Doğrusal regresyon modeli sonuçları

	Katsayılar	Standart Hata	t-ist	Pr(>|t|)
(Sbt. Trm.)	10.22687	0.07239	141.267	< 2e-16 ***
Alan	0.00808	0.00065	12.437	< 2e-16 ***
Oparkı	-0.12625	0.05774	-2.186	0.031147 *
Dogalgaz	0.14808	0.04096	3.616	0.000473 ***
Kapıcı	0.20502	0.04324	4.742	7.12e-06 ***
R^2= 0.7082 Sapma (Deviance) = 3.669647				

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Çalışmanın bu aşamasından sonra, uygun semiparametrik toplamsal modele ve doğrusal modele ek olarak iki semiparametrik model daha kurulmuştur: Semiparametrik$_{YAŞ}$ ve semiparametrik$_{ALAN}$ modelleri. Bu modellerden birincisi nonparametrik değişken olarak sadece "Yaş" değişkenini içerirken ikincisi ise sadece "Alan" değişkenini içermektedir ve bu modeller aşağıda ifade edilmiştir. Geriye doğru model seçim yöntemi uygulandıktan sonra elde edilen bu modellere ait sonuçlar, sırasıyla, Tablo5.10 ve Tablo5.11'de yer almaktadır.

$$Fiyat_i = \beta_0 + Alan_i\beta_1 + Asansör_i\beta_2 + Oparkı_i\beta_3 + Doga\lg az_i\beta_4 + Kapıcı_i\beta_5 + Opark_i\beta_6 + f(Yaş_i) + \varepsilon_i \quad (5.8)$$

$$Fiyat_i = \beta_0 + Oparkı_i\beta_1 + Doga\lg az_i\beta_2 + Kapıcı_i\beta_3 + f(Alan_i) + \varepsilon_i \quad (5.9)$$

Tablo5.10 incelendiğinde, semiparametrik$_{YAŞ}$ modelinde, "Kgaraj" dışındaki tüm değişkenlerin yanıt üzerinde istatistiksel olarak anlamalı etkilere sahip oldukları görülmektedir. Semiparametrik$_{YAŞ}$ modelinin sapma değeri (1.1404), uygun semiparametrik toplamsal regresyon modelinin sapma değerinden (0.7494) daha büyüktür. Bununla birlikte semiparametrik$_{YAŞ}$ modelinin R^2 değeri (0.95338), uygun semiparametrik toplamsal modelin R^2 değerinden (0.97097) daha düşüktür. Bu sonuçlardan yola çıkarak, uygun semiparametrik toplamsal regresyon modelinin semiparametrik$_{YAŞ}$ modelinden daha iyi olduğu sonucuna varılabilir.

Tablo5.10. Semiparametrik$_{YAŞ}$ regresyon modeli sonuçları

	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
(Sbt. Trm.)	10.1900	0.06564	155.24613	2.29312e-65
Alan	0.0101	0.00054	18.51332	3.17163e-23
Asansör	-0.1215	0.04085	-2.97395	4.62817e-03
Oparkı	-0.2009	0.04678	-4.29478	8.69905e-05
Dogalgaz	0.1249	0.03347	3.73077	5.13714e-04
Kapıcı	0.3204	0.03661	8.75238	1.96655e-11
Opark	-0.1290	0.03385	-3.81160	4.00795e-04
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(Yaş)	1	49	1.9959	0.009371**
	R^2=0.95338 Sapma (Deviance) = 1.1404			

Anlamlılık: 0 ‘***’0.001 ‘**’0.01 ‘*’0.05 ‘.’0.1 ‘ ’1

Tablo5.11, semiparametrik$_{ALAN}$ modelinin istatistiksel olarak anlamlı olan sadece üç parametrik değişkene sahip olduğunu göstermektedir: Oparkı, Dogalgaz, Kapıcı. Uygun semiparametrik toplamsal modelde istatistiksel olarak anlamlı etkilere sahip olan diğer parametrik değişkenler semiparametrik$_{ALAN}$ modelde anlamlı etkilere sahip değildir. Ek olarak, semiparametrik$_{ALAN}$ modelinin sapma değeri (0.7509), uygun semiparametrik toplamsal regresyon modelinin sapma değerinden (0.7494) daha büyüktür. Bununla birlikte semiparametrik$_{ALAN}$ modelinin R^2 değeri (0.88175), uygun semiparametrik toplamsal modelin R^2 değerinden (0.97097) daha düşüktür. Bu sonuçlar göstermektedir ki, uygun semiparametrik toplamsal regresyon modeli semiparametrik$_{ALAN}$ modelinden daha iyidir. semiparametrik$_{ALAN}$ modelinde, uygun semiparametrik toplamsal modelde yer alan bazı değişkenlerin olmayışı, daha önce kurulmuş olan doğrusal modeldekine benzer bir problemi ortaya çıkarmaktadır: Eğer uygun semiparametrik toplamsal regresyon modeli yerine semiparametrik$_{ALAN}$ modeli kurulursa, aslında yanıt üzerinde anlamlı etkilere sahip olan bazı değişkenler göz ardı edilmiş olabilir.

Tablo5.11. Semiparametrik$_{ALAN}$ regresyon modeli sonuçları

	Parametrik Kısım			
	Katsayılar	St. Hata	t-ist.	Pr ($>\|t\|$)
(Sbt. Trm.)	10.2355	0.06574	155.69931	3.07323e-111
Oparkı	-0.1358	0.05243	-2.59001	1.11938e-02
Dogalgaz	0.1291	0.03719	3.47030	8.00249e-04
Kapıcı	0.2367	0.03926	6.02922	3.57223e-08
	Nonparametrik Kısım			
	Sd Npar	Sd Npar	F	Pr (F)
s(Alan)	1	9	3.3398	0.001447**
	R^2=0.88175 Sapma (Deviance) = 0.7509			

Anlamlılık: 0 '***'0.001 '**'0.01 '*'0.05 '.'0.1 ' '1

Çalışmanın bu aşamasında, elde edilmiş olan uygun semiparametrik toplamsal modelin doğrusal, semiparametrik$_{YAŞ}$ ve semiparametrik$_{ALAN}$ modellerinden daha iyi olup olmadığını R^2 ve sapma sonuçları temel alınarak karşılaştırılmıştır. Bu karşılaştırmaya ilişkin sonuçlar Tablo5.12'de yer almakta.

Tablo5.12. Semiparametrik toplamsal model ve diğer modeller için bazı sonuçlar

Modeller	R^2	Sapma
Semiparametrik Toplamsal Model	**0.97097**	**0.74940**
Doğrusal Model	0.70820	3.66965
Semiparametrik$_{YAŞ}$ Modeli	0.95338	1.14040
Semiparametrik$_{ALAN}$ Modeli	0.88175	2.75090

Tablo5.12 göstermektedir ki, semiparametrik toplamsal modelin R^2 değeri, diğer tüm modellerin R^2 değerinden daha yüksektir. Buna karşın, sapma değeri ise diğer modellerden daha düşüktür. Bu sonuçtan yola çıkarak semiparametrik toplamsal modelin elde edilen diğer modellerden daha iyi olduğu sonucuna varılabilir.

Bu uygulamada, Eskişehir merkezde bulunan 105 evin satış fiyatları ile evlerin özellikleri arasındaki ilişkiler, önceki uygulamalarda olduğu gibi, parametrik, semiparametrik ve semiparametrik toplamsal regresyon modelleri ile

analiz edilmiştir. Diğer uygulamalardan farklı olarak, bu uygulamada model kurma aşamasında değişkenlerin belirlenmesinde geriye doğru model seçim (backward model selection) yöntemi uygulanmıştır. Yapılan analizde, tıpkı diğer iki uygulamada olduğu gibi, hem parametrik doğrusal bileşenleri hem de nonparametrik bileşenleri bulunduran uygun bir semiparametrik toplamsal regresyon modeli ile evlerin satış fiyatlarına ilişkin yapılan tahmin sonuçlarının diğer modellerden daha üstün olduğu görülmüştür.

5.1 alt bölümü için SONUÇ: Evlerin satış ve kira fiyatları ile ilgili problemlerde splayn düzeltme yöntemine dayalı olan tahminlerin geleneksel (parametrik regresyon) yöntemlerden daha iyi sonuçlar verdiği gözlenmiştir. Yapılan analizde, hem parametrik doğrusal bileşenleri hem de nonparametrik bileşenleri bulunduran *uygun bir semiparametrik toplamsal regresyon modeli* ile evlerin kira/satış fiyatlarına ilişkin yapılan tahmin sonuçlarının elde edilen diğer modellerden daha üstün olduğu görülmüştür. Bu nedenle daha önce de vurgulandığı gibi, birçok pratik problemlerin regresyon modellerinde bağımlı değişkeni etkileyen açıklayıcı değişkenlerin doğasını belirleyerek uygun bir semiparametrik toplamsal modelin bulunması çok önemlidir ve bu durumlarda splayn düzeltme yaklaşımı iyi sonuçlar vermektedir.

5.2. Hava Kirliliğinin Ölüm Oranı Üzerine Etkisinin Regresyon Splayn ile İncelenmesi

Hava, yeryüzünü çevreleyen ve atmosferi oluşturan gazların bir karışımıdır. Saf hava, hacimce %21 oksijen, %78 azot, eser miktarda karbon dioksit ve su buharı içerir.

Soluduğumuz hava saf değildir. Hava, doğal kaynaklar veya insan aktiviteleri ile atmosfere bırakılan kimyasal ve biyolojik binlerce maddelerle kirlenir. Bu kirleticiler atmosferde başka kirleticiler üretmek üzere reaksiyona girebilir. Havayı kirleten maddelere kirletici denir. Bunlar, partikül madde (PM), ozon, kükürt dioksit, karbon monoksit, azot oksitler, uçucu organik bileşikler, sülfür, sülfat ve nitrattır.

Şehir içi bölgelerde hava kirletici kaynaklarını;

1. Isınma ve sanayide kullanılan yakıtlar,

2. Sanayi tesisleri,
3. Ulaşımdaki motorlu taşıtlar,
4. Diğerleri (çöp depolama alanları, kanalizasyon, arıtma tesisleri v.b.),

diye sınıflandırmak mümkündür.

Hava kirlenmesine neden olan kirleticileri birincil ve ikincil kirleticiler diye iki grup altında toplamak mümkündür. Kaynaktan doğrudan atmosfere atılan kirleticilere birincil kirleticiler, atmosferde çeşitli reaksiyonlar sonucu oluşan kirleticilere ikincil kirleticiler denir. Kükürt dioksit, kükürt trioksit, partikül maddeler, karbon monoksit, azot monoksit, hidrojen klorür ve hidrojen flörür gibi kirleticiler birincil kirleticilerdir. Asit yağmuru, nitrat, sülfat, foto kimyasal smog (ozon, PAH ve azot dioksit) gibi kirleticiler ikincil kirletici grubuna girer.

Dünyada her yıl hava kirliliğinden 3 milyon insan ölmektedir. Bu değer dünyadaki toplam ölümün (ortalama 55 milyon) %5'ni oluşturmaktadır. Ölümlerin %90'nı gelişmekte olan ülkelerde görülmektedir.

Hava kirliliğinin sağlık üzerindeki olumsuz etkileri sonucu,

- Akciğer kanser vakalarında artış,
- Kronik astım krizi sıklığında artış,
- Göğüs daralması sıklığında artış,
- Öksürük/balgam sıklığında artış,
- Üst solunum sistemi akut bozukluğunda artış,
- Göz, burun ve boğaz tahribatında artış,
- Soluk alma kapasitesinde düşüş,
- Artan ölüm,
- İş veriminde ve üretimde düşüş,
- Sağlık tedavi masrafında artış,

olduğu gözlenmiştir.

Havayı kirleten maddelerin (kirleticiler), diğer bir ifadeyle hava kalitesini belirleyen maddelerin (değişkenlerin) bazılarının karakteristiklerine ve insan sağlığı üzerindeki etkilerine ilişkin bazı bilgiler aşağıda verilmiştir.

Kükürt Dioksit (SO_2)

Kükürt dioksit özellikle katı ve sıvı yakıtlarda bulunan kükürdün yanması sonucu oluşur.

Kükürt dioksit, renksiz, yanmaz ve patlamaz bir gazdır. Atmosferdeki konsantrasyonu 785 µg/m^3'e (300 ppb) ulaştığında (eşik değer) tadı, 1305 µg/m^3 (500 ppb) değerine (eşik değer) geldiğinde kokusu algılanır.

Kükürt dioksit suda oldukça fazla çözünür. Atmosferde kalış süresi 2 ile 4 gün arasında değiştiğinden çok uzun mesafelere taşınabilmektedir. Dolayısıyla kükürt dioksit sadece bulunduğu bölgelerde değil taşındığı yerlerde de önemli olumsuzluğa neden olmaktadır.

Kükürt dioksit konsantrasyonunun yüksek olduğu illerde ısınmada kullanılan kalitesiz katı ve sıvı yakıt satış yerleri, sülfürik asit üretim tesisleri, termik santraller, gübre üretim tesisleri, pigment üretim tesisleri, seramik ve briket tesisleri, sıcak asfalt üretim tesisleri, zararlı atık yakma tesisleri, klor alkali tesisleri, demir çelik sanayisi ve atık yakma tesisleri sıkı şekilde denetlenmelidir. Isınmada kalitesiz katı ve sıvı yakıt satışına izin verilmemelidir.

Kükürt dioksit asidik bir gazdır. Nemle birleşme meylindedir. Kükürt dioksitle kirlenmiş hava solunduğu zaman; kükürt dioksit burun, geniz ve boğazdaki nemle reaksiyona girerek solunum sistemindeki sinirleri tahrip eder. Solunun yolu tahriş edildiğinde, refleks öksürük krizleri, göğüs sıkışması olur. Özellikle astım, kronik akciğer hastalığı bulunan kişilerde solunum yollarının daralmasına ve kronik solunum hastalığına neden olur.

Partikül Madde (PM10, PM2.5)

Havadaki partikül maddeler (PM); ısınma, motorlu taşıtlar ve endüstriyel tesislerde katı/sıvı yakıtların yakılması ile bazı endüstriyel tesislerde üretim işlemi esnasında oluşur. Karbon içeren yakıtların tam yanmaması sonucu duman oluşur. Toz boyutu 100–0.1 µm, smog boyutu 0.5–0.001 µm ve gaz boyutu 0.01–0.00001 µm arasında değişir. Ayrıca SO_2 ve NO_x gazları atmosferde reaksiyona girerek sülfat ve nitrat partikülleri haline dönüşürler. Isınmada kullanılan katı ve sıvı yakıtların tam yanmaması sonucu partikül maddeler oluşur. Partikül maddeler, asit tozları (nitrat ve sülfat gibi), organik kimyasallar, metaller, toprak veya toz

partikülleri, bakteri, küf, mantar, deniz suyunun buharlaşması ile ortaya çıkan tuzlar ve alerjik polenlerden oluşmaktadır.

Sağlık üzerine etkisi partikül büyüklüğü ve konsantrasyonuna bağlıdır. PM10 (10 mm çapından küçük partiküller) ve PM2.5'un (2.5 mm çapından küçük partiküller) günlük dalgalanmalarına göre sağlık etkileri de değişir. Partikül madde çapı küçüldükçe sağlık üzerindeki olumsuz etkisi o kadar artmaktadır. Akut etkileri günlük ölümlerde artışa, solunum sistemi hastalıklarının alevlenmesine, hastane başvurularında artışa, bronkodilatatör kullanımı ve öksürük prevalansında artışa, solunum fonksiyonlarında azalmaya yol açmaktadır. Çok düşük değerlerde bile (100 mg/m3den az) kısa süreli maruz kalım sağlığı etkilemektedir. PM`nin düşük değerlerde uzun süreli etkileri de ölüm ve solunum sistemi hastalıklarında artış ve solunum fonksiyonlarında azalma gibi kronik etkilere yol açmaktadır.

Partikül maddeler kükürt dioksit ile birlikte insan sağlığı üzerinde daha ciddi olumsuz etkiye neden olmaktadır. Bu durum özellikle kış aylarında ısınma amacı ile kalitesiz katı ve sıvı yakıt kullanan il ve ilçelerde daha ciddi olarak görülmektedir.

Karbon Monoksit (CO)

Karbon monoksit, renksiz, kokusuz, zehirli, tatsız ve aşındırıcı olmayan bir gazdır. Karbon monoksit suda az çözünen ve normal şartlarda havadan daha az yoğun olan bir gazdır.

Karbon monoksit, yakıt içindeki karbonun eksik yanması ile yani yanma bölümünde yeterli hava olmadığı zaman meydana gelir. Trosferde bulunan iz maddelerden biridir. Atmosferdeki karbon monoksit konsantrasyonu genel olarak yaz aylarında daha düşük, kış aylarında ise daha yüksektir. Çünkü kış aylarında ilaveten ısınma amacı ile yakıt kullanılmaktadır. Şehir içi bölgelerde karbon monoksitin birincil ana kaynağı motorlu taşıtlar ve ısınma tesisleri, ikincil önemli kaynağı ise; katı atık depolama tesisleridir. Depolama tesislerinde oluşan metan gazları atmosferde okside olarak karbon monoksite dönüşürler. Karbon monoksitin atmosferde bozunma süresi takriben 2,5 aydır.

Karbon monoksit çok zehirli bir gazdır. Karbon monoksitle zehirlenmenin ilk belirtisi, gribe benzer. Baş ağrısı, uyuklama, yorgunluk, nefes kesilmesi,

bulantı ve baş dönmesi şeklinde de etkisini gösterebilir. Karbon monoksitten zehirlenen çoğu kişi grip olduğunu zannederek yanılır. Takip eden etkisi bilinçsizlik, solunum hastalığı ve ölümdür. Karbon monoksit, özellikle 0–18 yaş arası astımlı çocuklar üzerinde etkilidir. Karbon monoksitin sağlık üzerindeki en önemli etkisi; kalp ve beyin gibi canlı organizmalara oksijen verme kapasitesini azaltmasıdır. Kalp hastası kişiler, karbon monoksit kirliliğine özellikle hassastırlar. Karbon monoksit kansızlığa neden olur. Karbon monoksit hamile kadınlarda, çocuk düşürmeye, ölü çocuk doğurmaya, düşük ağırlıklı çocuk doğurmaya, erken çocuk ölümüne sebep olabilir.

Ozon (O_3)

Ozon 3 oksijen atomundan oluşan bir gazdır. Ozon, hem yer seviyesinde ve hem de üst atmosferde oluşur. Ozon bulunduğu yere göre faydalı veya zararlı olabilir.

Ozon doğal olarak, atmosferin üst tabakasında yer kürenin 6–30 mil üzerinde oluşur ve koruyucu bir tabaka olarak atmosferi güneşin zararlı ultraviyole ışınlarından korur. Faydalı olan bu ozon, insanlar tarafından yapılan kimyasal maddeler ile kademli olarak tahrip edilmektedir. Yeryüzünün bazı bölgelerinde koruyucu ozon katmanı tükenmiştir (örneğin, yeryüzünün kuzey ve güney kutuplarında ozon delikleri oluşmuştur).

Yer yüzeyine yakın seviyede; otomobiller, enerji santralleri, endüstriyel kazanlar, rafineriler, kimyasal fabrikalardan ve benzeri kaynaklardan atmosfere verilen kirleticiler, güneş ışınlarının mevcudiyetinde kimyasal olarak reaksiyona girerek ozonu oluşturur. Yer seviyesindeki ozon zararlı bir kirleticidir. Ozon kirliliği, özellikle yaz aylarında güneşli havalarda oluşur.

Çocuklar, dış ortamda aktif olan yetişkinler, astım gibi solunum hastalığı olan ve ozona karşı çok hassas olan kişiler; ozon maruziyeti için en hassas grubu oluşturur. Ozon maruziyetine karşı en yüksek risk gruplarından birisi aktif çocuklardır, çünkü yaz aylarının büyük bir kısmını dışarıda oynayarak geçirirler. Ancak tüm yaş grupları ve dışarıda aktif olan kişiler de risk altındadır. Çünkü fiziksel aktivite sırasında ozon, akciğerlerin derinliklerine kadar nüfuz ederek zararlı etkilerini gösterir ve kalıcı hasarlar yaratabilir.

Solunum rahatsızlığı olan kişilerde, astımlılar dahil, ozona maruz kalma sonucu, akciğerlerin etkilenmesi daha kolaydır. Diğer insanlara göre daha düşük ozon seviyelerinde de ozonun zararlı etkilerini hissedebilirler.

Bilim adamlarının henüz nedenini bilmemelerine rağmen, bazı sağlıklı insanlarda da ozona karşı duyarlı olabilir.

Ozon, öksürük, boğaz tahrişi ve/veya göğüste rahatsızlık hissine sebebiyet vererek solunum yollarını tahriş edebilir.

Ozon, akciğer fonksiyonunu azaltarak, derin ve kuvvetli nefes almayı güçleştirebilir. Solunum hızlanır ve normalden daha yüzeysel olur. Akciğer fonksiyonundaki bu azalma, kişinin dış ortamdaki aktivitelerini yerine getirmekten alıkoyabilir.

Ozon, astımı daha kötü hale getirebilir. Ozon seviyesi yüksek olduğunda, astımlı olan kişiler, bir doktora ve tedaviye ihtiyaç duyan, astım krizlerine girebilirler. Bunun nedenlerinden birisi de ozon, insanları; astım tetikleyicileri olan evcil hayvanlar, polenler ve ev tozu akarları gibi alerjenlere karşı daha hassas hale getirir.

Ozon, akciğerlerin iç yüzeyini iltihaplandırabilir ve zarar verebilir.

Yapılmış olan uygulama çalışmalarında, havayı kirleten kükürt dioksit (SO_2), partikül madde (PM10, PM2.5), karbon monoksit (CO) ve ozon (O_3) maddelerinin ölüm oranı üzerindeki etkileri incelenmiştir. Bu amaçla iki uygulama çalışması yapılmıştır: Birinci uygulamada Los Angeles'daki hava kirliliğinin ölüm oranı üzerindeki etkisi; ikinci uygulamada ise Chicago'daki hava kirliliğinin ölüm oranı üzerindeki etkisi regresyon splaynı kullanılarak ortaya konmuştur.

Uygulama1. Los Angeles'daki Hava Kirliliğinin İncelenmesi (Memmedli ve Omay, 2007)

Bu bölümde, Los Angeles'daki hava kirliliği ile kaydedilen ölüm oranı arasındaki ilişki incelenmiştir ve bu uygulamada kullanılan veriler www.ihapss.jhsph.edu adresinden alınmıştır. Dikkate alınan veriler, 01/01/1987-31/12/2000 yılları arasında, Los Angeles'da kaydedilen günlük ölüm oranı ve hava kalitesi değerleridir. Ölçümler her gün için yapılmıştır. Olması gereken veri

sayısı 5114 iken orijinal veri setinde toplam 15342 veri vardır. Bunun nedeni veri setinin `agecat` değişkeni vasıtası ile 3 gruba ayrılmış olmasıdır:

(1) yaş(age)<65

(2) 65≤yaş<75,

(3) yaş≥75.

İlk 14 yıllık veri (ilk 5114 veri) 65 yaşından küçük ölümler için, ikinci 14 yıllık veri 65-75 yaş grubu için ve son 14 yıllık veri ise 75 yaşından büyük ölümler içindir. Yapmış olduğumuz analizlerde, aynı günlerde ve aynı anda kaydedilmiş olan ölüm oranları toplanarak bu üç grup birleştirilmiş ve toplam 5114 veri ile çalışılmıştır. İlgilenilen yanıt değişkeni yıllar içerisinde Los Angeles'da meydana gelen günlük ölüm oranlarıdır (`death`–kazalar dışındaki tüm ölüm olayları). Gözlenen ölüm oranlarına karşılık gelen, ölüm üzerinde anlamlı etkilere sahip olması mümkün açıklayıcı değişkenler ise aşağıda verilmiştir:

`o3median` : ozon seviyeleri (PPB) (*O_3 Median*)

`comedian` : karbon monoksit seviyeleri (PPB) (*CO Median*)

`pm10median`: partikül seviyeleri (mg/m^3) (*PM_{10} Median*)

`pm25median`: partikül seviyeleri (mg/m^3) (*$PM_{2.5}$ Median*)

`tmpd` : ortalama sıcaklık (F°)

Yukarıda ifade edilen hava kalitesi değişkenlerine ek olarak ölüm oranı, `time` değişkeni ile de değişme eğilimindedir. `time` değişkeni yukarıda ifade edilen 5114 veriyi iki kısma bölmektedir ve şu şekilde tanımlanmaktadır:

Gün	1	2	…	2557	2558	…	5113	5114
`time`	-2556.5	-2555.5	…	-0.5	0.5	…	2555.5	2556.5

Hava kirliliğinin ölüm oranını nasıl etkilediğini ortaya çıkarmak amacıyla farklı regresyon modelleri kurulmuş ve onların performansları değerlendirilmiştir. Yapılan analizlerde `R` programının temel paketlerinin yanı sıra `gamair`, `mgcv` ve `NMMAPSdata` paketleri de kullanılmıştır. Ele alınan regresyon modelleri sırasıyla aşağıda ifade edilmiştir:

- Doğrusal Model

- Genelleştirilmiş Doğrusal Model (GLM)
- Toplamsal Modeller
- Genelleştirilmiş Toplamsal Model (GAM)
- İnce Tabakalı Splayn (TPS) ile GAM

İfade edilen bu modellerden, toplamsal model, genelleştirilmiş toplamsal model ve ince tabakalı splayn kullanılarak oluşturulan genelleştirilmiş toplamsal modellerde regresyon splaynı kullanılarak model uyumu yapılmıştır. Şimdi yukarıda ifade edilmiş olan tüm modeller tek tek incelensin:

Doğrusal Model

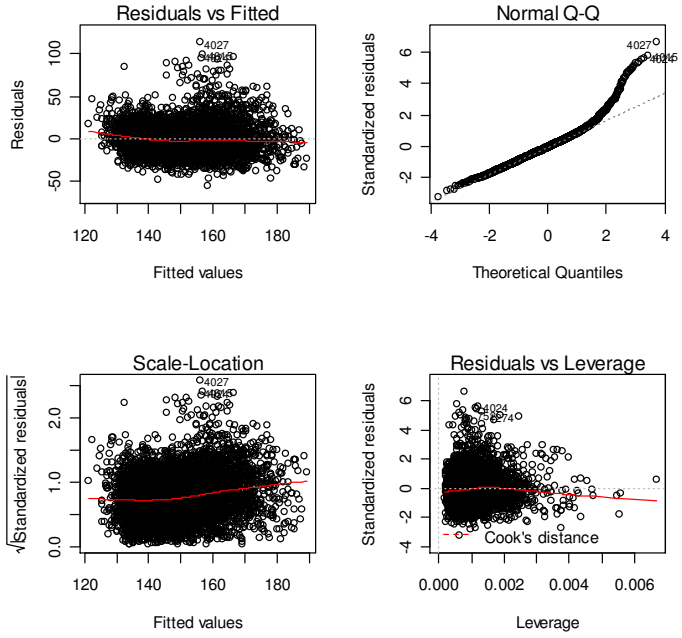
Tüm açıklayıcı değişkenlerin, ölüm oranı üzerinde doğrusal etkiye sahip olduğunu varsayarak, aşağıda ifade edilen doğrusal model kurulmuştur.

$$death_i = \beta_0 + time_i\beta_1 + tmpd_i\beta_2 + o3median_i\beta_3 + comedian_i\beta_4 + \varepsilon_i \qquad (5.10)$$

Bu modele ilişkin özet istatistikler Tablo5.13'de verilmiştir. Doğrusal modele ilişkin grafikler ise Şekil5.4'de gösterilmektedir.

Tablo5.13. Doğrusal model için özet istatistikler

```
> summary(LA1)

Call:
lm(formula = death ~ time + tmpd + o3median + comedian, data = la)

Residuals:
    Min      1Q  Median      3Q     Max
-55.265 -10.923  -1.150   9.077 113.125

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.049e+02  2.891e+00   70.87   <2e-16 ***
time        -2.178e-03  1.629e-04  -13.37   <2e-16 ***
tmpd        -8.680e-01  4.476e-02  -19.39   <2e-16 ***
o3median    -3.138e-01  3.025e-02  -10.38   <2e-16 ***
comedian     5.755e-03  3.411e-04   16.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.17 on 5109 degrees of freedom
Multiple R-Squared: 0.3169,     Adjusted R-squared: 0.3164
F-statistic: 592.6 on 4 and 5109 DF,  p-value: < 2.2e-16

> AIC(LA1)
[1] 43601.46

> deviance(LA1)
[1] 1506674
```



Şekil5.4. Doğrusal model için kontrol grafikleri.

Şekil5.4'ün sol üst panelinde yer alan grafik, $\hat{\mu}_i$ model uyum verilerine karşı, $\hat{\varepsilon}_i$ model artıklarını ifade eder. Burada $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ ve $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\boldsymbol{\mu}}$ 'dür. Sabit varyanslılık durumunda bu artıklar sıfır doğrusunun altında ve üstünde hemen hemen eşit olarak serpilmelidir. Artıkların ortalamasındaki bir trend genellikle hatalı model yapısının sonucudur ve böyle bir durumda sabit varyans varsayımı bozulmaktadır. Bu gibi durumlarda yanıt üzerinde yapılacak bir dönüşüm ya da genelleştirilmiş doğrusal model kullanılması farklı varyanslılık problemini ortadan kaldırabilir. Şekil5.4'de bazı noktaların sıfır etrafındaki eşit saçılımı bozdukları gözlenmektedir, dolayısıyla problemli bir durum mevcuttur.

Şekil5.4'ün sol alt panelinde yer alan grafik de sabit varyans varsayımını kontrol etmektedir ve bu grafikte, sapan gözlemlerden kaynaklanan bazı problemler göze çarpmaktadır.

Şekil5.4'ün sağ üst panelinde bir normal Q-Q grafiği yer almaktadır. Bu grafikte standardize artıklar sıralanır ve daha sonra standart normal dağılımın yüzdeliklerine karşı çizilir. Eğer artıklar normal dağılıyorsa, çizilen şekil bir doğru şeklinde görülecektir. Şekil5.4'deki grafik bu tanımlamaya tam olarak uymamaktadır, grafiğin üst kuyruğunda doğrusallıktan sapma gözlenmektedir.

Şekil5.4'ün sağ alt panelinde her bir verinin kuvvetine (leverage) karşı standardize edilmiş artıklarının çizimi yapılmıştır. Kuvvet (leverage) basit olarak **A** şapka matrisinin köşegen elemanlarını ifade etmektedir. Büyük (geniş) bir artık kombinasyonu ve yüksek bir kuvvete (leverage) karşılık gelen verinin, genel uyum üzerinde güçlü bir etkiye sahip olduğunu göstermektedir. Her bir verinin tam olarak ne kadarlık bir etkiye sahip olduğunun iyi bir özeti, onun Cook Mesafesi (Cook's Distance) ile sağlanır. Eğer $\hat{\mu}_i^{[k]}$, *i*. uyumu yapılan değer (i^{th} fitted value) ise, *k*. veri (k^{th} datum) uyumdan çıkarıldığında Cook mesafesi aşağıdaki gibi ifade edilir.

$$d_k = \frac{1}{(p+1)\hat{\sigma}^2} \sum_{i=1}^{n} \left(\hat{\mu}_i^{[k]} - \hat{\mu}_i \right)^2$$

Burada *p*, parametre sayısı, *n* ise veri sayısıdır. d_k'nın çok büyük bir değeri, model sonuçları üzerinde güçlü bir etkiye sahip noktayı ifade eder. Eğer Cook mesafesinin değeri, model tahminlerinin sadece bir ya da iki veriye çok duyarlı olabileceğini ifade ediyorsa, modelleme sonuçlarının güçlülüğünü (robustness) kontrol etmek amacıyla, sıkıntı yaratan bu noktalar olmaksızın herhangi bir analizi tekrarlamak genellikle mantıklı olabilir. Cook mesafesinin, kuvvet (leverage) ve standardize artıkların bir fonksiyonu olarak ifade edilebilmesi nedeniyle, Cook mesafesinin konturları grafik üzerinde gösterilebilir. Şekil5.4'de noktaların hiç biri net olarak çizgi (kontur) dışında (out of line) görünmemektedir.

Genelleştirilmiş Doğrusal Model (GLM)

Ölüm verilerini modellemekte kullanılan geleneksel yaklaşımlardan biri, gözlenen ölümün Poisson rassal değişken olması varsayımıdır. Bu varsayımdan hareketle ve Şekil5.4'ün bir sonucu olarak, doğrusal model yerine genelleştirilmiş doğrusal modelin kullanılması daha uygun görünmektedir. Elde edilen genelleştirilmiş doğrusal model aşağıda ifade edilmiştir.

$$\begin{aligned} \log(E[death_i]) = {} & \beta_0 + time_i\beta_1 + tmpd_i\beta_2 + pm10median_i\beta_3 + \\ & pm25median_i\beta_4 + o3median_i\beta_5 + comedian_i\beta_6 + \varepsilon_i \qquad (5.11) \\ death_i \;\square\; & Poi(E[death_i]) \end{aligned}$$

Bu modele ilişkin sonuçlar Tablo5.14'de yer almaktadır.

Yukarıda kurulmuş olan doğrusal ve genelleştirilmiş doğrusal modeller "AIC" ve "sapma" (deviance) kriterlerine göre karşılaştırıldığında, genelleştirilmiş doğrusal modelde hem AIC hem de sapma değerinde büyük bir düşüş gözlenmektedir. Fakat genelleştirilmiş doğrusal model için kontrol grafikleri, Şekil5.4'de doğrusal model için ifade edilmiş olan grafiklerle hemen hemen aynı görüntüye sahiptir. Dolayısıyla, genelleştirilmiş doğrusal modelde de mevcut problemlerin (özellikle sapan gözlem probleminin) hala devam ettiği görülmektedir.

Tablo5.14. Genelleştirilmiş doğrusal model için özet istatistikler

```
> summary(LA2)

Call:
glm(formula = death ~ time + tmpd + pm10median + pm25median +
    o3median + comedian, family = poisson, data = la)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.00837  -0.94433  -0.07325   0.72809   6.87172

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.355e+00  4.419e-02 121.171  < 2e-16 ***
time        -1.520e-05  2.271e-06  -6.691 2.22e-11 ***
tmpd        -5.497e-03  6.866e-04  -8.006 1.18e-15 ***
pm10median  -7.174e-04  3.351e-04  -2.141 0.032305 *
pm25median   1.289e-03  4.308e-04   2.992 0.002768 **
o3median    -1.738e-03  4.629e-04  -3.755 0.000173 ***
comedian     3.404e-05  6.549e-06   5.198 2.01e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1609.4  on 593  degrees of freedom
Residual deviance: 1081.8  on 587  degrees of freedom
AIC: 5157.5
```

Şekil5.4'ün sağ alt panelinde yer alan grafikte herhangi bir sapan gözlem mevcut görünmemektedir. Fakat buna rağmen, yukarıda ifade edildiği gibi, sözü geçen problemlerin hala devam etmesi nedeniyle, doğrusal ve genelleştirilmiş doğrusal modeller için sapan gözlem adayları veri setinden çıkarılarak yeniden modeller kurulmuş ve kurulan tüm doğrusal modeller R^2, hata kareler ortalaması (MSE), $\hat{\beta}$ ve $se(\hat{\beta})$ değerleri göz önünde bulundurularak karşılaştırılmıştır. Diğer taraftan, kurulan tüm genelleştirilmiş doğrusal modeller ise sapma, $\hat{\beta}$ ve

$se(\hat{\beta})$ değerleri dikkate alınarak karşılaştırılmıştır. Sapan gözlem adayı noktalar Şekil5.4'de de görüldüğü gibi, 4743., 4749. ve 4755. noktalardır ve bu noktaların veri setinden çıkarılması işlemi Tablo5.15'de yer alan sırayla yapılmıştır.

Tablo5.15. Kurulan modellerden çıkarılan sapan gözlem adayları

Çıkarılan Gözlemler		
• 4755. gözlem **(1)** • 4749. gözlem **(2)** • 4743. gözlem **(3)**	• 4743. ve 4749. gözlemler **(4)** • 4743. ve 4755. gözlemler **(5)** • 4749. ve 4755. gözlemler **(6)**	• 4743., 4749. ve 4755. gözlemler **(7)**

Tablo5.15 incelendiğinde, doğrusal ve genelleştirilmiş doğrusal modeller için yedişer adet yeni model kurulduğu dikkati çekmektedir ve bu tabloda parantez içinde yer alan sayılar modellere verilen numaraları göstermektedir. Bu modellere ilişkin bazı sonuçlar Tablo5.16'da yer almaktadır. Tablo5.16'nın ilk satırı gözlemler çıkarılmadan önce kurulan modellere ait sonuçları içermektedir.

Tablo5.16. Sapan gözlem adayları çıkarılarak kurulan modellere ait sonuçlar

	Doğrusal Model (LM)		Gen. Doğ. Model (GLM)
Modeller	R^2	MSE	Sapma
LM/GLM	0.3164	295	1081.8
1	0.3164	294	1049.0
2	0.3173	294	1055.0
3	0.3174	293	1034.0
4	0.3184	292	1007.0
5	0.3175	292	1007.0
6	0.3173	293	1022.0
7	0.3178	291	973.0

Tablo5.16 incelendiğinde doğrusal model için R^2 ve MSE değerleri açısından göze çarpan farklar yoktur aynı zamanda $\hat{\beta}$ ve $se(\hat{\beta})$ değerleri de çok yakın sonuçlar vermiştir. Genelleştirilmiş doğrusal model için sapma değerleri incelendiğinde, 7. model için (üç gözlem de çıkarıldığında) bu değerin diğerlerinden biraz farklı olduğu görülmektedir fakat bu modelin $\hat{\beta}$ ve $se(\hat{\beta})$ değerleri birbirlerine çok yakın sonuçlar vermiştir. Bu sonuçlar, Şekil5.4'ün sol alt panelinde yer alan sonucu destekler niteliktedir ve yukarıda ifade edilen sapan

gözlem aday noktalarının veri setinden çıkarılması gerekmemektedir. Fakat daha önce de bahsedildiği gibi hala mevcut problemler devam etmektedir.

Böyle bir problem ortaya çıktığında, modelin esnek olmayan halde olabilmesi şüphesi ile doğrusal regresyon modelleri yerine nonparametrik regresyon modellerinin (toplamsal ve genelleştirilmiş toplamsal modeller) incelenmesi yararlı olacaktır.

Toplamsal Model

Elde edilen toplamsal model aşağıda verilmiştir:

$$death_i = \beta_0 + pm25median_i\beta_1 + f_1(time_i) + f_2(tmpd_i) + f_3(pm10median_i) + f_4(o3median_i) + \varepsilon_i \quad (5.12)$$

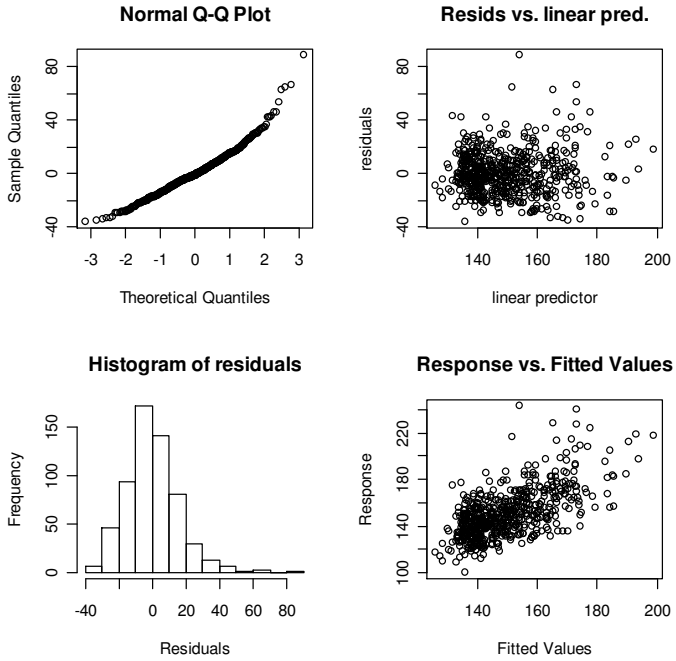
Bu model bir semiparametrik toplamsal modeldir. Modele ilişkin sonuçlar ise Tablo5.17'de yer almaktadır.

Tablo5.17. Semiparametrik toplamsal model için özet istatistikler

```
> summary(LA3)

Family: gaussian
Link function: identity

Formula:
death ~ s(time) + s(tmpd) + s(pm10median) + pm25median + s(o3median)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.60465    0.66348 225.485   <2e-16 ***
pm25median    0.20888    0.08692   2.403   0.0166 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank      F  p-value
s(time)       7.348    9.000  6.108 3.19e-08 ***
s(tmpd)       3.311    7.000  5.766 1.81e-06 ***
s(pm10median) 6.738    9.000  3.292 0.000632 ***
s(o3median)   3.015    7.000 10.203 4.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.381   Deviance explained = 40.4%
GCV score = 271.38   Scale est. = 261.14    n = 594

> AIC(LA3)
[1] 5015.326

> deviance(LA3)
[1] 149266.1
```

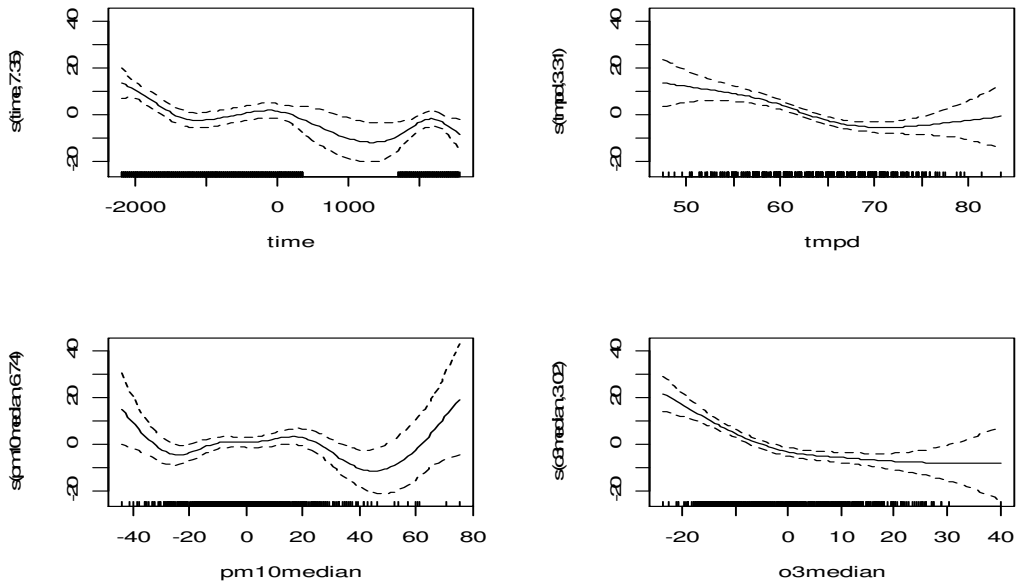
Tablo5.17 incelendiğinde kurulan semiparametrik toplamsal modelin genelleştirilmiş doğrusal modelden daha düşük AIC kriterine sahip olmasına karşın sapma değerinin genelleştirilmiş doğrusal modelin sapma değerinden oldukça yüksek olduğu görülmüştür.

Semiparametrik toplamsal modele ilişkin kontrol grafikleri ise Şekil5.5'de yer almaktadır ve bu kontrol grafikleri `R`'de `mgcv` paketinde yer alan `gam.check` komutu kullanılarak elde edilmiştir. Şekil5.5'in sol üst panelinde normal dağılım varsayımı için sağa çarpık (positive skew) bir durum (problemli bir durum) gözlenmektedir. Şekil5.5'in sol alt paneli de bu durumu destekleyen bir çizim sergilemektedir. Şekil5.5'in sağ üst paneli artıkların uyumu yapılan değerlere karşı (fitted value ya da linear predictor) bir çizimini göstermektedir ve bu grafikte sabit varyanslılık varsayımının tam olarak sağlanmadığı gözlenmektedir. Şekil5.5'in sağ alt panelinde ise yanıt değerlerinin uyumu yapılan değerlere (fitted value) karşı çizimi yer almaktadır ve bu grafikte de varyansın ortalamayla birlikte arttığı gözlenmektedir. Diğer bir ifadeyle, sabit varyanslılık varsayımı için problemli bir durum söz konusudur.



Şekil5.5. Semiparametrik toplamsal model için kontrol grafikleri.

Şekil5.6'da ise semiparametrik toplamsal modelden elde edilen pürüzsüz (smooth) fonksiyon tahminleri yer almaktadır.



Şekil5.6. Semiparametrik toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri

Mevcut problemlerin hala devam etmesi gerçeğinden ve ölüm değişkeninin poisson dağılan bir rassal değişken olması geleneksel yaklaşımından yola çıkarak, bu noktada, toplamsal modeller yerine genelleştirilmiş toplamsal modellerin kullanılmasına karar verilmiştir.

Genelleştirilmiş Toplamsal Model (GAM)

Bu bölümde kurulan modelde üstel aile dağılımı "Poisson", link fonksiyonu ise "log" alınarak, bir genelleştirilmiş toplamsal model kurulmuştur.

$$\begin{aligned}\log(E[death_i]) = {} & f_1(time_i) + f_2(tmpd_i) + f_3(pm10median_i) + \\ & f_4(pm25median_i) + f_5(o3median_i) + \\ & f_6(comedian_i) + \varepsilon_i \\ death_i \sim {} & Poi(E[death_i])\end{aligned} \tag{5.13}$$

Bu model ile ilgili özet istatistikler Tablo5.18'de görülmektedir. Bu tablo incelendiğinde, genelleştirilmiş toplamsal modelin AIC değerinin semiparametrik toplamsal modele göre çok küçük bir artışı gözlense de sapma değerinde gözle görülür bir düşüş gerçekleşmiştir. Elde edilen genelleştirilmiş toplamsal modelle

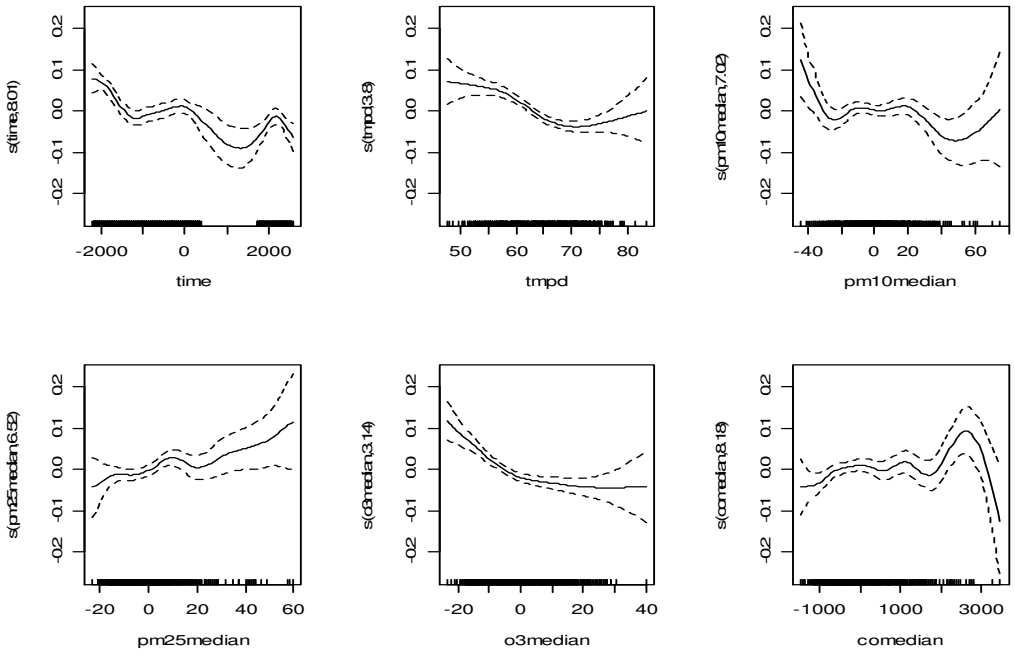
ilgili kontrol grafikleri, semiparametrik toplamsal modellin kontrol grafikleri ile hemen hemen aynı görüntüyü sergilemektedir. Bu modele ilişkin pürüzsüz fonksiyon tahminleri ise Şekil5.7'de yer almaktadır. Şekil5.7 incelendiğinde, tüm değişkenlerin yanıt üzerindeki etkisinin nonparametrik olduğu gözlenmektedir.

Tablo5.18. Genelleştirilmiş toplamsal model için özet istatistikler

```
> summary(LA4)

Family: poisson
Link function: log

Formula:
death ~ s(time) + s(tmpd) + s(pm10median) + s(pm25median) + s(o3median)
+
    s(comedian)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.003989   0.003367    1486   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank Chi.sq  p-value
s(time)       8.005    9.000  90.79 1.13e-15 ***
s(tmpd)       3.796    8.000  60.62 3.52e-10 ***
s(pm10median) 7.016    9.000  30.82 0.000318 ***
s(pm25median) 6.523    9.000  22.80 0.006671 **
s(o3median)   3.142    7.000  44.85 1.46e-07 ***
s(comedian)   8.182    9.000  29.92 0.000452 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.395   Deviance explained = 43.5%
UBRE score = 0.65883  Scale est. = 1         n = 594

> AIC(LA4)
[1] 5047.006

> deviance(LA4)
[1] 910.0135
```

Elde edilen bu genelleştirilmiş toplamsal modelin AIC değerindeki artış nedeniyle ve kontrol grafiklerinin hem semiparametrik toplamsal model için hem de genelleştirilmiş toplamsal model için hemen hemen aynı görüntüyü vermesi nedeniyle bu iki modelin hipotez testi ile karşılaştırılmasına karar verilmiştir. Hipotez testi sonuçları Tablo5.19'da yer almaktadır.

Tablo5.19 incelendiğinde, 2 numaralı modelin (genelleştirilmiş toplamsal modelin), istatistiksel olarak daha anlamlı sonuçlar verdiği, diğer bir ifadeyle yanıtı, semiparametrik toplamsal modele göre daha iyi açıkladığı görülmektedir.



Şekil5.7. Genelleştirilmiş toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri

Tablo5.19. Semiparametrik toplamsal ve genelleştirilmiş toplamsal model için hipotez testi

```
> anova(LA3,LA4.3,test="F")
Analysis of Deviance Table

Model 1: death ~ s(time) + s(tmpd) + s(pm10median) + pm25median +
s(o3median)
Model 2: death ~ s(time) + s(tmpd) + s(pm10median) +
s(pm25median) + s(o3median) +
    s(comedian)
  Resid. Df Resid. Dev      Df Deviance      F    Pr(>F)
1   571.589     149266
2   556.336        910  15.253   148356 9726.3 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

İnce Tabakalı Splayn (TPS) ile Oluşturulan Genelleştirilmiş Toplamsal Model

Bu aşamadan sonra, `pm10median`, `pm25median` ve `time` açıklayıcı değişkenlerini bir pürüzsüz fonksiyonda içeren, bununla birlikte `tmpd` ve `o3median` açıklayıcı değişkenlerini de diğer bir pürüzsüz fonksiyonda içeren bir ince tabakalı splayn modeli kurulmaya karar verilmiştir. Böyle bir karar verilmeden önce ince tabakalı splayn için bir çok alternatif model oluşturulmuş ve

oluşturulan tüm modeller R^2, AIC ve sapma değerlerine göre karşılaştırılarak, elde edilen modeller içinden bu kriterlere göre en iyi sonuçları veren, birinci ve ikinci terimi ince tabakalı splayn olan aşağıdaki genelleştirilmiş toplamsal model seçilmiştir:

$$\log(E[death_i]) = f_1(time_i, pm10median_i, pm25median_i) + f_2(tmpd_i, o3median_i) + f_3(comedian_i) + \varepsilon_i \qquad (5.14)$$
$$death_i \sim Poi(E[death_i])$$

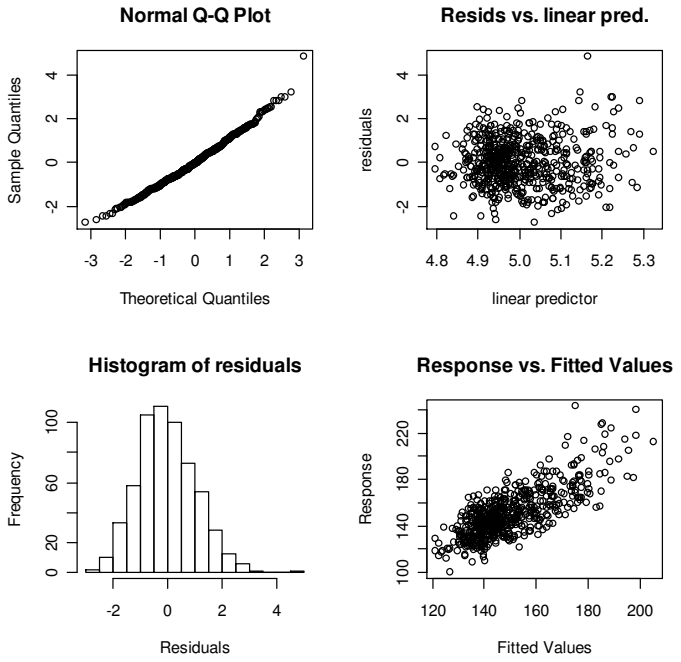
Bu modele ilişkin özet istatistikler Tablo5.20'de verilmiştir. Modelle ilgili kontrol grafikleri ise Şekil5.8'de yer almaktadır.

Tablo5.20. İnce tabakalı splayn modeli için özet istatistikler

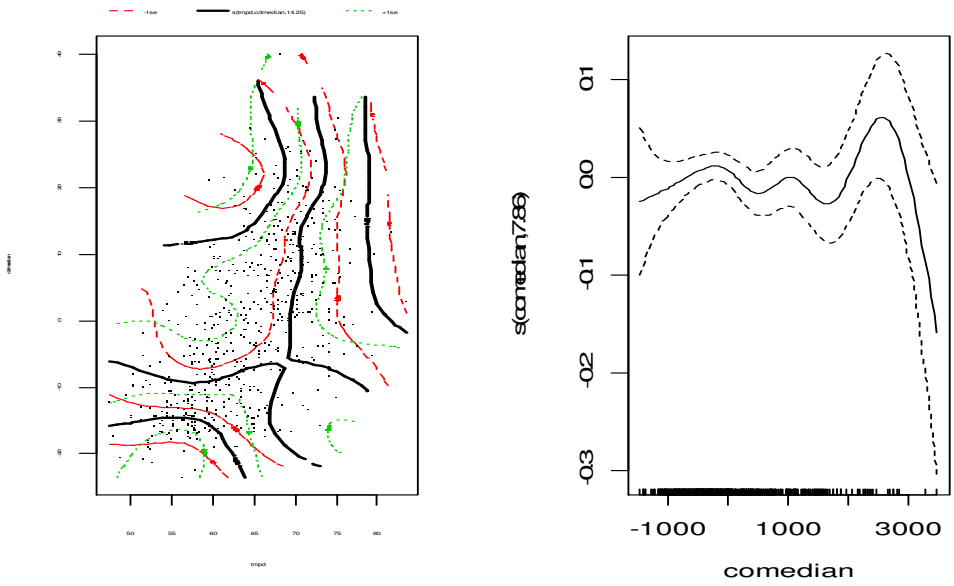
```
> summary(LA5)

Family: poisson
Link function: log

Formula:
death ~ s(time, pm10median, pm25median) + s(tmpd, o3median) +
    s(comedian)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 5.002970   0.003369    1485   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Est.rank Chi.sq  p-value
s(time,pm10median,pm25median) 74.202   65.000 254.88  < 2e-16 ***
s(tmpd,o3median)              14.250   29.000  76.06 4.31e-06 ***
s(comedian)                    7.864    9.000  22.98  0.00624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =    0.5   Deviance explained = 58.2%
UBRE score = 0.46138  Scale est. = 1         n = 594

> AIC(LA5)
[1] 4929.725

> deviance(LA5)
[1] 673.4297
```

Tablo5.20 incelendiğinde, *bu modele ait AIC ve sapma değerlerinin önceki tüm modellerden daha düşük olduğu gözlenmiştir*. Şekil5.8'de yer alan kontrol grafikleri incelendiğinde, *varsayımlarla ilgili problemlerin ortadan kalktığı* görünmektedir. Şimdiye kadar incelenen modellerin sonuçlarını karşılaştırarak, ince tabakalı splayn içeren son modelin diğer tüm modellerden daha iyi

performans gösterdiği sonucuna varılmıştır. *Diğer bir ifadeyle son model, Los Angeles hava kirliliği-ölüm verileri için daha uygun bir modeldir.* Bu modeldeki pürüzsüz (smooth) fonksiyonların tahminleri ise Şekil5.9'da yer almaktadır.



Şekil5.8. İnce tabakalı splayn içeren son model için kontrol grafikleri.



Şekil5.9. İnce tabakalı splayn içeren son modelden elde edilen pürüzsüz fonksiyon tahminleri

Yapılan uygulamada Los Angelos'da tesbit edilen hava kirliliğinin ölüm oranını nasıl etkilediğini ortaya çıkarmak amacıyla farklı regresyon modelleri kurulmuş ve onların performansları değerlendirilmiştir. Ele alınan regresyon modelleri sırasıyla doğrusal model, genelleştirilmiş doğrusal model (GLM), toplamsal modeller, genelleştirilmiş toplamsal model (GAM) ve ince tabakalı splayn (TPS) ile GAM'dir. İfade edilen bu modellerden, toplamsal model, genelleştirilmiş toplamsal model ve ince tabakalı splayn kullanılarak oluşturulan genelleştirilmiş toplamsal modellerde regresyon splaynı kullanılarak model uyumu yapılmıştır. Kurulan tüm modellerde model varsayımlarının sağlanıp sağlanmadığı incelenmiştir. Tüm modeller AIC ve sapma değerlerine göre karşılaştırıldığında, ince tabakalı splayn ile elde edilen GAM'e ait ait AIC ve sapma değerlerinin önceki tüm modellerden daha düşük olduğu gözlenmiştir. Diğer modellerde varsayımların tamamının sağlanmamasına karşın, bu modele ait kontrol grafikleri incelendiğinde, varsayımlarla ilgili problemlerin ortadan kalktığı görülmektedir. Bu durumda, ince tabakalı splayn içeren son modelin diğer tüm modellerden daha iyi performans gösterdiği, diğer bir ifadeyle son modelin, Los Angeles hava kirliliği-ölüm verileri için daha uygun bir model olduğu sonucuna varılmıştır.

Uygulama 2. Chicago'daki Hava Kirliliğinin İncelenmesi

Bu bölümde bir önceki bölüme benzer olarak, Chicago'daki hava kirliliği ile kaydedilen ölüm oranı arasındaki ilişki incelenmiştir ve kullanılan veriler www.ihapss.jhsph.edu adresinden alınmıştır. Dikkate alınan veriler, 01/01/1987-31/12/2000 yılları arasında, Chicago'da kaydedilen günlük ölüm oranı ve hava kalitesi değerleridir. Ölçümler her gün için yapılmıştır ve tıpkı Los Angeles verilerinde olduğu gibi, `agecat` değişkeni ile ayrılmış olan üç grup birleştirilerek 5114 veri dikkate alınmıştır. İlgilenilen yanıt değişkeni yıllar içerisinde Chicago'da meydana gelen günlük ölüm oranlarıdır (`death`–kazalar dışındaki tüm ölüm olayları). Gözlenen ölüm oranlarına karşılık gelen olası açıklayıcı değişkenler ise aşağıda verilmiştir:

`o3median` : ozon seviyeleri (PPB)

`so2median` : kükürt dioksit seviyeleri (PPB)

`pm10median`: partikül seviyeleri (mg/m^3)

`tmpd` : ortalama sıcaklık (F°)

Yukarıda ifade edilen hava kalitesi değişkenlerine ek olarak, Chicago verilerinde de ölüm oranı `time` değişkeni ile de değişme eğilimindedir.

Bu bölümde de hava kirliliğinin ölüm oranını nasıl etkilediğini ortaya çıkarmak amacıyla farklı regresyon modelleri kurulmuş ve onların performansları farklı kriterler göz önünde bulundurularak değerlendirilmiştir. Yapılan bu analizlerde de `R` programının temel paketlerinin yanı sıra `gamair` ve `mgcv` ve `NMMAPSdata` paketleri de kullanılmıştır. Ele alınan regresyon modelleri sırasıyla aşağıda ifade edilmiştir:

- Doğrusal Model
- Genelleştirilmiş Doğrusal Model (GLM)
- Toplamsal Modeller
    - **a.** Tam Toplamsal Model
    - **b.** Semiparametrik Toplamsal Model
- Genelleştirilmiş Toplamsal Model (GAM)
- İnce Tabakalı Splayn (TPS)

Şimdi yukarıda ifade etmiş olduğumuz modelleri tek tek inceleyelim:

Doğrusal Model

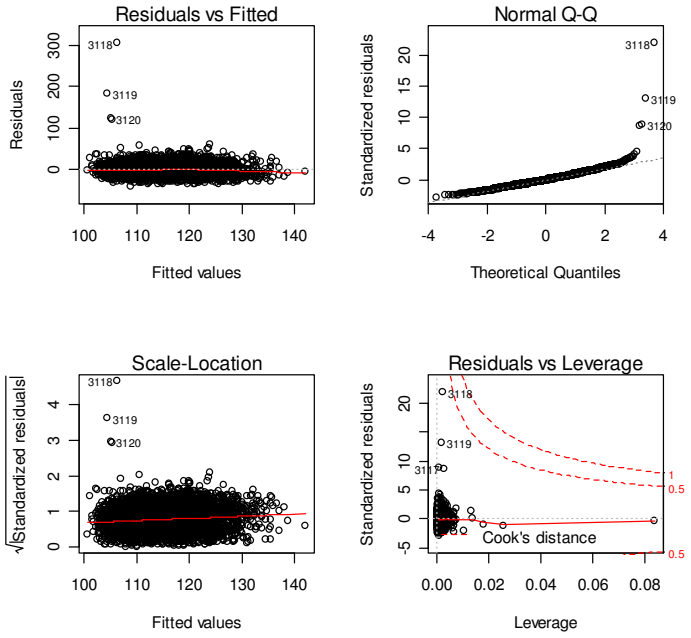
Tüm açıklayıcı değişkenlerin, ölüm oranı üzerinde doğrusal etkiye sahip olduğunu varsayarak, aşağıda ifade edilen doğrusal model kurulmuştur.

$$death_i = \beta_0 + time_i\beta_1 + o3median_i\beta_2 + so2median_i\beta_3 + pm10median_i\beta_4 + tmpd_i\beta_5 + \varepsilon_i \quad (5.15)$$

Bu modele ilişkin özet istatistikler Tablo5.21'de verilmiştir. Doğrusal modele ilişkin grafikler ise Şekil5.10'da gösterilmektedir.

Tablo5.21. Doğrusal model için özet istatistikler

```
> summary(ch1)

Call:
lm(formula = death ~ time + o3median + so2median + pm10median +
    tmpd, data = chicago)

Residuals:
     Min       1Q   Median       3Q      Max
-39.7253  -8.8785  -0.9023   7.7323 304.8217

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.323e+02  7.023e-01 188.394  < 2e-16 ***
time        -1.444e-03  1.379e-04 -10.474  < 2e-16 ***
o3median     5.882e-02  2.427e-02   2.424   0.0154 *
so2median    2.363e-01  7.843e-02   3.013   0.0026 **
pm10median   8.722e-02  1.283e-02   6.800 1.17e-11 ***
tmpd        -3.307e-01  1.298e-02 -25.472  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.95 on 4835 degrees of freedom
Multiple R-Squared:  0.17,      Adjusted R-squared: 0.1691
F-statistic:   198 on 5 and 4835 DF,  p-value: < 2.2e-16

> AIC(ch1)
[1] 39264.51

> deviance(ch1)
[1] 941222.5
```



Şekil5.10. Doğrusal model için kontrol grafikleri.

Şekil5.10'un sol üst panelinde bazı noktaların sıfır etrafındaki eşit saçılımı bozdukları gözlenmektedir, dolayısıyla problemli bir durum mevcuttur. Sabit varyans varsayımının kontrol edildiği Şekil5.10'un sol alt panelinde de sapan gözlemlerden kaynaklanan bazı problemler göze çarpmaktadır. Şekil5.10'un sağ üst panelinde yer alan normal Q-Q grafiğinin üst kuyruğunda doğrusallıktan sapma gözlenmektedir. Şekil5.10'un sağ alt panelinde yer alan Cook mesafesinin kontur grafiği üzerinde, noktaların hiç birinin net olarak kontur dışında (out of line) olmadığı görünmektedir.

Genelleştirilmiş Doğrusal Model (GLM)

Şekil5.10'un bir sonucu olarak ve daha önce Los Angeles örneğinde ölüm verileri için göz önünde bulundurulmuş olan varsayımdan hareketle, doğrusal model yerine genelleştirilmiş doğrusal modelin kullanılması daha uygun görünmektedir. Kurulmuş olan genelleştirilmiş doğrusal model aşağıda ifade edilmiştir.

$$\log(E[death_i]) = \beta_0 + time_i\beta_1 + o3median_i\beta_2 + so2median_i\beta_3 + pm10median_i\beta_4 + tmpd_i\beta_5 + \varepsilon_i \qquad (5.16)$$
$$death_i \sim Poi(E[death_i])$$

Bu modele ilişkin sonuçlar Tablo5.22'de yer almaktadır.

Yukarıda kurulmuş olan doğrusal ve genelleştirilmiş doğrusal modeller "AIC" ve "Sapma" kriterlerine göre karşılaştırıldığında, genelleştirilmiş doğrusal modelde özellikle sapma değerinde büyük bir düşüş gözlenmektedir. Fakat genelleştirilmiş doğrusal model için kontrol grafikleri, Şekil5.10'da doğrusal model için ifade edilmiş olan grafiklerle hemen hemen aynı görüntüye sahiptir. Dolayısıyla, genelleştirilmiş doğrusal modelde de mevcut problemlerin hala devam ettiği ve birbirine çok yakın olan 4 sapan gözlemin de hala var olduğu görülmektedir.

Veriler daha ayrıntılı incelendiğinde, sapan gözlemlerin verilerde meydana gelen günlük en yüksek ölüm oranları olduğu ve bunların ardışık günlerde meydana geldiği gözlenmiştir. Ölüm oranındaki bu yükselme, gözlenen oldukça yüksek sıcaklık ve ozon dönemi ile ilişkilidir. Aşağıda söz konusu durum gösterilmiştir:

```
> chicago$death[3114:3123]
 [1] 119 116 121 226 411 287 228 159 142 123
> chicago$o3median[3114:3123]
 [1] 18.1413783 24.1604249 36.7455450 29.7035454 28.1150909 21.1150086
 [7]  5.6497315  2.4096013  0.4478698 -5.2211149
> chicago$tmpd[3114:3123]
 [1] 79.0 84.5 92.0 91.5 86.0 83.0 78.5 74.0 75.5 73.5
```

Tablo5.22. Genelleştirilmiş doğrusal model için özet istatistikler

```
> summary(ch2)

Call:
glm(formula = death ~ time + o3median + so2median + pm10median +
    tmpd, family = poisson, data = chicago)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-4.09431  -0.84778  -0.08558   0.72004  22.39571

Coefficients:
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)  4.892e+00  4.583e-03 1067.449  < 2e-16 ***
time        -1.251e-05  9.204e-07  -13.594  < 2e-16 ***
o3median     4.870e-04  1.621e-04    3.004 0.002668 **
so2median    1.985e-03  5.134e-04    3.866 0.000111 ***
pm10median   7.453e-04  8.442e-05    8.828  < 2e-16 ***
tmpd        -2.831e-03  8.558e-05  -33.081  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9340.7  on 4840  degrees of freedom
Residual deviance: 7681.5  on 4835  degrees of freedom
AIC: 39542

Number of Fisher Scoring iterations: 4
```

Böyle bir problem ortaya çıktığında, model son derece esnek olmayan (inflexible) halde olabilir ve ölüm oranının sıcaklık ve ozonla ilgili bazı doğrusal olmayan yanıtlarına ihtiyaç duyulabilir. Bu durumda doğrusal regresyon modelleri yerine nonparametrik regresyon modellerinin incelenmesi yararlı olacaktır.

Toplamsal Modeller

İlk olarak tam toplamsal (tüm açıklayıcı değişkenler nonparametrik alınarak) model incelenmiştir. Daha sonra, etkilerinin doğrusal olduğu gözlenen değişkenler modelde parametrik alınarak semiparametrik toplamsal model oluşturulmuştur.

a) *Tam Toplamsal Model*

Kurulmuş olan tam toplamsal model aşağıda verilmiştir:

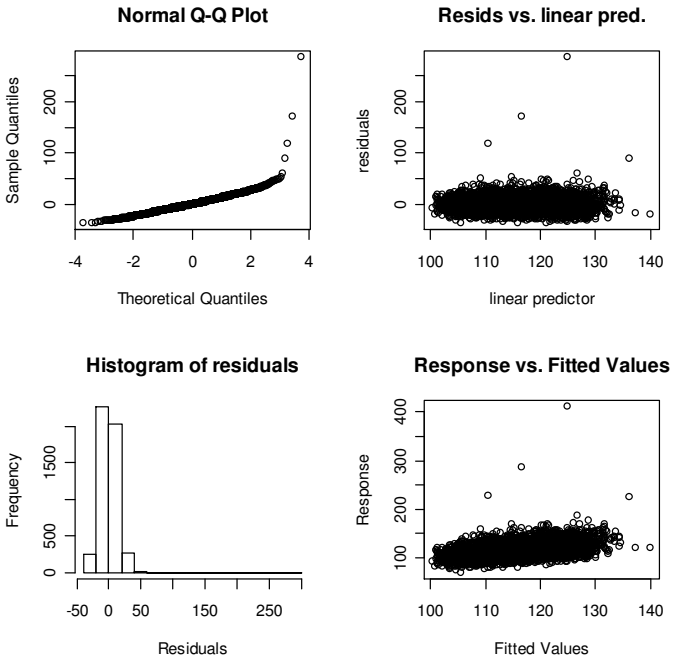
$$death_i = \beta_0 + f(time_i) + f(o3median_i) + f(so2median_i) + f(pm10median_i) + f(tmpd_i) + \varepsilon_i \quad (5.17)$$

Modele ilişkin sonuçlar ise Tablo5.23'de yer almaktadır. Bu tablo incelendiğinde kurulan tam toplamsal modelin genelleştirilmiş doğrusal modelden daha düşük AIC kriterine sahip olmasına karşın sapma değerinin genelleştirilmiş doğrusal modelin sapma değerinden oldukça yüksek olduğu görülmüştür.

Tam toplamsal modele ilişkin kontrol grafikleri ise Şekil5.11'de yer almaktadır. Tam toplamsal modele ait kontrol grafiklerinin doğrusal ve genelleştirilmiş doğrusal modeller ile hemen hemen aynı görünümde olduğu gözlenmiştir.

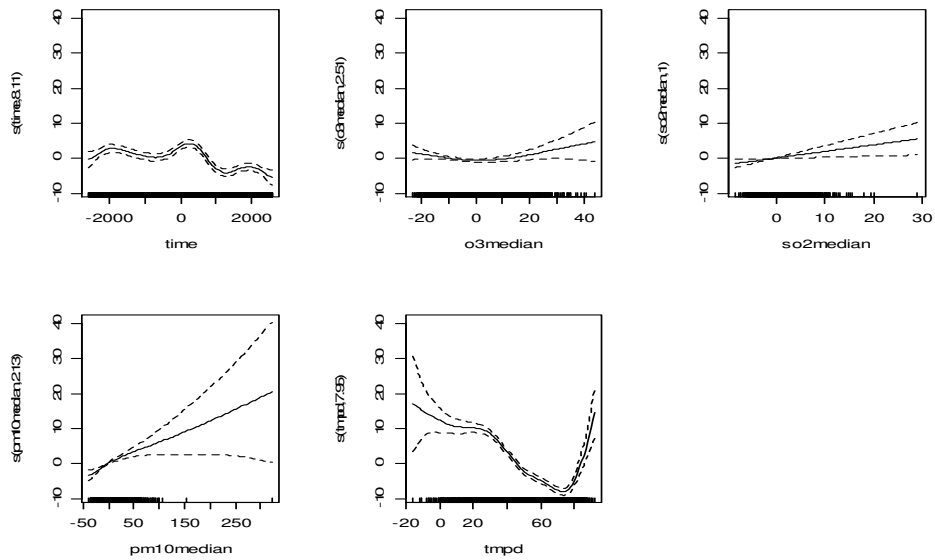
Tablo5.23. Tam toplamsal model için özet istatistikler

```
> summary(ch3)

Family: gaussian
Link function: identity

Formula:
death ~ s(time) + s(o3median) + s(so2median) + s(pm10median) +
    s(tmpd)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  115.330      0.195   591.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank      F  p-value
s(time)       8.114    9.000 22.168  < 2e-16 ***
s(o3median)   2.508    6.000  2.971  0.00676 **
s(so2median)  1.000    1.000  5.929  0.01493 *
s(pm10median) 2.129    5.000  7.511 4.93e-07 ***
s(tmpd)       7.948    9.000 89.976  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.214   Deviance explained = 21.8%
GCV score = 184.91   Scale est. = 184.05    n = 4841

> AIC(ch3)
[1] 39009.52

> deviance(ch3)
[1] 886787.7
```



Şekil5.11. Tam toplamsal model için kontrol grafikleri.

Şekil5.12'de ise tam toplamsal modelden elde edilen pürüzsüz (smooth) fonksiyon tahminleri yer almaktadır. Bu grafikler incelendiğinde `so2median` değişkeninin ölüm oranı üzerinde doğrusal bir etkiye sahip olabileceği görülmektedir. Bu nedenle `so2median` modelde parametrik olarak alınmalıdır.



Şekil5.12. Tam toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri

Diğer taraftan pm10median ve so2median değerlerinin dağılımı ile ilgili bir problem gözlenmektedir (bu probleme daha sonraki aşamalarda değinilecektir).

b) *Semiparametrik Toplamsal Model*

Yukarıda ifade edildiği gibi, tam toplamsal modelde so2median değişkeninin etkisinin doğrusal olarak ortaya çıkması nedeniyle, bu değişken parametrik olarak alınarak semiparametrik toplamsal model oluşturulmuştur. Kurulan semiparametrik model,

$$death_i = \beta_0 + so2median_i\beta_1 + f(time_i) + f(o3median_i) + f(pm10median_i) + f(tmpd_i) + \varepsilon_i \quad (5.18)$$

ve bu modele ilişkin özet istatistikler Tablo5.24'de verilmiştir.

Tablo5.24. Semiparametrik toplamsal model için özet istatistikler

```
> summary(ch4)

Family: gaussian
Link function: identity

Formula:
death ~ s(time) + s(o3median) + so2median + s(pm10median) + s(tmpd)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.44866    0.20099 574.398   <2e-16 ***
so2median     0.18852    0.07742   2.435   0.0149 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank      F  p-value
s(time)       8.114    9.000 22.168  < 2e-16 ***
s(o3median)   2.508    6.000  2.971  0.00676 **
s(pm10median) 2.129    5.000  7.511 4.93e-07 ***
s(tmpd)       7.948    9.000 89.976  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.214   Deviance explained = 21.8%
GCV score = 184.91   Scale est. = 184.05    n = 4841

> AIC(ch4)
[1] 39009.52

> deviance(ch4)
[1] 886787.9
```

Tablo5.24 incelendiğinde, tam toplamsal model ile semiparametrik toplamsal modellerin AIC ve sapma değerlerinde herhangi bir değişme olmadığı

ve bununla birlikte her iki modele ilişkin tahminlenen serbestlik dereceleri, F değerleri ve karşılık gelen olasılık değerlerinin hemen hemen aynı olduğu gözlenmiştir. Kontrol grafikleri de aynı görünümü sergilemektedir. Bu noktada, `pm10median` değişkeninin de parametrik olabileceği şüphesinden yola çıkarak bir diğer semiparametrik toplamsal regresyon modeli kurulmuştur. Fakat bu modelin diğer modelden daha kötü sonuçlar verdiği gözlenmiştir ve burada bu modele yer verilmemiştir.

Mevcut problemlerin hala devam etmesi gerçeğinden ve ölüm değişkeninin poisson dağılan bir rassal değişken olması geleneksel yaklaşımından yola çıkarak, bu noktada, toplamsal modeller yerine genelleştirilmiş toplamsal modellerin kullanılmasına karar verilmiştir.

Genelleştirilmiş Toplamsal Model (GAM)

Bu bölümde kurulan modelde üstel aile dağılımı "Poisson", link fonksiyonu ise "log" alınarak, genelleştirilmiş toplamsal model kurulmuştur.

$$\log(E[death_i]) = f_1(time_i) + f_2(o3median_i) + f_3(so2median_i) + f_4(pm10median_i) + f_5(tmpd_i) + \varepsilon_i \qquad (5.19)$$
$$death_i \sim Poi(E[death_i])$$

Bu model ile ilgili özet istatistikler Tablo5.25'de görülmektedir. Bu tablo incelendiğinde, genelleştirilmiş toplamsal modelin AIC değerinin, tam toplamsal ve semiparametrik toplamsal modellere göre bir artışı gözlense de sapma değerinde gözle görülür bir düşüş gerçekleşmiştir. Kurmuş olduğumuz genelleştirilmiş toplamsal modelle ilgili kontrol grafikleri, tam toplamsal ve semiparametrik toplamsal modellerin kontrol grafikleri ile hemen hemen aynı görüntüyü sergilemektedir. Bu modele ilişkin pürüzsüz fonksiyon tahminleri ise Şekil5.13'de yer almaktadır. Şekil5.13 incelendiğinde, tüm değişkenlerin yanıt üzerindeki etkisinin nonparametrik olduğu gözlenmektedir. Diğer taftan `pm10median` ve `so2median` değerlerinin dağılımı ile ilgili problem Şekil5.13'de dikkatimizi çekmektedir.

Tüm değişkenlerin yanıt üzerinde nonparametrik etkiye sahip olduğunun gözlenmesi nedeniyle, bu örnek uygulamada genelleştirilmiş semiparametrik toplamsal model incelenmemiştir. Fakat hala daha önce bahsedilen dört sapan

gözlemin olması, diğer taraftan so2median ve pm10median değerlerinin dağılımlarının bir problem içermesi sorunları çözülememiştir.

Tablo5.25. Genelleştirilmiş toplamsal model için özet istatistikler

```
> summary(ch5)

Family: poisson
Link function: log

Formula:
death ~ s(time) + s(o3median) + s(so2median) + s(pm10median) +
    s(tmpd)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.745893   0.001341    3540   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
               edf Est.rank  Chi.sq  p-value
s(time)       8.427    9.000  320.30  < 2e-16 ***
s(o3median)   7.837    9.000   37.64 2.02e-05 ***
s(so2median)  3.648    8.000   16.42   0.0368 *
s(pm10median) 8.132    9.000   69.79 1.68e-11 ***
s(tmpd)       8.191    9.000 1239.82  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.215   Deviance explained = 23.2%
UBRE score = 0.49655  Scale est. = 1         n = 4841

> AIC(ch5)
[1] 39093.68

> deviance(ch5)
[1] 7170.341
```

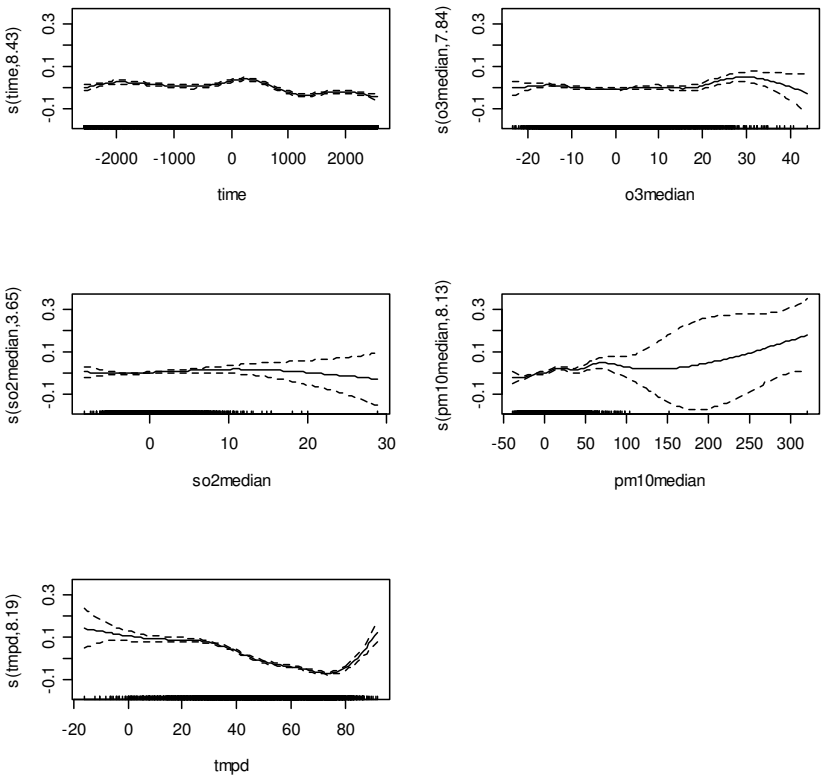
Bahsedilen bu 4 günlük ölüm dalgalanması etrafındaki verilerin daha detaylı olarak incelenmesi yapıldığında görülmüştür ki, yüksek sıcaklıklar yüksek ölüm oranlarından önceki birkaç günde kaydedilmiştir ve aynı zamanda bu dönemde yüksek ozon seviyeleri de gözlenmiştir.

Yüksek ölüm olayında, kestirici değişken için uygun birleştirme (aggregations) ve gecikme (lags) işlemleri önerilebilir. Hava kalitesi değişken değerleri farklı yollarla birleştirilebilir. İncelenen problemde sorunlu gündeki ve öncesindeki üç gündeki değişken değerlerinin toplamının alınması önerilmiştir (Wood, 2006a):

```
lag.sum<-function(a,l0,l1)
## l0 is the smallest lag, l1 the largest
{      n<-length(a)
       b<-rep(0,n-l1)
```

```
        for (i in 0:)(l0-l1)) b<-b+a[(i+1):(n-l1+i)]
        b
}
death<-chicago$death[4:5114]
time<-chicago$time[4:5114]
o3<-lag.sum(chicago$o3median,0,3)
tmp<-lag.sum(chicago$tmpd,0,3)
pm10<-lag.sum(log(chicago$pm10median+40),0,3)
so2<-lag.sum(log(chicago$so2median+10),0,3)
```

Verilerin ayrıntılı incelenmesi göstermiştir ki, ozon ve sıcaklığın her ikisinin birlikte aşırı artması yüksek ölüm oranlarına sebep olabilmektedir. Bu nedenle, bu iki açıklayıcı değişkeni bir pürüzsüz fonksiyonda içeren, bir ince tabakalı splayn modeli kurulmaya karar verilmiştir.



Şekil5.13. Genelleştirilmiş toplamsal modelden elde edilen pürüzsüz fonksiyon tahminleri

İnce Tabakalı Splayn (TPS)

İkinci terimi ince tabakalı splayn olan aşağıdaki genelleştirilmiş toplamsal modeli göz önüne alalım:

$$\log(E[death_i]) = f_1(time_i) + f_2(o3_i, tmp_i) + f_3(so2) + f_4(pm10) + \varepsilon_i$$
$$death_i \sim Poi(E[death_i]) \quad (5.20)$$

Bu modele ilişkin özet istatistikler incelendiğinde, so2 değişkeninin ölüm oranı üzerinde (hem doğrusal ve hem de doğrusal olmayan) anlamlı bir etkisi olmadığı gözlenmiştir. Bu nedenle so2 değişkeninin modelden çıkarılmasına karar verilmiştir. Kurulan yeni model,

$$\log(E[death_i]) = f_1(time_i) + f_2(o3_i, tmp_i) + f_3(pm10) + \varepsilon_i$$
$$death_i \sim Poi(E[death_i]) \quad (5.21)$$

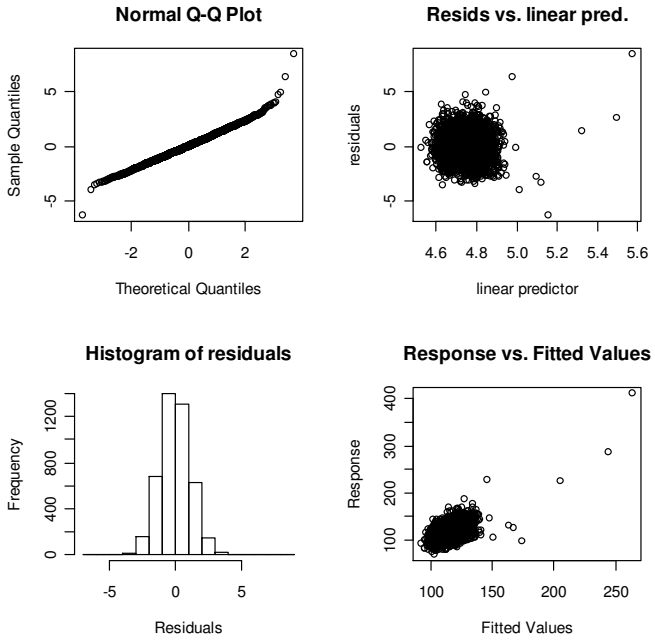
şeklinde ifade edilir. Bu modele ilişkin özet istatistikler Tablo5.26'de verilmiştir. Modelle ilgili kontrol grafikleri ise Şekil5.14'de yer almaktadır.

Tablo5.26. Yeni ince tabakalı splayn modeli için özet istatistikler

```
> summary(chi35)

Family: poisson
Link function: log

Formula:
death ~ s(time) + te(o3, tmp) + s(pm10)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.742981   0.001415    3352   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Est.rank Chi.sq  p-value
s(time)     8.169    9.000 262.84  < 2e-16 ***
te(o3,tmp) 23.831   15.000 683.80  < 2e-16 ***
s(pm10)     5.983    9.000  80.08 1.56e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.34   Deviance explained = 33.5%
UBRE score = 0.3346  Scale est. = 1         n = 4362

> AIC(chi35)
[1] 34509.26

> deviance(chi35)
[1] 5743.566
```

Tablo5.26 incelendiğinde, bu modele ait AIC ve sapma değerlerinin önceki tüm modellerden daha düşük olduğu gözlenmiştir. Şekil5.14'de yer alan kontrol grafikleri incelendiğinde, varsayımlarla ilgili problemlerin ortadan kalktığı görünmektedir. Şimdiye kadar incelenen modellerin sonuçlarını karşılaştırarak, ince tabakalı splayn içeren son modelin diğer tüm modellerden daha iyi

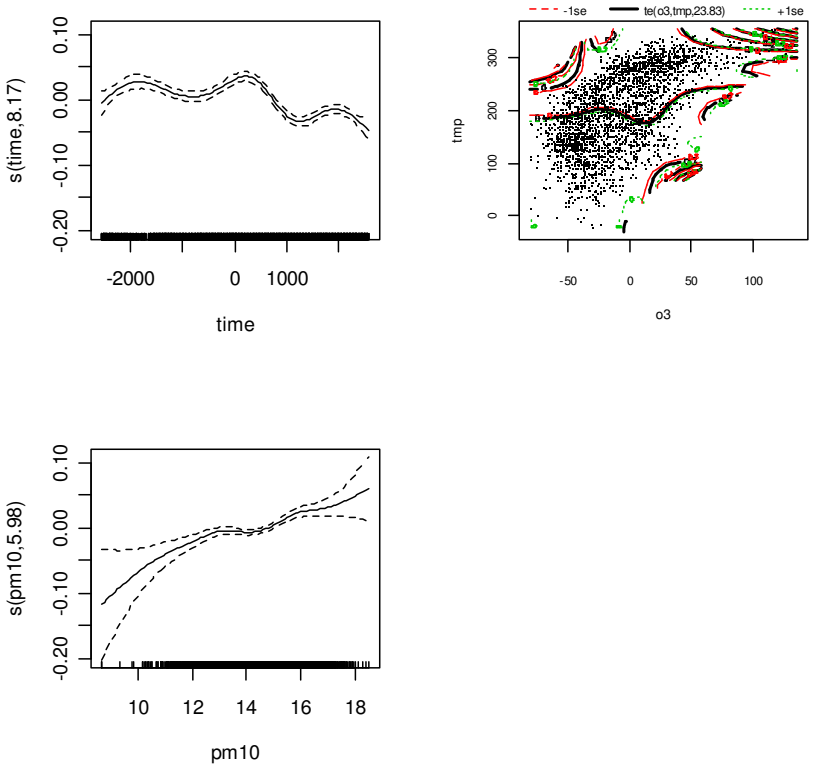
performans gösterdiği sonucuna varılmıştır. Diğer bir ifadeyle son model, Chicago hava kirliliği-ölüm verileri için daha uygun bir modeldir. Bu modeldeki pürüzsüz (smooth) fonksiyonların tahminleri işe Şekil5.15'de yer almaktadır.



Şekil5.14. İnce tabakalı splayn içeren son model için kontrol grafikleri.

Yapılan uygulamada Chicago'daki hava kirliliğinin ölüm oranını nasıl etkilediğini ortaya çıkarmak amacıyla farklı regresyon modelleri kurulmuş ve onların performansları değerlendirilmiştir. Ele alınan regresyon modelleri bir önceki uygulamada olduğu gibi, sırasıyla doğrusal model, genelleştirilmiş doğrusal model (GLM), toplamsal modeller, genelleştirilmiş toplamsal model (GAM) ve ince tabakalı splayn (TPS) ile GAM'dir. İfade edilen bu modellerden, toplamsal model, genelleştirilmiş toplamsal model ve ince tabakalı splayn kullanılarak oluşturulan genelleştirilmiş toplamsal modellerde regresyon splaynı kullanılarak model uyumu yapılmıştır. Kurulan tüm modellerde model varsayımlarının sağlanıp sağlanmadığı incelenmiştir ve bir sapan gözlem problemi saptanmıştır. Veriler daha ayrıntılı olarak incelenerek ve uygun birleştirme işlemi yapılarak problem aşılmıştır. Tüm modeller AIC ve sapma değerlerine göre karşılaştırıldığında, ince tabakalı splayn ile elde edilen GAM'e ait AIC ve sapma değerlerinin önceki tüm modellerden daha düşük olduğu gözlenmiştir. Diğer

modellerde varsayımların tamamının sağlanmamasına karşın, bu modele ait kontrol grafikleri incelendiğinde, varsayımlarla ilgili problemlerin ortadan kalktığı görülmektedir. Bu durumda, ince tabakalı splayn içeren son modelin diğer tüm modellerden daha iyi performans gösterdiği, ve Chicago hava kirliliği-ölüm verileri için daha uygun bir model olduğu sonucuna varılmıştır.



Şekil5.15. İnce tabakalı splayn içeren son modelden elde edilen pürüzsüz fonksiyon tahminleri

5.2 alt bölümü için SONUÇ: Bu alt bölümde hava kirliliğinin ölüm oranına etkisinin analizi için iki farklı veri seti ele alınmıştır. Her iki veri seti için ince tabakalı splayn içeren genelleştirilmiş toplamasal bir uygun regresyon modelinin, elde edilen modeller içinde en iyi sonuçları verdiği gözlemlenmiştir. ABD' in diğer şehirleri için de aynı problem incelendiğinde benzer sonuçlar elde edilmiştir. Bu yanıt değişkeninin açıklayıcı değişkenlerle karmaşık bağıntısı olan problemler için genelleştirilmiş regresyon modellerinin ve regresyon splaynı yönteminin ne derecede önem taşıdığını açıkça göstermektedir.

6. SONUÇ VE ÖNERİLER

Bu tez çalışmasında regresyonda pürüzlülük ceza yaklaşımı için splayn düzeltme ve regresyon splaynı yöntemleri incelenmiştir. Toplamsal modellerde splayn düzeltme yönteminin temelini cezalı en küçük kareler regresyonu oluşturur. Cezalı en küçük kareler yönteminde, sıradan en küçük karelerden farklı olarak hata kareler toplamına bir ceza fonksiyonu eklenir ve bu fonksiyon sayesinde tamamıyla esnek eğimli uyumlar ve sabit eğimli uyumlar arasında bir uzlaşma sağlanmış olur. Teorem 3.1'e göre, (3.2) cezalı hata kareler toplamını minimum yapan bu fonksiyon $t_1, t_2, ..., t_n$ düğüm noktaları ile bir doğal kübik splayndır ve gözlem noktalarının tamamı düğüm noktası olarak dikkate alınmaktadır. Doğal kübik splayn düzeltmeyi belirleyen **f** vektörünün (3.3) ifadesi ile veya toplamsal modellerde (3.28) denklemini doğrudan çözerek hesaplanması pratik açıdan elverişli değildir ve bu nedenle farklı nümerik yöntemlerin kullanımı faydalıdır: Backfitting algoritması, direkt metot, Speckman algoritması.

Pürüzlülük ceza yaklaşımı uygulanan regresyon modellerinde, iyi (optimum) düzeltme parametresinin seçimi önemli konulardan biridir. λ parametresinin optimum seçimi $\hat{f}$ tahmin fonksiyonunun, gerçek *f* fonksiyonuna mümkün olduğu kadar yakın olmasını sağlar. Düzeltme parametresinin seçimi için çeşitli seçim kriterleri dikkate alınabilir: CV kriteri, GCV kriteri, Mallows'un C_p kriteri, AIC kriteri ve AIC_C kriteri.

Düzeltme parametresinin otomatik seçiminin zor olduğu durumlarda (örneğin, splayn düzeltme ile elde edilen büyük boyutlu toplamsal modellerde) kullanılabilecek bir yöntem uygun serbestlik derecesinin seçimidir. Nonparametrik regresyonda tanımlanan serbestlik derecesi kavramları doğrusal regresyona benzer olarak verilebilir. Bu açıdan (3.45), (3.46) ve (3.47) denklemleri ile üç farklı serbestlik derecesi tanımı verilmiştir. İfade edilen bu serbestlik derecelerinden herhangi birisi düzeltme parametresinin değerinin belirlenmesinde uygulanabilir. Hesaplama açısından, $\mathbf{S}_\lambda$ matrisinin köşegen elemanlarının toplamı olan $df_3(\lambda) = tr(\mathbf{S}_\lambda)$ diğerlerine göre avantaja sahiptir.

Pürüzlülük ceza yaklaşımı için kullanılan regresyon splaynında, splayn düzeltme yönteminden farklı olarak bütün gözlem değerlerinin düğüm noktası

olarak seçilmesi yerine gözlem sayısından daha az düğüm noktası belirlenmektedir. Diğer bir ifadeyle, taban fonksiyonlar ile ifade edilen splayn fonksiyonlarında bu taban fonksiyonların sayısı gözlem sayısından çok az olabilir. Regresyon splaynında çok az düğüm noktasının kullanımı, hesaplamaların önemli ölçüde kolaylaşmasına ve modelin tahmin açısından daha da esnek olmasına neden olur. En popüler taban fonksiyonlarından biri kübik splayn tabandır ve farklı alternatif kübik splayn tabanlar mevcuttur.

Cezalı regresyon splaynı uygulanarak incelenen genelleştirilmiş toplamsal modellerde, düzeltici f_i fonksiyonlarının kestirimi için cezalı log-olabilirlik (log-likelihood) yöntemi kullanılabilir. Negatif cezalı log-olabilirliğin minimizasyonu problemi, ***IRLS*** *(Iteratively Re-weighted Least Squares)* algoritması yardımı ile gerçekleştirilebilir.

Genelleştirilmiş toplamsal modellerde düzeltme parametreleri sıfıra eşitlendiğinde serbestlik derecesinin, $\boldsymbol{\beta}$'nın boyutuna eşit olduğu açıktır. Diğer ekstrem durumda, yani düzeltme parametreleri çok büyük olduğunda, model aşırı esnektir ve bu nedenle de farklı serbestlik derecesine sahiptir. Kurulmuş esnek modelin serbestlik derecesinin ölçülmesi için bir yol, *etkin serbestlik derecesinin* $tr(\mathbf{A})$ olarak hesaplanmasıdır, burada $\mathbf{A}$ şapka matristir. Serbestlik derecesinin belirlenmesi için bir diğer yaklaşım $\boldsymbol{\beta}$ vektörünün her bir bileşeni için cezayı göz önüne almaktır, yani ayrı ayrı her bir düzeltme fonksiyonu için serbestlik derecesini kullanmaktır. Diğer taraftan, düzeltme parametresinin seçimi için etkili bir yaklaşım GCV kullanılarak verilebilir.

İnce tabakalı splayn (TPS), belirli bir açıklayıcı değişken grubunun fonksiyonunun, yani birden fazla değişkeni olan regresyon fonksiyonunun tahmini için kullanılan çok değişkenli bir splayndır. TPS yanıt değişkeni açıklayıcı değişkenlerin bileşiğinin etkilediği problemler için çok önemli modeldir. Bu türlü problemlere örnek olarak hava kirliliğinin ölüme etkisi verilebilir (bkz. bölüm 5). TPS bir değişkenli splayn fonksiyonların tenzor çarpımı olarak tanımlandığında, tenzor taban splayınlar kullanılarak, hesaplamalar kolaylaştırılabilir.

Tez çalışmasının 5. bölümünde yer alan uygulamaların ilk kısmında, kiralık ve satılık evlerin fiyatlarının, söz konusu evlerin özelliklerinden nasıl etkilendiğini ortaya koymak için kurulan regresyon modellerinde splayn düzeltme

yöntemi kullanılmıştır. Yapılan uygulamalarda sırasıyla, *i*) Kanada'daki evlerin özelliklerinin satış fiyatları üzerindeki etkisi, *ii*) Eskişehir'deki evlerin özelliklerinin satış fiyatları üzerindeki etkisi, *iii*) Eskişehir'deki evlerin özelliklerinin kira fiyatları üzerindeki etkisi splayn düzeltme ile incelenmiştir. Yapılan örneklerin üçünde de, nonparametrik regresyon modellerinin elde edilmesinde kullanılan splayn düzeltme yönteminin, düzeltme parametresi ve düzeltme matrisini bulundurması nedeniyle, sıradan en küçük kareler regresyonu modelinden çok daha iyi sonuçlar verdiği gösterilmiştir. Bu uygulamalarda, evlerin satış/kira fiyatları ile evlerin özellikleri arasındaki ilişkiler, genel olarak parametrik regresyon, semiparametrik regresyon, semiparametrik toplamsal regresyon ve toplamsal regresyon modelleri kullanılarak analiz edilmiştir. Analizler sonucunda, hem parametrik doğrusal bileşenleri hem de bir kaç nonparametrik bileşeni bulunduran uygun semiparametrik (toplamsal) regresyon modelin diğer modellerden daha iyi sonuçlar verdiği gözlenmiştir.

Dünyada her yıl hava kirliliğinden 3 milyon insan ölmektedir. Bu değer dünyadaki toplam ölümün (ortalama 55 milyon) %5'ni oluşturmaktadır ve ölümlerin %90'nı gelişmekte olan ülkelerde görülmektedir. Bu gerçeklerden yola çıkarak, 5. bölümün ikinci kısmında hava kirliliğinin ölüm oranı üzerindeki etkisi *kübik regresyon splaynı* kullanılarak araştırılmıştır. Hava kirleticilerinin, diğer bir ifadeyle hava kalitesini belirleyen maddelerin (değişkenlerin) temelini, Kükürt Dioksit (SO_2), Partikül Madde (PM10, PM2.5), Karbon Monoksit (CO), Ozon (O_3) maddeleri teşkil ediyor. Yapılan ikinci kısım uygulamalarda, havayı kirleten maddelerin ölüm oranı üzerindeki etkileri incelenmiştir. Bu amaçla tezde iki uygulama çalışmasının sonuçları verilmiştir: Birinci problem Los Angeles'daki hava kirliliğinin ölüm oranı üzerindeki etkisi; ikinci ise Chicago'daki hava kirliliğinin ölüm oranı üzerindeki etkisi regresyon splaynı kullanılarak ortaya konmuştur. Her iki problem için, genelleştirilmiş doğrusal (GLM), toplamsal, genelleştirilmiş toplamsal (GAM) ve ince tabakalı splayn (TPS) ile GAM modelleri incelenmiştir. Tüm modeller AIC ve sapma değerlerine göre karşılaştırılmış ve bunun yanı sıra kurulan her model için varsayımların sağlanıp sağlanmadığı tek tek incelenmiştir. İnce tabakalı splayn bileşenler içeren uygun bir model diğer tüm modellerden daha iyi performans göstermiştir. Bu nedenle,

hava kirliliği-ölüm problemi için ince tabakalı splayn bileşenler içeren GAM modellerinin en uygun model olduğu sonucuna varılmıştır.

Son olarak, aşağıdaki önerilerimiz yazılabilir:

- Bazı açıklayıcı değişkenlerin yanıt değişkenini doğrusal etkilemediği belirlenen problemlerde pürüzlülük ceza yaklaşımı ile splayn regresyonunun uygulanması çok yararlıdır;
- Veri seti çok büyük olmadığında, splayn düzeltme yöntemi ve model tahmini için backfitting algoritması kullanılabilir;
- Problemde yanıta doğrusal etkisi olan ve doğrusal etkisi olmayan açıklayıcı değişkenlerin belirlenmesi çok önemlidir ve bu durumda uygun semiparametrik regresyon modelinin kullanımı en iyi sonucu verir;
- Veri seti büyük ve çok boyutlu olan problemlerde regresyon splaynının kullanımı daha uygundur;
- GAM'da problemin doğasına bağlı olarak ince tabakalı splayn bileşenlerin seçimi, en iyi modelin belirlenmesinde büyük öneme sahiptir;
- Pürüzlülülük ceza yaklaşımında iyi bir model, optimum düzeltme parametresi veya uygun serbestlik derecesinin seçimi ile belirlenir.
- Bu tez çalışması splayn düzeltme ve regresyon splaynı yöntemleri ile regresyonda pürüzlülük ceza yaklaşımını öğrenmek isteyenler için rehber rol oynamakta ve bu konuda temel bilgileri kapsamaktadır.

KAYNAKLAR

Aydın, D. (2005), *Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı ile Tahmin ve Çıkarsama*, Doktora Tezi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

Aydın D., Omay R.E. (2006), "The Empirical Performances of the Selection Criteria for Nonparametric Regression Using Smoothing Spline", *The 5th WSEAS International Conference on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS AND CYBERNETICS*, Venice, Italy, November 20-22.

Aydın D., Omay R.E. (2007), "The Smoothing Parameter Selection Problem in Smoothing Spline Regression for Different Data Sets", *WSEAS Transactions on Mathematics*, **6(3)**, 477-482.

Breiman, L., Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion)", *J. Amer. Statist. Assoc.* **80**, 580-619.

Björk, A. (1996), *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics, Philadelphia, U.S.A..

Buja, A., Hastie, T., Tibshirani, R. (1989), "Linear Smoothers and Additive Models", *The Annals of Statistics*, **17(2)**, 453-555.

Craven, P., ve Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions," *Numerische Mathematik*, **31**, 377-403.

Dierckx, P. (1999), *Curve and Surface Fitting with Splines*, Oxford Science Publications, New York.

Dobson, A.J. (2002), *An Introduction to Generalized Linear Models*, Chapman&Hall/CRC, U.S.A..

Draper, N.R., Smith, H. (1998), *Applied Regression Analysis*, John Willey&Sons, New York.

Duchon, J. (1975), "Fonctions Splines et Vecteurs Aleatoires", *Tech. Report 213, Seminaire d'Analyse Numerique*, Universite Scientifique et Medicale, Grenoble.

Duchon, J. (1976), "Fonctions-Spline et Esperances Conditionnelles de Champs Gaussiens," *Ann. Sci. Univ. Clermont Ferrand II Math.*, **14**, 19 -27.

Duchon, J. (1977), "Splines Minimizing Rotation-Invariant Semi-Norms in Sovolev Spaces," *Constructive Theory of Functions of Several Variables,* eds. W. Schempp and K. Zeller, 85-100.

Eilers, P., Marx, B.D., 1996, "Flexible smoothing with B-splines and penalties (with comments and rejoinder)", *Statistical Science,* **11(2)**, 89-121.

Engle, R.F., Granger, C.W.J., Rice, C.A., Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weahter and Electricity Sales", *Journal of Amer. Statis. Assoc.*, **81**, 310-320.

Eubank,R.L., Kambour, E.L., Kim, T.C., Kipple, K, Reese, S.C., Schimek, M. (1998), "Estimation in Partially Linear Models", *Computational Statistics and Data Analysis*, **29**, 27-34.

Eubank,R.L. (1999), *Nonparametric Regression and Smoothing Spline*, Marcel Dekker, New York, U.S.A..

Eubank,R.L., (2000) "Spline Regression", *Smoothing and Regression: Approaches, Computation and Application* (Ed: Schimek, M.G.), Willey Series in Prob. And Stat., U.S.A., 1-18.

Friedman, J. H., Stuetzle, W. (1981). "Projection Pursuit Regression." *J. Amer. Statist. Assoc.* **76**, 817.

Gasser, T., Sroka, L., Jennen-Steinmentz, C. (1986), "Residual Variance and Residual Pattern in Nonlinear Regression", *Biometrika*, **73**, 625-633.

Golub, G.H., Loan, C.F.V. (1996), *Matrix Computations*, The John Hopkins University Pres, London.

Green, P.J., Jennison, C., Seheult, A. (1985), "Analysis of Field Experiments by Least Square Smoothing", *Journal of Royal. Statis. Soc.*, Ser. B, **47**, 299-315.

Green, P.J., Yandell, B. (1985), "Semiparametric Generalized Linear Models", *Proceedings 2nd International GLIM Conference*, Lancaster, Lecture notes in Statistics No.32, 44-55, Springer-Verlag, New York.

Green, P.J., Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman &Hall, London.

Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer-Verlag, New York.

Gu, C., and Kim, Y.J. (2002), "Penalized Likelihood Regression: General Formulation and efficient approximation", *The Canadian Journal of Statistics*, **30(4)**, 619-628.

Hardle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hastie, T.J., Tibshirani (1999), *Generalized Additive Models*, Chapman &Hall, London.

Hastie, T.J. (2006), The gam Package. http://cran.r-project.org/doc/packages/gam.pdf

Hurvich, C.M., Simonoff J.S., Tasi C.L. (1988), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion", *Journal of Royal. Statis. Soc.*, Ser. B, **60**, 271-293.

Hutchinson, M.F., Bishof, R.J. (1983), "A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Hunter Valley, New South Wales", *Australian Meteorological Magazine*, **31**, 179-184.

Mammadov, M.B., Yüzer,A.F., Aydın, D. (2005), "Splayn Düzeltme Regresyonu ve Düzeltme Parametresinin Seçimi", *4. İstatistik Kongresi bildiri özetleri kitabı*, Belek-Antalya, 148-149.

Marx, B.D., Eilers,P.H.C. (1998), "Direct Generalized Additive Modeling With Penalized Likelihood", *Computational Statistics and Data Analysis*, **28**, 193-209.

McCullagh, P., Nelder J.A. (1989), *Generalized Linear Models*, Chapman&Hall/CRC, NewYork.

Meinguet, J. 1979. "Multivariate interpolation at arbitrary points made simple", J. *Appl. Math. Phys*, **30**, 292–304.

Memmedli M., Omay R.E. (2007), "Hava Kirliliğinin Ölüm Oranı Üzerindeki Etkisinin Genelleştirilmiş Toplamsal Regresyon Modelleri ile İncelenmesi", *5. İstatistik Kongresi Bildiri Özetleri Kitabı*, 20-24 Mayıs, Belek-Antalya.

Montgomery, D.C., Peck, E.A., Vining, G.,G. (2001), *Introduction to Linear Regression Analysis*, *WILEY*, New York.

Myers, R.H. (1990), *Classical and Modern Regression with Applications*, Duxbury Classic Series, U.S.A..

Myers, R.H., Montgomery, D.C., Vining, G.G. (2002), *Generalized Linear Models with Applications in Engineering and the Sciences*, Wiley Series in Probability and Statistics, New York.

Nelder, J.A., Wedderburn, R.W.M. (1972), "Generalized Linear Models", *Journal of the Royal Statistical Society*, Series A, **135**, 370-384.

Omay, R.E., Aydın, D., Mammadov, M. (2006a), "Splayn Düzeltme Yöntemi ile Semiparametrik Additive Modellerin Kestirimi", *5. İstatistik Günleri Sempozyumu Bildiri Özetleri Kitabı*, Antalya.

Omay, R.E., Aydın D. (2006b), "Investigation of House Price in Eskisehir (Turkey) by Using Semiparametric Additive Regression Model", *The 5th WSEAS International Conference on COMPUTATIONAL INTELLIGENCE, MAN-MACHINE SYSTEMS AND CYBERNETICS*, Venice, Italy, November 20-22.

Omay, R.E., Aydın D. (2007a), "A Semiparamteric Additive Regression Model: Investigation House Price in Eskişehir(Turkey)", *WSEAS Transactions on Mathematics*,**6(3)**, 494-499.

Omay, R.E., Aydın D., Mammadov M. (2007b), "Semiparametrik Toplamsal Regresyon Modeli ile Tahmin: Eskişehir'deki Evlerin Kira Fiyatları ve Özellikleri Arasındaki İlişkilerin Analizi", *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, (Basımda).

Öztürk, M. (2005), Şehir İçi Bölgelerde Hava Kirliliğinin Sağlık Üzerine Etkileri. http://www.cevreorman.gov.tr/belgeler1/hk.doc

O'Sullivan, F., Yandell, B.S., Raynor, W.J. (1986), "Automatic Smoothing of Regression Functions in Generalized Linear Models", *Journal of American Statistical Association*, **81(393)**, 96-103.

Parker, R., Rice, J. (1985), "Discussion of Silverman", *Journal of the Royal Statistical Society*, Series B, **47(1)**, 43.

Peng, R.D., Welty, L.J. (2004), "The NMMAPSdata package", *R News*, **4(2)**, 10-14.

Reinsch, C.H. (1967), "Smoothing by Spline Functions", *Numerische Mathematik* **10**, 177-183.

Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press, Cambridge.

Schimek, M.G. (2000a), "Estimation and Inference in Patially Linear Models with Smoothing Splines", *Journal of Statistical Planing and Inference*, **91**, 525-540.

Schimek, M.G. (2000b), *Smoothing end Regression*, Wiley Series in Probability and Statistics, New York.

Schoenberg, I.J. (1964a), "Spline Functions and the Problem of Graduation", *Proceedings of the National Academy of Sciences*, **52**, 947-950.

Schoenberg, I.J. (1964b), "On Interpolation by Spline Functions and Its Minimum Properties", *Internat. Ser. Numer. Anal.*, **5**, 947-950.

Seaman, R., Hutchinson, M. (1985), "Comparative Real Data Tests of Some Objective Analysis Methods by Withholding", *Australian Meteorological Magazine*, **33**, 37-46.

Silverman, B.W. (1985), "Some Aspects of the Spline Smothing Approach to Nonparametric Regression Curve Fitting", *Journal of Royal Statistical Society*, Ser. B, **47**, 1-52.

Speckman, P. (1988), "Kernel Smothing in Partially Linear Model", *Journal of Royal Stat. Soc.* Ser. B., **50**, 413-436.

Wahba, G., Wendelberger J. (1980), "Some new mathematical methods for variational objective analysis using splines and cross-validation", *Monthly Weather Review*, **108**, 36-57.

Wahba, G. (1980), "Spline Bases, Regularization, and Generalized Cross-Validation for Solving Approximation Prpblems with Large Quatities of Noisy Data", *Proceeding of the International Conference on Approximation Theory in Honour of George Lorenz*, Jan 8-10, New York.

Wahba, G. (1990), *Spline Models for Observational Data*, Society for Industrial and Applied Mathematics, U.S.A..

Wood, S.N. (2000), "Modeling and Smothing Parameter Estimation With Multiple Quadratic Penalties", *Journal of Royal Statistical Sosiety*, Ser. B, **62**, 413-428.

Wood, S.N. (2001), "mgcv: GAMs and Generalized Ridge Regression for R", *R News*, **1/2**, 20-25.

Wood. S. Augustin, N.H. (2002), "GAMs with Integrated Model Selection Using Penalized Regression Splines and Applications to Environmental Modelling", *Ecological Modelling*, **157**, 157-177.

Wood, S.N. (2003), "Thin Plate Regresssion Splines", *Journal of Royal Statistical Sosiety*, Ser. B, **65**, 95-114.

Wood, S.N. (2004), "Stable and Efficient Multiple Smothing Parameter Estimation for Generalized Additive Models", *Journal of the American Statistical Association*, **99(467)**, ABI/INFORM Global, 673-686.

Wood, S.N. (2006a), *Generalized Additive Models: An Introduction with R*, Chapman&Hall/CRC, U.S.A..

Wood, S.N. (2006b), The gamair Package. http://cran.r-project.org/doc/packages/gamair.pdf

Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge University Pres, U.K..