SEMİPARAMETRİK REGRESYON MODELLEMEDE SPLAYN DÜZELTME YAKLAŞIMI İLE TAHMİN VE ÇIKARSAMALAR

Dursun AYDIN

Doktora Tezi

Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı
Kasım-2005

JÜRİ VE ENSTİTÜ ONAYI

Dursun AYDIN'ın **"Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı ile Tahmin ve Çıkarsamalar"** başlıklı **İstatistik Anabilim Dalındaki**, Doktora tezi 27.09.2005 tarihinde, aşağıdaki jüri tarafından Anadolu Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca değerlendirilerek kabul edilmiştir.

	Adı-Soyadı	**İmza**
Üye (Tez Danışmanı)	**Prof. Dr. Ali Fuat YÜZER**	**……**
Üye	**Doç. Dr. Mammadagha MAMMADOV**	**……**
Üye	**Prof. Dr. Embiya AĞAOĞLU**	**……**
Üye	**Prof. Dr. Musa ŞENEL**	**……**
Üye	**Doç. Dr. Zeki ÇAKMAK**	**……**

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

Enstitü Müdürü

ÖZET

Doktora Tezi

SEMİPARAMETRİK REGRESYON MODELLEMEDE SPLAYN DÜZELTME YAKLAŞIMI İLE TAHMİN VE ÇIKARSAMALAR

DURSUN AYDIN

Anadolu Üniversitesi
Fen Bilimleri Enstitüsü
İstatistik Anabilim Dalı

Danışmanlar: Prof. Dr. Ali Fuat YÜZER
İkinci Danışman: Doç. Dr. Mammadagha MAMMADOV
2005, 128 sayfa

Bu tezde semiparametrik regresyon modelinin kestirimi için kısmi splayn ve Speckman yaklaşımı adı altında iki farklı yaklaşım incelenmiştir. Adı geçen bu iki yaklaşım, bir uygulama üzerinde MATLAB ortamında yazılan bir programla gerçekleştirilmiş ve modelin hem parametrik hem de parametrik olmayan bileşeni hakkında çıkarsamalar yapılmıştır. Parametrik olmayan ve semiparametrik regresyon modellerinin kestiriminde ise, cezalı en küçük kareleri esas alan splayn düzeltme yöntemi kullanılmıştır. Bu yöntemin gerçekleştirilmesinde en önemli etmenlerden biri olan düzeltme parametresinin seçimiyle ilgili yaygın olarak kullanılan seçim kriterleri incelenmiştir. Söz konusu bu seçim kriterlerinden hangisinin daha iyi bir düzeltme parametresini seçtiğini belirlemek amacıyla, MATLAB ortamında yazılan bir program yardımıyla bir simülasyon çalışması yapılmıştır.

Anahtar Kelimeler: Parametrik olmayan regresyon, Semiparametrik regresyon, Cezalı en küçük kareler yöntemi, Splayn düzeltme, Düzeltme parametresi, Seçim kriterleri

ABSTRACT

PhD Thesis

ESTIMATIONS AND İNFERENCES IN SEMIPARAMETRIC REGRESSION MODELLING USING SMOOTHING SPLINE APPROACH

DURSUN AYDIN

Anadolu University
Graduate School of Sciences
Statistics Program

Supervisors: Prof.Dr. Ali Fuat YÜZER
Second Supervisor: Doç.Dr. Mammadagha MAMMADOV
2005, 128 pages

In this thesis, two different approaches called partial spline and Speckman approach are examined in order to estimate semiparametric regression model. The two approaches in question carried out using program that coded in MATLAB environment and the inferences are made for both the parametric and the nonparametric components of the model. Smoothing spline method based on penalized least square is used for estimation of nonparametric and semiparametric regression models. The selection criteria, one of the most important and commonly used factors in choosing smoothing parameter of implementation of that method are examined. In order to find out which is the best among those selection criteria, a simulation study is performed using the program that coded in MATLAB environment.

Keywords: **Nonparametric regression, Semiparametric regression, Penalized least squares method, Smoothing spline, Smoothing parameter, Selection criteria.**

TEŞEKKÜR

Bu tezin hazırlanmasında, bilgi ve birikimlerinden daima faydalandığım değerli hocalarım,

Prof. Dr. Ali Fuat YÜZER

ve

Doç.Dr. Mammadagha MAMMADOV'a,

yardımlarını hiç bir zaman esirgemeyen değerli hocam,

Prof. Dr. Embiya AĞAOĞLU'na

değerli hocam,

Prof. Dr. Musa ŞENEL

ve

Yard.Doç.Dr. Zakir POYRAZ'a

büyük özveri ile desteğini esirgemeyen

eşim Düriye

kızım İnci Gizem

oğlum Eren'e

teşekkür ederim...

Dursun AYDIN

Kasım-2005

İÇİNDEKİLER

Sayfa

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

AIC_c : Akaike bilgi kriteri

CEKK : Cezalı en küçük kareler

CV : Çapraz-geçerlilik

C_p : Mallows'un kriteri

df : Serbestlik derecesi

EDF : Eşdeğer serbestlik derecesi

EMSE : Hata kareler ortalamasının tahmini

$\mathbf{f}$: Splayn fonksiyonun tanımlı oduğu aralıkta aldığı değerler vektörü

GCV : Genelleştirilmiş çapraz-geçerlilik

GEKK : Genelleştirilmiş en küçük kareler

GOF : Uyum iyiliği

$\mathbf{H}$: Semiparametrik model için düzeltme matrisi (Şapka matrisi)

$\mathbf{H}_p$: Kısmi splayn için düzeltme matrisi

$\mathbf{H}_s$: Speckman yöntemi için düzeltme matrisi

LRS : Lokal risk kriteri

M : İkinci dereceden türev

MSE : Hata kareler ortalaması

N : Tekrarlanma matrisi

NCS : Doğal kübik splayn

OLS : Sıradan en küçük kareler

PS : Pürüzlülük cezası

RCP : Klasik Pilotları Kullanan Risk Tahmini

RSS : Hata kareler toplamı

S_λ : Splayn düzeltme yönteminin düzeltme matrisi

SE : Standart hata

$tr(S_\lambda)$: Düzeltme matrisinin izi

λ : Düzeltme parametresi

1.GİRİŞ

Son yıllarda istatistikte *parametrik olmayan* (*nonparametrik*) ve *semiparametrik* (*yarı parametrik*) *regresyon* alanında yapılan çalışmalara ilgi oldukça artmıştır. Bu konuyla ilgili yeteri kadar çalışmalar yapılmış ve farklı metot ve teknikler önerilmiş olmasına rağmen ülkemizde bu konuda şu ana kadar ayrıntılı bir çalışma yapılmamıştır. Bu çalışma böyle bir eksikliği gidermek amacıyla yapılmıştır. Bilindiği gibi, *regresyon analizi* bağımlı ve bağımsız değişkenler arasındaki ortalama ilişkinin matematiksel bir fonksiyonla ifade edilmesinde, bağımsız değişkenlere bağımlı değişkenin doğrusal bir ilişki içersinde olduğunu varsayar. Açıktır ki yöntem bazı varsayımlara dayanır ve bu varsayımların sağlanamaması durumunda yapılan tahminler "*iyi bir tahmin"* olma özelliğini kaybeder. Söz konusu varsayımlardan en önemlisi, bağımlı ve bağımsız değişkenler arasındaki ilişkinin şeklinin biliniyor olmasıdır. Oysa değişkenlerin doğası gereği ilişki şeklinin bilinmediği pek çok durumla karşılaşmak mümkündür. Bu durumda, *parametrik regresyonun* doğrusallık varsayımının esnetilmesine olanak sağlayan regresyon yöntemleri gereksinimi ortaya çıkar. Böyle bir gereksinim doğrulutusunda, pürüzsüz bir ana kütle regresyon fonksiyonunun doğrusallık varsayımını esneten *parametrik olmayan regresyon*, bir çözüm yöntemi olarak gözönüne alınabilir.

Bu çalışmanın temel amacı, *klasik* (*parametrik*) *regresyon* teknikleri ile uygun sonuç vermeyen bu tür regresyon problemlerine çözüm bulmak amacıyla yapılan bu araştırmada, öncelikle karşılaşılan bu tür problemlerin çözümü için, *modern regresyon analizinin* konusunu oluşturan, oldukça geniş bir uygulama alanına sahip ve parametrik olmayan bir yöntem olarak işlem gören *semiparametrik regresyon* tanıtılmıştır. Ayıca semiparametrik regresyon modelinin parametrelerine ilişkin tahmin ve çıkarsamalar yapılarak, semiparametrik regresyon modelinin istatistiksel açıdan değerlendirilmesine olanak sağlanmıştır.

Semiparametrik regresyon modeli, parametrik ve parametrik olmayan regresyon fonksiyonun toplamsal olarak birleşiminden oluşması nedeniyle böyle modellere, *kısmi doğrusal regresyon modelleri* de denir. Semiparametrik

regresyon, parametrik değişkenlerin etkilerinin sıfır olması ya da bu tür değişkenlerin analizde yer almadığı durumlarda parametrik olmayan regresyon olarak işlem görür. Ayrıca semiparametrik regresyon modelinde yer alan parametrik olmayan fonksiyon, önsel bir şekle sahip olmayan çok geniş ve esnek fonksiyonlar sınıfından seçilir. Başka bir anlatımla bu modeller, en az esneklikten (düşük dereceli bir polinom) en çok esnekliğe (yüksek dereceli bir polinom - interpolasyon) doğru bir genişletmeye sahip, bir düzeltme parametresi ile indeksli ve az sayıda parametreyle özetlenemeyen bir parametrik olmayan regresyon fonksiyonunu içermesi nedeniyle, parametrik regresyon modellerinden çok daha esnektir. Son yıllarda giderek popülerlik kazanan bu tür regresyon modellerinin kestirimde kullanılan yöntemlerden biri de, *splayn düzeltme yöntemi*dir. Splayn düzeltme yönteminin esasını, *cezalı en küçük kareler regresyonu* oluşturur. Parametrik olmayan ve semiparametrik regresyon modellerinin kestiriminde kullanılan cezalı en küçük kareler yönteminde, sıradan en küçük karelerden farklı olarak, hata kareler toplamına bir *düzeltme parametresine sahip ceza fonksiyonu* eklenir. Bu ceza fonksiyonu sayesinde, tamamiyle esnek eğimli uyumlar ve sabit eğimli uyumlar arasında bir uzlaşma sağlanır ve böylece söz konusu modelle çok daha tutarlı tahminler yapılır.

Bu konuda bu çalışmayı destekleyen çok sayıda parametrik olmayan ve semiparametrik regresyon çalışması vardır: Engle, Granger, Rice ve Weiss (1986) günlük ortalama hava sıcaklığı ile elektrik satışları arasındaki ilişkinin semiparametrik tahmini, Speckman (1988) kısmi doğrusal modelin bir uygulaması, Robinson (1988) semiparametrik regresyon modelinin kestirimi, Hurvich, Simonoff ve Tasi (1988) geliştirilmiş Akaike bilgi kriterini kullanarak parametrik olmayan regresyonda düzeltme parametresinin seçimi, Wahba (1990) gözleme dayalı veriler için splayn modelleri, Hardle (1991) uygulamalı parametrik olmayan regresyon konusu, Green ve Silverman (1994) genelleştirilmiş doğrusal modeller ve parametrik olmayan regresyonda pürüzlülük ceza yaklaşımı, Eubank, Kambour, Kim, Kiple ve Reese (1998) kısmi doğrusal modellerde tahmin, Yatchew (1988) iktisatta parametrik olmayan regresyon teknikleri, Eubank (1999) parametrik olmayan regresyon ve splayn düzeltme, Schamlesee ve Stoker (1999) Amerika Birleşik Devletlerinde hane halkı benzin

tüketiminin kısmi doğrusal modelle analizi, Michael G. Schimek (2000) Splayn düzeltme ile kısmi doğrusal modellerde tahmin ve çıkarsamalar, Kim, Park ve Kim (2002) semiparametrik regresyon modellerinde diagnostik etkiler, Lee (2002) splayn düzeltme için düzeltme parametresi seçimi konusunda bir simülasyon çalışması, Lee (2003) farklı pürüzsüzlük tahminlerini birleştirerek iyileştirilmiş splayn düzeltme regresyonu, David, Wand ve Carrol (2003) semiparametrik regresyon, Hardle, Müler, Sperlich ve Werwatz (2004) parametrik olmayan ve semiparametrik modeller ve burada yer verilmeyen v.b'leri.

Bu çalışma beş bölümden oluşmaktadır. Birinci bölüm olarak ele alınan giriş bölümünde, tezin konusu ve önemi, bu konuda yapılan çalışmalar, tezin içeriği ve uygulama alanı için yapılan incelemeler ana çizgileriyle gözden geçirilmiştir.

İkinci bölümde parametrik, parametrik olmayan ve semiparametrik regresyon yaklaşımı olmak üzere üç farklı regresyon yaklaşımı, regresyonda düzeltme kavramı, pürüzlülük ceza yaklaşımı ve splayn düzeltme yöntemi dışında kalan diğer parametrik olmayan regresyon tekniklerinin genel bir incelemesi yapılmıştır.

Üçüncü bölümde, splayn fonksiyonu ve parçalı kübik splayn'nın tanımlanmasından sonra, kübik splayn interpolasyonu için Lagrange ve lokal polinominal yöntem olmak üzere iki farklı yaklaşım incelenmiştir. Daha sonra splayn düzeltme yöntemine temel oluşturan doğal kübik splayn interpolasyonu ve doğal kübik splayn interpolantının optimumluk özellikleri, parametrik olmayan regresyon ortamında splayn düzeltme kestiricisinin elde edilmesi, splany düzeltme kestiricinin minimum olma özelliği ve son olarak da splayn düzeltme kestiricisi olan doğal kübik splaynı elde etmek için splayn düzeltme yöntemine bir alternatif olarak düşünülen Reinsch algoritması ele alınmıştır.

Dördüncü bölümde, esas itibariyle semiparametrik regresyon konusu ele alınmış olup, söz konusu modelin kestiriminde, kısmi splayn ve Speckman yaklaşımı adı altında iki farklı yaklaşım incelenmiştir. İzleyen bölümlerde, semiparametrik regresyon modeline ilişkin çıkarsamalar, hem parametrik bileşen hem de parametrik olmayan bileşen için gerçekleştirilmiştir. Teorik olarak

incelenen söz konusu semiparametrik regresyon çalışmasını uygulamayla desteklemek amacıyla, evlerin satış fiyatları ile evlerin karakteristiklerini gösteren değişkenler arasındaki ilişkinin araştırılmasına ilişkin bir uygulama yapılmıştır. Yapılan örnek uygulamada ki tüm sayısal değerlendirmeler, MATLAB ortamında geliştirilen bir programla gerçekleştirilmiştir. Söz konusu uygulamada parametrik regresyon modeli ile semiparametrik regresyon modelinin kestirimi için başvurulan Speckman yöntemi ve kısmi splayn yöntemlerinin performansları karşılaştırmalı bir biçimde incelenmiştir. Bunun yanı sıra, semiparametrik regresyon modelleri hakkında splayn düzeltme yöntemini esas alan kestirimlere ilişkin tahmin ve çıkarsamalara da yer verilmiştir.

Son olarak beşinci bölümde, splayn düzeltme yönteminin esasını oluşturan cezalı kareler toplamına eklenen ceza fonksiyonunun katsayısının diğer bir deyişle, düzeltme parametresinin seçimine konu olan seçim kriterleriyle ilgili bazı temel kavramlar ele alınmış, söz konusu düzeltme parametresinin seçimi için yaygın olarak kullanılan seçim kriterlerinden klasik ve risk tahmin yöntemleri adı altında, "*çapraz-geçerlilik, genelleştirilmiş çapraz geçerlilik, iyileştirilmiş Akaike bilgi kriteri, Mallows'un Cp kriteri, klasik pilotları kullanan risk tahmini ve lokal risk tahmini*" kriterleri incelenmiştir. Adı geçen bu kriterlerden hangisinin daha iyi bir tahmin sonucu veren düzeltme parametresini seçtiğini belirlemek amacıyla da bir Monte Carlo simülasyon deney çalışması yapılmıştır. Simülasyon çalışması bir önceki bölümde olduğu gibi, MATLAB programı ortamında geliştirilen bir program yardımıyla gerçekleştirilmiştir. Yapılan kestirimlerde, splayn düzeltme kestiricilerinin kalitesini değerlendirmek için MSE (*hata kareler ortalaması*) performans ölçü kriteri olarak alınmıştır. Adı geçen bu performans kriterine göre, yöntemler başarı sıralamasına tabi tutularak en iyi yöntem ve dolayısıyla en iyi düzeltme parametresi belirlenmiştir.

2. REGRESYON TÜRLERİNE GENEL BİR YAKLAŞIM VE REGRESYONDA DÜZELTME KAVRAMI

Regresyon, istatistikte en yaygın kullanım alanına sahip yöntemlerden biridir. Regresyon analizinin amacı, $z_1,...,z_k$ bağımsız (açıklayıcı) değişkenler (*covariates*) dizisine göre y bağımlı (açıklanan) değişkeninin (*response variable*) koşullu ortalamasının fonksiyonel bağımlılığını ortaya koymaktır. Bilinen *klasik doğrusal regresyon modeli,*

$$y_i = \beta_0 + \beta_1 z_{i1} + ... + \beta_k z_{ik} + \varepsilon_i,\ i = 1,...,n$$

şeklinde ifade edilir. Burada, ε_i, sıfır ortalamalı, sabit varyanslı ve açıklayıcı değişkenlerden bağımsız bir hata terimidir. Ayrıca, $z_1,...,z_k$ değişkenlerine göre y'nin koşullu ortalaması, $z_1,...,z_k$ değişkenlerinin bir doğrusal fonksiyonudur. Genellikle böyle bir doğrusal model çok yararlı olmasına karşın, verilerin doğrusal olmaması (*nonlinearity*) halinde, bu tür modelin uygun olmadığı açıktır. Bazen bu problem, bağımsız değişkenlerde dönüşüm yapılarak atlatılabilir. Bununla birlikte, uygulamada veriler gözönüne alınarak en uygun fonksiyonel şekli tahmin etmek zordur. Ancak parametrik olmayan regresyonla, bağımsız değişkenlere göre bağımlı değişkenin fonksiyonel bağımlılığını belirtmeksizin verilere uygun modeller oluşturulabilir.

Önsel bir parametrik modelin uygun olduğu düşünülebilen durumlarda bile, bağımsız değişkenlere göre bağımlı değişkenin fonksiyonel bağlılığını ortaya çıkarmada, parametrik olmayan regresyon yararlı bir tekniktir. Ancak, karşılaşılan birçok problemde, verilerin doğasına göre y bağımlı değişkeni, bağımsız değişkenlerden bazıları ile doğrusal, bazıları ile de doğrusal olmayan bir ilişki içerisinde olabilir. Bu gibi durumlarda, sade parametrik ya da parametrik olmayan modeller uygun değildir. Böyle durumlarda, hem parametrik hem de parametrik olmayan kısmı bünyesinde bulunduran ve toplamsal modellerin özel bir durumu olan semiparametrik regresyon daha uygun sonuçlar verir.

Yukarıda sözü edilenlerden de anlaşılacağı gibi, bağımsız değişkenlere göre bağımlı değişkeninin fonksiyonel bağlılığını ortaya çıkarmada parametrik,

parametrik olmayan ve semiparametrik regresyon olmak üzere üç farklı regresyon türü söz konusu olmaktadır. Bu bölümde, sırasıyla bu regresyon yaklaşımları ve regresyonda düzeltme kavramı ana çizgileriyle ele alınmıştır.

2.1. Parametrik Regresyon

Geleneksel parametrik regresyon yaklaşımı, uygun bağımlı ve bağımsız değişkenler ile bu değişkenler arasındaki ortalama ilişkinin matematiksel bir fonksiyonla ifade edilmesini ve bu fonksiyonda yer alan parametrelerinin açık bir şekilde belirtilmesini gerektirir. Genel olarak kullanılan fonksiyonel şekiller, Box-Cox dönüşümlerinin yanı sıra, doğrusal, yarı-logaritmik ve logaritmik-doğrusal olan modellerdir. Esasında doğrusal ve doğrusallaştırılmış modellerin bu ailesi için belirli varsayımlar, bağımsız değişkenler ile bağımlı değişken arasında açık ve kesin, deterministik bir ilişkiyi belirtir [1].

Parametrik regresyon ortamında regresyon fonksiyonunun şeklini ilgilendiren hipotezler oldukça sınırlayıcıdır. Söz konusu regresyon fonksiyonunun, bağımsız $\mathrm{Z} = (z_1,...,z_k)$ değişkenlerinin belirli bir fonksiyonu olarak, örneğin, Z'in logaritmik, parabolik veya doğrusal bir fonksiyonu olarak yazılabildiğini varsayar. Ayrıca, $E(\mathbf{y} \mid \mathrm{Z})$ şartlı beklenen değeri, Z'in bir fonksiyonudur. Bu durumda, $E(\mathbf{y} \mid \mathrm{Z}) = f(\mathrm{Z})$ koşullu ortalaması, örneğin, $f(\mathrm{Z})$ regresyon fonksiyonu Z değişkenlerinin doğrusal bir fonksiyonu olup, Z biliniyoriken $\mathbf{y}$'nin ortalama dağılımının Z ile fonksiyonel ilişkisini gösterir. Başka bir deyişle, $\mathrm{Z} = (z_1,...,z_k)$ bağımsız değişkenlerindeki değişime karşılık y bağımlı değişkenin ortalama tepkisini dile getirir. Bu fonksiyonel ilişkiyi gösteren "*klasik doğrusal regresyon*" modeli,

$$\begin{aligned} \mathbf{y} &= E(\mathbf{y} \mid \mathrm{Z}) + \boldsymbol{\varepsilon} \\ &= \mathrm{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned} \qquad (2.1)$$

şeklinde ifade edilir. (2.1) modelinde yer alan $\mathbf{y}$, bağımlı değişkene karşı gelen gözlem değerlerinin $(n \times 1)$ boyutlu vektörü, Z, bağımsız değişkenlerin gözlem değerlerini içeren $(n \times (k+1))$ boyutlu bir matrisi, $\boldsymbol{\beta}$, tahmin edilen $(k+1)$ tane parametrik regresyon katsayılarını ve $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y} \mid \mathrm{Z})$, gözlenemeyen fakat

pozitif ya da negatif değerler alabilen, sıfır ortalamalı ve sabit varyanslı bir rassal değişken olan hata terimlerini gösterir.

Parametrik regresyon yaklaşımında temel amaç, verilerin (2.1) regresyon modeline uyumuna göre, en iyi β değerlerini bulmaktır. Bu bakış açısından, (2.1) denkleminin önemli bir karakteristiği, onun parametrik şeklidir: Regresyon fonksiyonu, bilinmeyen β parametreleri tarafından yönetilir. Diğer bir ifadeyle, doğrusal regresyon fonksiyonunu belirlemek için, bilinmeyen β parametrelerini tahmin etmek zorunludur. Bu yaklaşım parametrik olarak adlandırılır, çünkü regresyon fonksiyonunun belirgin bir şekli bilinmekte ve az sayıda parametre ile (burada β regresyon katsayıları ile) regresyon fonksiyonu yazılabilir. Böylelikle, parametrik modeller tümüyle bir parametre vektörüyle belirtilirler. Uyum modelleri kolayca yorumlanabilir ve model altındaki varsayımlar sağlanırsa, parametrik regresyonla uygun tahminler elde edilir. Bununla birlikte, eğer varsayımlar bozulursa, parametrik tahminler uygun olmayabilir [2].

Parametrik regresyon, söz konusu regresyon fonksiyonunun parametrik bir şekle sahip olduğunu varsayarak, elde edilen modelin veriler tarafından desteklenip desteklenmediğini ortaya koyar. Ancak bu yaklaşımla ilgili bazı problemler de söz konusudur: Biricisi, karmaşık şekle sahip olan bir f fonksiyonunu parametrik bir fonksiyon ile modellemek zordur. İkincisi, bazen verilerin birkaç parametrik şekil (model) arasından hangisine uyduğunu görmek zor olabilir. Ayrıca, böyle bir modelleme stratejisi çok sayıda optimumluk özelliğine sahip olmasına rağmen, belirsizlik altında (risk ortamında) gerçekleştirilen model tahminleri yanlı veya uygun olmayan sonuçlar verebilir. Uygulamada karşılaşılan özel problem doğrultusunda, parametrik regresyon modeline alternatif parametrik modeller ve standart testler dikkate alınabilir, fakat dikkate alınan her hangi bir parametrik modelin doğru olacağının hiçbir garantisi yoktur. Açıkçası, gerçek fonksiyonların herhangi bir parametrik aileye ait olduğunun da hiçbir garantisi yoktur. Böylece, uyglamalarda parametrik tahmin için gerekli olan güçlü sınırlamaları (parametrik model varsayımları) yüklemenin maliyeti de dikkate alınmalıdır [3].

2.2. Parametrik Olmayan Regresyon

Parametrik kestiricilerle, parametreler ve değişkenleri ilişkilendiren bir fonksiyonel şekil ve verilen **x** bağımsız değişkenler dizisine göre y bağımlı değişkeninin koşullu beklenen değerini ortaya çıkarmaya çalışırlırken, parametrik olmayan kestiricilerle, önsel spesifik bir fonksiyonel şekilden bağımsız, doğrudan şartlı beklenenini keşfetmeye çalışılır [1]. Örneğin, (x_i, y_i), $i = 1,...,n,$ gözlem değerleri verildiğinde, $\mathbf{x} = (x_1,...,x_p)$ kestirici (bağımsız) değişkenlerinin ve y bağımlı (açıklanan) değişkeninin ölçümleri arasındaki ilişkiyi açıklayan en yaygın yöntem, $E(\mathbf{y}|\mathbf{x}) = f(\mathbf{x})$ koşullu beklenen fonksiyonunu tahmin etmektir. Böyle bir ilişkiyi açıklayan *genel parametrik olmayan regresyon modeli*,

$$\begin{aligned}\mathbf{y} &= E(\mathbf{y}|\mathbf{x}) + \boldsymbol{\varepsilon} \\ &= f(\mathbf{x}) + \boldsymbol{\varepsilon} = f(x_1,...,x_p) + \boldsymbol{\varepsilon}\end{aligned} \tag{2.2}$$

biçiminde ifade edilir. (2.2) modelinde belirtilen f, parametreleri belirginlenmemiş bir fonksiyonel ilişkiyi gösteren *bilinmeyen* bir fonksiyondur. Diğer bir ifadeyle, az sayıda parametre ile basit olarak özetlenemeyen, belirgin bir şekle sahip olmayan ve kestirici **x** değişkenlerinin bir fonksiyonudur. Ayrıca (2.2)'de yer alan $\boldsymbol{\varepsilon}$, rassal hata terimi olup $E(\boldsymbol{\varepsilon} | \mathbf{x}) = 0$, ve $Var(\boldsymbol{\varepsilon} | \mathbf{x}) = \sigma^2(x)$ koşulunu sağlayan bağımsız bir rassal değişkendir [4]. (2.2) modeline karşı gelen bir regresyon eğrisi, değişkenler arasındaki genel bir ilişkiyi tanımlar. Bu ilişki hakkında bazı bilgilere sahip olunması önemlidir. Regresyon fonksiyonunun şekli bize **x** değişkenlerinin belirli x_i değerleri için daha yüksek y_i gözlemlerinin hariç tutulacağını veya iki değişken arasında bağlılığın özel bir türünü gösterip göstermediğini anlatır [5].

Parametrik olmayan regresyonun amacı parametreleri tahmin etmekten çok, doğrudan f regresyon fonksiyonunu tahmin etmektir. (2.2) modelinin önemli özel bir durumu, yalnızca bir x açıklayıcı değişkenin yer aldığı *"basit" parametrik olmayan regresyon modeli*,

$$y_i = f(x_i) + \varepsilon_i \ , i = 1,...,n$$

biçiminde olup, bu model bir x değişkenine karşı y değişkenin aldığı değerlerin dağılımının grafiksel olarak gösterilmesinde kullanılan önemli bir uygulamadan dolayı, genellikle *"dağılım grafiğini düzeltme ya da düzgünleştirme"* (*scatterplot smoothing*) olarak adlandırılır [6].

Model (2.2) ile belirtilen bir regresyon eğrisini tahmin edecek parametrik olmayan yaklaşımın dört ana amacı vardır. Birincisi, iki değişken arasında genel bir ilişkiyi açıklayan bir model sağlamak. İkincisi, sabit bir parametrik modeli referans almaksızın yapılacak gözlemlerin bir kestirimini vermek. Üçüncüsü, izole edilen noktaların etkisini bulmak için bir araç sağlamak ve dördüncüsü de, komşu x değerleri arasında interpolasyon veya kayıp değerleri yerine getiren esnek bir metot oluşturmaktır.

Parametrik olmayan modelleme, birleşik dağılımların şekli üzerine kısıtlamalar getirir. Bu nedenle fonksiyonel şekile ilişkin çok şey anlatmaz. Bununla birlikte, regresyon eğrisi kestiricisinin uygunluğu, parametrik modellemeden çok daha genel şartlar altında ortaya konulur. Kuramda Rust [7] doğrusal olmayan modellerde, normal olmayan hataların ve heterokadastikliğin (değişen varyans) otomatik olarak barındırıldığını vurgular. Ayrıca, fonksiyonel şekli kısıtlamanın esnetilmesi ile de esnek bir etkileşim sağlanmaya çalışılır. Diğer taraftan, esnekliğin artması da modele bir maliyet getirir. Parametrik olmayan kestiricinin yakınsama oranı, genellikle parametrik kestiricilerden daha yavaştır. Böylece, parametrik olmayan çok boyutlu yüzeylerin doğru olarak tahmini çok büyük örneklem hacimleri gerektirir. Ancak, parametrik modellerde olduğu gibi, bu tür modellerde de istatistiksel doğruluk, örneklem hacminin değil, kestiricilerin varyans ve kovaryanslarının bir fonksiyonudur [3].

Buraya kadar sözü edilenlerden de anlaşılacağı gibi, parametrik olmayan regresyon yaklaşımı ile y değişkeni için farklı x değerlerine uygun olan bir beklenen değer elde edilmeye çalışılır. Bu yapısıyla da parametrik olmayan regresyon, regresyon fonksiyonunun önsel bir şekline (örneğin doğrusal, karesel v.b gibi) sahip olunmaksızın her bir x için ortalama y değeri tahminde kullanılarak, parametrik regresyonun bir alternatifi olur. Böyle bir regresyon fonksiyonu belirli bir tanımlamaya sahip olmadığı ve az sayıda parametre ile

özetlenemediği için, model parametrik olmayan (*nonparametric*) olarak adlandırılır ve bu modeller problemedeki f regresyon fonksiyonuna pürüzsüzlük gibi kısıtlamalar yükleyerek, bilinen parametrik modelleri destekler ve esneklikleri ile de yeni modellere yol açarlar.

Parametrik olmayan regresyonla, çok geniş esnek fonksiyonlar sınıfından amaca uygun bir modelin seçimi gerçekleştirilir. Bu fonksiyonlar en az esneklikten (düşük dereceli bir polinom) en çok esnekliğe (interpolasyon) süren bir genişletme ve bir düzeltme parametresi ile indekslenir. Düzeltme düzeyine bağlı olan bu fonksiyonlar karmaşık bir parametrik şekle sahip olabilirler. Diğer yandan, söz konusu probleme ilişkin ele alınan çok sayıda açıklayıcı değişken varsa, parametrik olmayan regresyon modelleri yanlış tahminler üretebilir ve onları yorumlamak zor olabilir [2].

2.3. Semiparametrik Regresyon

Semiparametrik (*semiparametric*) regresyon modelleri bağımlı değişkenin bazı açıklayıcı değişkenlerle *doğrusal*, fakat diğer bazı açıklayıcı değişkenlerle *doğrusal olmayan* ilişki içerisindeki regresyon modelleridir. Gerçekte bu modeller, standart regresyon tekniklerini genelleştirdiğinden toplamsal modellerinin özel bir durumunu oluştururlar.

Semiparametrik modeller, doğrusal *parametrik* bileşen ve doğrusal olmayan *parametrik olmayan* bileşenlerin her ikisini de içerdiğinden aynı zamanda, *kısmi doğrusal* modellerdir. Bu modeller "boyutluluğun verdiği sıkıntı" nedeniyle tamamıyla parametrik olmayan regresyona tercih edilebilen ve her bir değişkenin etkisinin daha kolay yorumlanmasına olanak sağlayan regresyon modelleridir. Bu yapısıyla *"semiparametrik regresyon modelleri"* hem parametrik hem de parametrik olmayan bileşenleri birleştirdiklerinden dolayı standart doğrusal modellerden çok daha esnektir [8].

Semiparemetrik toplamsal modellerdeki çıkarsamalar son zamanlarda dikkate değer bir ilgi görmektedir. Bu modellerde, bağımlı değişken bir ya da daha fazla açıklayıcı değişkenle doğrusal olarak ilişkili fakat diğer ek değişken ya da değişkenlerle ilişkisinin kolayca parametreleştirilmediği varsayılır.

Değişkenlerin skaler olduğu ve f fonksiyonunun parametrik bir aile içerisinde bulunmadığı bilinen *semiparametrik regresyon modeli*,

$$\begin{aligned} \mathbf{y} &= E(\mathbf{y} \mid \mathrm{Z}, \mathbf{x}) + \boldsymbol{\varepsilon} \\ &= \mathrm{Z}\boldsymbol{\beta} + f(\mathbf{x}) + \boldsymbol{\varepsilon} \end{aligned} \tag{2.3}$$

şeklinde ifade edilir. (2.3) modelinde belirtilen f, ikinci mertebeden sürekli türeve sahip olan fonksiyonlar uzayının bir elamanıdır. Diğer bir deyişle, $f \in C^2[a,b]$ olup modelin düzeltme kısmı olarak bilinir ve parametreleri belirginleştirilmemiş bir fonksiyonel ilişkiyi gösterir. Ayrıca $E(\boldsymbol{\varepsilon} \mid \mathrm{Z}, \mathbf{x}) = 0$ *ve* $Var(\boldsymbol{\varepsilon} \mid \mathrm{Z}, \mathbf{x}) = \sigma_\varepsilon^2$ olduğu başka bir anlatımla, $\boldsymbol{\varepsilon}$ hata terimlerinin sıfır ortalamalı ve σ^2 varyansı ile istatistiksel bağımsız ve normal olarak dağıldığı (*normally and independently distributed*-NID) ve $\mathrm{Z} = (z_1, ..., z_k)$ bağımsız değişkenlerinin koşullu ortalaması $\mathbf{x} = (x_1, ..., x_p)$ kestirici değişkenlerinin pürüzsüz bir fonksiyonu olduğu varsayılır [9]. (2.3) regresyon modelleri, $\boldsymbol{\beta}$ *parametre vektörünü* ve $E(\mathbf{y} \mid \mathrm{Z}, \mathbf{x}) = \mathrm{Z}\boldsymbol{\beta} + f(\mathbf{x})$ *ortalama vektörünü* etkin olarak tahmin etmesinin yanı sıra, uyumun sayısal olarak özetlemesi ve grafiksel olarak görüntülemesine de olanak sağlarlar.

Karşılaşılan bir problemde çok sayıda açıklayıcı değişkene bağlı olan bir y değişkeni gözlemlendiğini varsayalım. Çoklu doğrusal regresyonda bu bağlılığın doğrusal olduğu varsayılır ve genel doğrusal modellerin bilinen teorileri tahminde kullanılır. Semiparametrik modellerin felsefesi, en basit şekliyle, x olarak adlandırılan açıklayıcı değişkenlerin sadece biri üzerinde doğrusallık varsayımını esnetmek ve z olarak adlandırılan, kalan açıklayıcı değişkenler vektörü üzerinde doğrusal bağlılığı korumaktır.

Semiparametrik modeller parametrik ve parametrik olmayan modellerin bileşenlerini birleştirdiği için (2.3) modelindeki ikinci terimin esnekliğe sahip olması birinci terimin kolay yorumlanabilirliğini bozmaz. Böylece, bu tür regresyon yaklaşımı bir kısım değişken için parametrik olmayan regresyonun esneklik avantajına ve kalan tüm diğer değişkenler için de parametrik regresyonun etkinlik avantajına sahiptir [3].

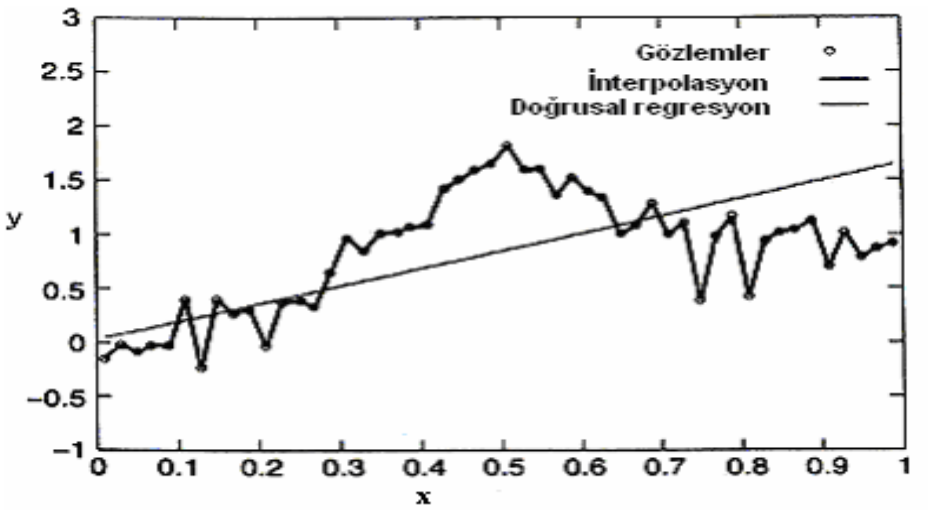
2.4. Regresyonda Düzeltme Kavramı

Birçok durumda, yüzlerce noktadan oluşan eğriler çok karmaşık bir şekle sahip olması nedeniyle, bunları yüksek dereceden polinomiyal modellerle bile temsil etmek olanaksız hale gelir. Düzeltme fikrinin esasını, verileri bir eğriye uydurmak ve daha basit fonksiyonların birleşimi olabilen esnek fonksiyonları kullanmak oluşturur. Düzeltici (düzeltilmiş eğri) ise, $x_1,...,x_p$ gibi kestirici değişkenlerinin (*predictor variable*) bir fonksiyonu olarak bir y bağımlı değişkeninin trendini özetlemek için bir araç oluşturur. Düzelticinin en önemli özelliği onun parametrik olmayan doğasıdır: y ve $x_1,...,x_p$ kestirici değişkenleri arasındaki bağımsızlık için katı kuralları dikkate almaz [10].

Regresyon analizinin amacı, bilinmeyen regresyon fonksiyonu için uygun bir tahmin üretmektir. Bağlı olarak gözlemsel hataları azaltarak y'nin x'e göre ortalama bağımlılığının önemli ayrıntılarını vermek, yorumu kolaylaştırır. Böyle bir eğri yaklaştırma (tahmin) işlemi genel olarak "*düzeltme (smoothing)*" olarak adlandırılır [5]. f regresyon fonksiyonunu tahmin edecek en popüler klasik yöntemlerden biri doğrusal regresyon yaklaşımıdır. Kullanılan bu teknik, (x_i, y_i), $i = 1,...,n,$ veri dizisi için $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ fonksiyonuna göre

$$RSS(f) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 \qquad (2.4)$$

şeklinde ifade edilen *artık* (*hata*) *kareler toplamını* minimum yaparak, f fonksiyonunu tahmin eder. (2.2) denklemindeki f doğasına göre yaklaşık olarak doğrusalsa tahmin için bu yaklaşım etkin olabilir. Fakat f fonksiyonunun doğası doğrusal değilse bu yaklaşım başarısız da olabilir. Böyle bir başarısızlık örneği Şekil 2.1'de verilen örnekte görülmektedir. Söz konusu şekilde görülen doğrusal regresyon uyumu, sadece değişkenler arasında bir ilişkinin varlığını önermeye hizmet eder.



Şekil 2. 2: Benzetim veri dizisi için doğrusal regresyon ve interpolasyon

Şekil 2.1'de görüldüğü gibi, f regresyon fonksiyonun bir diğer kestiricisi de (x_i, y_i), $i = 1,...,n,$ verilerinin interpolasyonu sayesinde elde edilebilir. Bu kestirici için doğrularla birleştirilen gözlemlerde, bireysel eğimler ile esasen sabit eğimlilik sağlanır. Böyle bir kestirici için $RSS(f) = 0$ olur. Verideki bilginin çok azını kullanan, doğrusal regresyon uyumundan farklı olarak bu uyum gerçekte, çok daha fazla bilgi içerir. Ancak, verilerin yararlı bir özetini sağlamada başarısız olur ve daha da önemlisi, modelin rassal-gürültü (hata) kavramından kaynaklanan ve (2.2) denkleminde regresyon fonksiyonuna bağlı olabilen verilerdeki özelliklerin ya da esas eğilimin (trendin) doyurucu bir biçimde açıklanmasını gerçekleştirmez.

Açıktır ki bu durumda, doğrusal regresyon ve interpolasyon uyumlarının esnetilmesi sağlayan daha esnek yöntemlerin kullanması zorunlu hale gelir. Açıkça görülüyor ki, veri uydurmada başarılı olmak için, başlangıç tahmin koşullarını değiştirmeli ve tasarımlarda, değişen eğimli f fonksiyonları üzerinde hata kareler toplamının minimizasyonunu dikkate alınmalıdır.

2.4.1. Pürüzlülük Ceza Yaklaşımı

Genel olarak, en basit şekliyle *"pürüzlülük ceza yaklaşımı"* polinom regresyondan çok az bir farkla regresyon doğrusu boyunca klasik doğrusal regresyonda ki model varsayımlarını esneten bir yöntem olarak bilinir. Pürüzlülük ceza yaklaşımının esas amacı hızlı olarak dalgalanan bir eğrinin eğilimini ölçmek

ve daha sonra eğri tahmininde oldukça farklı iki amaç olan sabit eğimli uyumlar ve esnek eğimli uyumlar arasında gerekli uzlaşmayı gerçekleştirecek bir şekilde tahmin problemini ortaya koymaktır [11]. Basit doğrusal regresyonun Şekil 2.1'deki veriler için uygun olmadığı görülmektedir. Söz konusu Şekil 2.1'deki gözlem değerleri, iki *"aşırı"* uyumlar diğer bir deyişle, tamamıyla esnek eğimli uyumlar ve sabit eğimli uyumlar arasında bir uzlaşmaya gereksinimiz olduğu izlenimi vermektedir. Bunu sağlamanın bir yolu da modele, regresyon fonksiyonunun eğimiyle ilişkili olarak ceza kısmının eklenmesi olabilir.

$f \in C^2[a,b]$ olmak üzere bir f eğrisi verilsin, f eğrisinin ne kadar *"pürüzlü veya dalgalı"* olduğunu ölçmenin birkaç yolu vardır. Bunlardan en yaygını, iki kez sürekli türeve sahip bir f fonksiyonunun eğiminin değişim oranı, f'' yardımıyla tanımlanır ve uyum fonksiyonunun eğimindeki toplam değişim ölçüsü, aşağıdaki gibi hesaplanır:

$$J(f) = \int_a^b f''(x)^2 dx \qquad (2.5)$$

Böylece, ikinci mertebeden sürekli türeve sahip tüm fonksiyonlar üzerinde minimum olabilen, eğimdeki hızlı değişimler için cezayı birleştiren başlangıç tahmin kriterinin değiştirilmiş biçimi,

$$RSS(f) + \lambda J(f), \quad \lambda > 0 \qquad (2.6)$$

olur. (2.6)'daki λ'ya *düzeltme parametresi* denir. λ düzeltme parametresinin değeri, bir uyumun eğiminin esneklikten yoksunluğunu veya esneklik üzerine koyulan önemin bir ölçüsü olarak görülebilir. λ büyükse, (2.6) eşitliğindeki ana bileşen $J(f)$ pürüzlülük ceza terimi olacak ve böylece minimum $\hat{f}$ çok az eğrilik gösterecektir. Limit durumunda ise, λ sonsuza yöneleceğinden $\int f''^2$ terimi sıfıra yaklaşır ve $\hat{f}$ eğrisi Şekil 2.1'deki gibi bir doğrusal regresyon uyumuna yaklaşır. Diğer taraftan, λ oldukça küçükse, (2.6) eşitliğine ana katkı hata kareler toplamı olur ve $\hat{f}$ tahmin eğrisi, veriyi yakından izleyecektir. Limit durumunda ise, λ sıfıra yaklaştığı için, $\hat{f}$ eğrisi Şekil 2.1'de görülen interpolasyon eğrisine yaklaşacaktır. Açıkça görüldüğü gibi bu noktada temel sorun, verileri en iyi temsil

eden bir eğri tahmini elde etmek için en uygun λ değerinin seçimi etrafında yoğunlaşır. Bu konu ile ilgili tüm ayrıntılar üçüncü ve beşinci bölümlerde ele alınmıştır.

2.5. Parametrik Olmayan Regresyonda Düzeltme Teknikleri

Düzeltme teknikleri ile, farklı ölçümler arasında fonksiyonel bir ilişki bulmaya çalışılır. Standart (parametrik) regresyon ifadesinde olduğu gibi, verilerin bir ya da daha fazla açıklayıcı değişken ve bir bağımlı değişken ölçümlerinden oluştuğu varsayılır. Standart regresyon teknikleri, kestiriciler ve bağımlı değişkenler arasındaki ilişkiyi tanımlamak için bir fonksiyonel şekil (bir doğru denklemi gibi) belirtirler. Düzeltme teknikleri ise, uyum eğrisinin şeklini belirtmek için kendileri veri noktaları sağlayan daha esnek yaklaşımlardır [12].

Uygulamada *splayn düzeltme* yönteminin dışında en yaygın kullanılan düzeltme teknikleri aşağıda sıralanmıştır (bak: [5, 9]):

I.) k. En Yakın Komşu Düzeltmesi
II.) Kernel (Çekirdek) Düzeltmesi
III.) Ortagonal Serilerin Kestiricileri
IV.) Lokal Regresyon Kestiricileri
V.) Splayn Regresyonu

Tüm bu tekniklerde, bir düzeltme düzeyinin belirtilmesi gerekir. Bu düzey düzeltme parametresi veya düğüm sayılarının bir fonksiyonudur. Yukarıda sıralanan tekniklerden herhangi birine uygun uyum eğrisi, aşağıda tanımlanan vektörle belirlenir:

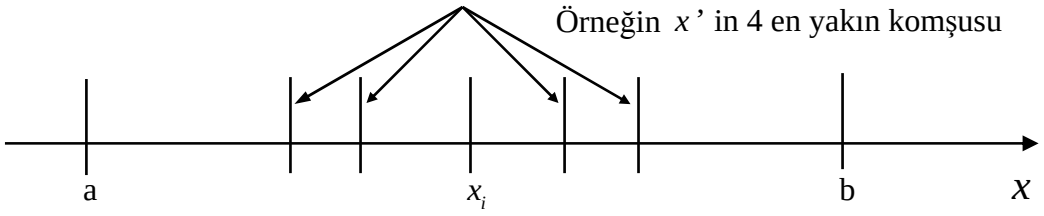
$$\hat{\mathbf{f}}_{\lambda}(x) = S_{\lambda}\mathbf{y}$$

Burada belirtilen S_{λ}, pozitif bir düzeltme düzeyi ve x değişkenine bağlı olan, fakat y değişkenine bağlı olmayan $(n \times n)$ tipinde bir düzeltme ya da şapka matrisi olarak bilinir. Tüm bu düzeltme teknikleri bu şekilde ki yazılıştan dolayı doğrusal düzelticiler olarak adlandırılırlar. Doğrusal modellerin birçok özelliği bu kestiricilerle elde edilmiştir.

Bu tekniklerin temel amacı, ortalama fonksiyonu için parametrik bir şekil belirtmek değil, verileri sağlayacak bir fonksiyonel şekil belirlemektir. İzleyen bölümde adı geçen düzeltme teknikleri ana çizgileriyle ele alınmıştır. Ancak, esas konuyu oluşturması nedeniyle "*splayn düzeltme"* tekniği, üçüncü bölümde ayrıntılı olarak ele alınmıştır.

2.5.1. k- En Yakın Komşu Düzeltme

$y_i = f(x_i) + \varepsilon_i,\ i = 1,...,n$ modeli ile ilgili $(x_1, y_1),...,(x_n, y_n)$ gözlem vektörleri gözönüne alınsın. Burada $\{x_i\}_{i=1}^{n}$, x değişkeninin [a,b] aralığında aldığı faklı değerleri gösterir. Ayrıca, x_i noktasındaki f'in kestiricisi şekilde de görüldüğü gibi, x'e en yakın k tane komşuya ait olan y_i değerlerinin ortalamasıdır. Diğer bir deyişle,



k-en yakın komşu $(k-NN)$ tahmini, değişen komşuluklarda ağırlıklı bir ortalamadır. Bu komşuluk, öklid uzaklığındaki x_i'nin k.en yakın komşular arasında bulunan x değişkenleri aracılığı ile tanımlanır. Kuramda, $k-NN$ ağırlık dizisi, Loftsgaarden ve Quesenbery [13] tarafından tanıtılmış ve sınıflandırma amaçları için Cover ve Hart [14] tarafından kullanılmıştır. Biçimsel olarak, "*k-en yakın komşu kestiricisi",*

$$\hat{f}_k(x) = n^{-1}\sum_{i=1}^{n} W_{ki}(x) y_i \qquad (2.7)$$

şeklinde tanımlanır. Bu eşitlikte yer alan $\{W_{ki}(x)\}_{i=1}^{n}$ ifadesi,

$$W_{ki}(x) = \begin{cases} n/k & i \in J_x \text{ ise;} \\ 0 & \text{diğer durımlarda} \end{cases} \qquad (2.8)$$

ve

$$J_x = \{i : x_i,\ x\text{'en yakın k tane gözlemden biri}\}$$

şeklinde belirtilen indeks dizisi ile tanımlanan bir ağırlık dizisidir. k düzeltme parametresi olup tahmin edilen eğrinin pürüzsüzlüğünün derecesini düzenler ve kernel düzelticilerinin bant genişliğine (düzeltme parametresi) benzer bir rol oynar. Bu yöntemle elde elden f fonksiyonu süreklidir. Eğer k büyükse, f fonksiyonu daha yavaş değişir. $k = n$ ise, f fonksiyonu y'nin ortalamasına eşit ve sabittir. k küçükse, f daha hızlı değişir [15].

Ağırlıkların oluşturmasında bilgi vermek için aşağıdaki örneği dikkate alalım. $\{(x_i, y_i)\}_{i=1}^{5} = \{(1,5),(7,12),(3,1),(2,0),(5,4)\}$ olduğunu varsayalım. $x = 4$ ve $k = 3$ için $\hat{f}_k(x)$'in $k - NN$ tahminini hesaplayalım. x'e en yakın k gözlemleri son üç veri noktasıdır, bu yüzden $J_x = J_4 = \{3,4,5\}$ ve böylece

$$W_{k1}(4) = 0,\ W_{k2}(4) = 0,\ W_{k3}(4) = 1/3,$$
$$W_{k4}(4) = 1/3 \text{ ve } W_{k5}(4) = 1/3$$

olur. Buna göre $k - NN$ tahmini,

$$\hat{f}_3(4) = (1 + 0 + 4)/3 = 5/3$$

olarak hesaplanır.

2.5.2. Kernel Düzeltmesi

Bu bölümde, (2.2) parametrik olmayan regresyon modeli kestiricisinin daha genel bir şekli dikkate alınmaktadır. Böyle bir parametrik olmayan regresyon kestiricisi, ilk kez Nadarya [16] ve Watson [17] tarafından önerilmiştir:

$$\hat{f}(x) = \sum_{i=1}^{n} W_i(x) y_i \qquad (2.9)$$

Bu durumda x'e bağlı $W_i(x)$ ağırlıkları yardımıyla, y_i'nin bir ağırlıklı toplamı olarak x noktasında regresyon fonksiyonunu tahmin edilir. Lokal ortalama ağırlıkları elde etmek için kavramsal olarak en uygun yol, bir ölçek parametresi tarafından kontrol edilen, her iki yönde azalan ve merkezi sıfırda bulunan tek moda sahip bir dağılımı kullanmaktır. Bu amaçla uygulamada, genellikle

"kernel" olarak bilinen olasılık yoğunluk fonksiyonlar kullanılır [9]. Sürekli ve sınırlı kernel fonksiyonu, aşağıda belirtildiği gibi integrali 1'e eşit olan simetrik reel bir K fonksiyonudur:

$$\int K(u)du = 1 \tag{2.10}$$

Kernel düzeltme için ağırlık dizisi

$$W_i(x) = \frac{\frac{1}{\lambda n} K\left(\frac{x_i - x}{\lambda}\right)}{\frac{1}{\lambda n}\sum_1^n K\left(\frac{x_i - x}{\lambda}\right)} = \frac{\frac{1}{\lambda n} K(u)}{\frac{1}{\lambda n}\sum_1^n K(u)} \tag{2.11}$$

biçiminde ifade edilir. Burada,

n : Gözlemlerin sayısı

K : Seçilen kernel fonksiyonu

λ : Düzeltme penceresinin yarıçapına karşı gelen, bir düzeltme parametresi

$W_i(x) = W(x, x_i)$: $x - x_i$ uzaklığına bağlı ve i .gözlem y_i 'ye atanan ağırlık

olarak tanımlanır.

Bu durumda, i .gözlemin w ağırlığı, $x - x_i$ uzaklığının bir fonksiyonudur. Genellikle, uzaklık küçükse ağırlık yüksek ve uzaklık büyükse ağırlık düşük olur. Ağırlıklar K tarafından belirtilir ve *bant genişliği* (düzeltme parametresi) olarak bilinen, λ tarafından kontrol edilir. Diğer bir deyişle, ağırlıkların hacmi λ tarafından parametreleştirilir [5]. Buna göre, $W_i(x)$ ağırlık dizisi (2.9)'da yerine yazıldığında, kernel tahmini aşağıdaki şekli alır:

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

(2.12) tahmin genellikle, *"Nadarya –Watson kestiricisi"* olarak adlandırılır.

Uygulamada kullanılan farklı tipte kernel fonksiyonları vardır. Ancak kernel fonksiyonunun seçimi bant genişliğinin seçiminden daha az önemsizdir. Bazı kernel fonksiyonları aşağıda verilmiştir:

1. Normal kernel $K(u)=\frac{1}{\sqrt{2\pi}}e^{(-\frac{1}{2}u^2)}, u\in[-\infty,\infty]$

2. Düzgün kernel (dikdörtgen veya kutu) : $K(u)=\frac{1}{2},\ u\in[-1,1]$

3. Üçgensel kernel : $(1-|u|),\ u\in[-1,1]$

4. Epanechnikov kernel: $\frac{3}{4}(1-u^2),\ u\in[-1,1]$

5. Dördüncü dereceden kernel (Quartic): $\frac{15}{16}(1-u^2)^2,\ u\in[-1,1]$

6. Altıncı dereceden kernel (triweight) $\frac{35}{32}(1-u^2)^3,\ u\in[-1,1]$

Kernel fonksiyonları içinde en basit olanı [-1,1] aralığında $\frac{1}{2}$ ve diğer durumlarda sıfır değerini alan, düzgün kernel fonksiyonudur. Fakat normal ve diğer kernel fonksiyonları da yaygın olarak kullanılır [18]. (2.12) kernel fonksiyonlarında, λ parametresinin önemli bir rol oynar: *Büyük λ değeleri* için eğri çok yavaş değişir ve düzeltme önemlidir. Bu durumda, tahminin varyansı sınırlı, fakat tahmin oldukça sapmalıdır. *Küçük λ değerleri için* eğri oldukça düzensizdir ve sapmalar sınırlı fakat tahminin varyansı büyüktür. Bu yüzden, λ parametresi sapmalar ve tahminin doğruluğu arasında bir uzlaşma sağlar [15].

2.5.3. Ortogonal Serilerin Kestiricileri

Regresyon fonksiyonunun aşağıdaki şekilde bir *Fourier serisi* olarak gösterilebildiği varsayılsın:

$$f(x)=\sum_{j=0}^{\infty}\beta_j\varphi_j(x) \qquad (2.13)$$

Burada $\{\varphi_j\}_{j=0}^{\infty}$ bilinen bir taban oluşturan fonksiyon ve $\{\beta_j\}_{j=0}^{\infty}$ bilinmeyen Fourier katsayılarıdır. Taban oluşturan fonksiyonların iyi bilinen örnekleri *Laguerre ve Lengendre polinomlarıdır*. Önce taban oluşturan fonksiyon belirlenir, daha sonra β_j fourier katsayıları tahmin edilerek f fonksiyonu tahmin edilir. (2.13)'de sonsuz sayıda sıfırdan farklı β_j katsayılarının olabileceğinden, verilen

sınırlı n örneklem hacmi için, sadece katsayıların bir alt kümesi etkin olarak tahmin edilebilir [5]. $\left\{\hat{\beta}_j\right\}_{j=0}^{t(n)}$ en küçük kareler parametre tahminleri olmak üzere, (2.13)'de $t(n)$ terim dikkate alınırsa, f fonksiyonunun "*ortogonal seri kestiricisi*" aşağıdaki gibi tahmin edilir:

$$\hat{f}_{t(n)}(x) = \sum_{j=0}^{t(n)} \hat{\beta}_j \varphi_j(x) = \sum_{i=0}^{n} W_{ni}(x) Y_i \qquad (2.14)$$

Burada $W_n(x) = (W_{n1}, ..., W_{nn})^T$, $W_n(x) = \varphi_{tx}^T (\mathbf{\Phi_t^T \Phi})^{-1} \mathbf{\Phi}_t^T$ ile elde edilir ve ayrıca $\varphi_{tx} = (\varphi_0(x), ..., \varphi_n(x))^T$ *ve* $\mathbf{\Phi}_t = (\varphi_{t1}, ..., \varphi_{tn})^T$ dir.

Bu kestiriciler kolaylıkla hesaplanabilir. Ayrıca, toplamsal yapılara ve semiparametrik modellere uzantısı uygundur. Ortogonal seriler kestiricileri tekniğinin en önemli dezavantajı $t(n)$ düzeltme parametresi ve temel sistemin nasıl seçileceği konusunda oldukça az sayıda çalışma olmasıdır [4].

2.5.4. Lokal Regresyon Kestiricileri

Lokal ortalamanın doğal bir uzantısı, *lokal regresyon* düşüncesidir. Prensip olarak regresyon fonksiyonu, herhangi bir x noktasının komşuluğunda düşük dereceden bir polinom (doğrusal bir fonksiyon) ile çok iyi tahmin edilebilir. λ, lokal komşulukların hacmi (büyüklüğü) olsun. O zaman verilen bir x noktası için veriler aşağıdaki gibi modellendirebilir:

$$y_i = a(x) + b(x) x_i + hata,\ x - \lambda \le x_i \le x + \lambda$$

Burada $a(x)$ ve $b(x)$ iki lokal parametredir. Ele alınan gözlemler eşit ağırlıklı ise, doğal olarak bu model aşağıdaki gibi *lokal regresyon* problemine götürür:

$$\min_{a,b} \sum_{i=1}^{n} \{y_i - a(x) - b(x) x_i\}^2 \qquad (2.15)$$

(2.15) ifadesini minimum yapan $\hat{a}(x)$ *ve* $\hat{b}(x)$ bulunarak, bir x noktasında elde edilen regresyon fonksiyonunun kestiricisi,

$$\hat{f}(x) = \hat{a}(x) + \hat{b}(x) x \qquad (2.16)$$

olarak elde edilir [19]. Bağımsız değişkeninin tüm olası değerlerinin kümesindeki noktaların dizisinde bu işlem tekrar edilerek, f regresyon fonksiyonunun parametrik olmayan kestiricisi elde edilir.

Yukarıdaki fikrin geliştirilmiş bir durumu, x noktasından uzak bir x_i gözleminden başlayarak katkı yükleyen, bir pürüzsüz ağırlık fonksiyonunu dahil etmektir. Diğer bir ifadeyle, daha uzak gözlemlere daha düşük ve daha yakın gözlemlere daha yüksek ağırlık atayan ağırlıklı regresyon yapılabilir. Bunu elde etmenin doğal bir yolu λ bant genişliği parametresi tarafından kontrol edilen ve bir kernel fonksiyonu tarafından belirtilmiş olan ağırlıkları sağlamaktır. Bu durumda *lokal ağırlıklı regresyon*,

$$\min_{a,b} \sum_{i=1}^{n} \left[y_i - a(x) - b(x) x_i \right]^2 K\left(\frac{x_i - x}{\lambda} \right) \qquad (2.17)$$

ifadesinin minimum problemine dönüşür. Eşitlik (2.15)'de olduğu gibi, (2.17)'yi minimum yapan $\hat{a}(x) ve \hat{b}(x)$ bulunur. Bu işlemden bazen kernel regresyonu olarak söz edilir çünkü lokal regresyona kernel ağırlıklarını uygular. (2.17)'deki doğrusal fonksiyon bir polinom ile yer değiştirerek, işlem lokal polinom regresyona genelleştirilebilir.

2.5.5. Splayn Regresyonu

Splayn düzeltme tahmini, n tane polinomial parçadan oluştuğundan tahmini tanımlamak için parametre sayısı çok fazladır. Sadece bir pürüzlülük cezası konularak tahminin varyansı kontrol altına alınabilir. Varyansı kontrol altına almanın alternatif bir yolu da modeldeki polinomial parça sayısını (böylece, parametre sayısını) azaltmaktır. Böylece, her bir x_i veri noktasında (splayn düzeltmede yapıldığı gibi) bir düğüm koymak yerine, daha az sayıda $k_1 <, ..., < k_m$ düğüm noktaları kullanılabilir.

Spalyn düzeltme her bir x_i, $i = 1, ..., n,$ veri noktasında bir düğüm noktası koyar. Daha az sayıda düğüm noktasının kullanımı, bu düğüm noktalarını koyacak yer sorununu ortaya çıkarır. Splayn eğrisi $[x_i, x_{i+1}]$ aralıklarıyla sınırlandırılan basit bir kübik polinom olduğundan, $f(x)$ fonksiyonunun daha karmaşık bir

şekle sahip olduğu bölgelerde çok sayıda düğüme sahip olması gerekir. Düğüm noktaları için basit düğüm seçme şemaları genellikle yeterlidir.

Parametrik olmayan eğri tahmininde splayn düzeltme yönteminin alternatif yaklaşımı olan splayn regresyon düzeltmede, splayn regresyonu uyumunu elde etmek için birkaç yöntem vardır. Bu yöntemlerden biri de, regresyon yoluyla splayn interpolasyonudur. Burada amaç pürüzsüz olduğu varsayılan (2.2) modelindeki f fonksiyonunun tahminidir. Özel olarak, aşağıda verilen m sayıda düğüm noktası içeren r.dereceden splayn fonksiyonu (*budanmış üstel temel fonksiyon*) ile f fonksiyonu çok iyi tahmin edilebilir:

$$f(x) = b_0 + b_1 x + ... + b_r x^r + \sum_{j=1}^{m} \beta_j (x - k_j)_+^r \qquad (2.18)$$

Burada r, splayn regresyonunun derecesi (genellikle önceden seçilir) ve k_j, j.düğüm noktası, $\beta = (\mathrm{b}_0, ..., \mathrm{b}_\mathrm{r}, \beta_1, ..., \beta_m)^T$ regresyon katsayılarının dizisi ve $(a)_+ = \max(0, a)$ Ayrıca $\min(x_i) < k_1, ..., < k_m < \max(x_i)$ ve $\{k_1, ..., k_m\}$ dizisi $\{x_1, ..., x_n\}$'nin bir alt kümesi oluğu varsayılır [20].

Eşitlik (2.18) ile verilen splayn regresyon fonksiyonunun tahmini olan $\hat{f}$ fonksiyonu, r, m, $\mathbf{k} = (k_1, ..., k_m)^T$ ve $\boldsymbol{\beta} = (b_0, ..., b_r, \beta_1, ..., \beta_m)^T$ parametrelerinin tahmini yoluyla elde edilebilir:

$$\hat{f}(x) = \hat{b}_0 + \hat{b}_1 x + ... + \hat{b}_{\hat{r}} x^{\hat{r}} + \sum_{j=1}^{m} \hat{\beta}_j (x - \hat{k}_j)_+^{\hat{r}} \qquad (2.19)$$

Burada yer alan $\hat{r}, \hat{m}, \hat{\mathbf{k}}$ *ve* $\hat{\boldsymbol{\beta}}$ sabitleri sırasıyla $r, m, \mathbf{k}$ *ve* $\boldsymbol{\beta}$ sabitlerinin tahminleridir. Eğer $\hat{r}, \hat{m}$ *ve* $\hat{\mathbf{k}}$ önceden belirtilirse, o zaman $\boldsymbol{\beta} = (\mathrm{b}_0, ..., \mathrm{b}_\mathrm{r}, \beta_1, ... \beta_\mathrm{m})^\mathrm{T}$ katsayılarının kestiricileri,

$$\mathrm{X} = \begin{bmatrix} 1 & x_1 & . & . & . & x_1^r & (x_1 - k_1)_+^r & . & . & . & (x_1 - k_m)_+^r \\ . & . & . & & & & . & . & & & \\ . & . & & . & & & . & & . & & \\ . & . & & & . & & . & & & . & \\ 1 & x_n & . & . & . & x_n^r & (x_n - k_1)_+^r & & & & (x_n - k_m)_+^r \end{bmatrix}$$

şeklinde verilen bağımsız değişkenler matrisinin tanımlandığı en küçük kareler regresyonu yoluyla tahmin edilebilir:

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \mathbf{y} .$$

(2.18) kullanılarak $f(x)$'i elde etmek için, β parametrelerinin tahminin yanı sıra, düğümlerin yerini ve sayısını seçmek gerekir. Bu işlemi gerçekleştirmek için iki genel strateji vardır. Birinci strateji, sıradan en küçük kareler kullanılarak β parametrelerinin tahmini ve oldukça az sayıda düğüm noktalarını seçmektir. Bu strateji ile, düğümlerin seçimi son derece önemlidir. İkinci strateji ise, oldukça çok sayıda düğüm kullanmaktır. Ancak, β parametrelerinin tahmini için sıradan en küçük kareler kullanılmaz. Birinci stratejinin tersine, ikici strateji için düğümlerin seçiminin önemi daha azdır: Önemli olan β parametrelerinin nasıl tahmin edileceğidir (örneğin, cezalı en küçük kareler ile) [21].

Model (2.2)'de belirtilen f fonksiyonunun tahmini için (2.18)'deki budanmış üstel taban oluşturan fonksiyonlara (*truncated power basis functions*) alternatif olarak, $k_1,...,k_m$ düğüm noktalı, $B_1,...,B_{m+r}$ *kübik* $B-Splaynları$ kullanılabilir. $k = k(j),\ j = 1,...,m$, azalmayan bir dizi olak üzere, k. düğüm dizisi için r. dereceden. j. $B-Splayn$ (*basis spline*), $B_{j,r,k}$ ile gösterilir ve

$$B_{j,r,k}(x) = (k_{j+k} - k_j)[k_j,...,k_{j+r}](k-x)_+^{r-1},\ x \in \square \qquad (2.20)$$

ile tanımlanır. Genel olarak $B_{j,r,k}$ notasyonu yerine B_j kullanılır [22].

Daha genel olarak, $k_1,...,k_m$ düğüm noktaları ile $r \geq 0$ dereceden [*a*, *b*] aralığındaki $B-splayn$ *fonksiyonları* aşağıdaki gibi tanımlanan ardışık tekrar bağıntılarıdır:

$$k_1 = k_2 = k_3 = k_4 = \mathrm{a},\ k_j < k_{j+r}\ (j = 1,...m)\ ve\ k_{5+\mathrm{r}} = \mathrm{k}_{6+\mathrm{r}} = \mathrm{k}_{7+\mathrm{r}} = \mathrm{k}_{8+\mathrm{r}} = \mathrm{b}$$

olsun. Bu durumda, r. *dereceden* j. $B-splayn$

$$B_{j,r}(x) = \frac{x-k_j}{k_{j+r-1}-k_j} B_{j,r-1}(x) + \frac{k_{j+r}-x}{k_{j+r}-k_j} B_{j+1,r-1}(x),\ k_j \leq x < k_{j+r}\ ,\ j = 1,...,m+4 \qquad (2.21)$$

şeklinde ifade edilir. Ayrıca burada, $r = 1$ için:

$$B_{j,1}(x) = \begin{cases} 1, \; k_j \le x < k_{j+1} \\ 0, \; \textit{diğer durumlarda} \end{cases}$$

olarak tanımlanır [22]. Buradan hareketle, *kübik* $B-$*splayn*, $B_j(x) = B_{j,3}(x)$ biçiminde ifade edilir. Böylece, seçilen düğümler ile bir kübik splayn

$$f(x) = \sum_{j=1}^{m+4} \beta_j B_j(x) \tag{2.22}$$

olarak ifade edilir. β_j, $j = 1,...,m+4$ bilinmeyen regresyon katsayıları, en küçük kareler regresyonu yoluyla aşağıdaki gibi tahmin edilir:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m+4} \beta_j B_j(x_i) \right)^2 \tag{2.23}$$

Buna göre, $\hat{\beta}_j$, $j = 1,...,m+4,$ tahmin edilen regresyon katsayılarını göstermek üzere, splayn regresyonuna göre parametrik olmayan regresyon fonksiyonu,

$$\hat{f}(x) = \sum_{j=1}^{m+4} \hat{\beta}_j B_j(x) \tag{2.24}$$

şeklinde tahmin edilir.

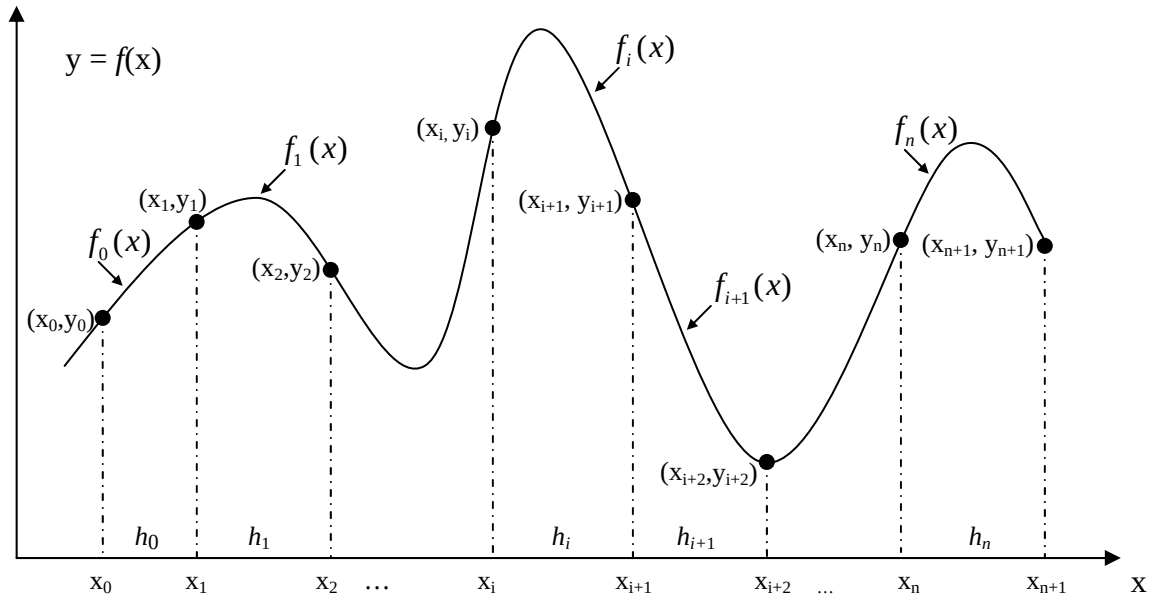
3. KÜBİK SPLAYN İNTERPOLASYONU VE PARAMETRİK OLMAYAN REGRESYONDA SPLAYN DÜZELTME YÖNTEMİ

Kelime anlamı ağaç parçasının ince bir şeridi olan splayn (*spline*), bir dizi veri noktalarına polinomial bir eğri uydurma ya da bu noktalar arasından pürüzsüz olarak geçen ve birçok parçadan oluşan esnek bir eğridir. Splayn fonksiyonlar bu fikrin uygulaması olan yeni bir matematiksel araçtır. Bu fonksiyonların temel düşüncesi, tanımlanan aralığı bağımsız (tasarım) değişkenlerin gözlem değerleri yardımıyla alt aralıklara bölerek, her bir alt aralıkta farklı bir polinomial fonksiyon ile bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellendirerek istenilen mertebeden türevi olan sürekli bir fonksiyon elde etmektir.

Modern istatistik teorisi, verilere parametrik olmayan modellerin uyumu ile başlamıştır. Splaynlar parametrik olmayan fonksiyonu tahmin eden yöntemlerden biri olarak sunulur ve genellikle parametrik olmayan regresyon ortamında ele alınır. Bu fonksiyonlar klasik parametrik çıkarsamanın bir gelişimi olup parametrik ve parametrik olmayan modeller arasındaki boşluğu doldururlar. Splaynlar parametrik bir fonksiyon şeklinde değilken, çoğu durumlarda bir polinomial gösterime sahip olan temel fonksiyonların bir birleşimi olarak yazılabilir ve böylece bir anlamda parametrik olur. Bu bölümde, genel splayn fonksiyonunun tanımının yanı sıra, kübik splayn ve doğal kübik splayn'nın bir incelemesi ile parametrik olmayan regresyon ortamında splayn'nın kullanımı açıklanacaktır.

3.1. Splayn Fonksiyonu ve Parçalı Kübik Splayn

$(x_i, y_i)_{i=1}^{n}$ gözlem değerleri olsun. Herhangi bir $[a,b]$ aralığında $a = x_0 < x_1 < ... < x_n < x_{n+1} = b$ koşulunu sağlayan x_i reel sayılarının $(a, x_1), (x_1, x_2), ..., (x_i, x_{i+1}), ..., (x_{n-1}, x_n), (x_n, b)$ alt aralıklarına bölündüğünü ve her bir $x_i \le x \le x_{i+1},\ i = 0,1,...,n$ alt aralığında $k.$dereceden bir $f_i(x)$ polinom fonksiyonu tanımlandığını varsayalım. Buradaki x_i sayıları, *"düğüm noktaları"* olarak adlandırılır. Polinomlardan oluşan ve splayn adı verilen $f(x)$ fonksiyonu, Şekil 3.1'deki gibi alt aralıklarda tanımlanan $f_0(x), f_1(x), ..., f_i(x), ..., f_n(x)$



Şekil 3.1: f_i (x) polinom fonksiyonlar ve $h_i = x_{i+i} - x_i$, i. alt aralığın uzunluğu ($i = 0,1,...,n$) olmak üzere bir f splayn fonksiyonunun grafiği.

biçimindeki $k.$dereceden $(k \geq 1)$ polinom fonksiyonların birleşimi olarak tanımlanır:

$$f(x) = f_i(x), \quad x_i \leq x \leq x_{i+1}, i = 0,1,...,n$$
$$ve$$
$$f_i(x_{i+1}) = f_{i+1}(x_{i+1}), \quad i = 0,1,...,n$$

Burada $f_i(x), \quad x_i \leq x \leq x_{i+1}, \quad i = 0,1,...,n.,$ olmak üzere $k.$dereceden bir polinomdur. $k = 1$ için $f(x)$ *parçalı-doğrusal splayn,* $k = 2$ için *karesel splayn,* $k = 3$ olduğunda *kübik splayn,* ve benzer şekilde farklı dereceden splayn fonksiyonları tanımlanabilir. Buna göre, söz konusu $[a,b]$ aralığında tanımlanan ve aşağıdaki özelliklere sahip bir $f(x)$ fonksiyonu ***kübik splayn*** olarak adlandırılır [23-26]:

I. $f(x) = f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x_i \leq x \leq x_{i+1}, \quad i = 0,...,$n.

II. $f(x_i) = y_i$, i = 1,...,n. Splayn her bir veri noktasından geçer.

III. $f_i(x_{i+1}) = f_{i+1}(x_{i+1}), \quad i$ = 0,1,...,n-1. Splayn sürekli bir fonksiyon şeklindedir.

IV. $f_i'(x_{i+1}) = f_{i+1}'(x_{i+1}), \quad i$ = 0,1,...,n-1. Splayn pürüzsüz bir fonksiyon şeklindedir.

V. $f_i''(x_{i+1}) = f_{i+1}''(x_{i+1}),\quad i = 0,1,...,\text{n}-1.$ İkinci türevi düğüm noktalarında süreklidir.

Yukarda belirtilen özelliklerinden de anlaşılacağı gibi, $[a,b]$ aralığı üzerinde tanımlı bir f fonksiyonu aşağıdaki iki koşulu sağlıyorsa *bir kübik splayn*'dır: Birincisi, $(a,x_1),(x_1,x_2),...,(x_i,x_{i+1}),...,(x_n,b)$ alt aralıklarının her birinde f kübik polinomdur; ikincisi f 'in kendisi, birinci ve ikinci türevleri her bir x_i düğüm noktalarında, böylece $[a,b]$ aralığının tümü üzerinde süreklidir. Bu şekilde, polinomial parçalar x_i düğüm noktalarında kesintisiz uyum sağlar.

Kübik splaynı elde etmenin birkaç yolu vardır. Her bir kübik parçanın dört polinom katsayısı ile belirlenmesi kübik splaynın en açık şekillerinden biridir:

$$f(x) = f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3,\ x_i \le x \le x_{i+1},\ i = 0,...,\text{n} \qquad (3.1)$$

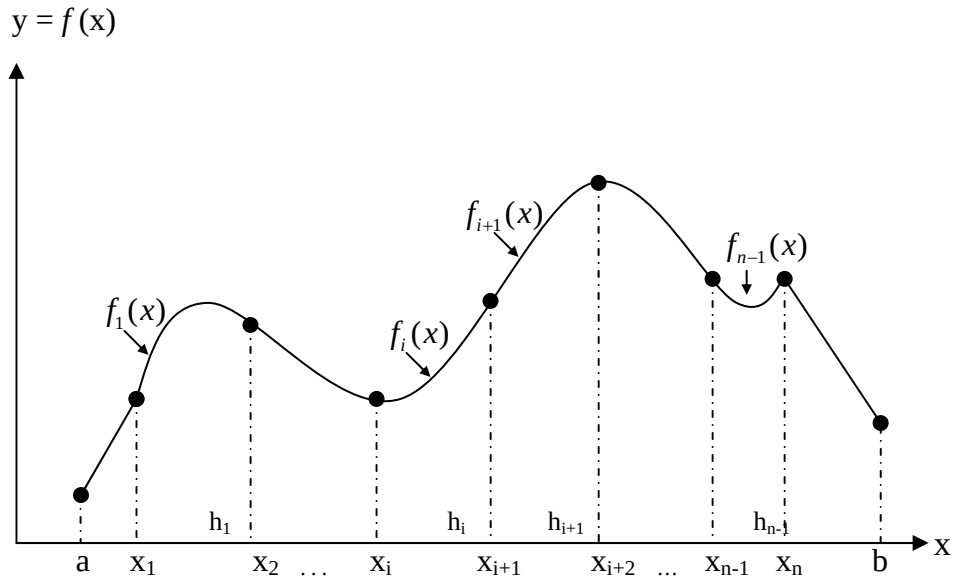
Burada a_i, b_i, c_i ve d_i, $i = 0,1,...,n$ sabitleri i.polinomun katsayılarıdır. $x_0 = a$ ve $x_{n+1} = b$ olarak tanımlanır. f ve ilk iki türevi üzerindeki süreklilik koşulları, katsayılar arasındaki değişik ilişkileri ifade eder. Örneğin, $i = 0,1,...,n-1$ için x_{i+1}.düğüm noktasındaki f fonksiyonunun sürekliliği aşağıdaki eşitlikleri verir:

$$f_{i+1}(x_{i+1}) = d_{i+1}(x_{i+1} - x_{i+1})^3 + c_{i+1}(x_{i+1} - x_{i+1})^2 + b_{i+1}(x_{i+1} - x_{i+1}) + a_{i+1} = a_{i+1}$$

ve $f_i(x_{i+1}) = f_{i+1}(x_{i+1})$ eşit olduğundan

$$f_i(x_{i+1}) = d_i(x_{i+1} - x_i)^3 + c_i(x_{i+1} - x_i)^2 + b_i(x_{i+1} - x_i) + a_i = a_{i+1}.$$

$[a,b]$ aralığında bir kübik splayn, a ve b uç noktalarında ikinci ve üçüncü türevlerinde sıfır değerini alırsa buna *doğal kübik splayn* (*natural cubic spline – NCS*) denir. Bu ek koşullar *doğal sınır koşulları* olarak adlandırılır. Buna göre, $d_0 = c_0 = d_n = c_n = 0$ olduğu bulunur. Bu durum, Şekil 3.2'de gösterilen doğal kübik splaynın grafiğinde görüldüğü gibi $[a, x_1]$ ve $[x_n, b]$ aralıklarında $f(x)$ fonksiyonunun doğrusal olmasını sağlar.



Şekil 3.2: $f_i(x)$ kübik parçalar, $h_i = x_{i+1} - x_i$, i. alt aralığın uzunluğu ve x_i, $i = 1,...,n$, düğüm noktaları olmak üzere bir doğal kübik splayn fonksiyonu grafiği.

3.1.1. Kübik Splayn İnterpolasyonu: Lagrange Yöntemi

$f(x)$ parçalı kübik polinom olduğundan, onun $f''(x)$, ikinci mertebeden türevi $[a,b]$ aralığında parçalı doğrusaldır. $f''(x) = f_i''(x)$ fonksiyonu için $[x_i, x_{i+1}]$ $(i = 0,1,...,n)$ aralığında *doğrusal lagrange interpolasyon* formülü aşağıdaki gibidir:

$$f_i''(x) = f''(x_i)\frac{x - x_{i+1}}{x_i - x_{i+1}} + f''(x_{i+1})\frac{x - x_i}{x_{i+1} - x_i},\ x_i \le x \le x_{i+1},\ i = 0,1,...n, \qquad (3.2)$$

$f(x)$ fonksiyonunun x_i düğüm noktasındaki ikinci mertebeden türevi $M_i = f''(x_i)$ ile, i.aralığın uzunluğu $h_i = x_{i+i} - x_i$ olarak gösterilirse, (3.2) formülünü aşağıdaki gibi yazılabilir:

$$f''(x) = f_i''(x) = \frac{M_i}{h_i}(x_{i+1} - x) + \frac{M_{i+1}}{h_i}(x - x_i), \quad x_i \le x \le x_{i+1},\ i = 0,1,...,n. \qquad (3.3)$$

$f(x)$ fonksiyonunu elde etmek için (3.3) eşitliğinin iki kez integrali alınır. Buna göre p_i ve q_i gibi iki integral sabiti ile $f(x)$ fonksiyonu aşağıdaki gibi olur:

$$f_i(x) = \frac{M_i}{6h_i}(x_{i+1} - x)^3 + \frac{M_{i+1}}{6h_i}(x - x_i)^3 + p_i(x_{i+1} - x) + q_i(x - x_i),\ x_i \le x \le x_{i+1},\ i = 0,1,...,n, \qquad (3.4)$$

$y_i = f_i(x_i)$ *ve* $y_{i+1} = f_i(x_{i+1})$ olduğuna göre, p_i ve q_i, $i = 0,1,...,n$ sabitlerinin belirlenmesi için (3.4)'de x yerine sırasıyla x_i ve x_{i+1} yazılarak aşağıdaki denklemler elde edilir.

$$y_i = \frac{M_i}{6} h_i^2 + p_i h_i \quad ve \quad y_{i+1} = \frac{M_{i+1}}{6} h_i^2 + q_i h_i, \quad i = 0,1,...,n, \tag{3.5}$$

(3.5)'den p_i ve q_i sabitleri bulunarak (3.4) denkleminde yerine yazıldığında, sonuç olarak $f_i(x)$ kübik fonksiyonu için aşağıdaki ifade yazılabilir:

$$f_i(x) = \frac{M_i}{6h_i}(x_{i+1} - x)^3 + \frac{M_{i+1}}{6h_i}(x - x_i)^3 + \left(\frac{y_i}{h_i} - \frac{M_i h_i}{6}\right)(x_{i+1} - x) + \left(\frac{y_{i+1}}{h_i} - \frac{M_{i+1} h_i}{6}\right)(x - x_i), \tag{3.6}$$
$$x_i \le x \le x_{i+1},\ i = 0,1,...,n$$

(3.6) ifadesinde sadece M_i, $i = 0,1,...,n$ sayıları bilinmeyenlerdir. Bu değerleri bulmak için parçalı kübik splayn'nın tanımındaki IV. özelliği uygulamak gerekir. Buna göre, (3.6) fonksiyonunun türevi aşağıdaki gibi olur:

$$f_i'(x_i) = -\frac{M_i}{2h_i}(x_{i+1} - x)^2 + \frac{M_{i+1}}{2h_i}(x - x_i)^2 - \left(\frac{y_i}{h_i} - \frac{M_i h_i}{6}\right) + \frac{y_{i+1}}{h_i} - \frac{M_{i+1} h_i}{6} \tag{3.7}$$

(3.7)'de x yerine x_i yazılırsa, sonuç olarak

$$f_i'(x_i) = -\frac{M_i}{3} h_i - \frac{M_{i+1}}{6} h_i + d_i, \qquad d_i = \frac{y_{i+1} - y_i}{h_i} \tag{3.8}$$

ifadesini elde edilir. $f_{i-1}'(x)$ için benzer bir ifade elde etmek amacıyla (3.7) eşitliğindeki i indisi $(i-1)$ indisiyle değiştirilebilir:

$$f_{i-1}'(x_i) = \frac{M_i}{3} h_{i-1} + \frac{M_{i-1}}{6} h_i + d_{i-1}, \qquad d_{i-1} = \frac{y_i - y_{i-1}}{h_{i-1}} \tag{3.9}$$

M_i, $i = 0,1,...,n$ sayılarının bulunması için kübik splaynın tanımındaki IV. özelliği, başka bir anlatımla $f_i'(x_i) = f_{i-1}'(x_i)$, $i = 1,...,n$ eşitliği uygulanır. Böylece, (3.8) ve (3.9) ifadelerini kullanarak, M_i, $i = 0,1,...,n$ katsayıları için aşağıdaki lineer denklemler sistemi elde edilir:

$$h_{i-1}\mathrm{M}_{i-1} + (2h_{i-1} + 2h_i)\mathrm{M}_i + h_i\mathrm{M}_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right), \quad i = 1,..,n. \qquad (3.10)$$

(3.10) denklemleri $(n+2)$ sayıda M_0, M_1, ... , M_{n+1} bilinmeyenlerini içeren n tane lineer denklemler sistemidir. Değişkenlerin sayısı denklemlerin sayısından iki fazladır. Farklı kübik splaynların tanımında a ve b uç noktalarında farklı koşullar verilerek değişkenler ve denklemlerin sayısı arasındaki eşitsizlik kaldırılır. Aşağıdaki Tabloda sınır (uç) noktalarındaki koşullara bağlı olarak çeşitli kübik splaynlar gösterilmiştir [23]: Tablo 3.1'de yer alan stratejilerdeki bağıntılar ile sırasıyla kısaca şu ifadeler anlatılmaktadır: *Birleştirilen (kenetli) splayn*, $f'(a) = d_0$ *ve* $f'(b) = d_{n+1}$ birinci türev sınır koşullarını içeren bir kübik splaynı belirtir ve sınırlardaki eğimleri içerir. *Doğal kübik splayn*, $f''(a) = 0$ *ve* $f''(b) = 0$ serbest sınır koşullarına sahip bir kübik splayn olup düğüm noktaları arasından

Tablo 3.1: Kübik Splayn İçin Sınır Noktalarının Kısıtlamaları

Stratejinin Tanımı	M_0 ve M_{n+1}'e İlişkin Denklemler
(i) *Birleştirilen (kenetli) kübik splayn*: $f'(x_0),\ f'(x_{n+1})$ türevlerini içerir.	$\mathrm{M}_0 = \frac{3}{h_0}\left[d_0 - f'(x_0) - \frac{\mathrm{M}_1}{2}\right],$ $\mathrm{M}_{n+1} = \frac{3}{h_n}\left[f'(x_{n+1}) - d_n\right] - \frac{\mathrm{M}_n}{2}$
(ii) *Doğal kübik splayn (esnek bir eğri)*: $f''(x_0) = f''(x_{n+1}) = 0$	$\mathrm{M}_0 = 0,\ \mathrm{M}_{n+1} = 0$
(iii) *Extrapole edilen kübik splayn*: Sınır noktalarına göre $f''(x)$'in extrapolasyonu.	$\mathrm{M}_0 = \mathrm{M}_1 - \frac{h_0(\mathrm{M}_2 - \mathrm{M}_1)}{h_1},$ $\mathrm{M}_{n+1} = \mathrm{M}_n + \frac{h_n(\mathrm{M}_n - \mathrm{M}_{n-1})}{h_{n-1}}$
(iv) *Parabol olarak sonuçlanan kübik splayn*: Sınır noktaları çevresinde $f''(x)$ sabittir.	$\mathrm{M}_0 = \mathrm{M}_1,\ \mathrm{M}_{n+1} = \mathrm{M}_n$
(v) *Ayarlanmış kübik splayn*: Her bir sınır noktasında $f''(x)$ bilinir.	$\mathrm{M}_0 = f''(x_0),\ \mathrm{M}_{n+1} = f''(x_{n+1})$

geçen esnek bir eğridir. *Extrapole edilen splayn,* $f''(a)$'yı belirlemek için x_1 *ve* x_2'deki iç düğümlerden, $f''(b)$'yi belirlemek için x_n *ve* x_{n-1}'deki düğümlerden elde edilen extrapolasyonu kullanan bir kübik splayndır. Extrapole edilen splayn, son kübiği komşu kübiğin bir uzantısı olması varsayımına eşdeğerdir. Diğer bir deyişle, splayn şekilleri $[x_0, x_2]$ aralığında tek bir kübik eğri ve bir diğeri $[x_{n-1}, x_{n+1}]$ aralığında tek kübik eğridir. *Parabolik olarak sonuçlanan splayn,* $[x_0, x_1]$ aralığında $f''(x) \equiv 0$ ve $[x_n, x_{n+1}]$ aralığında $f''(x) \equiv 0$ varsayımlarını kullanan bir kübik splayndır. $[x_0, x_1]$ aralığında $f''(x) \equiv 0$ varsayımı kübiği $[x_0, x_1]$ aralığında karesel olmaya zorlar ve benzer bir durum $[x_n, x_{n+1}]$ aralığında da gerçekleşir. *Ayarlanmış splayn,* $f''(a)$ ve $f''(b)$ ikinci türev sınır koşulları açıkça belirtilen bir kübik splayndır. $f''(a)$ ve $f''(b)$ ikinci türev değerleri, her bir uç noktadaki eğrilikte ayarlama yaparlar.

3.1.2. Kübik Splayn İnterpolasyonu : Lokal Polinomial Yöntem

Bu yaklaşımda doğrudan kübik splayn tanımındaki I-IV özelliklerini kullanarak, interpolyasyon kat sayılarının bulunması için denklemlerin yazılması amaçlanmaktadır. I.özelliğe dayanarak splayn fonksiyonunu şöyle yazılabilir:

$$f(x) = f_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad x_i \le x \le x_{i+1}, \; i = 0,1,...,n. \tag{3.11}$$

II – III özelliği ve (3.11) ifadesine göre

$$y_i = f_i(x_i) = a_i, \; i = 0,1,...,n \tag{3.12}$$

ve

$$y_{i+1} = f_i(x_{i+1}) = a_i + b_i h_i + c_i h_i^2 + d_i h_i^3, \; i = 0,1,...,n \tag{3.13}$$

olarak yazılır. Burada $h_i = x_{i+1} - x_i$ i.nci alt aralığın uzunluğunu göstermektedir. Şimdi (3.11)'de birinci ve ikinci türevleri hesaplanırsa,

$$f_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \tag{3.14}$$

$$f_i''(x) = 2c_i + 6d_i(x - x_i) \tag{3.15}$$

olarak elde edilir. $f(x)$ fonksiyonunun x_i düğüm noktasındaki ikinci mertebeden türevi $\mathrm{M}_i = f''(x_i)$ ve $i.$aralığın uzunluğu $h_i = x_{i+i} - x_i$, $i = 0,1,...,n$ ile gösterilirse, kübik splayn tanımının II, III özellikleri ve (3.11) – (3.15) formüllerinin yardımıyla katsayılar aşağıdaki gibi hesaplanabilir:

$$\left\{\begin{array}{l} a_i = y_i \\ b_i = \dfrac{y_{i+1} - y_i}{h_i} - \dfrac{h_i(M_{i+1} + 2M_i)}{6} \\ c_i = M_i / 2 \\ d_i = (M_{i+1} - M_i)/6h_i \end{array}\right\}, i = 0,1,...,n \qquad (3.16)$$

Böylece splayn eğri uydurma problemi M_i, $i = 0,1,...,n$ ikinci türev değerlerini bulmaya indirgenir. Kübik splaynın tanımındaki IV-özelliği uygulanırsa, önce (3.14)'deki i indisi $(i-1)$'le değiştirilerek,

$$f'_{i-1}(x) = b_{i-1} + 2c_{i-1}(x - x_{i-1}) + 3d_{i-1}(x - x_{i-1})^2$$

olduğu bulunur. IV'e göre $f'_{i-1}(x_i) = f'_i(x_i)$, $i = 1,2,...,n$ ve (3.16) formülleri bu eşitliklerde göz önüne alınırsa, bilinmeyen $\mathrm{M_i}$, $i = 0,1,...,n$ katsayılarını içeren aşağıdaki denklemler sistemi elde edilir:

$$h_{i-1}\mathrm{M}_{i-1} + (2h_{i-1} + 2h_i)\mathrm{M}_i + h_i\mathrm{M}_{i+1} = 6\left(\frac{y_{i+1} - y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right), \quad i = 1,..,n. \qquad (3.17)$$

(3.17) lineer denklemler sisteminin (3.10)'un aynısı olduğu görülmektedir.

3.2. Doğal Kübik Splayn'nın Temel Özellikleri

Kübik splayn'nın tanımına ek olarak, $[a, b]$ gözlem aralığının uç noktalarında f fonksiyonunun ikinci mertebeden türevlerinin sıfır olması koşulu eklenirse kübik splayn, *doğal kübik splayn* olarak adlandırılır. Bu durumda $[a, x_1]$ ve $[x_n, b]$ aralıklarında $f(x)$ splayn fonksiyonu doğrusal olur. Bu nedenle x_1 ve x_n düğüm noktalarında ikinci mertebeden türevler sıfırdır. Böylelikle, doğal kübik splayn için (3.17) ifadesinde $\mathrm{M}_0 = \mathrm{M}_1 = \mathrm{M}_n = \mathrm{M}_{n+1} = 0$ olur. Böylece, belirlenecek M 'lerin sayısı (n-2)'ye indirilir. Bu durumda (3.17) denklemlerinde

(x_1, x_n) aralığındaki düğüm noktalarının ele alınması yeterli olur ve uygun i indisi $i = 2,...,n-1$ arasında değiştirilebilir. Böylece $i = 2,...,n-1$ değerleri için (3.17) denklemler sistemi ($\mathrm{M}_1 = \mathrm{M}_n = 0$ koşuluna göre) sonuç olarak şöyle yazılır:

$$\frac{1}{6}h_{i-1}\mathrm{M}_{i-1} + \frac{1}{3}(h_{i-1}+h_i)\mathrm{M}_i + \frac{1}{6}h_i\mathrm{M}_{i+1} = \left(\frac{y_{i+1}-y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}\right), \quad i = 2,3,..,n-1. \qquad (3.18)$$

(3.18) denklemleri açık şekilde aşağıdaki gibi yazılabilir:

$$\left.\begin{array}{ll}
i=2, & 0 + \frac{1}{3}(h_1+h_2)\mathrm{M}_2 + \frac{1}{6}h_2\mathrm{M}_3 = \frac{1}{h_1}y_1 - (\frac{1}{h_1}+\frac{1}{h_2})y_2 + \frac{1}{h_2}y_3 \\
i=3, & \frac{1}{6}h_2M_2 + \frac{1}{3}(h_2+h_3)\mathrm{M}_3 + \frac{1}{6}h_3\mathrm{M}_4 = \frac{1}{h_2}y_2 - (\frac{1}{h_2}+\frac{1}{h_3})y_3 + \frac{1}{h_3}y_4 \\
. & \\
. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . & \\
. & \\
i=n-2, & \frac{1}{6}h_{n-3}\mathrm{M}_{n-3} + \frac{1}{3}(h_{n-3}+h_{n-2})\mathrm{M}_{n-2} + \frac{1}{6}h_{n-2}\mathrm{M}_{n-1} = \frac{1}{h_{n-3}}y_{n-3} - (\frac{1}{h_{n-3}}+\frac{1}{h_{n-2}})y_{n-2} + \frac{1}{h_{n-2}}y_{n-1} \\
i=n-1, & \frac{1}{6}h_{n-2}\mathrm{M}_{n-2} + \frac{1}{3}(h_{n-2}+h_{n-1})\mathrm{M}_{n-1} + 0 = \frac{1}{h_{n-2}}y_{n-2} - (\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}})y_{n-1} + \frac{1}{h_{n-1}}y_n
\end{array}\right\} \qquad (3.19)$$

(3.19) (veya (3.18)) sisteminin matris formundaki şekli de kısaca aşağıdaki gibi olur:

$$R\mathrm{M} = Q^T\mathbf{y} \qquad (3.20)$$

Burada R, $(n-2)\times(n-2)$ simetrik matris, M, $(n-2)$ sütun vektör, Q, $n\times(n-2)$ matris ve y, n boyutlu sütun vektördür. $\mathrm{M}^T = (\mathrm{M}_2, \mathrm{M}_3, ..., \mathrm{M}_{n-1})$, $\mathbf{y}^{\mathbf{T}} = (y_1, y_2, ..., y_n)$ olup $R(r_{ij})$ ve $Q(q_{ij})$ bant matrislerinin elemanları ise aşağıdaki gibi hesaplanır:

$$r_{\mathrm{ij}} = \begin{cases} \frac{1}{6}\mathrm{h}_{\mathrm{j}-1}, & \mathrm{i} = \mathrm{j}-1, \\ \frac{1}{3}\left(\mathrm{h}_{\mathrm{j}-1} + \mathrm{h}_{\mathrm{j}}\right), & \mathrm{i} = \mathrm{j}, \\ 0, & \left|\mathrm{i}-\mathrm{j}\right| \geq 2 \\ \frac{1}{6}\ \mathrm{h}_{\mathrm{j}}, & \mathrm{i} = \mathrm{j}+1 \end{cases} \qquad i = 2,3,...,n-1 \text{ ve } j = 2,3,...,n-1 \qquad (3.21)$$

ve

$$q_{ij} = \begin{cases} \dfrac{1}{h_{j-1}}, & i = j-1, \\ -\left(\dfrac{1}{h_{j-1}} + \dfrac{1}{h_j}\right), & i = j, \\ 0, & |i-j| \geq 2 \\ \dfrac{1}{h_j}, & i = j+1 \end{cases} \qquad i = 1,2,...,n \text{ ve } j = 2,3,...,n-1 \qquad (3.22)$$

Burada $h_i = x_{i+1} - x_i$, $i = 1,2,...,n-1$ dir. (3.20) denkleminde belirtilen RM ve $Q^T\mathbf{y}$ matrislerinin açık bir şekilde gösterilişi aşağıdaki şeklide yazılır:

$$R\mathrm{M} = \begin{bmatrix} \frac{1}{3}(h_1+h_2) & \frac{1}{6}h_2 & 0 & . & . & . & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \frac{1}{6}h_3 & 0 & . & . & . \\ 0 & \frac{1}{6}h_3 & \frac{1}{3}(h_3+h_4) & \frac{1}{6}h_4 & 0 & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & 0 \\ . & . & . & 0 & \frac{1}{6}h_{n-3} & \frac{1}{3}(h_{n-3}+h_{n-2}) & \frac{1}{6}h_{n-2} \\ 0 & . & . & . & 0 & \frac{1}{6}h_{n-2} & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix} \begin{bmatrix} \mathrm{M}_2 \\ \mathrm{M}_3 \\ . \\ . \\ . \\ \mathrm{M}_{n-2} \\ \mathrm{M}_{n-1} \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{h_1} & -(\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & 0 & . & . & . & . & 0 \\ 0 & \frac{1}{h_2} & -(\frac{1}{h_2}+\frac{1}{h_3}) & \frac{1}{h_3} & 0 & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & 0 & \frac{1}{h_{n-3}} & -(\frac{1}{h_{n-3}}+\frac{1}{h_{n-2}}) & \frac{1}{h_{n-2}} & 0 \\ 0 & . & . & . & 0 & 0 & \frac{1}{h_{n-2}} & -(\frac{1}{h_{n-2}}+\frac{1}{h_{n-1}}) & \frac{1}{h_{n-1}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_{n-1} \\ y_n \end{bmatrix} = Q^{\mathrm{T}}\mathbf{y}$$

R ve Q matrisleri üç köşegen (*tridiagonal*) matrislerdir, başka bir deyişle indisleri $|i-j| \geq 2$ koşulunu sağlayan elemanlar sıfırdır. Buna göre de $R\mathrm{M} = \boldsymbol{b}$ (burada $\boldsymbol{b} = Q^T y$) denklemi R^{-1} bulunmadan daha kolay (doğrusal işlem

kullanarak) çözülebilir. $Q^T y$ vektörü ise (3.18) sisteminin sağ kısmındaki vektördür.

R matrisi her bir i için $|r_{ii}| > \sum_{i \neq j} |r_{ij}|$ anlamında kesin köşegen dominanttır. Doğrusal cebirdeki standart inceleme bu özellik durumunda R matrisinin kesin pozitif tanımlı matris olduğunu göstermektedir [11]. Bu yüzden, R matrisinin tersi vardır ve bu durumda,

$$K = QR^{-1}Q^T \tag{3.23}$$

şeklinde belirtilen yeni bir K matrisi tanımlanabilir.

$f(x)$ splayn fonksiyonunun x_i, $i = 1, 2, ..., n$ düğüm notalarında aldığı değerler vektörünü **f** ile gösterelim. Kübik splayn tanımındaki II özelliğine göre **f =(** $y_1, y_2, ..., y_n$**)** verilen y_i gözlem değerlerinin vektörü olur. Sonraki bölümlerde pürüzlük ceza fonksiyonu içeren parametrik olmayan regresyon modellerinde kullanılan $f(x)$ doğal kübik splayn fonksiyonu için **f** vektörünün λ düzeltme parametresine bağlı olarak farklı bir şekilde belirtildiği gösterilecektir. Buna göre, splayn düzeltme yönteminde y gözlem vektörü yerine **f** vektörünü kullanacaktır. (3.18)'deki denklemin sağ kısmını göz önüne alarak $Q^T\mathbf{f}$ vektörünün koordinatları aşağıdaki gibi hesaplanabilir:

$$(Q^T\mathbf{f})_i = \frac{f_{i+1} - f_i}{h_i} - \frac{f_i - f_{i-1}}{h_{i-1}} \tag{3.24}$$

İzleyen bölümlerde ele alınacak olan splayn düzeltme yöntemiyle tahmin yapmada, aşağıdaki teoremlerden yararlanılacaktır.

Teorem 3.1. **f** ve M vektörlerinin doğal kübik splayn belirtmesi için

$$R\mathrm{M} = Q^T\mathbf{f} \tag{3.25}$$

ifadesi gerek ve yeter koşuldur. (3.25)'in sağlanması durumunda pürüzlülük cezası şöyle yazılabilir [11]:

$$\int_a^b f''^2(x)dx \;=\; \mathrm{M}^T R\mathrm{M} \;=\; \mathbf{f}^T K\mathbf{f} \qquad (3.26)$$

İspat: Teoremin birinci kısmının doğruluğu doğal kübik splaynın yukarıda gösterilen yapısından doğrudan alınır. Buna göre, (3.25)'nin sağlandığını kabul ederek, (3.26) ifadesinin doğru olduğunu ispat edelim. Kısmi integral kuralı uygulanarak, (3.26) eşitliği aşağıdaki gibi yazılabilir:

$$\int_a^b f''^2(x)dx \;=\; \int_a^b f''(x)f''(x)dx \;=\; f''(x)f'(x)\Big|_a^b \;-\; \int_a^b f'''(x)f'(x)dx$$

$f''(a)=f''(b)=0$ ve $f'''(x)$ fonksiyonunun her bir (x_i, x_{i+1}) aralığında sabit ve $[x_1, x_n]$ parçasının dışında sıfır olduğundan dolayı

$$\int_a^b f''^2(x)dx \;=\; -\int_a^b f'''(x)f'(x)dx \;=\sum_{i=1}^{n-1} f'''(x_i^+) \int_{x_i}^{x_{i+1}} f'(x)dx \;.$$

(x_i, x_{i+1}) aralığında $f'''(x) \;=\; \dfrac{\mathrm{M}_{i+1}-\mathrm{M}_i}{h}$ ve $f(x_i)=f_i$ olduğu için

$$\int_a^b f''^2(x)dx \;=\; \sum_{i=1}^{n-1} \frac{\mathrm{M}_{i+1}-\mathrm{M}_i}{h_i}(f_i - f_{i+1})$$

olur. Diğer taraftan, $\mathrm{M}_1 = \mathrm{M}_n = 0$ olduğu için

$$\int_a^b f''^2(x)dx \;=\; \sum_{i=2}^{n-1} \mathrm{M}_i \Big(\frac{f_{i+1}-f_i}{h_i} \;-\; \frac{f_i - f_{i-1}}{h_{i-1}} \Big)$$

Bu son ifadede sırasıyla (3.24), (3.25), (3.23) eşitliklerini göz önüne alırsak sonuç olarak aşağıdakileri yazılabilir:

$$\int_a^b f''^2(x)dx = \mathrm{M}^T Q^T\mathbf{f} \;=\; \mathrm{M}^T R\mathrm{M} = \mathbf{f}^T QR^{-1}Q^T\mathbf{f} \;=\; \mathbf{f}^T K\mathbf{f}$$

Bu ifade ile teoremin ispatı tamamlanmış olur.

Bu noktada, doğal kübik splayn interpolyasyonunun önemli özelliklerinden biri olan *optimumluk özelliği* ele alınacaktır. $C^2[a,b]$ simgesi $[a,b]$ aralığında ikinci mertebeden sürekli türeve sahip fonksiyonlar uzayını

göstermek üzere, doğal kübik splayn interpolantı; verileri interpole eden $C^2[a,b]$ sınıfında bulunan tüm pürüzsüz fonksiyonlar arasında, $\int f''(x)^2$ pürüzlülük cezasının minimum yapan bir özelliğe sahiptir.

Teorem 3.2. $n \geq 2$ ve $f(x)$ fonksiyonunun $[a,b]$ aralığındaki (f_i, x_i), $i = 1,2,...,n$, $a < x_1 < ... < x_n < b$ koşulunu sağlayan gözlem değerlerine uygun doğal kübik splayn interpolantı olduğunu varsayalım. Bu durumda $g(x_i) = f_i$, $i = 1,2,...,n$ koşulunu sağlayan her hangi bir $g(x) \in C^2[a,b]$ fonksiyonu için

$$\int_a^b f''^2(x)\,dx \leq \int_a^b g''^2(x)\,dx\ . \tag{3.27}$$

(3.27) ifadesindeki eşitlik durumu ancak ve ancak $f(x) \equiv g(x)$ olduğunda sağlanır.

İspat: $[a,b]$ aralığında $h(x) = g(x) - f(x)$ fonksiyonunu tanımlayalım. $h(x_i) = 0$, $i = 1,2,...,n$ olur. $f''(x)$ doğal kübik splaynının a ve b uç noktalarında türevinin sıfır olduğu göz önüne alınarak, kısmı integral kuralı uygulanırsa, aşağıdaki eşitlik yazılabilir:

$$\int_a^b f''(x)h''(x)\,dx = f''(x)h'(x)\Big|_a^b = -\int_a^b f'''(x)h'(x)\,dx = -\int_a^b f'''(x)h'(x)\,dx$$

$[x_i, x_{i+1}]$ aralığında $f'''(x) = f'''(x_i^+)$ sabit olduğundan,

$$\begin{aligned}\int_a^b f''(x)\,h''(x)\,dx = -\int_a^b f'''(x)h'(x)\,dx &= \sum_{i=1}^{n-1} f'''(x_i^+)\int_{x_i}^{x_{i+1}} h'(x)\,dx \\ &= \sum_{i=1}^{n-1} f'''(x_i^+)\{h(x_{i+1}) - h(x_i)\} = 0.\end{aligned} \tag{3.28}$$

sonucuna ulaşılır. Şimdi (3.28)'i kullanarak aşağıdaki integralı değerlendirelim:

$$\begin{aligned}\int_a^b g''^2(x)dx &= \int_a^b (f''(x) + h''(x))^2\,dx = \int_a^b f''^2(x)dx + 2\int_a^b f''(x)h''(x)dx + \int_a^b h''^2(x)dx \\ &= \int_a^b f''^2(x)dx + \int_a^b h''^2(x)dx \geq \int_a^b f''^2(x)dx\end{aligned} \tag{3.29}$$

Bu durumda (3.27) eşitsizliğini elde etmiş oluruz. Buna göre (3.29)'da eşitlilik durumu ancak $\int_a^b h''^2(x)dx = 0$ için sağlanır. Diğer bir ifadeyle, söz konusu eşitlik $h(x)$ fonksiyonunun $[a,b]$ aralığında lineer olması durumunda sağlanabilir. Bu durumda, $h(x)$ fonksiyonu x_i, $i = 1,2,...,n$ noktalarında sıfır değerlerini alan bir lineer fonksiyon ve $n \geq 2$ olduğunda $h(x) \equiv 0$ olur. Diğer bir deyişle, $f(x) \equiv g(x)$ sonucu ortaya çıkar. Böylece, teoremin ispatı tamamlanmış olur.

3.3. Parametrik Olmayan Regresyonda Splayn Düzeltme Yöntemi

Cezalı en küçük kareler regresyonu ve splayn düzeltme yöntemi son yıllarda popülerlik kazanan esnek veri uydurma metodolojileri için sıkça kullanılan tekniklerdir. Temel splayn düzeltme kavramının başlangıcı Whittaker (1923)'e dayanırken, splayn düzeltme ve onun türlerinin modern gelişimine daha çok Grace Wahba'nın katkısı olmuştur. Bu bağlamda, splayn modellerle ilgili çok daha ayrıntılı bilgi Wahba [27]'de bulunabilir.

Bir y bağımlı değişkeni ve bu bağımlı değişkenle ne tür bir ilişki içerisinde olduğu bilinmeyen bir x açıklayıcı değişkenin yer aldığı *parametrik olmayan regresyon modeli* aşağıdaki şekilde tanımlanır:

$$y_i = f(x_i) + \varepsilon_i, \quad a < x_1 < ... < x_n < b, \quad \varepsilon_i \sim N(0, \sigma^2) \tag{3.30}$$

Burada,

$f \in C^2[a,b]$: Bilinmeyen bir pürüzsüz fonksiyon

$(y_i)_{i=1}^n$: Bağımlı değişkenine ait gözlem değerleri

$(x_i)_{i=1}^n$: Parametrik olmayan açıklayıcı değişkene ait gözlem değerleri

$(\varepsilon_i)_{i=1}^n$: Bağımsız ve özdeş olarak dağılan, σ^2 ortak varyanslı ve sıfır ortalamalı rassal hata terimleridir.

Parametrik olmayan regresyonda amaç bilinmeyen gerçek $f(x)$ fonksiyonunu tahmin etmektir [28]. Geleneksel olarak, dikkate alınacak esas problem, gözlenen verilerden (3.30) modeline uygun f'in tahminidir. f fonksiyonunu tahmin etmek için kullanılan en popüler yöntemlerden biri doğrusal

regresyonudur. Doğrusal regresyonda $f(x)$ fonksiyonu $\hat{f}(x) = a + bx$ şeklinde tahmin edilir. Sırasıyla, a *ve* b sabit ve eğim kestiricileri, $f(x) = \alpha + \beta x$ şeklindeki tüm fonksiyonlar için

$$\mathrm{RSS}(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 \qquad (3.31)$$

ile verilen *hata kareler toplamını* minimum yaparak elde edilir. (3.30) modeline uygun gözlem verileri için f fonksiyonu yaklaşık olarak doğrusalsa tahmin için doğrusal regresyon yaklaşımı etkin olabilir. Eğer (3.30) modeline uygun f doğrusal değilse, birinci bölümde belirtildiği gibi sabit eğim koşulu bozulmaktadır. Buna göre, doğrusal regresyon böyle durumlarda uygun sonuçlar vermez. Bu nedenle, veri uydurmada başarılı olabilmek için, doğrusal regresyon modelinde öne sürülen tahmin koşulları değiştirmeli ve tasarımlarda değişen eğimli f fonksiyonları üzerinde $RSS(f)$ ifadesinin, diğer bir deyişle hata kareler toplamının minimizasyonunu dikkate alınmalıdır.

Parametrik (doğrusal ve doğrusal olmayan) regresyonda bir f fonksiyonu tümüyle az ve sonlu sayıda parametre ile belirtilir (yukarıdaki doğrusal regresyon örneğinde α ve β parametreleri gibi). Böylece, en uygun bir parametre vektörünün bulunması için (3.31)'in minimizasyonu sonlu ve düşük boyutlu bir uzayda araştırılır. Dolayısıyla, parametrik regresyonda tahmin problemleri sınırlı ya da sonlu boyutlu bir uzayda araştırılır. Bunun yerine tüm olası fonksiyonlara göre (3.31) ile verilen hata kareler toplamını minimum yapmaya çalışıldığını varsayalım. Bu durumda sonsuz boyutlu uzayda minimizasyon problemi ele alınmış olur. Bu problem, önceki parametrik regresyon problemi ile kıyaslandığında daha zor ve karmaşıktır. Böyle bir problem, dikkate alınacak fonksiyonlar sınıfını belirtilerek biraz da olsa kolaylaştırılabilir. Örneğin, birinci ve ikinci türevleri [a,b] aralığı üzerinde sürekli olan bütün f fonksiyonlarının oluşturduğu $C^2[a,b]$ kümesinde (uzayında) hata fonksiyonunun minimizasyonu dikkate alınabilir. Bununla birlikte, doğal kübik splaynların adı geçen fonksiyonlar sınıfında minimumu gerçekleştirmesi, parametrik olarak ifade edilemeyen problemi bir anlamda parametrik hale getirilmiş olur. Bu durumda

sonsuz boyutlu problemin sonlu boyutlu bir probleme indirgenmesiyle çözümün bulunması yeterince kolaylaşır.

Bu durumda, hata kareler toplamını minimum yapacak fonksiyonlar kümesine katı parametrik kısıtlamalar yüklemeksizin, eğrinin pürüzlülüğü için bir ceza uygulanabilir. İki kez türevi alınan bir f eğrisinin pürüzlülüğünü ölçmenin sezgisel olarak çekici bir yolu, $\int \{ f''(x) \}^2 dx$ kareli ikinci türev integralini hesaplamaktır: Bu ölçüme göre sadece doğrusal $f(x)$ fonksiyonları sıfır pürüzlülüğe sahipken, diğer $C^2[a,b]$ sınıfındaki tüm fonksiyonlar, pozitif bir pürüzlülüğe sahiptirler.

$C^2[a,b]$'de verilen herhangi bir f fonksiyonu ve λ pozitif bir skaler ($\lambda > 0$) olsun. ***Splayn düzeltme*** yönteminin esası, $f \in C^2[a,b]$ uzayındaki tüm f fonksiyonları arasında,

$$S(f) = \sum_{i=1}^{n} \{ y_i - f(x_i) \}^2 + \lambda \int_a^b \{ f''(x) \}^2 dx \qquad (3.32)$$

eşitliği ile belirtilen $S(f)$ *"cezalı en küçük kareler kriterini"* minimum yapmaktır. Diğer bir ifadeyle, *splayn düzeltme kestiricisi* olan $\hat{f}$ *tahmin eğrisi* $C^2[a,b]$ uzayının fonksiyonları arasında $S(f)$ *cezalı kriterini minimum yapan* eğri olarak tanımlanır.

Eşitlik (3.32) ifadesindeki ilk terim, *hata kareler toplamını* (RSS) gösterir ve bu ifade uyumdan yoksunluğu cezalandırır. Diğer bir deyişle, uyumun verilere yakınlığını ölçer. İkinci terim *pürüzlülük* (PS) *cezasını* gösterir ve bu pürüzlülüğe bir ceza yükler. Diğer bir deyişle, fonksiyondaki eğriliği cezalandırır. Cezalı kriterde yer alan λ ise, birinci bölümde de açıklandığı gibi *düzeltme parametresini* belirtir ve bu parametre $\int_a^b \{ f''(x) \}^2 dx$ ile ölçümlenen eğrinin pürüzlülüğü ve $\sum_{i=1}^{n} \{ y_i - f(x_i) \}^2$ ile ölçümlenen verilere uyumunu dengeler. Ayrıca, λ parametresi 0 dan $+\infty$'a değişirken, çözüm interpolasyondan basit bir doğrusal modele değişir. Eğer $\lambda = \infty$ alınırsa, o zaman (3.32) denklemi sabit

eğimli doğrusal regresyon uyumu üretir, buna karşılık $\lambda = 0$ alınırsa tümüyle esnek eğimli bir interpolasyon uyumuna karşı gelir [29].

Problem (3.32) için çözüm *splayn düzeltme kestiricisi* olarak adlandırılır ve $x_1,...,x_n$ düğümleri ile bir "*doğal kübik splayn"* olarak bilinir. İzleyen bölümde $\hat{f}$ 'in nasıl elde edildiği gösterilecektir.

3.3.1. Splayn Düzeltme Kestiricisinin Tahmini

Bölüm 3.1.1'de anlatılan kübik splayn interpolasyonu $\hat{f}$ tahmininin önemli özelliklerinin elde edilmesine olanak sağlar. Bunlardan en önemlisi, $\hat{f}$ tahmininin x_i noktalarındaki düğümlerle bağlı bir doğal kübik splayn olmasıdır. $\hat{f}$ 'nin doğal kübik splayn olarak bilinmesi önemli bir ilerlemedir. Çünkü bu durumda sonsuz boyutlu $C^2[a,b]$ pürüzsüz fonksiyonlar kümesinde minimizasyon problemi, (x_i, y_i) gözlem noktaları ve λ düzeltme parametresiyle bağlı olarak tanımlanan doğal kübik splaynlar dikkate alınarak, $S(f)$ minimizasyon problemiyle değiştirilmiş olur. Bu durumda $\hat{f}$ tahmin fonksiyonu splayn şeklinde belirtilebilir.

Bölüm 3.2' de belirtildiği gibi, f fonksiyonunun $\mathbf{f}$, $\mathbf{M}$ vektörleri ve Q, R matrisleri ile tanımlanan bir doğal kübik splayn olduğu varsayılsın. S(*f*) cezalı kareler toplamı, matris ve vektör terimleriyle ifade edilerek onu minimum yapan $\hat{f}$ aşağıdaki şekilde tahmin edilebilir.

$\mathbf{y} = (y_1,...,y_n)^{\mathrm{T}}$ verilen gözlem değerleri vektörü olsun. x_i düğüm noktalarında $f(x_i)$ değerlerinin vektörü, $\mathbf{f} = (f_1,..., f_n)^T = (f(x_1),..., f(x_n))^T$ olduğundan f 'e göre (3.31)' deki hata kareler toplamı,

$$\sum \{ y_i - f(x_i) \}^2 = (\mathbf{y} - \mathbf{f})^T (\mathbf{y} - \mathbf{f})$$

olarak yazılabilir. (3.26) ifadesine göre $\int f''^2$ pürüzlülük ceza terimi, $\mathbf{f}^{\mathrm{T}} K \mathbf{f}$ değerine eşittir. Buna göre, $\int_a^b f''(x)^2 dx = \mathbf{f}^{\mathrm{T}} K \mathbf{f}$ olduğunu göz önünde bulundurarak, cezalı en küçük kareler toplamı aşağıdaki şekilde yazılabilir:

$$\begin{aligned} S(f) &= \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b f''(x)^2 dx \\ &= (\mathbf{y}-\mathbf{f})^{\mathrm{T}} (\mathbf{y}-\mathbf{f}) + \lambda \mathbf{f}^{\mathrm{T}} \mathrm{K} \mathbf{f} \quad . \end{aligned}$$

Sonuç olarak cezalı kareler toplamı,

$$\mathrm{S}(f) = \mathbf{f}^{\mathrm{T}} (\mathrm{I} + \lambda \mathrm{K}) \mathbf{f} - 2\mathbf{y}^{\mathrm{T}} \mathbf{f} + \mathbf{y}^{\mathrm{T}} \mathbf{y} \qquad (3.33)$$

şeklinde ifade edilir. f splayn düzeltme kestiricisi (3.32) veya (3.33)'de belirtilen cezalı en küçük kareler toplamını minimum yapan eğri olarak belirtilmektedir. (3.23) formülü ile tanımlanan K matrisi yarı-pozitif tanımlı ve $\lambda > 0$ olduğundan, λK matrisi de yarı-pozitif tanımlı bir matris olacaktır. Dolayısıyla $(I + \lambda K)$ kesin pozitif tanımlı matristir. Bu yüzden (3.33) kare formu tek bir minimumuma sahiptir. Buna göre, (3.33)'e minimum değer veren $\mathbf{f}$ vektörü, (3.33)'ün $\mathbf{f}$ 'e göre türev fonksiyonuna sıfır değerini veren vektördür:

$$\begin{aligned} \mathrm{S}'(f) &= 2\mathbf{f}(\mathrm{I} + \lambda K) - 2\mathbf{y} = 0 \\ 2\mathbf{f}(I + \lambda K) &= 2\mathbf{y}. \end{aligned}$$

Buna göre $\mathbf{f}$ vektörü,

$$\mathbf{f} = (I + \lambda K)^{-1} \mathbf{y} \qquad (3.34)$$

olarak elde edilir. Böylece, düğüm noktalarındaki değer vektörü (2.34) formülü ile belirlenen splayn düzeltme kestiricisi, (3.32) $S(f)$ cezalı kriterini minimum yapar. (3.34) ifadesinde yer alan $(I + \lambda K)^{-1}$ matrisi, *düzeltme matrisi* olarak adlanır (benzer matrise doğrusal regresyonda şapka matrisi denir) ve bu matris, (3.23) ve (3.34) formüllerine esasen, sadece verilen bir λ >0 düzeltme parametresi ve $\boldsymbol{x} = (x_1, ..., x_n)$ düğüm noktaları vektörü ile belirlenir. Böylece $\mathbf{y}$ değerlerini $\mathbf{f}$ vektörüne görüntüleyen $n \times n$ boyutlu *düzeltme matrisi*,

$$S_\lambda = (I + \lambda K)^{-1} \qquad (3.35)$$

eşiliği ile tanımlanır. $\mathbf{f} = (f(x_1),..., f(x_n))$ splayn düzeltme kestiricisi, (3.35)'de verilen düzeltme matrisinin yardımıyla $\mathbf{y} = (y_1,..., y_n)$ vektörünün bir doğrusal dönüşümü olarak da tanımlanabilir:

$$\mathbf{f} = \begin{pmatrix} f_\lambda(x_1) \\ f_\lambda(x_2) \\ . \\ . \\ . \\ f_\lambda(x_n) \end{pmatrix}_{(n\times 1)} = \left(S_\lambda\right)_{(n\times n)} \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}_{(n\times 1)} \quad \textit{yada kısaca, } \mathbf{f}_\lambda = S_\lambda \mathbf{y}. \qquad (3.36)$$

Burada f_λ, $\lambda > 0$ sabit düzeltme parametresi için $x_1,..., x_n$ düğümlü doğal kübik splayn ve S_λ, (3.35)'te verilen λ' ya bağlı bilinen pozitif tanımlı bir düzeltme matrisidir. Splayn düzeltme fonksiyonu x'e ait gözlem değerlerine uygun olan bir *doğal kübik splayndır*. Bu parametreleri belirginleştirilmemiş bir model olarak da görülebilir. Ancak ceza terimi, splayn katsayılarının doğrusallığa doğru çekilmesine olanak sağlar, böylece sınırlayıcı serbestlik derecesi yaklaşımı kullanılır.

Splayn düzeltme kestiricisi,

$$\mathbf{f}_\lambda(x) = \sum_{i=1}^{n} S_\lambda(x) y_i$$

şeklinde her bir x_i, $i = 1,...,n$ için hesaplanabilen S_λ sabitlerinin var olması anlamında doğrusaldır.

Teorem 3.1'e göre her bir $\mathbf{f} = (f(x_1),..., f(x_n))$ vektörü, tek bir $f(x)$ kübik splayn interpolasyon fonksiyonunu belirtir. Söz konusu $\mathbf{f}$ vektörü, (3.33) $S(f)$ cezalı kriterini minimum yapan değerler vektörü olması nedeniyle, cezalı kareler toplamının minimum problemi için temel bir özellik oluşturur. Bu temel özellik aşağıda verilmiştir [11]:

Teorem 3.3 $n \geq 3$ ve $x_i, i = 1,...,n,$ düğüm noktaları için $a < x_1 < ... < x_n < b$ koşulunun sağlandığı varsayılsın. Verilen $\mathbf{y} = (y_1,..., y_n)^T$ gözlem değerleri ve

$\lambda > 0$ düzeltme parametresi için $f(x)$, x_i, $i = 1,...,n,$ düğüm noktaları ile $\mathbf{f} = (I + \lambda K)^{-1}\mathbf{y}$ vektörüne uygun doğal kübik splayn olsun. Bu durumda, her hangi bir $g \in C^2[a,b]$ fonksiyonu için

$$S(f) \leq S(g)$$

olup eşitlilik durumu yalnız ve yalnız $g = f$ olması durumunda sağlanır.

Doğal kübik splayn düzeltmeyi belirleyen $\mathbf{f}$ vektörünün (3.34) formülü ile direkt olarak hesaplanması pratik açıdan verimli değildir. Bu amaçla farklı sayıda etkili algoritmalar işlenmiştir. Onlardan biri olan Reinsch algoritması bir sonraki altbölümde açıklanmıştır.

3.3.2. Splayn Düzeltme Kestiricisinin Reinsch Algoritması ile Bulunması

Bu alt bölümde düğüm noktalarında splayn fonksiyonunun değerler vektörünü belirlemek için Reinsch [30] tarafından ortaya konulan bir algoritma ele alınmıştır. Reinsch algoritmasının esas fikri, f'nin ikinci türevinin $x_i, i = 1,...,n$ düğüm noktalarındaki M_i değerleri için tekil olmayan bir doğrusal denklemler sistemi kurmaktır. Bu denklemler bantlı bir yapıya sahiptir ve $O(n)$ cebirsel işlemde çözülebilir. Algoritma M_i ve y_i veri değerlerine dayanarak belirgin formülle f_i değerlerini verir. Tartışmada nümerik doğrusal cebirden, bir bant matrisinin *Cholesky ayrıştırması* konusunda değişik görüşler kullanılacaktır [23-26].

Bir matrisin sıfırdan farklı tüm girişleri az sayıda köşegen üzerinde yer alırsa "*bant matrisi"* olarak bilinir ve sıfırdan farklı köşegen sayısı matrisin "*bant genişliği (bandwith)*" olarak adlandırılır. Böylece B, $2k+1$ bant genişliği olan simetrik bir bant matris ise, $|i-j| > k$ için b_{ij} elemanı sıfırdır. Bant matrisleri, sıfırdan farklı köşegenleri dikkate alması nedeniyle işlemlerde kolaylık sağlarlar. Köşegen ve üç köşegen matrisler sırasıyla 1 ve 3 bant genişlikli bant matrislerdir. Bölüm 3.2'de tanımlanan Q ve R matrislerinin her ikisi de 3 bant genişliğine sahiptir.

Bölüm 3.2'deki gibi (n-2)- M vektörünü tanımlayalım. (3.23) ve (3.34)'den

$$(I + \lambda QR^{-1}Q^T)\mathbf{f} = \mathbf{y} \tag{3.37}$$

olarak yazılabilir. (3.37)' den

$$\mathbf{f} = \mathbf{y} - \lambda QR^{-1}Q^T\mathbf{f}$$

elde edilir. Şimdi (3.25) ifadesine göre $Q^T\mathbf{f}$ yerine RM yazarak, $\mathbf{f}$ için M ve $\mathbf{y}$'ye dayalı aşağıdaki formül elde edilir:

$$\mathbf{f} = \mathbf{y} - \lambda Q\mathrm{M}\,. \tag{3.38}$$

(3.38) ifadesinin her iki kısmını soldan Q matrisine çarpıp, yine $Q^T\mathbf{f}$ yerine RM yazarak

$$Q^T\mathbf{y} - \lambda Q^TQ\mathrm{M} = R\mathrm{M}$$

ifadesi elde edilir. Buradan hareketle, M için aşağıdaki denklem bulunur:

$$\left(R + \lambda Q^TQ\right)\mathrm{M} = Q^T\mathbf{y} \tag{3.39}$$

Bu denklem Reinsch algoritmasının çekirdeğidir. Bu denklem $\mathbf{f}$ vektörünün bulunmasında (3.37) denklemine kıyasla, bant teknikleri kullanılarak doğrusal bir zamanda çözülebilir. $\left(R + \lambda Q^TQ\right)$ matrisinin 5 bant genişliğine sahip olduğu görülebilir, ayrıca R kesin pozitif tanımlı ve $\lambda > 0$ olduğundan bu matris simetrik ve kesin pozitif tanımlıdır. Bu yüzden bu matris

$$R + \lambda Q^TQ = LDL^T$$

şeklinde bir *Cholesky ayrışımına* sahiptir. Burada D kesin pozitif köşegen matris ve L, elemanları $j < i-2$ *ve* $j > i$ için $l_{ij} = 0$ ve köşegenleri $l_{ii} = 1$ şeklinde olan bir alt üçgen bant matristir. Bu durumda Q ve R matrisleri, sıfırdan farklı köşegenleri saklaması koşuluyla, $O(n)$ cebirsel işlemde elde edilebilir. Böylece L ve D matrislerinin hesaplanma süresi onların boyutları ile orantılıdır. Diğer bir deyişle, L ve D matrisleri, hesaplanmaları için sadece doğrusal bir zamana gereksinim duyar [11]. (3.34)' de belirtilen $\mathbf{f}$ vektörü hesaplandıktan sonra splayn

düzeltme fonksiyonu bölüm 3.2'de gösterildiği gibi bulunabilir. Buna göre splayn düzeltme algoritması aşağıdaki gibi verilebilir:

Splayn Düzeltme için Algoritma

Adım 1: (3.24) formülünü kullanarak, $Q^T\mathbf{y}$ vektörü elde edilir.

Adım 2: $\left(R+\lambda Q^T Q\right)$ matrisinin sıfırdan farklı köşegenleri ve böylece L ve D Cholesky ayrışımı çarpanları bulunur.

Adım 3: $LDL^T M = Q^T\mathbf{y}$ gibi (3.39) denklemi M için ileri ve geriye doğru yerine koyma ile çözülür.

Adım 4: (3.38)'de verilen,

$$\mathbf{f} = \mathbf{y} - \lambda Q\mathrm{M}$$

ifadesinin yardımıyla $\mathbf{f}$ vektörü bulunur.

Adım 1'in her bir veri seti için sadece bir kez yapılması gerekir; düzeltme parametresi λ'nın yeni bir değeri kullanılırsa adım 1'in tekrar edilmesi gerekmez. Ayrıca, tasarım noktaları değişmeden kalırsa, yeni $\mathbf{y}$ veri değerleri kullanıldığında **adım 2** ihmal edilebilir.

4. SEMİPARAMETRİK REGRESYONDA SPLAYN DÜZELTME YÖNTEMİNE DAYALI ÇIKARSAMALAR VE BİR UYGULAMA

Şu ana kadar tek bir kestirici (açıklayıcı) x değişkenine göre y gözlemlerinin bağımlılığı ile ilgilenilmiştir. Bu durum uygulamada karşılaşılan çok sayıda problem için yeterliyken, bağımlı değişkenlerin birçok bağımsız değişken tarafından eş zamanlı olarak etkilendiği durumlar da söz konusudur. Birden çok açıklayıcı değişkenlere göre bağımlılığın istatistiksel analizi genellikle *çoklu regresyonu* kullanmaya yöneltir. Açıklayıcı değişkenler ya kantitatif (sayısal) yada kalitatif (kategorik veya sayısal olarak ifade edilmeyen) olabilirler ve bağımlı değişkenle olan ilişki yapısı aşağıdaki modelle ifade edilirler:

$$y_{\mathrm{i}} = \mathbf{z}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta} + hata\,. \qquad (4.1)$$

Burada $\mathbf{z}_i$ i.gözlem için açıklayıcı değişkenler vektörü, $\boldsymbol{\beta}$ tahmin edilecek, regresyon katsayılarına karışı gelen vektördür. Genelde, $\mathbf{z}_i$ vektörü sabit terim için 1 girişli bir sabit içerir.

İkinci bölümde, (2.1) doğrusal regresyon modelinin (2.2) modelindeki gibi eğri uydurmak için genelleştirilmesine benzer şekilde, (4.1) modelinin dayandığı varsayımlardan doğrusallık varsayımının değiştirilmesi gerekir. Doğal bir benzerlikle,

$$y_i = f(\boldsymbol{x}_i) + hata \qquad (4.2)$$

modeli içinde bir genelleştirme dikkate alınabilir. Burada f bir vektör değişkeninin gerçek-değerli, keyfi fakat pürüzsüz (smooth) bir fonksiyonudur. Bununla birlikte, üçüncü bölümde gösterilen ve bir açıklayıcı değişkenin parametrik olmayan bir yapıda işlem görmesine olanak sağlayan kübik splaynlar ve cezalı en küçük kareler mekanizması, (4.1) modelini genelleştirmek için yeterlidir.

Yatchew [9] tarafından yapılan çalışma, bir y bağımlı değişkeni, x ve z açıklayıcı değişkenleri arasındaki regresyon ilişkisi için önemli bir referanstır. Buna göre, f'in parametrik olmayan bir fonksiyon olduğu, $y = f(x,z) + \varepsilon$ parametrik olmayan regresyon modeli ile çalışmak istenilebilir. Ancak, çok sayıda

açıklayıcı değişken varsa, o zaman parametrik olmayan metotlar *"boyutluluğun yarattığı sıkıntı"* nedeniyle zorlaşır. Bu nedenle, son zamanlara x değişkeninin düşük boyutlu olduğu, sözde kısmi doğrusal model, $y = \mathrm{z}^{\mathrm{T}}\beta + f(x) + \varepsilon$ şeklindeki, semiparametrik regresyon modeline ilgi artmaktadır. Yatchew [31], semiparametrik model için kuramsal açıdan birçok anahtar sonuç elde eden çalışmasında çok sayıda örnek sunmaktadır. Bu bölümde, söz konusu semiparametrik regresyon modelin katsayılarının elde edilmesinde kullanılan tahmin ve yöntemler incelenmiştir.

4.1. Semiparametrik Regresyon Modeli

Semiparametrik regresyon modeli kompleks (karmaşık) regresyon problemlerin çözümünde oldukça önemli bir yere sahiptir. Gerçekte dünya, insan aklının bütün detayları ile kavrayamayacağı kadar karmaşıktır. "Semiparametrik regresyon modelleri" karmaşık verileri anlaşılabilir özet bir duruma indirger. Uygun bir şekilde uygulandığında bu modeller, verilerin gerekli olanlarını içerirken önemsiz detayları dışlar ve böylece sağlıklı karar verilmesine yardımcı olur.

Semiparametrik regresyon modelleri bağımlı değişkenin bazı açıklayıcı değişkenlerle *doğrusal*, fakat diğer bazı açıklayıcı değişkenlerle *doğrusal olmayan* ilişki içerisindeki regresyon modelleridir. Semiparametrik regresyon modeller standart regresyon tekniklerini genelleştiren toplamsal modellerinin özel bir durumudur. Semiperametrik modeller *doğrusal parametrik bileşen* ve *parametrik olmayan bileşenlerin* her ikisini de içerdiğinden, bu modellere *kısmi doğrusal* modeller de denir. Daha önce de değinildiği gibi bu modeller, *"boyutluluğun verdiği sıkıntı"* nedeniyle tamamıyla parametrik olmayan regresyona tercih edilebilen ve her bir değişkenin etkisinin daha açık bir şekilde yorumlanmasına olanak sağlayan toplamsal regresyon modellerinin özel bir durumu şeklindedir. Diğer taraftan, semiparametrik regresyon modelleri bağımlı değişkenin birkaç değişkenle doğrusal bir şekilde fakat diğer bağımsız değişkenlerle doğrusal olmayan ilişki içerisinde olduğuna inanıldığında, hem parametrik hem de parametrik olmayan bileşenleri birleştirdiklerinden dolayı kullanım açısından standart doğrusal modellerden çok daha esnektir.

Semiparametrik regresyon modeli

$$y_i = \mathbf{z}_i^T \boldsymbol{\beta} + f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2), \quad 1 \le i \le n \tag{4.3}$$

biçiminde ifade edilir. Burada,

y_i: y bağımlı değişkeninin i.gözlem değeri

$\mathbf{z}_i$: Parametrik kısma karşı gelen k boyutlu bağımsız değişkenlerin i.gözlemler vektörü

$\boldsymbol{\beta}$: k boyutlu regresyon katsayıları vektörü

$f \in C^2[a,b]$: Modelin parametrik olmayan kısmına karşılık gelen bilinmeyen bir pürüzsüz fonksiyon

x_i: Parametrik olmayan kestirici (açıklayıcı) değişkenin i.gözlem değeri

ε_i, $i = 1,...n$, bağımsız ve aynı dağılımlı (i.i.d), σ^2 ortak varyanslı ve sıfır ortalamalı rassal hata terimleridir.

(4.3) modeli matris-vektör terimi ile

$$\mathbf{y} = \mathrm{Z}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon} \tag{4.4}$$

şeklinde de ifade edilebilir. Burada değişkenler aşağıdaki şekilde tanımlanır:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad \mathrm{Z} = \begin{bmatrix} z_{11} & . & . & . & z_{1k} \\ z_{21} & & & & z_{2k} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ z_{n1} & . & . & . & z_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ . \\ . \\ . \\ f(x_n) \end{bmatrix} \quad ve \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

$\mathbf{y}$ gözlemlerin bir $(n \times 1)$ boyutlu vektörü, Z parametrik kısma karşı gelen bağımsız değişkenlerin $(n \times k)$ boyutlu gözlem matrisi, $\boldsymbol{\beta}$ regresyon katsayılarının $(k \times 1)$ boyutlu vektörü, $\mathbf{f}$ splayn düzletme kestiricisine karşı gelen $(n \times 1)$ boyutlu vektörü ve $\boldsymbol{\varepsilon}$ normal dağılan rassal hataların $(n \times 1)$ boyutlu vektörüdür. (4.3) modelinde yer alan tüm açıklayıcı değişkenlerin belirlenmiş olduğu varsayılır ve söz konusu modelde, $f(x_i)$ sabit terimi kapsadığından

parametrik kısım sabit terim içermez [32]. Uygulamada genellikle, y bağımlı değişkeninin açıklayıcı değişkenlerin çoğu üzerinde bağımlılık şekli deneyimlere göre bilinir ve uygun model yazılır. Böyle bir ilişkiyi yansıtan (4.3) modelinde, y bağımlı değişkeni açıklayıcı değişkenlerden Z ile parametrik (doğrusal) bir şekilde ilişkili fakat diğer açıklayıcı değişkenlerden x ile doğrusal olmayan ilişki içerisindedir. Buna göre, Z matrisini oluşturan z değişkenleri doğrusal değişken olarak adlandırılırken x değişkeni parametrik olmayan değişken olarak adlandırılır [11].

Bu bölümde, (4.3) semiparametrik regresyon modeli dikkate alınacaktır Amaç *model uyumunu elde etmektir*. Başka bir anlatımla, $\boldsymbol{\beta}$ *parametre vektörünü*, $f \in C^2[a, b]$ *fonksiyonu* ve $\boldsymbol{\mu} = Z\boldsymbol{\beta} + \mathbf{f}$ *ortalama vektörünü etkin olarak tahmin etmek, uyumu sayısal olarak özetlemek ve ayrıca uyumu grafiksel olarak görüntülemektir*. Bunun için farklı düzeltme tekniklerine dayalı birkaç yaklaşım önerilmiştir: Green ve ark. [33], Engle ve ark. [34], Wahba [27] ve Green ve Silverman [11], (4.3) modeline *splayn düzeltme yöntemini* uygulamıştır. Speackman [35] ve Robinson [36] *kernel* (*çekirdek*) *düzeltmesini* ve Cheen [37] ise parçalı polinom yaklaşımını önermiştir. Araştırmanın izleyen bölümlerinde, Green ve ark. [33]'e dayalı ve Green ve Silverman [11]'de ayrıntılı şekilde incelenen pürüzlülük ceza yaklaşımı (*kısmi splayn yöntemi*) ve Speckman [35] tarafından geliştirilen *geleneksel splayn düzeltme yaklaşımı* ele alınmıştır.

4.2. Kısmi Splayn Yöntemi

Dikkate alınan $\{y_i, z_i, x_i\}_{i=1}^{n}$ gözlem ya da ölçüm değerleri, (4.3)'de belirtilen bir semiparametrik regresyon modeline uygulanmak istenirse, aşağıda verilen

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

hata kareler toplamını minimum yapan f fonksiyonunu ve $\boldsymbol{\beta}$ parametreleri en küçük kareler (*EKK*) yöntemine göre tahmin edilebilir. Ancak, f üzerine konulan kısıtlama azlığı bu yaklaşımı başarısız edecektir. Burada $\{x_i\}$ değişkeninin z'den

farklı olmasının önemi gözönüne alınmalıdır: $\boldsymbol{\beta}$ ve f 'nin herhangi bir değeri $f(x_i) = y_i - \mathbf{z}_i^T\boldsymbol{\beta}$ interpolasyonu yoluyla elde edilebilir. Ancak $\boldsymbol{\beta}$ bilinmeyen bir değer olduğundan bu yaklaşımla $\boldsymbol{\beta}$ belirlenemez. Bu problem Engle ve ark. [34] tarafından tanıtılan,

$$S(\beta, f) = \sum_{i=1}^{n} \left\{ y_i - \mathbf{z}_i^T\boldsymbol{\beta} - f(x_i) \right\}^2 + \lambda \int_a^b \{ f''(x) \}^2 \, dx \qquad (4.5)$$

cezalı en küçük kareler (*CEKK*) toplamını minimum yapan f fonksiyonu ve $\boldsymbol{\beta}$ parametreleri seçilerek çözüme kavuşturulabilir. Böyle bir işlem gerçekte, (4.3) modeli için ikinci bölümde incelenen kübik splayn düzeltmenin gerçek avantajını kullanan ve splayn düzeltme yöntemini esas alan bir çözümdür. (4.5) cezalı kriterinde yer alan λ parametresinin seçimi dördüncü bölümde ayrıntılı olarak ele alınan düzeltme parametresi olarak bilinen pozitif bir sabittir.

4.2.1. Düzeltme Matrisinin Hesaplanması

Aşağıdaki tartışmada (4.4) denkleminde ifade edildiği gibi $\mathbf{y}$, i.bileşeni y_i olan n boyutlu vektörü ve Z, satırları $\mathbf{z}_i^{\mathrm{T}}$ olan $(n \times k)$ tipindeki model matrisini göstermektedir. Üçüncü bölümde, $\{x_i\}, i = 1,...,n,$ düğüm noktalarının farklı ve sıralı olduğunu varsayıldı. Bu bölümde ele alınan çoklu regresyonda, $\{y_i, \mathbf{z}_i, x_i\}, i = 1,...,n,$ değişkenlerine ait gözlem değerlerinin tümünün yeniden düzenlenmesi uygun olmayabilir ve $\{x_i\}$ noktaları arasındaki düğümlerin şekli, genel olarak $\{\mathbf{z}_i\}$'lerin arasındaki ilişkiden farklıdır. Bu durumu daha dikkatli bir yaklaşımla ele almak iyi olur.

$x_1,...,x_n$ düğüm noktalarının farklı ve sıralı değerleri $s_1,...,s_q$ ile gösterilsin. $x_1,...,x_n$ ve $s_1,...,s_q$ arasındaki bağlantı, $x_i = s_j$ ise $\mathrm{N}_{ij} = 1$, değilse 0 girişli olan bir $(n \times q)$ tipinde gözlem değerlerinin *tekrarlanma matrisi* (*incidence matrix*) olarak adlandırılan N matrisi yardımıyla gerçekleştirilir. Diğer bir ifadeyle, N tekrarlanma matrisinin n_{ij} elamanları,

$$\mathrm{n}_{ij} = \begin{pmatrix} 1, & x_i = s_j \ \ ise \\ 0, & diğer\ \ durumlarda \end{pmatrix}$$

ile gösterilir. N-matrisinin satırları x_i düğümlerine, sütunları ise s_j düğümlerine uygun olarak belirlenir. N-tekrarlanma matrisinin her satırında ancak bir elaman 1, kalanları ise 0'dır. Sütunlarda ise, birkaç elaman 1 olabilir. i. sütunda, örneğin 3., 5., ve 7. elamanlar 1 ise x_3, x_5 *ve* x_7 değerlerinin s_i ile aynı olması demektir. Eğer N-tekrarlanma matrisinde n_{ij}=1 ise, $N\mathbf{f}$ çarpımında $\mathbf{f}$ vektörünün f_j koordinatı *i.* sırada yazılırsa,

$$\tilde{\mathbf{f}} = N\mathbf{f} \Rightarrow \tilde{f}_i = f_j,\ i = 1,...,n\,.$$

olur. Böylelikle, N matrisinin yardımıyla $f(x_j), j = 1,...,n$ değerleri $s_1 < ... < s_q$ sırasına uygun yazılmış olur. x_i düğüm noktalarının hepsinin aynı (özdeş) olmadığı varsayımından da $q \geq 2$ olduğu sonucu çıkar [11].

$\mathbf{f} = (a_1,...,a_n) = (f(x_1),..., f(x_n))$ vektörü, $a_i = f(x_i)$ değerlerinin vektörü olsun. $\mathbf{f}$ vektörü bütünüyle $f(x_i)$ değerlerinin vektörü olduğundan (4.5) cezalı kareler toplamı, $x_1 < ... < x_n$ koşulu sağlandığında,

$$S(\beta, f) = \left(\mathbf{y} - \mathrm{Z}\boldsymbol{\beta} - \mathbf{f}\right)^{\mathrm{T}} \left(\mathbf{y} - \mathrm{Z}\boldsymbol{\beta} - \mathbf{f}\right) + \lambda \int f''(x)^2\, dx$$

olarak yazılabilir. Kavramsal olarak $S(\beta, f)$ ifadesinin $\boldsymbol{\beta}$ ve $\mathbf{f}$'e göre minimum problemi iki adımda dikkate alınabilir: Biricisi $a_i = f(x_i), i = 1,...,n$ ifadesine bağlı minimum, ikincisi $\mathbf{f}$ ve $\boldsymbol{\beta}$'nın seçimi üzerine minimumdur.

$x_1 < ... < x_n$ koşulunu sağlayan ve $f(x_j) = a_j,\ j = 1,...,n$ noktalarını veren f fonksiyonunun interpolasyonuna bağlı, $\int f''(x)^2 dx$ fonksiyonunun minimum problemi ikinci bölümde tartışılmıştır. Bu durumda, minimum veren f eğrisi x_j düğümlü doğal kübik splayndır. Bölüm 3.2'deki gibi R, Q matrisleri ve $K = QR^{-1}Q^T$ tanımlanır. Eğer $x_1 < ... < x_n$ koşulu sağlanmıyorsa, $x_1,...,x_n$ düğüm noktaları ile yer değiştiren $s_1,...,s_q$ düğüm noktaları yardımıyla söz konusu

matrisler hesaplanır. Üçüncü bölümde, teorem 3.1 ile $\int f''(x)^2 dx$ fonksiyonunun minimum değerinin $\mathbf{f}^T K \mathbf{f}$ olduğu ispatlanmıştır. Buna göre, genelde (4.5) ile belirtilen $S(\beta, f)$ cezalı kareler toplamı

$$S(\beta, f) = (\mathbf{y} - \mathrm{Z}\boldsymbol{\beta} - \mathrm{N}\mathbf{f})^{\mathrm{T}} (\mathbf{y} - \mathrm{Z}\boldsymbol{\beta} - \mathrm{N}\mathbf{f}) + \lambda \mathbf{f}^T K \mathbf{f} \tag{4.6}$$

şeklinde yazılabilir. Pürüzlülük ceza yaklaşımını olarak adlandırılan cezalı en küçük karelerin esası geleneksel doğrusal regresyon modelinin çözümünde kullanılan sıradan en küçük karelere benzer olarak, (4.5) modelini minimum yapan $\mathbf{f}$ fonksiyonunu ve $\boldsymbol{\beta}$ parametrelerinin kestirimidir. Bunun için basit hesaplamalar yapılarak (4.6) denkleminin sırasıyla $\boldsymbol{\beta}$ ve $\mathbf{f}$ 'e göre türevleri alınıp sıfıra eşitlenirse,

$$\begin{aligned} \mathrm{S}'(\boldsymbol{\beta}) &= -\mathrm{Z}^T (y - Z\boldsymbol{\beta} - N\mathbf{f}) + (y - Z\boldsymbol{\beta} - N\mathbf{f})^T (-Z) \\ &= -Z^T y + Z^T Z\boldsymbol{\beta} + Z^T N\mathbf{f} - y^T Z + \boldsymbol{\beta}^T Z^T Z + \mathbf{f}^T N^T Z \\ &= -2Z^T y + 2Z^T Z\boldsymbol{\beta} + 2Z^T N\mathbf{f} = 0 \end{aligned}$$

elde edilir. Buradan da

$$Z^T Z\boldsymbol{\beta} + \mathrm{Z}^{\mathrm{T}} N\, f = Z^T y \tag{4.7}$$

denklemleri bulunur. Daha sonra,

$$\begin{aligned} \mathrm{S}'(\mathbf{f}) &= -\mathrm{N}^T (y - Z\boldsymbol{\beta} - N\mathbf{f}) + (y - Z\boldsymbol{\beta} - N\mathbf{f})^T (-N) \\ &= -N^T y + N^T Z\boldsymbol{\beta} + N^T N\mathbf{f} - y^T N + \boldsymbol{\beta}^T Z^T N + \mathbf{f}^T N^T N + 2\lambda K\mathbf{f} \\ &= -2N^T y + 2N^T Z\boldsymbol{\beta} + 2N^T N\mathbf{f} + 2\lambda K\mathbf{f} = 0 \end{aligned}$$

elde edilir. Buradan,

$$N^T Z\boldsymbol{\beta} + N^T N\mathbf{f} + \lambda K\mathbf{f} = N^T y \tag{4.8}$$

olarak elde edilir. (4.7) ve (4.8) denklemleri birleştirilerek, aşağıdaki blok matris denklemi şeklinde yazılabilir:

$$\begin{bmatrix} \mathrm{Z}^{\mathrm{T}}\mathrm{Z} & \mathrm{Z}^{\mathrm{T}}\mathrm{N} \\ \mathrm{N}^{\mathrm{T}}\mathrm{Z} & \mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{f} \end{bmatrix} = \begin{bmatrix} \mathrm{Z}^{\mathrm{T}} \\ \mathrm{N}^{\mathrm{T}} \end{bmatrix} \mathbf{y} \tag{4.9}$$

(4.4) modelinin parametrik kısmı göz ardı edildiğinde $(\beta = 0)$ (4.8) denklemi

$$\left(\mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K\right)\mathbf{f} = \mathrm{N}^{\mathrm{T}}\mathbf{y} \tag{4.10}$$

denklemine indirgenir. Böylece, $\mathrm{N}\,\mathbf{f}$ uyum değerleri vektörünü elde etmek için $\mathbf{y}$ vektörüne uygulanan ve verilen bir $\lambda > 0$ sabitine bağlı *düzeltme matrisi*,

$$S_\lambda = \mathrm{N}(\mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K)^{-1}\mathrm{N}^{\mathrm{T}} \tag{4.11}$$

ifadesi ile elde edilir. Ayrıca, x_i düğüm noktaları farklı ve önceden sıralıysa, $\mathrm{N} = \mathrm{I}$ (birim matris) olması nedeniyle, S_λ düzeltme matrisi (3.35)'de ifade edildiği gibi,

$$S_\lambda = (I + \lambda K)^{-1}$$

biçimine indirgenir.

Parametrik terim dikkate alındığında, söz konusu $\boldsymbol{\beta}$ parametrik katsayılar ve bilinmeyen f fonksiyonunun tahmin vektörlerinin elde edilmesinde, izleyen alt başlıklarda incelenen yaklaşımlar ortaya çıkmaktadır.

4.2.2. Backfitting Süreci

Eşitlik (4.3)'de belirtilen semiparametrik regresyon modelini oluşturan $\boldsymbol{\beta}$ ve $\mathbf{f}$ vektörleri (4.9) matris denkleminin çözümü olarak elde edilir. Genellikle (4.9) matris sistemi büyük olduğundan söz konusu sistemin direkt çözümü elverişli değildir. Bu nedenle, (4.9) matris denklemine çözmek için farklı iteratif süreçler önerilebilir. Bu yöntemlerden en yaygın olanı *"backfitting"* sürecidir [33]. (4.9) matris sistemi (4.7) ve (4.8) gibi iki sistemin yardımıyla verilebilir. Söz konusu (4.7) ve (4.8) denklemleri yardımıyla,

$$Z^{\mathrm{T}}Z\boldsymbol{\beta} = Z^{\mathrm{T}}\left(\mathbf{y} - N\,\mathbf{f}\right) \tag{4.12}$$

$$\left(N^{\mathrm{T}}N + \lambda K\right)\mathbf{f} = N^{\mathrm{T}}\left(\mathbf{y} - Z\boldsymbol{\beta}\right). \tag{4.13}$$

sistemleri yazılabilir. (4.12) ve (4.13) denklemleri, backfitting süreci için sezgisel olarak bir fikir ileri sürmeye olanak sağlar. Sabit bir $\mathbf{f}$ vektörü için (4.12)'den $y - N\,\mathbf{f} = y^*$ farkına uygun $\boldsymbol{\beta}$ parametreler vektörü, en küçük karelere regresyonuyla tahmin edilir:

$$\hat{\boldsymbol{\beta}} = \left(Z^T Z\right)^{-1} Z^T y^* . \qquad (4.14)$$

Diğer taraftan sabit $\boldsymbol{\beta}$ vektörü için (4.13) denklemi, $\tilde{y} = \mathbf{y} - Z\boldsymbol{\beta}$ farkına uygun bir kübik splayn interpolasyonu yapma imkânı sağlar. Bu durumda f splayn fonksiyonunun $\hat{\mathbf{f}}$ tahmin vektörü,

$$\hat{\mathbf{f}} = N\mathbf{f} = \left(N^{\mathrm{T}}N + \lambda K\right)^{-1} N^{\mathrm{T}}\left(\mathbf{y} - z\boldsymbol{\beta}\right) \text{ veya } \hat{\mathbf{f}} = \mathrm{S}_\lambda \tilde{y} \qquad (4.15)$$

olarak elde edilir. (4.12), (4.13) denklemleri $\hat{\mathbf{f}}$ ve $\boldsymbol{\beta}$ vektörlerini tahmin etmek için backfitting olarak adlandırılan bir iteratif süreç oluşturması imkanı sağlar [38 ve 39]. (4.12), (4.13) denklemlerine uygun

$$\begin{aligned} \hat{\mathbf{f}}^{(n)} &= S_\lambda(y - Z\beta^{(n-1)}) \\ \beta^{(n)} &= (Z^T Z)^{-1} Z(y - \hat{\mathbf{f}}^{(n)}),\ n = 1,2,.. \end{aligned} \qquad (4.16)$$

iteratif sürecini ele alınım. Burada $\hat{\mathbf{f}} = N\mathbf{f}$, $(n \times 1)$ boyutlu vektör ve $S_\lambda = \mathrm{N}(\mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K)^{-1}\mathrm{N}^{\mathrm{T}}$ düzeltme matrisidir. (4.9) ifadesinin sol kısmındaki blok matris pozitif tanımlı olduğundan, (4.9) matris denkleminin tek bir çözümü olur ve (4.16) süreci her hangi bir $\beta^{(0)}$ başlangıç vektörü için yakınsak olur. Diğer taraftan,

$$\begin{bmatrix} \mathrm{Z}^{\mathrm{T}}\mathrm{Z} & \mathrm{Z}^{\mathrm{T}}\mathrm{N} \\ \mathrm{N}^{\mathrm{T}}\mathrm{Z} & \mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K \end{bmatrix}$$

blok matrisinin ise pozitif tanımlı olması için gerek ve yeter koşullar aşağıdaki gibi ifade edilir [11]:

a) Z matrisinin sütunları lineer bağımsızdır.

b) Tüm $i = 1,2,...,n$ için $\delta_1 + \delta_2 x_i$ doğrusal şekline eşit $\mathbf{z}_i^T\boldsymbol{\beta}$ lineer kombinasyonu yoktur.

Gerçekte bu koşullar, açıklayıcı değişkenleri $(\mathbf{z}_i^T, 1, x_i)^T$ olan tam parametrik doğrusal modelde en küçük kareler tahminlerinin tek olması koşullarıdır.

4.2.3. Green ve Silverman'nın Doğrudan Yaklaşımı

Backfitting süreci, ilgili matrisin öz değerlerinin mutlak değerleri 1'den küçük olduğu için teorik olarak her zaman yakınsaktır. Ancak pratikte en büyük öz değer 1'e yakın olduğunda yakınsama hızı çok yavaş olur. (4.9) denklemlerini $O(n)$ zamanda çözmek için alternatif (iteratif olmayan) yaklaşımlar da mevcuttur. (4.15) formülünü $\mathbf{f}$ 'i yok etmek amacıyla (4.12)'de kullanarak,

$$\mathrm{Z}^{\mathrm{T}}\left(I - S_{\lambda}\right) Z\,\boldsymbol{\beta} = \mathrm{Z}^{\mathrm{T}}\left(I - S_{\lambda}\right)\mathbf{y} \qquad (4.17)$$

$(p \times p)$ lineer (doğrusal) denklemler sistemini elde edilir. Burada $S_{\lambda} = \mathrm{N}(\mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K)^{-1}\mathrm{N}^{\mathrm{T}}$ düzeltme matrisidir. (4.17) eşitliği köşegen olmayan $\left(I - S_{\lambda}\right)$ matrisi ile genelleştirilmiş en küçük karelerin (GEKK) normal denklemleridir. Bu denklemler sisteminin çözümünün hesaplanması açısından karmaşık bir problem olarak görülebilir. Fakat S_{λ} matrisinin özel yapısı "*Reinsch algoritması*" yardımıyla ortaya çıkarılabilir. Sabit bir p için $Z^{\mathrm{T}}\left(I - S_{\lambda}\right)Z$ ifadesi $O(n)$ işlemlerde hesaplanabilir. $\mathrm{Z}^{\mathrm{T}}\left(I - S_{\lambda}\right)\mathbf{y}$ 'de aynı zaman oranında hesaplanabilir. Buna göre, $(p \times p)$ (4.17) lineer denklemler sistemi standart metotlar, örneğin Cholesky ayrıştırması metodu uygulanarak, $O(p^2)$ işlemde çözülebilir. Böylece, (4.15) formülünde bulunan $S_{\lambda}y$ ve $S_{\lambda}Z$ vektörleri önceden elde edilmiş olur. Sonuç olarak (4.17)'den ve (4.15)'den tahmin edilmesi gereken,

$$\hat{\boldsymbol{\beta}} = \left[\ \mathrm{Z}^{\mathrm{T}}\left(I - S_{\lambda}\right)Z\right]^{-1} Z^{\mathrm{T}}\left(I - S_{\lambda}\right)\mathbf{y} \qquad (4.18)$$

$$\hat{\mathbf{f}} = S_{\lambda}\left(\mathbf{y} - z\boldsymbol{\beta}\right) \qquad (4.19)$$

vektörleri bulunmuş olur.

Eşitlik (4.4) ile verilen semiparametrik regresyon modelinin kısmi splayn yaklaşımına göre,

$$\mu_p = Z\,\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \mathbf{H}\,\mathbf{y}$$

ortalama vektörleri (uyumları) elde etmek için **y** vektörlerine uygulanması gerekli olan **H** düzeltme matrisi aşağıdaki şekilde hesaplanır: S_λ düzeltme matrisi yardımıyla,

$$\tilde{Z} = (\mathrm{I} - S_\lambda) Z$$

dönüşümü yapılır. Z parametrik değişkeni, onun $\tilde{Z}$ dönüşümü ve S_λ matrisi vasıtasıyla söz konusu düzeltme matrisi,

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

olarak elde edilir. Buna göre, Schimek [40] ile Eubank ve ark. [41] tarafından da belirtildiği gibi, düzeltme matrisi,

$$\mathbf{H}_{\mathrm{p}} = S_\lambda + \tilde{Z}(\tilde{Z}^T Z)^{-1} \tilde{Z} \qquad (4.20)$$

şeklinde yazılır.

Kuramda, parametik değişken Z'nin bileşenleri x'e bağlı olduğunda optimum λ seçimi için Green ve Silverman'nın kısmi splayn kestiricilerinin genellikle sapmalı olduğu Rice [42] tarafından gösterilmiştir. Speckman [35] tarafından geliştirilen yöntemde, bu sapmalar önemli ölçüde indirgenebilir. İzleyen bölümde, (4.4) semiparametrik regresyon modelinin kestirimi için Green ve Silverman [11] tarafından verilen (4.18) ve (4.19) kestiricilerine alternatif bir yaklaşım, Speckman [35] tarafından önerilmiş olup, söz konusu yaklaşım izeleyen bölümde ele alınmıştır.

4.3. Speckman Yöntemi

Bu bölümde, Speckman [35] tarafından verilen, kısmi splayn yöntemine alternatif bir yaklaşım incelenecektir. Yaklaşımı açıklamak için,

$$y_i = \mathrm{z}_i^T \boldsymbol{\beta} + f\left(x_i\right) + \varepsilon_i, \;\; i = 1, \ldots, n \qquad (4.21)$$

semiparametrik regresyon modelindeki z_i açıklayıcı değişkenlerinin x_i parametrik olmayan kestirici değişkenine göre bir regresyon bağlılığı olarak düşünülen,

$$z_{\mathrm{i}} = m(x_{\mathrm{i}}) + \eta_{\mathrm{i}},\ \mathrm{i} = 1,...,n \tag{4.22}$$

şeklindeki model dikkate alınır. Burada belirtilen $m = \left(m(x_1),...,m(x_n)\right)$, x'in pürüzsüz vektör-fonksiyonu ve η_{i}, hata terimlerinin vektörleridir. (4.22) modeli (4.21)'de yerine yazılarak,

$$y_i = \left(m(x_{\mathrm{i}}) + \eta_{\mathrm{i}}\right)^{\mathrm{T}} \boldsymbol{\beta} + f\left(x_i\right) + \varepsilon_i = m(x_{\mathrm{i}})^{\mathrm{T}}\ \boldsymbol{\beta} + f\left(x_i\right) + (\eta_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i) \tag{4.23}$$

ifadesi elde edilir. (4.23) ifadesinde, $(\eta_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i) = hata$ ve $m(x_{\mathrm{i}})^{\mathrm{T}}\ \boldsymbol{\beta} + f\left(x_i\right) = f_0(x_i)$ olarak belirtelim. Bir anlamda yeni modelin

$$y_i = f_0(x_i) + hata \tag{4.24}$$

şeklinde olması için aşağıdaki gibi bir f_0 fonksiyonu tanımlanır:

$$f_0(x_i) = m(x_i)^{\mathrm{T}}\ \boldsymbol{\beta} + f(x_i) \tag{4.25}$$

(4.21) ve (4.25) ifadelerinin farkı alınırsa, aşağıdaki ifade elde edilir:

$$y_i - f_0(x_i) = \left\{\mathbf{z}_i - m(x_i)\right\}^{\mathrm{T}}\ \boldsymbol{\beta} + hata\ . \tag{4.26}$$

Bu eşitlik gösteriyor ki, $\boldsymbol{\beta}$ parametre vektörü söz konusu açıklayıcı değişkenlere göre y_i değişkeninin artıklarının regresyonundan elde edilir. S_λ verilen herhangi bir λ düzeltme parametresi için splayn düzeltme yönteminin şapka matrisi olsun. Ayrıca $\mathbf{y}$ i.gözlem değeri y_i ile gösterilen bağımlı değişken vektörü, Z i.satırları $\mathbf{z}_i^{\mathrm{T}}$ ile belirtilen bağımsız değişkenlerin gözlem değerlerinin matrisi ve θ i.satırları $m(x_i)^T$ ile belirtilen matris olduğu gösterilirse (4.26) ilişkisi gereği aşağıdaki işlemler gerçekleştirilir:

i.) (4.22) ve (4.24) modellerinden sırasıyla $S_\lambda Z$ ve $S_\lambda \mathbf{y}$ tahminlerini veren $\theta^T = \left(m(x_1),...,m(x_n)\right)$ ve $f_0 = \left(f_0(x_1),...,f_0(x_n)\right)^T$ vektörlerini tahmin etmek için splayn düzeltme kullanılır.

ii.) (4.26) ilişkisinden aşağıdaki ifadeler yazılır:

$$\mathbf{y} - f_0 = \mathbf{y} - S_\lambda \mathbf{y} = (\mathrm{I} - S_\lambda)\mathbf{y}$$

ve

$$\mathrm{Z} - \theta = Z - S_\lambda Z = (\mathrm{I} - S_\lambda) Z$$

Buradan hareketle, artıklar $\tilde{\mathbf{y}} = (\mathrm{I} - S_\lambda)\mathbf{y}$ ve $\tilde{Z} = (\mathrm{I} - S_\lambda) Z$ olarak tanımlanır.

iii.) (4.25) ilişkisinden hareketle $\tilde{\mathbf{y}}$ 'nin $\tilde{Z}$ 'ye göre regresyon denklemi,

$$\tilde{\mathbf{y}} = \tilde{Z}\boldsymbol{\beta} + hata$$

olarak yazılır. $\tilde{\mathbf{y}}$ 'nin $\tilde{Z}$ 'ye göre regresyon denklemine bilinen en küçük kareler yöntemi uygulanarak, (4.21) modelinin parametrik bileşeni için regresyon katsayıları aşağıdaki gibi bulunur:

$$\hat{\boldsymbol{\beta}} = \{\tilde{Z}^{\mathrm{T}}\tilde{Z}\}^{-1}\tilde{Z}^{\mathrm{T}}\tilde{\mathbf{y}} \qquad (4.27)$$

iv.) $\hat{\boldsymbol{\beta}}$ tahmini (4.21) modelinde yerine yazılarak, $y_{\mathrm{i}} - \mathbf{z}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{y}_i^*$ olarak belirtilir. Buna göre (4.21) semiparametrik regresyon modeli,

$$\mathbf{y}_i^* = f(x_i) + hata$$

şeklindeki modele dönüşür. $\mathbf{y}_i^*$ değerlerine uygulanan splayn düzeltme yöntemine göre, (4.21) modelinin parametrik olmayan bileşeni için aşağıdaki gibi bir $\hat{\mathbf{f}}$ tahmini elde edilir:

$$\hat{\mathbf{f}} = (I - \lambda K)^{-1}\mathbf{y}_i^* = S_\lambda \mathbf{y}_i^* \qquad (4.28)$$

Kısmi splayn yaklaşımına benzer olarak, (4.4) semiparametrik regresyon modelinin Speckman yaklaşımına göre,

$$\mu_p(\hat{\mathbf{y}}) = Z\hat{\boldsymbol{\beta}} + \hat{\mathbf{f}} = \mathbf{H}\mathbf{y}$$

uyumlarını (ortalama vektörleri) elde etmek için kullanılan **H** düzeltme matrisi (şapka matrisi) aşağıdaki şekilde hesaplanır: S_λ matrisi yardımıyla,

$$\tilde{Z} = (\mathrm{I} - S_\lambda) Z$$

dönüşümü yapılır. Z parametrik değişkeni, onun $\tilde{Z}$ dönüşümü ve S_λ matrisi yardımıyla söz konusu düzeltme matrisi,

$$\mathbf{H}_s = S_\lambda + \tilde{Z}(\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}\left(\mathrm{I} - S_\lambda\right) \qquad (4.29)$$

şeklinde elde edilir.

4.4. Düzeltme Parametresinin Seçimi

Spline düzeltme yöntemine dayalı semiparametrik modeli değerlendirmek için bir λ düzeltme parametresinin seçilmesi gerekir. Dördüncü bölümde ayrıntılı olarak incelenen ve düzeltme parametresinin seçiminde kullanılan *Çapraz-Geçerlilik, Genelleştirilmiş Çapraz-Geçerlilik, Geliştirilmiş Akaike Bilgi Kriteri ve Mallow'un Cp Kriteri* olan klasik yöntemlerin yanı sıra, *Klasik Pilotları Kullanan Risk Tahmini ve Lokal (bölgesel) Risk Tahmin kriterleri* olarak adlandırılan, risk tahmin metotları yardımıyla λ düzeltme parametresinin seçimi yapılabilir. Semiparametrik model için λ düzeltme parametresinin belirlenmesinde kullanılan kriterlerin, parametrik olmayan bir model için kullanılan kriterlerden farkı (3.35)'te verilen S_λ matrisi yerine (4.20) ve (4.29)'de verilen $\mathbf{H}$ matrislerinin hesaplanması gerekir.

Kısmi splayn yaklaşımını kullanıldığında, (3.35)'de verilen ve simgesel olarak $tr\left(S_\lambda\right)$ ile gösterilen matrisin izinin yerine, (4.20)'de ifade edilen matrisin izi hesaplanır. Diğer bir ifadeyle $tr\left(\mathbf{H}_p\right)$ hesaplanır. Benzer olarak, Speckman yaklaşımı benimsendiğinde, söz konusu $tr\left(S_\lambda\right)$ yerine (4.29)'da verilen matrisin izi olan $tr\left(\mathbf{H}_s\right)$ hesaplanır. Burada ifade edilen $tr\left(\mathbf{H}_s\right)$ ve $tr\left(\mathbf{H}_p\right)$ izleri $O(n)$ adımda hesaplanabilir. Buna göre, λ düzeltme parametresi söz konusu bu matrislerin boyutları ile orantılı bir zamanda elde edilir [40].

4.5. Varyans ve Kovaryans Tahmini

Semiparametrik regresyon modelinin varyans ve kovaryansların tahminleri aşağıda belirtilen amaçlar için gereklidir:

- Düzeltme parametresinin seçiminde kullanılan yansız risk kriteri ve risk tahmin kriterlerinin hesaplanması.
- Modelin parametrik bileşeni hakkındaki çıkarsamalar.
- Modelin parametrik olmayan fonksiyonu hakkındaki çıkarsamalar.

Burada, Gasser, Sroka ve Jenner [43] tarafından verilen fark esaslı varyans kestiricisinin bir uyarlaması Schimek [40] tarafından verilmiştir. Bu varyans kestirici kernel ve splayn düzeltme regresyonunda başarılı bir şekilde uygulanmıştır.

Semiparametrik modellerde sadece parametrik olmayan bir kısım (f fonksiyonu) değil aynı zamanda $\beta \neq 0$ olan bir parametrik kısıma da (burada $\mathbf{z}\boldsymbol{\beta}$) söz konusudur. Bu nedenle $\mathbf{z}\boldsymbol{\beta}$ ile belirtilen parametrik terimi hesaplamak için geliştirilen varyans kestiricisi, aşağıdaki gibi elde edilir:

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}-\boldsymbol{P})\mathbf{A}\boldsymbol{y}}{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}(\mathbf{I}-\boldsymbol{P})\mathbf{A}\right)} \tag{4.30}$$

burada,

$\boldsymbol{P} = \mathbf{A}\mathbf{z}(\mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{z})^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$.

Buradaki $(n-2)\times n$ boyutlu $\mathbf{A}$ matrisi aşağıdaki elemanlardan oluşmaktadır: Girişlerin tümü sıfır fakat i.inci giriş $a_i c_i$ ile, (i+1).inci giriş $-c_i$ ile ve (i+2).inci giriş de $b_i c_i$ ile tanımlanmıştır. Burada,

$$a_i = \frac{x_{i+1}-x_i}{x_{i+1}-x_{i-1}}, \quad b_i = \frac{x_i - x_{i-1}}{x_{i+1}-x_{i-1}}$$

ve

$$c_i = \left(a_i^{\,2} + b_i^{\,2} + 1\right)^{-\frac{1}{2}}, \quad i = 1,\ldots,n-1.$$

Semiparametrik regresyonun parametrik katsayılarının varyanslarının tahmininde kullanılan varyans kestiricileri için Schimek [40] tarafından yapılan çalışmada iki yaklaşım ele alınmıştır. Bu yaklaşımlar: *Kısmi splayn (partial spline) ve Speckman yaklaşımıdır*. Sırasıyla, kısmi splayn (β_p) ve Speckman kestiricileri (β_s) için varyans-kovaryans matrisleri, aşağıdaki gibidir:

$$\mathrm{Var_P} = \sigma^2 \left(\tilde{Z}^T Z\right)^{-1} \tilde{Z}^T \tilde{Z} \left(Z^T \tilde{Z}\right)^{-1} \quad (4.31)$$

ve

$$\mathrm{Var_S} = \sigma^2 \left(\tilde{Z}^T \tilde{Z}\right)^{-1} \tilde{Z}^T \left(I - S_\lambda\right)^2 \tilde{Z} \left(\tilde{Z}^T \tilde{Z}\right)^{-1} \quad (4.32)$$

Burada $\tilde{Z} = (\mathrm{I} - S_\lambda) Z$ ve S_λ düzeltici matris olup, varyans-kovaryans matrislerinin asal köşegenleri üzerindeki elamanlar varyansları, diğer elamanlar ise kovaryansları gösterir.

4.6. Semiparametrik Modele İlişkin Çıkarsamalar

Eşitlik (4.3)'de verilen semiparametrik regresyon modelini değerlendirmek için hem parametrik hem de parametrik olmayan bileşen üzerinde hipotez testleri yapmak gerekir. Semiparametrik regresyon modelleri hakkında yapılan çıkarsamalar aşağıdaki varsayımlara dayanır:

- Bağımlı ve bağımsız değişkenler sürekli ölçekle ölçümlenir.
- Hata varyansı σ^2'nin uygun tahminidir.
- Bağımsız değişkenler arasında korelasyon yoktur.
- Parametrik olmayan regresyon kestiricisi **f**, incelenen y bağımlı değişkeni bakımından doğrusaldır.
- y bağımlı değişkeni bağımsız ve normal olarak dağılır.

Semiparametrik modelle tutarlı tahminlerin gerçekleştirilebilmesi için yukarıda bahsedilen varsayımların sağlanması gerekir. İzleyen alt başlıklarda, semiparametrik modelle yapılan kestirimleri istatistiksel açıdan değerlendirebilmek amacıyla, modelin hem parametrik bileşeni hem de parametrik olmayan bileşeni hakkında yapılan çıkarsamalar konusu ele alınmıştır.

4.6.1. Parametrik Bileşen İçin Çıkarsama

Regresyon analizi, örneklem verileriyle yapıldığından, elde edilen $\hat{\beta}$ tahmini vektörleri, β parametrelerine ilişkin testlerde başka bir deyişle, regresyon modelinin anlamlılığının sınanmasında kullanılır. Dolayısıyla parametrik katsayılara ilişkin test, modelin anlamlılığını da test eder. Asimtotik olarak normal

olduğu verilen parametrik kısmın katsayıları için sırasıyla, güven aralıkları ve test istatistiklerini hesaplamak için β_p *(kısmi spline yaklaşımı)* ve β_s (*Speckman yaklaşımı)'nin* varyans-kovaryans matrislerine başvuralabilir. Var_p ve Var_s matrislerinin her biri $\hat{\boldsymbol{\beta}}$ tahmini katsayıların standart hatalarını (*standard errors-SE*) vermektedir. Böylece, parametrik katsayıların istatistiksel açıdan anlamlı olup olmadığını test edilmek istendiğinde;

$$\mathrm{H}_0 : \hat{\beta}_j = 0$$
$$\mathrm{H}_1 : \hat{\beta}_j \neq 0$$

hipotezleri

$$\mathrm{t}_{\mathrm{df}} = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_j\right)}, \quad j = 1, \ldots, k \tag{4.33}$$

formülü ile verilen, $\mathrm{df} = n - tr\left(S_\lambda\right) - k$ (serbestlik derecesi) ile t-dağılımına sahip bir test istatistiği yardımıyla test edilir. Ayrıca yukarıdaki varsayım bir F testini doğrular (burada Speckman'ın yaklaşımı söz konusudur). Speckman [35], varyansın $\left(\sigma^2\right)$ bir kestiricisi olarak,

$$MSE = \frac{Hata\ KarelerToplamı}{Serbestlik\ Derecesi} = \frac{RSS}{tr\left(\mathrm{I} - \mathbf{H}_s\right)^{\mathrm{T}}\left(\mathrm{I} - \mathbf{H}_s\right)} \tag{4.34}$$

şeklinde ifade edilen hata kareler ortalamasını kullanmayı önermiştir. (4.34)'ün payında yer alan hata (artık) kareler toplamı

$$RSS = n^{-1}\left\|\left(\mathrm{I} - \mathbf{H}_s\right)\mathbf{y}\right\|^2$$

formülü ile tanımlanır. (4.34)'deki varyans kestiricisi pozitif ama asimtotik olarak önemsiz bir yana (sapmaya) sahiptir. Parametrik katsayıların toplu olarak istatistiksel açıdan anlamlı olup olmadığı test edilmek istendiğinde,

$$\mathrm{H}_0 : \hat{\boldsymbol{\beta}} = 0 \quad \text{veya} \quad \mathrm{H}_0 : \beta_1 = \ldots = \beta_k = 0$$
$$\mathrm{H}_1 : \hat{\boldsymbol{\beta}} \neq 0 \quad \text{veya} \quad \mathrm{H}_1 : \beta_1 \neq \ldots \neq \beta_k \neq 0 \ (\textit{en az bir } \beta_j \neq 0)$$

hipotezleri aşağıdaki (4.35) ile verilen F testi yardımıyla test edilir. F testi için parametrik bileşenin;

$$SS_{par} = \hat{\boldsymbol{\beta}}_s^T \left(q^T q \right)^{-1} \hat{\boldsymbol{\beta}}_s$$

formülü ile verilen kareler toplamı gereklidir. Burada, $q^T = \left(\tilde{Z}^T \tilde{Z} \right)^{-1} \tilde{Z}^T \left(\mathrm{I} - S_\lambda \right)$ olarak ifade edilir. Diğer yandan, parametrik bileşenin kareler toplamının ortalaması,

$$MSS_{par} = n^{-1} \left(SS_{par} \right)$$

olarak tanımlanır. Bu durumda, $\mathrm{df}_1 = k$ *ve* $\mathrm{df}_2 = tr \left(\mathrm{I} - \mathbf{H}_s \right)^T \left(\mathrm{I} - \mathbf{H}_s \right)$ serbestlik dereceleri ile F test istatistiği;

$$\mathrm{F}_{\mathrm{df}_1, \mathrm{df}_2} = \frac{MSS_{par}}{MSE} = \frac{SS_{par} / n}{RSS / tr \left(\mathrm{I} - \mathbf{H}_s^{\,T} \right) \left(\mathrm{I} - \mathbf{H}_s \right)} \qquad (4.35)$$

eşitliği ile elde edilir. Ayrıca, Speckman [35] tarafından önerilen (4.34) varyans kestiricisi yerine değiştirilmiş farka-dayalı (4.30) varyans kestiricisi de kullanılabilir.

İstatistiksel çıkarsamalarda yapılan kestirimlerin, gerçek değerlerle genellemesi aralık kestirimiyle yapılır. Şu ana kadar kısmi splayn ve Speckman yaklaşımları yardımıyla elde edilen parametrik $\hat{\boldsymbol{\beta}}$ katsayılarına ilişkin kestirimler tek değer veya nokta kestirimlerdir. (4.18) ve (4.27) ile hesaplanan parametrik $\hat{\boldsymbol{\beta}}$ katsayılar vektörünün *100(*$1-\alpha$*)% güven aralığını*,

$$P \left(\hat{\beta}_i - t_\alpha SE(\hat{\beta}_i) \le \beta \le \hat{\beta}_i + t_\alpha SE(\hat{\beta}_i) \right) = 1 - \alpha, \quad i = 1, ..., k \qquad (4.36)$$

şeklinde verilebilir. Buradaki $SE(\hat{\beta}_i)$, parametrik kısma karşı gelen katsayıların standart hataları olup, (4.32)'de belirtilen varyans- kovaryans matrisinin köşegen elamanlarının elde edilir. t_α, α anlam düzeyi ve $\mathrm{df} = n - tr \left(S_\lambda \right) - k$ serbestlik derecesindeki t-Tablo değerdir.

4.6.2. Parametrik Olmayan Bileşen İçin Çıkarsama

Amaç f eğrisinin biçimsel olarak şeklini değerlendirmektir. Test edilmek istenen sıfır ve alternatif hipotezler:

$$H_0 : E(y_i) = \mu\,(\textit{doğrusal fonksiyon})$$
$$H_1 : E(y_i) = f(x_i)\,(\textit{pürüzsüz fonksiyon})$$

Şeklinde ifade edilirler. Böyle bir test sade parametrik bir modelle karşılaştırılan semiparametrik modelin bir anlam ifade edip etmeyeceğine karar verilmesine olanak sağladığı için ilgi odağıdır. Yukarıda bahsedilen sıfır hipotezinin testinde kullanılan testlerden bazılarına aşağıda değinilmiştir.

Hastie and Tibshirani [44], semiparametrik ortama uygulanabilen, bir parametrik olmayan uyum $\hat{f}_1$'e karşın bir doğru denklem uyumu $\hat{f}_0$'ı için bir F testi önermiştir. Böyle bir test için gerekli varsayımlar:

- Sıradan en küçük karelerden (OLS) elde edilen $\hat{f}_0$, sapmasız (yansız) kestiricidir.
- Speckman yaklaşımından elde edilen $\hat{f}_1$, sapmasız kestiricidir (kısmi splayn'dan elde edilmeyen).
- Seçilen düzeltme parametresi optimumdur.

Speckman'ın tahmin kavramıyla sapmasız sonuçlar elde edilir. Bu da çıkarsamayı ilerletmeye yöneltir. Yukarıda adı geçen F -test istatistiği,

$$F_{\mathrm{df}_1-\mathrm{df}_0,\,n-\mathrm{df}_1} = \frac{\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2 - \sum_{i=1}^{n}\hat{\nu}_i^2\right)/\left(\mathrm{df}_1 - \mathrm{df}_0\right)}{\sum_{i=1}^{n}\hat{\nu}_i^2/\left(n-\mathrm{df}_1\right)} \tag{4.37}$$

formülü ile verilir. Burada $\hat{\varepsilon}_i = (y_i - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{OLS})$ ve $\hat{\nu}_i = \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_s + \hat{f}(t_i) - \mathbf{z}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{OLS}$. Serbestlik dereceleri df_0, sıradan en küçük kareler modelindeki parametrelerin sayısına (k tane) ve $\mathrm{df}_1 = tr\left(2\mathbf{H}_s - \mathbf{H}_s\mathbf{H}_s^{\mathrm{T}}\right)$ ifadesine eşit olup Speckman yaklaşımından elde edilir. Diğer taraftan belirtmek gerekir ki, (4.37) F -test istatistiği σ^2'nin tahminini gerektirmeyen bir avantaja sahiptir.

Hastie ve Tibshirani [44], öne sürüdükleri F testinin yeterliliğini ortaya koymuşlardır. Fakat bazı araştırmacılar küçük örneklem sonuçlarına şüphe ile bakmaktadırlar (örneğin, Bowman ve Azzalini [45], normal dağılıma uygunluk için *olabilirlik oran testi* olarak adlandırılan bir diğer yaklaşımı önermişlerdir). Ayrıca, Raz [46] tarafından öne sürülen *mutlak bir permutasyon testi*, farklı ortamlarda önerilmesine rağmen, sınırlı örneklem bilgisine dayalı bir karar için çok daha uygun olduğu öne sürülmektedir.

Son zamanlarda Hong ve White [47] tarafından alternatif bir test öne sürülmüştür. Bu test yaklaşık olarak standart normal dağılan test istatistiği

$$\mathrm{M} = \frac{n\hat{m}_n / \hat{\sigma}_\varepsilon^2 - k_n}{\sqrt{2k_n}},$$

ile verilir. Burada; $\hat{m}_n = n^{-1}\sum_{i=1}^{n} \hat{\nu}_i \hat{\varepsilon}_i$ olup, $\hat{\varepsilon}_i$ ve $\hat{\nu}_i$ ifadeleri (4.37) eşitliğinde tanımlandığı gibidir, k_n bir boyut indirgeme sabiti ve $\hat{\sigma}_\varepsilon^2$ H_0 modeli altında artık (hata) varyansıdır.

Bu kesimde, parametrik olmayan bileşeninin güven aralığı ele alınacaktır. Amaç doğrusal olup olmamamsından başka, f eğrisinin şekli hakkında sağlam bir karar verebilmektir. Açıktır ki, örnek değişkenliğinden dolayı parametrik olmayan tahminlerde bazı eğriliklerin her zaman beklenmesi gerekir. Güven aralıkları önceden belirlenen bir olasılık düzeyinde ve örneklem sonuçlarına göre şekillenirler. Bununla birlikte, $yan\left(sapma(x)\right)$ bilinmeksizin böyle bir aralığı oluşturmak mümkün değildir. Bu yüzden, uygulamada sapma (yan) düzeltmesinden kaçınılır ve sözde (*so-called*) değişkenlik bantları (sınırları) geniş ölçüde kullanılır. $\hat{f}(x)$ bilinmeyen pürüzsüz fonksiyonun splayn düzeltme tahmini olmak üzere, *parametrik olmayan bileşenin güven aralığı*,

$$\hat{f}(x) \pm 2SE(\hat{f}(x)) = \hat{f}(x) - 2SE(\hat{f}(x)) \le f(x) \le \hat{f}(x) + 2SE(\hat{f}(x)) \qquad (4.38)$$

olarak ifade edilir. Buradaki $SE(\hat{f}(x))$, eğrinin tahmini standart hatasıdır. (4.38) güven aralığı $f(x)$'den ziyade $E\left(\hat{f}(x)\right)$ için aralık noktalarından oluşur. Elbette ki, dikkatli bir yorumlama, bandın doğasından dolayı zorunludur.

4.7. Müstakil Evlerin Satış Fiyatları ile Evlerin Özellikleri Arasındaki İlişkilerin Araştırılması Konusunda Bir Uygulama

Bilindiği gibi yaşadığımız evlerin fiyatlarının belirlenmesinde evlerin özellikleri etkin bir rol oynar. Söz konusu bu özelliklerin evlerin satış fiyatlarını nasıl etkilediklerinin bilinmesi, önemli bir etkendir. Bu nedenle, evlerin satış fiyatları üzerinde etkili olan değişkenlerin iyi tespit edilmesi gerekir. Söz konusu evlerin satış fiyatları üzerinde etkili olan değişkenlerin belirlemesi için başvurulan regresyon analizinde, evlerin satış fiyatları ile fiyatı etkileyen açıklayıcı değişkenler doğası gereğince hem parametrik hem de parametrik olmayan bir ilişki içerisinde olduğu göz ardı edilemeyen bir gerçektir. Örneğin, evlerin satış fiyatları ile evlerin bahçe dahil brüt kullanım alanları arasında kesin bir doğrusal ilişki vardır denilemez. Bu nedenle, böyle bir ilişkiyi analiz etmede, hem parametrik kısmı hem de doğrusallık varsayımını esneten, parametrik olmayan kısmı bir bütün olarak ele alan semiparametrik regresyon yöntemi kullanılmalıdır. Bu bölümde, yukarıda adı geçen ilişkiyi analiz etmek için hem bilenen parametrik regresyon hem de semiparametrik regresyon yöntemleri kullanılarak, söz konusu yöntemler, karşılaştırmalı bir biçimde incelenmiştir.

4.7.1 Veri ve Değişken Tanımları

Veriler Kanada'nın başkenti olan Ottawa'da, 1987 yılında satılan 92 müstakil evin satış fiyatı ve evlerin karakteristiklerini gösteren değişkenlere ilişkin gözlem değerlerinden oluşmaktadır. Söz konusu veriler Ho (1995) tarafından yazılan "essay on the housing market" adlı doktora tez çalışmasından alınmış olup, veriler Tablo 4.1'de özet olarak verilmiştir. Örnek uygulamada kullanılan değişkenler aşağıdaki gibi tanımlanır:

SF: *Evin satış fiyatı (dolar olarak)*
AU: *Evin anayola uzaklığı*
FD: *Şömine için yapay* (*dummy*) *değişken*
GD: *Garaj için yapay değişken*
KG: *Söz konusu semtte yaşayan insanların ortalama geliri(dolar olarak,*
NKA: *Evin net kullanım alanı (square feet)*
BKA: *Evin bahçe dahil brüt kullanım alanını (square feet) gösterir*

Tablo 4.1: Değişkenlere ilişkin gözlem değerlerinin özeti

Değişkenler	Ortalama	Standart Sap.	Max	Min
SF	146.0049	33.4323	240	75
AU	0.8399	0.4477	1.8643	0.1032
FD	0.6957	0.4627	1	0
GD	0.6413	0.4822	1	0
KG	49.2817	11.6262	79.4580	25.9540
NKA	1.1419	0.2862	1.9997	0.6247
BKA	5.2823	1.1551	9.9000	1.8910

4.7.2. Deneysel Değerlendirmeler

Yapılan örnek uygulamada semiparametrik regresyonun farklı iki yaklaşımı olan Speckman yöntemi, Green ve Silverman'nın kımsi splayn yöntemi ve parametrik regresyon yöntemi karşılaştırmalı bir biçimde incelenmiştir. MATLAB ortamında tarafımızdan yazılan bir program kullanılarak, söz konusu regresyon modellerinden elde edilen sonuçlar yorumlanmış ve splayn düzeltme yöntemini esas alan semiparametrik regresyon modelinin kestirimine ilişkin tahmin ve çıkarsamalar yapılmıştır.

***Parametrik Regresyon Modelinin Ayrıntıları*:** Örnek uygulamada kullanılan değişkenlere göre, (2.1) modeli ile verilen parametrik doğrusal regresyon modeli,

$$SF_i = \beta_0 + \beta_1 AU_i + \beta_2 FD_i + \beta_3 GD_i + \beta_4 KG_i + \beta_5 NKA_i + \beta_6 BKA_i + \varepsilon_i,\ i = 1,2,...,n. \quad (4.39)$$

şeklinde yazılabilir. (4.39) modelindeki $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$ regresyon katsayıları vektörü olup, söz konusu bu katsayılar sıradan en küçük kareler yöntemine (OLS) göre elde edilmiştir. Ayrıca, regresyon katsayılarının standart hataları, $t-test$ istatistikleri ve %95 güven aralıkları Tablo 4.2'de verilmiştir.
Sıradan en küçük kareler yöntemine göre elde edilen $\boldsymbol{\beta}$ regresyon katsayısı vektörü, (4.39) modelinde yerine yazılarak, parametrik regresyon modeli aşağıdaki gibi ifade edilmiştir:

$$SF = 62.67 - 8.2151 AU + 6.0548 FD + 14.468 GD + 0.56001 KG + 39.27 NKA + 0.81454 BKA + \varepsilon \quad (4.40)$$

Tablo 4.2: Parametrik Regresyon Sonuçları

Değişkenler	Katsayılar	Standart hat.	t-test istatist.	%95 Güven aralıkları
Sabit	62.67	20.451	3.0644	[(22.586) - (102.75)]
AU	-8.2151	6.6935	-1.2273	[(-21.334) - (4.9042)]
FD	6.0548	7.5919	0.79754	[(-8.8253) - (20.935)]
GD	14.468	6.3102	2.2929	[(2.1005) - (26.836)]
KG	0.56001	0.27959	2.003	[(0.012023) - (1.108)]
NKA	39.27	11.855	3.3124	[(16.033) - (62.506)]
BKA	0.81454	2.6181	0.31112	[(-4.3169) - (5.946)]
$R^2 = 0.33287$, $RSS = 67855$, $df = 5$ ve $t_{0.05}(5) = 2.571$				

Tablo 4.2 veya (4.40) modeli incelendiğinde, evlerin ana yola olan uzaklığında bir birimlik artış, evlerin satış fiyatlarında 8.2151 birimlik bir azalışa neden olmakta veya bunun tersi geçerli olmaktadır. Başka bir deyişle, söz konusu evler ana yoldan uzaklaştıkça fiyatlar düşerken, ana yola yaklaştıkça fiyatlar artmaktadır. Buna karşılık, diğer tüm değişkenler ile evlerin satış fiyatları arasında aynı yönlü bir ilişki gözlenmektedir. Örneğin, evlerin net kullanım alanlarındaki bir birimlik bir artış, evlerin satış fiyatlarında 39.27 birimlik bir artışa neden olmaktadır. Ancak (4.32) formülü ile belirlenen ve Tablo 4.2'de verilen $t-test$ istatistikleri incelendiğinde, AU, FD, GD, KG ve BKA değişkenlerine ilişkin katsayılarının $t-test$ istatistikleri, $t-tablo$ değerlerinden küçük olduklarından söz konusu değişkenlere ilişkin katsayılar istatistiksel açıdan anlamsızdır. Diğer taraftan, katsayıların %95 güven aralıklarına bakıldığında, regresyon katsayıların aralık tahminleri oldukça geniş bir aralığı kapsadığı görülmektedir. Ayrıca, modelin R^2 belirlilik katsayısına göre, evlerin satış fiyatlarındaki değişmelerin ancak %33.287'si söz konusu açıklayıcı değişkenler tarafından açıklanabilmektedir. Bunun yanı sıra, evlerin satış fiyatlarına ilişkin tahminlerde, yapılan hataların oldukça yüksek olduğu görülmektedir[1]. Diğer bir deyişle, Tablo 4.2'de verilen ve $RSS = 67855$ olarak hesaplanan hata kareler toplamı oldukça

[1] Doğrusal regresyon için H=Z (Z' Z)$^{-1}$ Z' şapka matrisi yardımı ile, RSS= **y**' (I-H) **y** formülü ile verilen hata kareler toplamı hesaplanır. Diğer taraftan, **y** ve **ŷ** değerleri arasındaki korelasyon katsayısının karesi R^2 değerini diğer bir deyişle, modelin belirlilik katsayısını vermektedir.

yüksek bir değerdir. Bu durum, parametrik regresyon modeli ile yapılan tahminlerinin yeterince tutarlı olmadığı biçiminde yorumlanabilir.

Semiparametrik Regresyon Modelinin Ayrıntıları: Semiparametrik regresyon, daha öncede vurgulandığı gibi ele alınan türden problemler için parametrik regresyonun daha üstün özellikli ve kullanışlı bir seçeneğini oluşturur. Bu uygulama için (4.3) ile tanımlanan *"semiparametrik regresyon modeli"*,

$$SF_i = \beta_1 AU_i + \beta_2 FD_i + \beta_3 GD + \beta_4 KG_i + \beta_5 NKA_i + f(BKA_i) + \varepsilon_i,\ i = 1,2,...,n \quad (4.41)$$

biçiminde ifade edilebilir. Semiparametrik regresyonda yapay değişkenler her zaman parametrik kısma dahil edilirler. Diğer bir deyişle, yapay değişkenler fonksiyonun eğriliğini etkilemedikleri için modelin parametrik olmayan kısmına dahil edilmezler. Bu uygulamada yer alan değişkenlerden BKA değişkeni parametrik olmayan değişken olarak ele alınırken, diğer tüm değişkenler parametrik değişken olarak ele alınmıştır.

Model (4.41)'de, β parametrik regresyon katsayıları ve $f(BKA)$ bilinmeyen fonksiyonu, bölüm 4'te incelenen *"kısmi splayn yöntemine"* göre tarafımızdan yazılan programla elde edilmiş olup, sonuçlar Tablo 4.3'de verilmiştir. Kısmi splayn yöntemine göre tahmin edilen semiparametrik modelin parametrik kısım katsayıları (4.41)'de yerine yazılarak, semiparametrik regresyon modeli aşağıdaki gibi ifade edilmiştir:

$$SF = -8.688AU + 5.6203FD + 14.122GD + 0.58242KG + 38.043NKA + f(BKA) + \varepsilon \quad (4.42)$$

Tablo 4.3 veya (4.42) incelendiğinde, evlerin ana yola olan uzaklığında bir birimlik bir artış, evlerin satış fiyatlarında 8.688 birimlik bir azalışa neden olur veya bunun tersi geçerlidir. Buna karşılık, diğer tüm değişkenler ile evlerin satış fiyatları arasında aynı yönlü bir ilişki olduğu görülmektedir. Örneğin, evlerin net kullanım alanlarındaki bir birimlik bir artış evlerin satış fiyatlarında 38.043 birimlik bir artışa neden olmaktadır.

Model (4.41) ile belirtilen semiparametrik regresyon modelinin kestiriminde kullanılan (4.5) formülü ile tanımlanan cezalı kareler kriterinde yer alan ceza katsayısının diğer bir deyişle, λ düzeltme parametresinin seçiminde, örneklem hacmine uygun olarak en iyi performansı sağlayan klasik yöntemlerden,

Tablo 4.3: Semiparametrik Regresyon Sonuçları (Kısmi Splayn Yöntemi)

Değişkenler	Katsayılar	Standart hat.	t-test istatist.	%95 Güven Aralıkları
Sabit	-	-	-	-
AU	-8.688	0.70338	-12.352	[(-10.067) - (-7.3094)]
FD	5.6203	0.79823	7.041	[(4.0558) - (7.1848)]
GD	14.122	0.66292	21.303	[(12.823) - (15.422)]
KG	0.58242	0.029402	19.809	[(0.52479)- (0.64004)]
NKA	38.043	1.2526	30.372	[(35.588) - (40.498)]
BKA	-	-	-	-
$R^2 = 0.96546$, $RSS = 718.33$, $F = 53.3967$, $df_1 = 5$ *ve* $df_2 = 83,4778$, $F_{\alpha}(df_1, df_2) = 2,29$ $df = 84,9972$ ve $t_{\alpha}(df) = 1,96$				

genelleştirilmiş çapraz geçerlilik (GCV) kriteri kullanılmıştır. Söz konusu seçim kriterlerin performanslarının değerlendirilmesi ayrıntılı olarak izleyen bölümde incelenmiştir.

Bölüm 4'te incelenen Speckman yöntemine göre (4.41) modelindeki, β parametrik regresyon katsayıları ve $f(BKA)$ bilinmeyen fonksiyonu MATLAB ortamında yazılan programla elde edilmiş olup, sonuçlar özetle Tablo 4.4'de verilmiştir. Speckman yöntemine göre tahmin edilen parametrik katsayılar (4.41)'de yerine yazılarak, semiparametrik regresyon modeli aşağıdaki gibi ifade edilebilir:

Tablo 4.4 :Semiparametrik Regresyon Sonuçları (Speckman Yöntemi)

Değişkenler	Katsayılar	Standart hat.	t-ist.	%95 Güven Aralıkları
Sabit	-	-	-	-
AU	-8.947	0.7044	-12.702	[(-10.328) - (-7.5664)]
FD	5.2692	0.80091	6.579	[(3.6994) - (6.839)]
GD	13.933	0.66395	20.984	[(12.631) - (15.234)]
KG	0.59775	0.029568	20.216	[(0.5398) - (0.65571)]
NKA	37.657	1.2581	29.931	[(35.191) - (40.123)]
BKA	-	-	-	-
$R^2 = 0.96555$, $RSS = 717.53$ $F = 39.2996$, $df_1 = 5$ *ve* $df_2 = 84,4432$, $F_{\alpha}(df_1, df_2) = 2,29$ $df = 84,9972$ ve $t_{\alpha}(df) = 1,96$				

$$SF = -8.974AU + 5.2692FD + 13.933GD + 0.59775KG + 37.657NKA + f(BKA) + \varepsilon \qquad (4.43)$$

Tablo 4.4 veya (4.43) incelendiğinde, (4.42) modeli sonuçlarına benzer olarak, evlerin ana yola olan uzaklığındaki bir birimlik bir artışın, evlerin satış fiyatlarında 8.974 birimlik bir azalışa neden olduğu görülür. Buna karşılık, yine (4.42)'de olduğu gibi, diğer tüm değişkenler ile evlerin satış fiyatları arasında aynı yönlü bir ilişki olduğu görülmektedir.

4.7.3 Parametrik ve Semiparametrik Regresyon Modellerinin Karşılaştırılması

Semiparametrik regresyon modelinin "kısmi splayn yöntemi" ile kestirimine ilişkin özet sonuçları gösteren Tablo 4.3 incelendiğinde, semiparametrik modelde bağımlı değişkendeki değişimlerin çok önemli bir kısmı açıklayıcı değişkenler tarafından açıklandığı görülmektedir. Parametrik regresyon modeli ile karşılaştırıldığında, parametrik model bağımlı değişkendeki değişimlerin %33.287'sini açıklarken, semiparametrik model %96.546'sını açıklamaktadır[2]. Diğer taraftan, parametrik modelin hata kareler toplamı 67855 olurken, semiparametrik modelin, Tablo 4.3'de görüldüğü gibi 718.33 olmaktadır. Bunun yanı sıra, regresyon katsayılarının %95 güven aralıklarına bakıldığında, parametrik modelin her bir bağımsız değişkenin etkisini gösteren katsayılarının güven aralıkları, semiparametrik modelin katsayılarından oldukça fazla geniş bir aralığı kapsadıkları görülmektedir. Ayrıca parametrik modelde bazı açıklayıcı değişkenlere ilişkin regresyon katsayılarının anlamsız olduğu görülürken, semiparametrik modelde regresyon katsayılarının anlamlı oldukları görülmektedir. Bu durum semiparametrik modelin parametrik modelden çok daha üstün olduğunun bir göstergesidir.

Tablo 4.4 incelendiğinde, Speckman yöntemine ilişkin semiparametrik modelde de bağımlı değişkendeki değişimlerin, kısmi splayn yönteminde olduğu gibi, çok önemli bir kısmının açıklandığı görülmektedir. Diğer bir ifadeyle Speckman yöntemi ile kestirimi yapılan semiparametrik model, bağımlı

---

[2] Semiparametrik modelin R^2 belirlilik katsayısı, $R^2 = \dfrac{\hat{\mathbf{y}}^T\hat{\mathbf{y}}}{\mathbf{y}^T\mathbf{y}}$ formülü ile hesaplanır.

değişkendeki değişimlerin %96.555'ini açıklamaktadır. Tablo 4.4'de görüldüğü gibi, semiparametrik modelin hata kareler toplamı 717.33 olmakta ve bu modelde regresyon katsayılarının da anlamlı oldukları görülmektedir. Ayrıca, parametrik açıklayıcı değişkenlerin etkisini gösteren regresyon katsayılarının %95 güven aralıkları parametrik regresyon modeline göre oldukça dar bir aralığı kapsadığı görülmektedir. Bu durumda, bir önceki kısmi splayn yönteminde olduğu gibi, Speckman yöntemini esas alan semiparametrik modelin de parametrik modelden çok daha üstün olduğunun göstergesi olarak görülebilir.

Speckman yönetemi kısmi splayn yöntem ile kıyaslandığında, her iki yöntemin de benzer sonuçlar verdiği görülmüştür. Buna rağmen, bu iki yöntem arasında bazı farklılıklar da vardır, Örneğin, Speckman yöntemine göre kestirimi yapılan semiparametrik model, bağımlı değişkendeki değişimlerin %96.555'ini açıklarken, kısmi splayn yöntemini esas alan semiparametrik model %96.546'sını açıklamaktadır. Ayrıca Speckman yöntemi için (4.35) formülü ile tanımlanan hata kareler toplamı 717.53 olurken, kısmi splayn yöntemi için 718.33 olmuştur. Bu durumda, az da olsa Speckman yönteminin daha başarlı olduğu söylenebilir.

4.7.4. Semiparametrik Regresyon Modeline İlişkin Çıkarsamalar

Esas itibariyle amaç, (4.41) ile verilen semiparametrik modelin parametrik bileşeni hakkında çıkarsamalar yapmaktır. İstatistikte gerçek değerler hakkında yapılan çıkarsamalar tahmini değerlere göre yapılır. Bunun için Speckman ve kısmi splayn yöntemine göre MATLAB ortamında elde edilen (4.41) modelinin parametrik katsayıların tahmini değerleri ve standart sapmaları, $t-test$ istatistikleri, parametrik katsayıların güven aralıkları ve F-test istatistiği bir önceki bölümde verilmiştir. İzleyen paragraflarda ele edilen bu sonuçlara ilişkin yorum ve çıkarsamalar yapılmıştır.

***Parametrik Bileşen Hakkında Çıkarsamalar*:** Eşitlik (4.41) ile verilen semiparametrik modelinin kısmi splayn yöntemine göre kestirilen parametrik katsayılarının anlamlı olup olmadığını belirlemek amacıyla, Tablo 4.2'de verilen t-test istatistikleri incelendiğinde, örneğin, net kullanım alanı değişkeninin evlerin satış fiyatları üzerindeki etkisinin anlamlı olup olmadığını test edebilmek için öne sürülen hipotezler,

$$H_0 : \hat{\beta}_5 = 0$$
$$H_1 : \hat{\beta}_5 \neq 0$$

şeklindedir. Söz konusu $\hat{\beta}_5$ katsayısına ilişkin hesaplanan $t-test$ istatistiği, %5 anlam düzeyi ve $df = n - tr(S_\lambda) - k$ = 84,9972 serbestlik derecesine karşı gelen $t-tablo$ değerinden diğer bir deyişle, $t_{0.05}(84.9972) = 1.96$'dan büyük olduğundan sıfır hipotez reddedilir. Buna göre, $\hat{\beta}_5$ katsayısı istatistiksel açıdan anlamlıdır. Böylece, evlerin net kullanım alanlarının satış fiyatları üzerindeki etkisinin istatistiksel açıdan anlamlı olduğu söylenebilir. Benzer şekilde, aynı değişken için Tablo 4.3'de verilen Speckman yöntemine ait $t-test$ istatistiği $t-tablo$ değeri ile karşılaştırıldığında, yine $\hat{\beta}_5$ katsayısının istatistiksel açıdan anlamlı olduğu görülmektedir. Bu bilgilerden hareketle, her iki yöntem için de, kalan tüm parametrik açıklayıcı değişkenlerin bağımlı değişken üzerindeki etkisini gösteren parametrik katsayıların istatistiksel açıdan anlamlı oldukları görülmektedir (bak. Tablo 4.2 ve 4.3).

F-testinde ise, bütün parametrik açıklayıcı değişkenlerin bağımlı değişken üzerindeki etkisini test edebilmek için hipotezler,

$$H_0 : \hat{\beta}_1 = ... = \hat{\beta}_5 = 0$$
$$H_1 : \hat{\beta}_1 \neq ... \neq \hat{\beta}_5 \neq 0 \,(\text{en az bir } \hat{\beta}_j \neq 0)$$

şeklinde oluşturulur. Semiparametrik modelin her iki kestirim yöntemine göre, anlamlı olup olmadığını belirlemek amacıyla Tablo 4.2 ve 4.3'te verilen *F*-istatistikleri %5 anlam düzeyindeki *F*-Tablo değerinden büyük olduğundan sıfır hipotezi reddedilir. Dolayısıyla her iki yönteme göre de kestirilen (4.3) semiparametrik modelinin, t-testi sonucunda olduğu gibi, *F*- test istatistiğine göre de istatistiksel açıdan anlamlı olduğu söylenebilir.

Eşitlik (4.41) ile verilen semiparametrik regresyon modelinin her iki yönteme göre kestiriminden elde edilen ve parametrik açıklayıcı değişkenlerin etkisini gösteren katsayıların nokta tahminlerinin anlamlı oldukları yukarda yapılan testler sonucunda ortaya çıkmıştır. Bunun yanı sıra, söz konusu regresyon katsayılarının (4.36) güven aralıkları da incelenmiş olup, bu sonuçlar Tablo 4.2 ve

4.3'de verilmiştir. Örneğin, Speckman yöntemine göre kestirimi yapılan ve *NKA* parametrik açıklayıcı değişkeninin etkisini gösteren β_5 katsayısı için %95 güven aralığı,

$$P\left(35.191 \le \beta_5 \le 40.123\right) = 0.95$$

biçimde elde edilmiştir. Buna göre, ele alınan örneklemin alındığı ana kütlenin β_5 katsayısı %95 güvenle 35.191 ile 40.123 arasında bir değerdir. Bu aralıkların dışında olma olasılığı %5 dir.

Benzer olarak, kısmi splayn yöntemine göre kestirimi yapılan β_5 katsayısının %95 güven aralığı,

$$P\left(\ 35.588 \ \le \ \beta_5 \ \le \ 40.498\ \right) = 0.95$$

olarak hesaplanmıştır. Buna göre, ele alınan örneklemin çekildiği ana kütlenin β_4 regreson katsayısı, %95 güvenle 35.588 ile 40.498 arasında bir değerdir. Diğer bir ifadeyle, söz konusu ana kütle regresyon katsayısının 35.588 ile 40.498 aralığını kapsama olasılığı %95 ve bu aralığın dışında olma olasılığı ise %5 dir. Kalan diğer parametrik açıklayıcı değişkenlerin etkisini gösteren regresyon katsayılarının %95 güven aralıklarına ilişkin benzer yorumlar yapılabilir.

Parametrik Olmayan Bileşen Hakkında Çıkarsamalar: Eşitlik (4.41) ile verilen semiparametrik modelin parametrik olmayan bileşeni sayısal olarak özetlenemediğinden, grafiksel olarak görüntülenmektedir. Söz konusu grafikler Şekil 4.1 ve 4.2'de görülmektedir. Az sayıda parametre ile özetlenemeyen parametrik olmayan bileşeni biçimsel olarak değerlendirmek diğer bir ifadeyle, parametrik bir modelle karşılaştırılması bakımından semiparametrik modelin bir anlam ifade edip etmeyeceğini belirmek amacıyla (4.37)'de belirtilen *F*-testi yapılmıştır. Bu test, semiparametrik modelin parametrik olmayan bileşeninin parametrik regresyon modelindeki gibi bir doğrusal fonksiyonla mı yoksa parametrik olmayan bir fonksiyonla mı temsil edilmesi gerektiği konusun bilgi vermesi bakımından analizin önemli bir kısmını oluşturmaktadır.

Bölün 4.6.2'de incelendiği gibi, Speckman yaklaşımından elde edilen $\hat{f}$, bilinmeyen f fonksiyonun yansız bir kestirici olması bakımdan sadece Speckman yöntemine ilişkin olarak elde edilen $f(BKA)$ parametrik olmayan fonksiyonu teste alınmış ve hipotez testi aşağıdaki şekilde ifade edilmiştir:

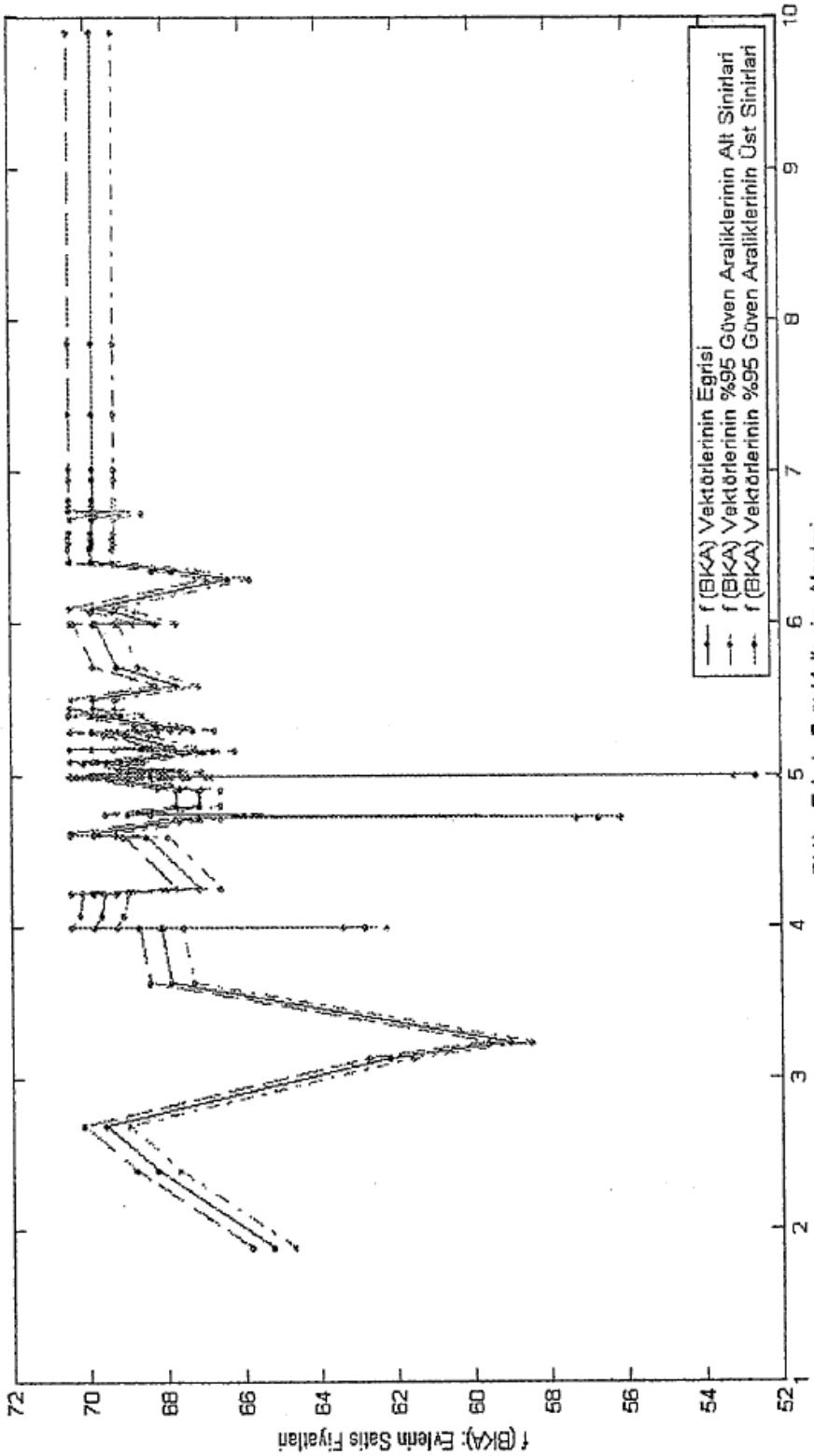
$H_0 : E(SF) = \beta_0 + \beta_1 AU + \beta_2 FD + \beta_3 GD + \beta_4 KG + \beta_5 NKA + \beta_6 BKA$ (*doğrusal fonksiyon*)
$H_1 : E(SF) = f(BKA)$ *diğer bir ifadeyle, pürüzsüz fonksiyon*
$\alpha = 0.05$ (Anlam düzeyi)

$F_{hes}(v_1 = 3.5568, v_2 = 83.4432) = 2870.2$

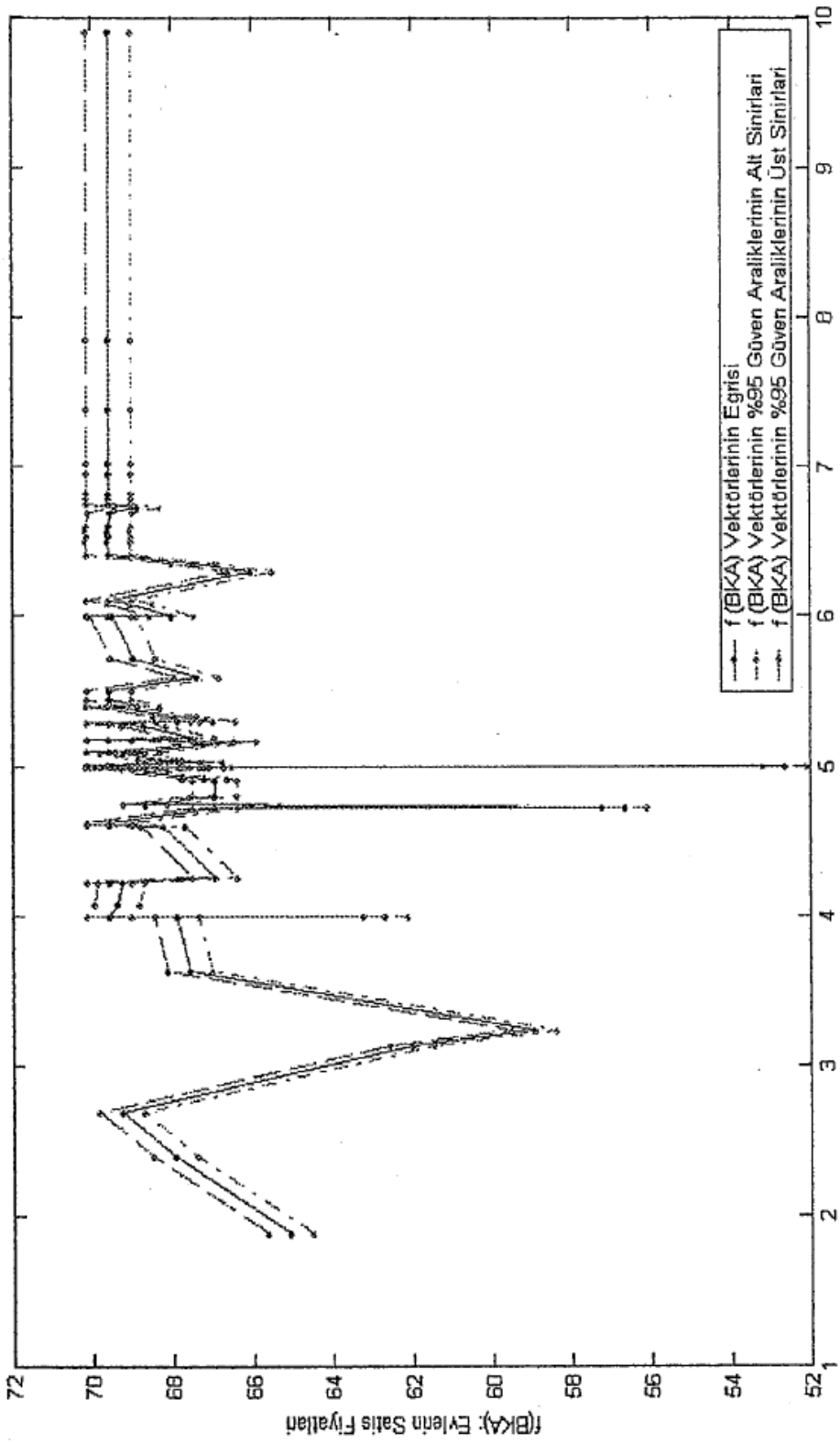
$F_{t(0.05)}(v_1 = 3.5568, v_2 = 83.4432) = 2.45$

Burada öne sürülen sıfır hipotezini test etmek amacıyla v_1, pay ve v_2, payda için hesaplan serbestlik dereceleri olmak üzere, (4.37) förmülü ile verilen *F*-test istatistiği, $F_{hes}(v_1 = 3.5568, v_2 = 83.4432) = 2870.2$ olarak hesaplanmıştır. Yukarıda görüldüğü gibi, hesaplanan *F*-test istatistiği, *F*-Tablo değerinden büyük olduğundan sıfır hipotezi reddedilir. Dolaysıyla semiparametrik modelin parametrik olmayan bileşeninin pürüzsüz bir fonksiyon olduğuna kara verilir ve söz konusu parametrik olmayan fonksiyon istatistiksel açıdan anlamlıdır. Böylece, parametrik katsayıların anlamlı olmasına ilaveten parametrik olmayan bileşenin de istatistiksel olarak anlamlı bir pürüzsüz fonksiyon olması nedeniyle semiparametrik modelin de istatistiksel açından anlamlı olduğu sonucuna varılır.

Parametrik olmayan fonksiyon ve bu fonksiyonun %95 güven aralıklarının Speckman ve kısmi splayn yöntemlerine göre elde edilen grafiksel görüntüsü aşağıdaki 4.1 ve 4.2 şekillerinde verilmiştir. Şekil 4.1 speckman yöntemine göre kestirimi yapılan evlerin satış fiyatlarının evlerin brüt kullanım alanlarına (BKA) göre değişimi ve % 95 güven aralıklarının grafiğini görüntülemektedir. Benzer olarak Şekil 4.2, kısmi splayn yöntemine göre kestirimi yapılan evlerin satış fiyatlarının evlerin BKA değişkenine göre değişimi ve 95 güven aralıklarının grafiğini vermektedir. Her iki grafikte, bölüm 4.2.1'de incelenen N -tekrarlanma matrisi yardımıyla hesaplanan $\mathbf{f}(BKA)$ vektörünün % 95 güven aralıkları dar bir aralıkta $\mathbf{f}(BKA)$ vektörünü izlediği görülmektedir. Bu durum yapılan tahminlerin kalitesinin iyi olduğunun bir göstergesidir.



Şekil 4.1: Speckman yöntemi için splayn düzeltme kestiricisi ve %95 güven sınırlarının grafiği



Şekil 4.2: Kısmi splayn yöntemi için splayn düzeltme kestiricisi ve %95 güven sınırlarının grafiği

5. SPLAYN DÜZELTME REGRESYONUNDA DÜZELTME PARAMETRESİNİN SEÇİMİ VE SİMÜLASYON ÇALIŞMASI

Üçüncü ve dördüncü bölümlerden de anlaşılacağı gibi, parametrik olmayan ve semiparametrik regresyon modellerinin kestiriminde kullanılan splayn düzeltme yönteminde hata kareler toplamına, $\lambda > 0$ düzeltme parametresi ile çarpılan ve ceza fonksiyonu olarak bilinen, parametrik olmayan kestirici değişkeninin kareli ikinci türev fonksiyonunun integrali eklenir. Ceza fonksiyonunun eklenme amacı, modelle daha iyi tahmin yapılmasını sağlamaktır. Bu durumda, parametrik olmayan ve semiparametrik regresyon modellerinin kestirimleri, önceki bölümlerde adı geçen cezalı kareler toplamını minimum yaparak elde edilirler. Bu nedenle, splayn düzeltme yönteminde (cezalı en küçük kareler yönteminde) bilinmeyen bir parametrik olmayan fonksiyonun tahmini genellikle değerleri verilere dayalı olarak belirlenen bir düzeltme parametresi gerektirir. Söz konusu $(0,\infty)$ aralığından seçilen optimum düzeltme parametresinin belirlenmesi regresyonda önemli bir problem olarak ortaya çıkmaktadır.

Bu bölümde, düzeltme parametresinin seçimine konu olan seçim kriterleri ile ilgili bazı temel kavramlar ele alınmış ve söz konusu parametrenin seçimi için yaygın olarak kullanılan seçim kriterleri incelenmiş ve bu kriterlerden hanginsin daha iyi bir düzeltme parametresi seçtiğini belirlemek amacıyla bir Monte Carlo simülasyon deneyi çalışması yapılmıştır.

5.1. Düzeltme Parametresi Seçimine İlişkin Bazı Temel Kavramlar

Bu bölümde, öncelikle model oluşturmada bazı anahtar kavramlara yer verilmiştir. Gözlemlere yeteri derecede *iyi uyum sağlayan* modelin belirlenebilmesi için gözlemlerin modele ne kadar iyi uyduğunun ölçülmesi gereklidir. Bunun için genellikle uyum iyiliği (*goodness of fit-GOF*) kullanılır. Uyum iyiliği bir modele karar verdikten sonra, tahmin için kullanılan bir kriterdir. Diğer taraftan, bir modelin zorluğunu (güçlüğünü) yada karmaşıklığının da ölçülmesi gerekir. Bir parametrik modelin karmaşıklığının genel ölçüsü, *serbestlik derecesi* (*degress of freedom -DF*) olarak adlandırılan, modeldeki parametre sayısıdır. Bir parametrik olmayan regresyon modelinin karmaşıklığının genel

ölçüsü ise, düzeltme parametresine bağlı olarak hesaplanan bir düzeltme matrisinin izidir. Diğer bir ifadeyle, parametrik olmayan regresyon modelinin karmaşıklığının ölçüsü, simgesel olarak $trace(S_\lambda)$ (yada kısaca $tr(S_\lambda)$) şeklinde ifade edilen *serbestlik derecesidir*. Bu konu izleyen bölümlerde daha ayrıntılı bir şekilde incelenmiştir.

5.1.1. Serbestlik Derecesi

Parametrik regresyonda mevcut gözlemlere dayanarak tahminim edilen parametre sayısının k olduğu varsayılsın. Bu durumda uygun modelin **H** şapka matrisinin izi, modelin serbestlik derecesine eşit olup söz konusu modelin serbestlik derecesi k'ya eşittir. Bu durumda gürültü (hata) serbestlik derecesi, $(n-k)$'ya eşit olup bu aynı zamanda (I-**H**) matrisinin izidir.

Parametrik olmayan regresyonda da parametrik duruma benzer olarak, hataları tahmin etmek için *eşdeğer serbestlik derecesi* (EDF) tanımlanır. EDF'nin herhangi bir tanımı bir tahmin metodu ve bir dağılım varsayımı ile ilgilidir. Gürültü(hata) için *eş değer serbestlik derecesi*,

$$EDF = tr\left(I - S_\lambda\right) \tag{5.1}$$

formülü ile verilir. (5.1) ifadesinde belirtilen S_λ, (3.35)'de tanımlanan matris olup splayn düzeltme regresyonu ile ilişkili düzeltme matrisidir. Splayn düzeltme regresyonunda serbestlik derecesi kavramı bir parametrik modeldeki parametre sayısının yaklaşık bir genelleştirilmesidir. Hastie ve Tibshirani [44] ve Wahba [27]'de belirtilen (3.36) splayn düzeltme uyumu, **y** vektörünün bir doğrusal dönüşümü olarak hesaplanır. Uyum değerleri $\hat{\mathbf{f}}_\lambda = \hat{\mathbf{y}}_\lambda = (I + \lambda K)^{-1}\mathbf{y}$ olarak belirtilir. Ayrıca λ düzeltme parametresine bağlı olan *etkin serbestlik derecesi*,

$$DF_\lambda = tr(S_\lambda) = \sum_{i=1}^{n} \frac{1}{1+\lambda d_i} \tag{5.2}$$

ile ifade edilir. (5.2)'de belirtilen d_i, K matrisinin *i*.öz değeridir. DF_λ, splayn uyumu için kullanılan yaklaşık parametre sayısını belirtir. Tanımdan da anlaşılacağı gibi, parametrik olmayan regresyon modellerinde DF_λ serbestlik

derecesi, doğrusal regresyondaki **H** şapka matrisine benzer bir rol oynayan, S_λ matrisinin köşegen elemanlarının yardımıyla hesaplanır.

Ekin serbestlik derecesi, aynı veri setine uygulanan modeller ve farklı tahminleri, özellikle bir doğrusal (lineer) modelle parametrik olmayan pürüzsüz modeli karşılaştırmak için kullanılan F-testinde yer alır.

5.1.2. Hata Kereler Ortalaması

Uyum iyiliği ve modelin karmaşıklığı (*complex*) bir modelin zıt iki görünümüdür. Amacımız bu iki zıt görünüm arasında bir denge kuran "*en iyi modeli"* bulmaktır. Modele daha fazla parametre ekleyerek her zaman veriler için iyi uyumlar elde edilebilir. Fakat az sayıda parametreye sahip modeller yorumlanabilir ve basit olmaları nedeniyle daha çekicidirler. Ayrıca, tahmin değişebilirliğine daha az bağlı olduklarından daha doğru kestirim üretebilirler [48]. Diğer taraftan, modeldeki parametre sayısı (zorluk) arttıkça tahmin hataları azalır. Böylelikle, en iyi kelimesini anlamlı kılmak için, modelin performansını ölçen bir hedef kriter gereklidir. Uyum iyiliğinin hedef kriter olmayacağı açıktır, çünkü onun iyileştirilmesi modelin daha karmaşık bir hal almasına yol açabilir. Şu ana kadar evrensel olarak kabul görmüş bir ölçü geliştirilememiştir. Ancak uygulamada geniş ölçüde kullanılan bazı kriterler mevcuttur. Bu bağlamda, regresyon modelleri için yaygın olarak kullanılan performans kriterlerden biri de *hata kareler ortalamasıdır*.

Nonparamertrik regresyonda amaç, bilinmeyen gerçek $f(x)$ fonksiyonunu tahmin etmektir. Söz konusu fonksiyon $f \in C^2[a, b]$ uzayındaki tüm f fonksiyonları arasında (3.32) $S(f)$ "*cezalı en küçük kareler kriterini"* minimum yapmaktadır. (3.32)'de verilen cezalı kriterdeki λ düzeltme parametresi eğrinin $\int_a^b \{f''(x)\}^2 dx$ ile ölçümlenen pürüzlülüğü ve $\sum_{i=1}^{n} \{y_i - f(x_i)\}^2$ ile ölçümlenen uyumun verilere yakınlığını dengeler. Bu nedenle, λ parametresinin iyi bir değerini seçmek gerekir. Reinsch [30], σ^2 (varyans) biliniyorsa

$$\frac{1}{n}\sum_{i=1}^{n}\left(f_\lambda(x_i) - y_i\right)^2 = \sigma^2 \qquad (5.3)$$

denklemi ile uyumdan yoksunluğun (uyumun verilere yakınlığı) belirlenebilmesi için λ parametresinin seçilmesini önermektedir. "***Optimum*** **λ"** tüm veri noktalarına göre gerçek hata kareler ortalamasını minimum yapan λ olarak tanımlanır. Bu *gerçek hata kareler ortalaması*

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(f_\lambda(x_i) - f(x_i)\right)^2 = \frac{1}{n}\left\|(S_\lambda)\mathbf{y} - \mathbf{f}\right\|^2 \qquad (5.4)$$

formülüyle verilen $MSE(\lambda)$ olarak tanımlanır. Wahba [49], λ parametresinin (5.3) eşitliğinin sol tarafı ile tanımlanan uyumdan yoksunluğu, gerçekte σ^2 (varyansı) değerini mümkün olduğu kadar küçültmesi için seçilmesi gerektiğini göstermiştir. Ancak, genel olarak varyans bilinmeyebilir. Diğer taraftan f fonksiyonu da bilinmemektedir. Bu durumda (5.3) denklemi için uygun örneklem hacmini (n) belirlemek çok zor ve bu anlamda sonuç pratik değildir [50].

Amacımız (5.4) *MSE* değerini minimum yapan λ düzeltme parametresini tahmin etmektir. (5.4)'deki $MSE(\lambda)$ ifadesi bilinmeyen f fonksiyonuna bağlı olduğundan hata kareler ortalamasının tahmini doğrudan yapılamaz. Bu nedenle f'in gerçek değeri yerine tahmini değeri kullanılması gerekir. f fonksiyonunun x_i düğüm noktalarındaki gerçek değerlerinin vektörü, $\mathbf{f} = \left(f(x_1),..., f(x_n)\right)$ ve bu fonksiyonun bir λ parametresine bağlı tahmin değerlerinin vektörü de $\hat{\mathbf{f}}_\lambda = \left(\hat{f}_\lambda(x_1),..., f_\lambda(x_2)\right)$ olsun. Buna göre, *tahmini hata kareler ortalaması*,

$$EMSE(\lambda) = E\left(\frac{1}{n}\left\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\right\|^2\right) = E\left(\frac{1}{n}\left\|(S_\lambda\mathbf{y}) - \mathbf{f}\right\|^2\right) \qquad (5.5)$$

ile belirtilen $EMSE(\lambda)$ fonksiyonu ile elde edilir. Mümkün olduğu kadar gerçek f fonksiyonuna yakın olacak $\hat{f}_\lambda$ fonksiyonunu tahmin etmek gerekir. (5.5)'deki *EMSE*, tahminler ile gerçek değerler vektörleri arasındaki L_2-öklid (*Euclidean*) uzaklığının beklenen değeridir. (5.5) ifadesinde bir $E\hat{\mathbf{f}}_\lambda$ eklenip-çıkartılması ile *EMSE* aşağıda gösterildiği gibi iki bileşene ayrıştırılabilir:

$$
\begin{aligned}
EMSE(\lambda) &= \frac{1}{n} E \left\| (E\hat{\mathbf{f}}_\lambda - \mathbf{f}) + (\hat{\mathbf{f}}_\lambda - E\hat{\mathbf{f}}_\lambda) \right\|^2 \\
&= \frac{1}{n} E \left\| E\hat{\mathbf{f}}_\lambda - \mathbf{f} \right\|^2 + \frac{2}{n} E (E\hat{\mathbf{f}}_\lambda - \mathbf{f})^T (\hat{\mathbf{f}}_\lambda - E\hat{\mathbf{f}}_\lambda) + \frac{1}{n} E \left\| (\hat{\mathbf{f}}_\lambda - E\hat{\mathbf{f}}_\lambda) \right\|^2 \\
&= \frac{1}{n} \left\| E\hat{\mathbf{f}}_\lambda - \mathbf{f} \right\|^2 + \frac{1}{n} E \left\| (\hat{\mathbf{f}}_\lambda - E\hat{\mathbf{f}}_\lambda) \right\|^2 \qquad (5.6) \\
&= \frac{1}{n} \left\| (I - S_\lambda)\mathbf{f} \right\|^2 + \frac{\sigma^2}{n} trS_\lambda^2 \\
&= (Yan(Bias))^2 + (Varyans(Variance)).
\end{aligned}
$$

Yan^2, modelin gerçek f fonksiyonuna ne kadar iyi yaklaştığını ölçerken, *varyans*, fonksiyonun ne kadar iyi tahmin edildiğini ölçer. Genellikle çok büyük modeller uzayı daha küçük Yan^2 fakat daha büyük *varyans*'a neden olur. Böylece *EMSE*, Yan^2 ve *varyans* arasında bir denge gösterir [51].

5.2. Varyans Tahmini

Hataların varyansı, σ^2 ölçek faktörünün tahmini splayn uyumunun diagnostikleri için gerekli ve şu ana kadar görüldüğü gibi, gereken düzeltme derecesinin belirlenmesi için önemlidir. Uygulamada hataların (artıkların) varyansının bilindiği çok nadirdir. Yukarıda belirtilen nedenlerden dolayı onun tahmini ile ilgilenebilir. Bu bölümde, literatürde bulunan σ^2 tahmini için iki yaklaşım ele alınmıştır.

5.2.1. Yerel Fark Alma Yaklaşımı

Birinci muhtemel yaklaşım f trend fonksiyonunu elimine edecek bir şekilde gözlemleri dönüştürmektedir. Örneğin Rice [52], verilerin birinci farklarını kullanarak varyansı,

$$\hat{\sigma}_R^2 = \frac{1}{2}(n-1)^{-1} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \qquad (5.7)$$

formülü ile tahmin etmeyi önermiştir. Bu kestiricinin arkasındaki temel düşünce, pürüzsüz bir f eğrisi için $\{f(x_i) - f(x_{i-1})\}^2$ kareli ortalamasına sahip olan $(y_i - y_{i-1})$ birinci farklılığı $2\sigma^2$ varyansına göre küçük olmasıdır. Bu f fonksiyonunun eğiminin büyük olmaması koşuluyla uygun bir yaklaşım olacaktır.

Trendi elimine etmek için fark almanın kullanımı genellikle zaman serileri analizde kullanılmaktadır.

Rice [52], ayrıca verilerin ağırlıklı ikinci farkları üzerine dayalı ikinci bir kestirici önermiştir. Buna göre, f fonksiyonuna bir doğrusal fonksiyonun eklenmesi durumu değiştirmeyecektir. Kestirici, bir polimal değerlerle lokal olarak tahmin edilen f fonksiyonunun değerleri ile hata kareler toplamına dayalı olup, bir birini takip eden üçlü noktalara en küçük kareler denklemini uygulayarak elde edilir. Bu kestirici, bir lokal tahmin olarak bir doğru denklemini kullanan, Gasser, Sroka ve Jennen [43] tarafından önerilmiştir:

$$\hat{\sigma}^2_{GSJ} = (n-2)^{-1}\sum_{i=2}^{n} c_i^2 \hat{\varepsilon}_i^2 \qquad (5.8)$$

Burada,

$$\hat{\varepsilon}_i^2 = \left[y_{i-1} + \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}(x_i - x_{i-1}) - y_i \right]^2$$

ve

$$c_i^2 = \left[\left(\frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} \right) + \left(\frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} \right) + 1 \right]^{-1}.$$

Gasser, Sroka ve Jenner [43] tarafından önerilen bu varyans kestiricisi, Schimek [40] ve Eubank ve ark. [41] tarafından yapılan çalışmalarda da verildiği gibi aşağıdaki şekilde de yazılabilir:

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{y}}{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)} \qquad (5.9)$$

Buradaki **A,** (4.30)'da verilmiş olan girişlerin tümü sıfır fakat i.girişi a_ic_i, (i+1).girişi $-c_i$ ve (i+2).girişi b_ic_i olan $(n-2)\times n$ boyutlu matristir.

Bu kestirici kernel ve splayn düzeltme regresyonunda başarılı bir şekilde uygulanmıştır. Bu teknik, kestirici üzerinde öncelikle regresyon fonksiyonunun etkisini önemli ölçüde kaldırır ve ayrıca $\beta = 0$ olduğu sürece model varyansı σ^2 için bir uygun-$\sqrt{n}$ kestiricisi meydana getirir.

5.2.2. Hata Kareler Yaklaşımı

Bazı temel yaklaşımlar sınıfı eğri uydurma hakkında artık kareler toplamına göre varyansın (σ^2) bir tahminini esas alır. Parametrik regresyon uygulamasında örneklem için hata kareler toplamı,

$$\hat{\sigma}^2(MSE) = \frac{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}{DF} = \frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-k} \tag{5.10}$$

formülündeki gibi serbestlik derecesine bölünerek, varyansın yansız bir kestirici elde edilir. Varyansın bu yansız kestiricisi aynı zamanda MSE değerine eşittir. Doğrusal regresyona benzer olarak, splayn düzeltme yönteminde de hata kareler toplamı da (5.1)'de belirtilen eş değer serbestlik derecesine bölünür. λ düzeltme parametresini ile bu yaklaşım aşağıdaki gibi bir varyans kestiricisi üretir:

$$\hat{\sigma}^2 = \hat{\sigma}_\lambda^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{tr\left(I - S_\lambda\right)} = \frac{\left\|(S_\lambda - I)\boldsymbol{y}\right\|^2}{tr\left(I - S_\lambda\right)} \tag{5.11}$$

Burada $\hat{f}_\lambda$, AIC_c, CV veya GCV kriteriyle seçilen λ düzeltme parametresi için hesaplanan splayn düzeltme tahmini ve S_λ, aynı λ ile hesaplanan düzeltme matristir. Gerçek f regresyon eğrisinin bir doğru denklemi olduğu özel durumlarda tüm λ'lar için $\hat{\sigma}_\lambda^2$ kestiricisi σ^2'nin yansız bir kestiricisidir (bakınız örneğin, Buckley, Eagleson ve Silverman [53]).

Bazı yazarlar tarafından önerilen ve λ parametresine bağlı uyumdan yoksunluk yöntemi olarak bilinen bir diğer varyans formülü,

$$\sigma^2 = \frac{\left\|\left(I - S_\lambda\right)\boldsymbol{y}\right\|^2}{n} \tag{5.12}$$

ile verilir [27].

Lokal fark almaya dayalı kestiricilerin daha yüksek hata kareler ortalaması için sezgisel nedenlerde biri, yüksek frekans etkileri üzerinde yer alan öneme göre yanları elimine etmesidir. Lokal fark almaya dayalı kestiriciler düzeltme parametresi seçimi gerektirmeyen bir avantaja sahiptirler (5.11) ile belirtilenden

çok daha küçük yana sahiptirler ve büyük örneklerde yok denecek kadar azdır. Bununla birlikte verilerdeki sıra korelasyon (5.11)'i hemen hemen hiç etkilenmezken diğer lokal farka dayalı (5.7), (5.9) ve (5.10) kestiricilerini önemli ölçüde etkilerler [11].

5.3. Düzeltme Parametresi Seçim Kriterleri: Klasik ve Risk Tahmin Metotları

Düzeltme parametresinin seçimi problemi eğri tahminlerinde kesinlikle yer almaktadır. Örneğin, polinom regresyon ile eğri uydurmada, uyum polinomunun derecesinin seçimi esas olarak düzeltme parametresinin seçimine eşdeğerdir. Splayn düzeltme metodunda düzeltme parametresi kesin olarak yer alır. Düzeltme parametresinin seçim problemi için iki farklı yaklaşım vardır.

Birinci yaklaşım, düzeltme parametresinin değerinin araştırmacılar tarafından bireysel olarak belirlenmesidir. Böyle sübjektif bir yaklaşım gerçekte yararlı olabilir. Fakat söz konusu parametrenin seçimi kişiden kişiye farklılık göstereceğinden bilimsel açıdan tutarlılık sağlamaz. Diğer bir yaklaşım, verilere dayalı olarak elde edilen düzeltme parametresinin seçimidir. Söz konusu bu yaklaşıma göre düzeltme parametresinin değeri, seçim kriteri olarak adlandırılan yöntemler yardımıyla elde edilir.

Splayn düzeltme kestiricisi değişen eğimli uyumlara olanak sağlayan problemleri çözer, fakat yeni bir problem de yaratır. Diğer bir ifadeyle, *verilen bir veri seti için* λ *düzeltme parametresinin yaklaşık değerinin nasıl belirleneceği problemini ortaya çıkarır*. Aynı λ değerinin her veri setiyle aynı derecede iyi çalışması beklenemez. Bu nedenle, en iyi seçilen düzeltme parametresi (5.5) ile verilen hata kareler ortalamasını minimum yapandır. Bu bağlamda, araştırmada adı geçen düzeltme parametresinin seçimi ile ilgili olarak en yaygın kullanılan yöntemlerden alt tanesi karşılaştırmalı bir biçimde ele alınmıştır. Adı geçen parametrenin seçimi için kullanılan kriterlerden dördü klasik metot olarak bilinirken, kalan ikisi risk tahmin metodu olarak bilinir. Herhangi bir seçim kriterini minimum yapan λ, uygun düzeltme parametresi olarak seçilir.

Düzeltme parametresinin seçimi için kullanılan klasik metotlar: *Çapraz-Geçerlilik* (*Cross-validation*-CV), *Genelleştirilmiş Çapraz-Geçerlilik*" (*Generalized cross-validation*-GCV), *Geliştirilmiş Akaike Bilgi Kriteri* (*İmproved Akaike information criterion* -AIC_c) *ve Mallow'un Cp Kriteri* (*Mallows' Cp Criterion*) olurken, kalan iki risk tahmin metotu: *Klasik pilotları kullanan risk tahmini (Risk estimation using classical pilots-RECP)* ve *lokal risk tahmin (Local risk estimation-LRS)* kriterleridir. Söz konusu düzeltme parametresinin seçiminde kullanılan klasik ve risk tahmin metotları, yukarda belirtildiği sırada izleyen alt bölümlerde ele alınmıştır.

5.3.1. Çapraz Geçerlilik

Verilere uygun λ düzeltme parametresini seçen çapraz geçerlilik yöntemi Wahba ve Wold [54] tarafından önerilmiştir. Çapraz geçerliliğin esas düşüncesi, veri noktalarından herhangi birini atmak ve kayıp veri noktaları altında geri kalan veriler tarafından en iyi kestirilen λ parametresinin değerini seçmektir. Diğer bir deyişle, çapraz-geçerlilik $\{x_i, y_i\}_{i=1}^{n}$ gözlem noktalarından herhangi birini atarak, kalan $(n-1)$ veri noktasına dayalı olarak, x_i'de bir pürüzsüz fonksiyon için kareli artıkları tahmin etmeyi ve kareli artıklarının toplamını minimum yapan düzeltme parametresini seçmeye çalışır.

$\mathbf{y}^{(-i)}$, orijinal $\mathbf{y}$ bağımlı değişken vektöründen y_i gözlemini attıktan sonra kalan n-1 gözlemden oluşan vektörü olsun. Ayrıca, söz konusu $\mathbf{y}^{(-i)}$ vektörüne uygun tahmin fonksiyonu $\hat{f}_{\lambda}^{(-i)}(x)$ olsun. Diğer bir ifadeyle, verilen λ düzeltme parametresi için $\hat{f}_{\lambda}^{(-i)}(x)$ tahmin fonksiyonu,

$$\sum_{j \neq i} (y_j - f(x_j))^2 + \lambda \int f''^2$$

ifadesini minimum yapan bir eğridir. Bu durumda, kestirim hatasının *çapraz geçerlilik (CV) tahmini*,

$$CV(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ y_i - \hat{f}_{\lambda}^{(-i)}(x_i) \right\}^2 \qquad (5.13)$$

fonksiyonunun değeri ile ölçülebilir. λ parametresinin çapraz geçerlilik tahmini (5.13) kriterini minimum yapan değerdir. Çapraz geçerlilik metodu, splayn düzeltme ortamında Wahba ve Wold [54]) ve doğrusal regresyon ortamında (ayrıca PRESS olarak adlandırılan) ise, Allen [55] tarafından tanıtılmıştır.

Çapraz geçerliliğin temel fikri, CV(λ) fonksiyonunu minimum yapan λ parametresini seçmektir. İlk bakışta, (5.13) ifadesinden görülüyor ki, λ düzeltme parametresine uygun CV(λ) fonksiyonunun değerini bulmak için n tane $\hat{f}^{(-i)}$ eğrisi bulmak ve bu karşı gelen n farklı düzeltme problemini çözmek gerekir. (5.13) problemini basitleştirme bakımından ilk adım (3.35) eşitliğinde belirtilen ve y_i gözlem değerleri vektörünü $\hat{y}_i$ veya $\hat{f}(x_i)$ kestirim değerlerine görüntüleyen S_λ şapka matrisini kullanmaktır. Çapraz-geçerlilik değerinin en ekonomik biçimde hesaplanması için ilk anahtar sonuç aşağıdaki teoremle verilir.

Teorem 5.1: Düzeltme parametresi λ ile tüm $\{x_i, y_i\}_{i=1}^{n}$ veri setinden hesaplanan splayn düzeltici $\hat{f}$ fonksiyonunun yer aldığı *çapraz geçerlilik fonksiyonu*,

$$\mathrm{CV}(\lambda)=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left\{y_i-\hat{f}_\lambda^{(-i)}(x_i)\right\}^2 \equiv \mathrm{CV}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i-\hat{f}_\lambda(x_i)}{1-(S_\lambda)_{ii}}\right\}^2 \qquad (5.14)$$

ifadesini sağlar ve λ parametresinin CV tahmini CV(λ) değerini minimum yapan değer olarak belirlenir [27].

İspat: Teorem (5.1)'in ispatı için aşağıda verilen bir lemma'dan yararlanılır. **Lemma 5.1 (*Leaving-Out-One Lemma*):** Verilen λ ve i sabitleri için, $\mathbf{f}^{(-i)}$ *ve* $\mathbf{y}^*$ vektörleri aşağıdaki gibi tanımlansın:

$$\mathbf{f}^{(-i)}=\left(f_1^{(-i)},\ldots,f_n^{(-i)}\right)=\left(\hat{f}_\lambda^{(-i)}(x_1),\ldots,\hat{f}_\lambda^{(-i)}(x_n)\right) \quad \text{ve} \quad \mathbf{y^*}=\begin{pmatrix} y_j^*=y_j,\ j\neq i \quad ve \\ y_i^*=\hat{f}^{(-i)}(x_i) \end{pmatrix}$$

vektörleri tanımlansın. O zaman (3.35)'de verilen S_λ matrisi yardımıyla,

$$\mathbf{f}^{(-i)}=S_\lambda\mathbf{y}^*. \qquad (5.15)$$

eşitliği sağlanır.

İspat: Herhangi bir pürüzsüz f eğisi için $y_i^* = \hat{f}^{(-i)}(x_i)$ ve $\hat{f}^{(-i)}$ tanımına göre, ardışık olarak aşağıdaki ifadeler yazılabilir:

$$\sum_{j=1}^{n}\left\{y_j^* - f(x_j)\right\}^2 + \lambda\int f''^2 \geq \sum_{j\neq i}^{n}\left\{y_j^* - f(x_j)\right\}^2 + \lambda\int f''^2 \geq \sum_{j\neq 1}^{n}\left\{y_j^* - \hat{f}^{(-i)}(x_j)\right\}^2 + \lambda\int f^{(-i)''^2}$$

$$= \sum_{j=1}^{n}\left\{y_j^* - \hat{f}^{(-i)}(x_j)\right\}^2 + \lambda\int f^{(-i)''^2}$$

Burada $\hat{f}_\lambda^{(-i)}(x)$ fonksiyonunun $\sum_{j=1}^{n}\left\{y_j^* - f(x_j)\right\}^2 + \lambda\int f''^2$ ifadesini minimum yaptığı görülür ve buna göre de $\mathbf{f}^{(-i)} = S_\lambda \mathbf{y}^*$ olmaktadır. Bu durum teoremin ispatını tamamlar. Lemma (5.1)2 esasen, S_λ yerine S yazılarak, $y_i - \hat{f}^{(-i)}(x_i)$ silinmiş artıkları için aşağıdaki gibi bir ifade elde edilir:

$$\hat{f}^{(-i)}(x_i) - y_i = \sum_{i=1}^{n} S_{ij} y_j^* - y_i = \sum_{j\neq i} S_{ij} y_j + S_{ii}\hat{f}^{(-i)}(x_i) - y_i$$

$$= \sum_{j=1}^{n} S_{ij} y_j + S_{ii}\left\{\hat{f}^{(-i)}(x_i) - y_i\right\} = \hat{f}(x_i) - y_i + S_{ii}\left\{\hat{f}^{(-i)}(x_i) - y_i\right\} \qquad (5.16)$$

(5.16) ifadesinden

$$y_i - \hat{f}^{(-i)}(x_i) = \frac{y_i - \hat{f}(x_i)}{1 - S_{ii}(\lambda)} \qquad (5.17)$$

olduğu görülür. (5.17) eşitliğinin karesi ve toplamı alınarak, her iki taraf $\frac{1}{n}$ ile çarpılarak,

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}}\right\}^2 \qquad (5.18)$$

elde edilir. Böylece teorem 5.1 ispatlanmış olur.

5.3.2. Genelleştirilmiş Çapraz Geçerlilik

Genelleştirilmiş çapraz geçerlilik (GCV), düzeltme parametresinin seçimi için bir popüler yöntem olan çapraz geçerliliğin değişik bir şeklidir. (5.17) denklemine göre, sıradan artıkların $\left(1-(S_\lambda)_{ii}\right)$ faktörlerine bölünerek elde edilen

çapraz geçerlilik değerinin hesaplanması için silinmiş artıklar gereklidir. GCV kriterinin ana düşüncesi, $\left(1-(S_\lambda)_{ii}\right)$ faktörü ile $(1-n^{-1}trS_\lambda)$ ortalama değerlerini yer değiştirmektedir. Bu durumda, $(1-n^{-1}trS_\lambda)$ faktörünün karesi ile düzeltilmiş artıkların kareler toplamı alınarak, sırdan çapraz geçerliliğe benzer olarak, *GCV* değeri aşağıdaki şekilde elde edilir:

$$\mathrm{GCV}(\lambda)=\frac{1}{n}\frac{\sum_{i=1}^{n}\left\{y_i-\hat{f}_\lambda(x_1)\right\}^2}{\left\{1-n^{-1}tr(S_\lambda)\right\}^2}=\frac{n^{-1}\left\|\left(I-S_\lambda\right)\mathbf{y}\right\|^2}{\left[n^{-1}tr\left(I-S_\lambda\right)\right]^2} \tag{5.19}$$

Sıradan çapraz geçerlilikteki gibi, λ parametresinin GCV tahmini $\mathrm{GCV}(\lambda)$ fonksiyonunu minimum yapan değer olarak belirlenir [27].

5.3.2a. Düzeltme Parametresinin GCV Tahminin Özellikleri

GCV, bir tahmini hata kareler ortalaması kriteridir. λ düzeltme parametresinin GCV tahmini, (3.4)'de verilen $MSE(\lambda)$ gerçek hata kareler ortalamasını minimum yapan bir tahmindir. Genelleştirilmiş çapraz geçerlilik fonksiyonu birkaç durumda hata kareler ortalamasına benzerdir. Bu durumla ilgili olarak Craven and Wahba [56] aşağıdaki asimtotik sonucu vermektedir:

$$MSE(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left(f_\lambda(x_i)-f(x_i)\right)$$

Burada belirtilen f fonksiyonu (3.32)'deki tahmin edilen gerçek fonksiyon ise, hem $MSE(\lambda)$ hemde $\mathrm{GCV}(\lambda)$ hata terimlerinin (ε_i) rassal bir fonksiyonu olarak ele alınır. λ^* düzeltme parametresi (5.5)'de belirtilen $MSE(\lambda)$'nın beklenen değeri $EMSE(\lambda)$'yi minimum yapan ve λ^+ düzeltme parametresi $\mathrm{GCV}(\lambda)$'nın beklenen değeri $\mathrm{EGCV}(\lambda)$'yi minimum yapan değer ise,

$$\lim_{n\to\infty}\frac{EMSE(\lambda^+)}{EMSE(\lambda^*)}\downarrow 1$$

olmakta veya λ parametresi ile tahmin edilen hata kareler ortalaması, herhangi bir λ parametresi ile elde edilen minimum hata kareler ortalamasına beklen değerde eşit olmaktadır.

5.3.3. Geliştirilmiş Akaike Bigi Kriteri

Akaike Bilgi Kriteri (AIC) ilk olarak Kullback –Leibler bilgisinin beklenen değerinin yaklaşık yansız bir kestiricisi olarak parametrik modeller için geliştirilmiştir. Hurvich ve ark. [57], doğrusal regresyon ve zaman serileri için küçük örneklerde AIC kriterinin yanının (sapmasının) oldukça büyük olabildiğini göstermişler ve AIC kriterinden çok daha düşük sapma içeren, geliştirilmiş AIC_c kriterini önermişlerdir. .Hurvich ve ark. [57] tarafından elde edilen geliştirilmiş AIC_c kriteri, parametrik olmayan düzelticiler için düzeltme parametresinin seçimi için kullanılmıştır. Söz konu geliştirilen *AIC_C kriteri*,

$$\mathrm{AIC_c}(\lambda) = \log(\hat{\sigma}^2) + \frac{1+tr(S_\lambda)/n}{1-\{tr(S_\lambda)+2\}/n} = \log(\hat{\sigma}^2) + 1 + \frac{2\{tr(S_\lambda)+1\}}{n-tr(S_\lambda)-2} \qquad (5.20)$$

formülüyle verilir. (5.20) formülünde de görüldüğü gibi bu kriter S_λ düzeltme matrisinin yalnızca izini kullandığından, düzeltme parametresinin seçimi için uygulaması kolay olan bir kriteridir. (5.12)'de belirtilen varyans fonksiyonu (5.20)'de yerine yazılarak, basit cebirsel işlemlerden sonra, Lee [58] ve [59] tarafından da belirtildiği gibi, (5.20)'de tanımlanan *AIC_c bilgi kriteri*,

$$\mathrm{AIC_c}(\lambda) = \log\frac{\sum\left\{y_i - \hat{f}_\lambda(x_i)\right\}^2}{n} + 1 + \frac{2\{tr(S_\lambda)+1\}}{n-tr(S_\lambda)-2} = \log\frac{\left\|(S_\lambda - I)\mathbf{y}\right\|^2}{n} + 1 + \frac{2\{tr(S_\lambda)+1\}}{n-tr(S_\lambda)-2} \qquad (5.21)$$

şeklini alır. Diğer yöntemlerde olduğu gibi, $\mathrm{AIC}_C(\lambda)$ kriterini minimum yapan λ değeri düzeltme parametresi olarak seçilir.

5.3.4. Mallows'un Cp Kriteri

Mallows'un kriterinin beklenen değerinin izleyen bölümde verilecek olan risk tahmin metoduna eşit olduğu görülmektedir. Başka bir deyişle, $E\left\{C_{\mathrm{p}}(\lambda)\right\} = R\left(\mathbf{f}, \hat{\mathbf{f}}_\lambda\right)$ olmaktadır. C_{p} yöntemi, splayn düzeltme literatüründe yansız risk yöntemi (*unbiased risk method -UBR*) olarak bilinir. σ^2 bilindiğinde λ için bir yansız risk tahmini mevcuttur. Regresyonda böyle bir tahmin Mallows [60] tarafından önerilmiş ve Craven ve Wahba [56] tarafından splayn düzeltmeye

uygulanmıştır. Hata kareler toplamının yansız bir tahminini, minimum yapan λ parametresini seçmeyi amaçlayan C_{p} *kriteri,*

$$C_{\mathrm{p}}(\lambda)=\frac{1}{n}\left\{\left\|(S_\lambda - I)\mathbf{y}\right\|^2 + 2\sigma^2 tr(S_\lambda)+\sigma^2\right\}=\frac{1}{n}\left\{\left\| \mathbf{y}-\hat{\mathbf{f}}_\lambda\right\|^2 + 2\sigma^2 tr(S_\lambda)+\sigma^2\right\} \quad (5.22)$$

ile verilir. C_{p} istatistiği, aynı zamanda $E\left\|\hat{\mathbf{f}}_\lambda - \mathbf{f}\right\|^2$ kareli kestirim hatasının da yansız bir tahminidir [61]. Craven ve Wahba [56] tarafından yapılan deneysel sonuçlar göstermiştir ki, (5.22)'de aynı σ^2 kullanıldığında, büyük örneklemlerde GCV tahmini ve yansız risk kriteri aynı sonucu vermektedir. Uygulamada çok iyi çalışan bu yöntemi yapmak için varyansın bir tahmininin gerekli olduğu açıkça görülmektedir. Uygulamada σ^2 bilinmiyorsa, varyans kestiricisi genellikle (5.11) ile elde edilir. Böylece, $C_{\mathrm{p}}(\lambda)$ kriterini minimum yapan λ seçilmek istendiğinde, (5.22)'de yer alan σ^2'yi uygun bir $\hat{\sigma}^2$ kestiricisiyle yer değiştirmek gerekir.

5.3.5. Klasik Pilotları Kullanan Risk Tahmini (RCP)

Risk fonksiyonu, tahmin ile gerçek regresyon fonksiyonu arasındaki uzaklığı ölçer. Gerçekte iyi bir tahmin düşük riskli olmalıdır. (5.5) formülü ile belirtilen tahmini hata kareler ortalamasına benzer bir biçimde, doğrudan bir hesaplama $R(\mathbf{f},\hat{\mathbf{f}}_\lambda)$ için *yan-varyans ayrışımına götürür*:

$$R(\mathbf{f},\hat{\mathbf{f}}_\lambda)=\frac{1}{\mathrm{n}}E\left\|\mathbf{f}-\hat{\mathbf{f}}_\lambda\right\|^2=\frac{1}{n}\left\{\left\|(S_\lambda - I)\mathbf{f}\right\|^2+\sigma^2 tr(S_\lambda S_\lambda^T)\right\} \quad (5.23)$$

Ancak (5.23) ile verilen formülde $\mathbf{f}$ bilinmediğinden, $R(\mathbf{f},\hat{\mathbf{f}}_\lambda)$ doğrudan elde edilemez. Bunun yerine riskin tahmin edilmesi gerekir. Bu nedenle, (5.23)'deki σ^2 ve $\mathbf{f}$'in uygun pilot tahminleriyle, $R(\mathbf{f},\hat{\mathbf{f}}_\lambda)$ riskini tahmin etmek ve meydana gelen risk kestiricisini minimum yapan λ parametresini seçmek gerekir. Pilot tahminlerin seçimi için, $\mathbf{f}$ ve σ^2'nin pilot tahminleri olan $\hat{\sigma}^2_{\lambda_p}$ ve $\hat{f}_{\lambda_p}$'yi hesaplamada pilot λ_{p} değeri kullanılır [62,63]. Söz konusu λ_{p} pilotunun seçimi için, klasik seçim metotlarından herhangi bir seçim kriteri (örneğin, CV kriteri)

kullanılabilir. Sonuç olarak, $R(\mathbf{f},\hat{\mathbf{f}}_\lambda)$ değeri için iyi bir kestirici diğer bir deyişle, *RECP'nin kestiricisi*,

$$R(\hat{\mathbf{f}}_{\lambda_p},\hat{\mathbf{f}}_\lambda) = \frac{1}{n}E\left\|\hat{\mathbf{f}}_{\lambda_p}-\hat{\mathbf{f}}_\lambda\right\|^2 = \frac{1}{n}\left\{\left\|(S_\lambda - I)\hat{\mathbf{f}}_{\lambda_p}\right\|^2 + \hat{\sigma}^2_{\lambda_p} tr(S_\lambda S_\lambda^T)\right\} \qquad (5.24)$$

formülü ile verilebilir. Benzer olarak diğer yöntemlerde olduğu gibi, $R(\hat{\mathbf{f}}_{\lambda_p},\hat{\mathbf{f}}_\lambda)$ ifadesini minimum yapan λ, düzeltme parametresi olarak seçilir.

5.3.6. Lokal Risk Tahmini

Önceki bölümlerde de bahsedildiği gibi her bir x_i noktasında hesaplanan farklı splayn düzeltme tahminleri, $\hat{f}_{\lambda_1}(x_i),...,\hat{f}_{\lambda_m}(x_i)$ olsun. Lee [59] tarafından önerilen metot x_i düğüm noktalarında hesaplanan aşağıdaki bölgesel riski minimum yapan $\hat{f}_\lambda(x_i)$ değerini seçmeyi amaçlar:

$$R_\lambda(x_i) = E\left\{f(x_i) - \hat{f}_\lambda(x_i)\right\}^2.$$

Burada ifade edilen, $R_\lambda(x_i)$ bilinmeyen bir miktar olduğundan pratikte minimum yapılamaz. Bu problemin üstesinden gelmek için, $R_\lambda(x_i)$ için bir kestirici hesaplanır ve meydana gelen kestiriciyi minimum yapan $\hat{f}_\lambda(\mathrm{x}_i)$ değeri seçilir. Bu işlem tüm x_i için tekrar edillerek, f için nihai bir tahmin elde edilir.

$S_\lambda\mathbf{f}$ vektörünün i. inci köşegen elemanı $(S_\lambda\mathbf{f})(x_i)$ ve $S_\lambda S_\lambda^T$ kare matrisinin i. inci köşegen elemanı $s_\lambda(x_i)$ olarak gösterilmek üzere, $R_\lambda(x_i)$ için doğrudan bir hesaplama aşağıdaki yan-varyans ayrışımını verir:

$$R_\lambda(x_i) = \left\{(S_\lambda\mathbf{f})(x_i) - f(x_i)\right\}^2 + \sigma^2 s_\lambda(x_i). \qquad (5.25)$$

(5.25)'deki bilinmeyen $\mathbf{f}$ ve σ^2 pilot tahminleriyle yer değiştirilerek $R_\lambda(x_i)$ tahmin edilir. Bu durumda, her bir i için, $\hat{R}_\lambda(\mathrm{x_i})$ 'yi minimum yapan $\hat{f}_\lambda(\mathrm{x}_i)$ ile, $f(x_i)$ tahmin edilebilir. $\mathbf{f}$ fonksiyonunun pilot tahmini için, splayn düzeltme tahmini $\hat{f}_{\lambda_p}$ ve σ^2'nin pilot tahmini için, (5.11)'de belirtilen $\hat{\sigma}_{\lambda_p}$ kullanılır. Bu

pilot tahminlerin hesaplanmasında kullanılan λ_p, klasik metotlardan herhangi birisi ile seçilebilir. Söz konusu bu pilot tahminler ile (5.25) ifadesindeki bilinmeyen miktarları yer değiştirilerek, $R_\lambda(x_i)$ *lokal risk kriterinin kestiricisi*,

$$\hat{R}_\lambda(x_i) = \left\{ \left(S_\lambda \hat{\mathbf{f}}_{\lambda_p} \right)(x_i) - \hat{f}_{\lambda_p}(x_i) \right\}^2 + \hat{\sigma}^2_{\lambda_p} s_\lambda(x_i) \qquad (5.26)$$

formülü ile elde edilir. Burada, $\left(S_\lambda \hat{\mathbf{f}}_{\lambda_p} \right)(x_i)$ ifadesi $S_\lambda \hat{\mathbf{f}}_{\lambda_p}$ vektörünün i. inci elamanıdır. Önerilen metot pratik olarak, aşağıdaki adımlar ile yerine getirilebilir:

i) Önceden verilen $\lambda_1 < ... < \lambda_m$ düzeltme parametrelerine karşı gelen bir dizi splayn düzeltme tahminleri hesaplanır: $F = \left\{ \hat{f}_{\lambda_1}, ..., \hat{f}_{\lambda_m} \right\}$.

ii) F kümesindeki elemanlar kullanılarak (5.21) ile verilen AIC_c kriterini minimum yapan λ_p değeri seçilir.

ii) Söz konusu λ_p için $\hat{f}_{\lambda_p}$ ve (5.11) kullanılarak $\hat{\sigma}_{\lambda_p}$ tahminleri hesaplanır.

iv) $\hat{f}_{\lambda_p}$ ve $\hat{\sigma}_{\lambda_p}$ pilotları (5.26) ifadesindeki yerine yazılarak $\hat{R}_\lambda(x_i)$ elde edilir.

v) Her bir x_i için $\hat{R}_\lambda(x_i)$ kriterini minimum yapan λ bulunur. F kümesinden uygun (minimum λ parametresine karşı gelen) $\hat{f}_\lambda(x_i)$ değeri $f(x_i)$ için nihai tahmin olarak kabul edilir [59].

5.4. Monte Carlo Simülasyon Deneyi

Bu bölümde, orijinali Profesör Steve Marrona tarafından geliştirilen deneysel çalışma düzeni yardımıyla, bölüm 5.3'te incelenen 6 seçim kriterinden hangisinin daha iyi bir düzeltme parametresi seçtiği belirlemek amacıyla, söz konusu seçim kriterlerinin bir karşılaştırılması yapılmıştır. Bu işlemlerde, MATLAB ortamında tarafımızdan yazılan bir programla, her bir faktör düzeyi için farklı büyüklükte 6 örneklem oluşturulmuş ve oluşturulan her bir örneklem için farklı sayılarda tekrarlamalar yapılmıştır. Tekrarlanan her bir örneklem veri dizisi için herhangi bir $\hat{f}$ tahmin eğrisinin kalitesini değerlendirmek için (5.4)'de verilen hata kareler ortalaması (MSE) kullanılmıştır. Herhangi iki metodun MSE değerleri meydanı arasındaki farkın anlamlı olup olmadığı *Wilcoxon işaretli sıra sayıları*

testi ile test edilmiştir. Böylece, yapılan Monte Carlo simülasyon çalışmasında adı geçen kriterler değerlendirilerek, hangi kriterin daha iyi bir tahmin sonucu veren düzeltme parametresini seçtiği belirlenmiştir.

5.4.1. Veriler ve Deneysel Düzeneğin Oluşturulması

Ele alınan deney planı, esas itibariyle orijinali Profesör Steve Maron'a ait olup Lee [61] tarafından da kullanılmıştır. Ele alınan deneysel düzenek, bağımsız ve etkili bir biçimde değişen şu üç faktörün etkisini çalışmak için tasarlanmıştır:

- Gürültü düzeyi (*Noise level*)
- Uzaysal değişim (*Spatial variation*)
- Varyans fonksiyonudur (*Variance function*).

Deneyde örneklem oluşturmada kullanılan veriler, Tablo 5.1'de genel biçimde verilen modellerden elde edilmiştir. Simülasyon deneyi, MATLAB ortamında yazılan bir programla gerçekleştirilmiş olup, *deneyin planı ve yürütülmesi ise şu şekilde tasarlanmıştır:*

- Her bir faktör düzeyi için yöntemlerin küçük ve büyük örneklem performanslarını görebilmek amacıyla, 25, 50, 100, 150, 200 ve 350 hacimlerinde 6 farklı örneklem oluşturulmuştur.
- Oluşturulan her bir örneklem,100, 200, 350 ve 500 kez tekrar edilmiştir.
- Her bir faktör düzeyi için tekrar edilen tüm örneklemlerde adı geçen her bir seçim kriterini minimum yapan λ düzeltme parametresi seçilmiştir.
- Her bir seçim kriterinden seçilen λ düzeltme parametresine uygun (3.36)'da ifade edilen $\hat{f}_\lambda$ splayn düzeltme kestiricileri hesaplannıştır.
- Her bir seçim kriterine göre hesaplanan $\hat{f}_\lambda$ splayn kestiricileri için (5.4) förmülü ile verilen MSE değerleri hesaplanmıştır.
- Her hangi iki seçim yönteminin performans ölçüsü olarak dikkate alınan MSE değerlerinin faklı olup olmadığı Wilcoxon testi ile test edilmiştir.
- Tablo 5.1'de belirtilen 3 faktörün etkisini belirlemek amacıyla faktör düzeyleri $r = 1, 2, 3, 4$ olmak üzere, 4 kez değiştirilmiştir.
- 3 faktör, 4 faktör düzeyi, 6 örneklem ve 4 farklı sayıda tekrarlama olmak üzere, toplam 288 sayısal deney yapılmıştır.

Tablo 5.1: Simülasyon düzeneğinin ayrıntıları

Faktör	Genel biçimi	Belirli seçenekler
Gürültü Düzeyi	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$\sigma_r = 0.02 + 0.04(r-1)^2,\ i = 1,...,n$
Uzaysal Değişim	$y_{ir} = f_r(x_i) + \sigma\,\varepsilon_i$	$\sigma = 0.2,\quad f_r(x) = \sqrt{x(1-x)}\sin\left[\frac{2\pi\{1+2^{(9-4r)/5}}{x+2^{(9-4r)/5}}\right]$
Varyans Fonk.	$y_{ir} = f(x_i) + \sqrt{v_r(x_i)}\,\varepsilon_i$	$v_r(x) = [0.15\{1+0.4(2r-7)(x-0.5)\}]^2$
$r = 1,...,4;\ x_i = \frac{i-0.5}{n};\ \varepsilon_i \sim iid\ N(0,1);\ f(x) = 1.5\theta\left(\frac{x-0.35}{0.15}\right) - \theta\left(\frac{x-0.8}{0.04}\right);\ \theta(u) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-u^2}{2}\right)$		

5.4.2. Deneysel Değerlendirmeler ve Sonuçlar

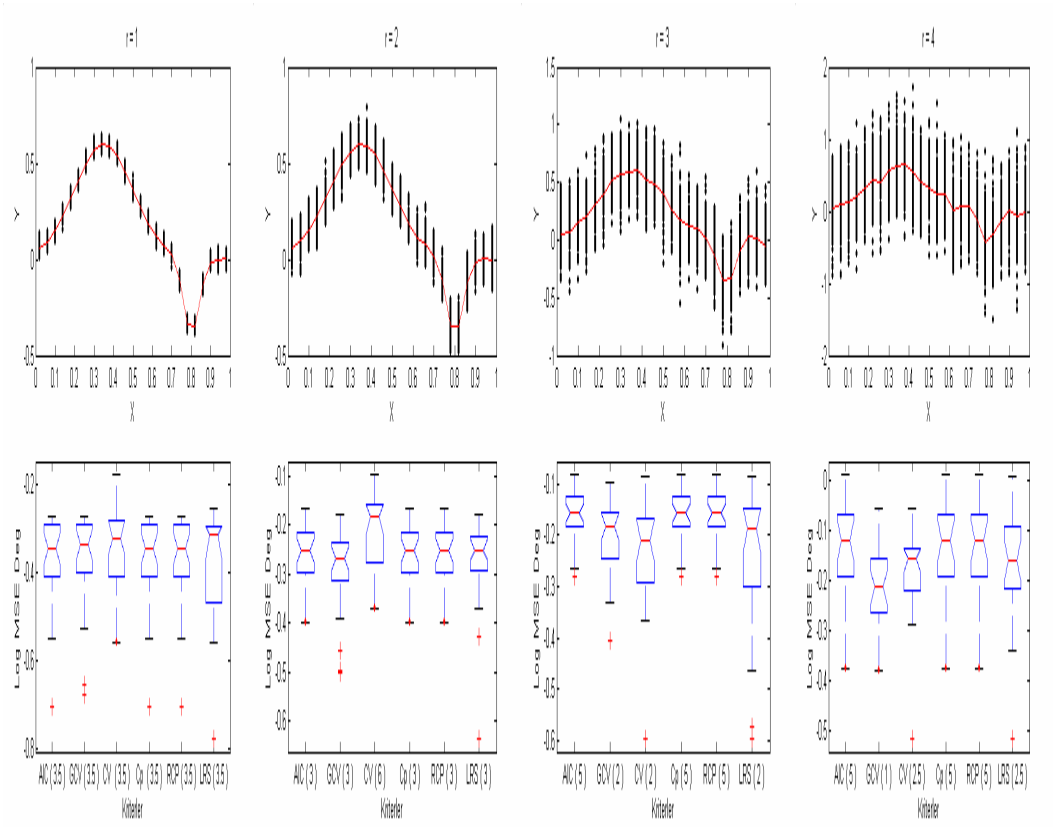
Simülasyon deneyleri sonucunda oluşturulan toplam 288 sayısal deneylerin her birinde, bölüm 5.3'de dikkate alınan 6 tane seçim kriterlerinden hangisinin daha iyi bir düzelte parametresi seçtiğini ve seçilen bu düzeltme parametreleri yardımıyla hesaplanan (3.36) $\hat{f}$ splayn düzeltme kestiricilerinin iyi bir kestirici olup olmadığını belirlemek amacıyla, MSE değerleri performas ölçü kriteri olarak kullanılmıştır.

Herhangi iki seçim metodun MSE değerleri meydanı arasındaki farkın anlamlı olup olmadığı %5 anlam düzeyinde eşleştirilmiş Wilcoxon işaretli sıra testi ile test edilmiştir. Ayrıca seçim metotları Wilcoxon işaretli sıra testine göre şu şekilde sıralanmıştır: *Bir metodun MSE değeri medyanı anlamlı bir biçimde kalan 5 yöntemde daha az ise, 1 sırası atanacak, bir metodun MSE değeri medyanı anlamlı olarak birinden daha büyük fakat dört metottan az ise, 2 sırası atanacak ve benzer biçimde 3-6 sıraları atanacak. Farklı medyan değerlerine sahip fakat anlamlı olmayan metotlar aynı ortalamalı sırayı paylaşacaklar. Bu sıralamada en küçük sırayı alan yöntem ya da paylaşan yöntemler daha üstündür.*

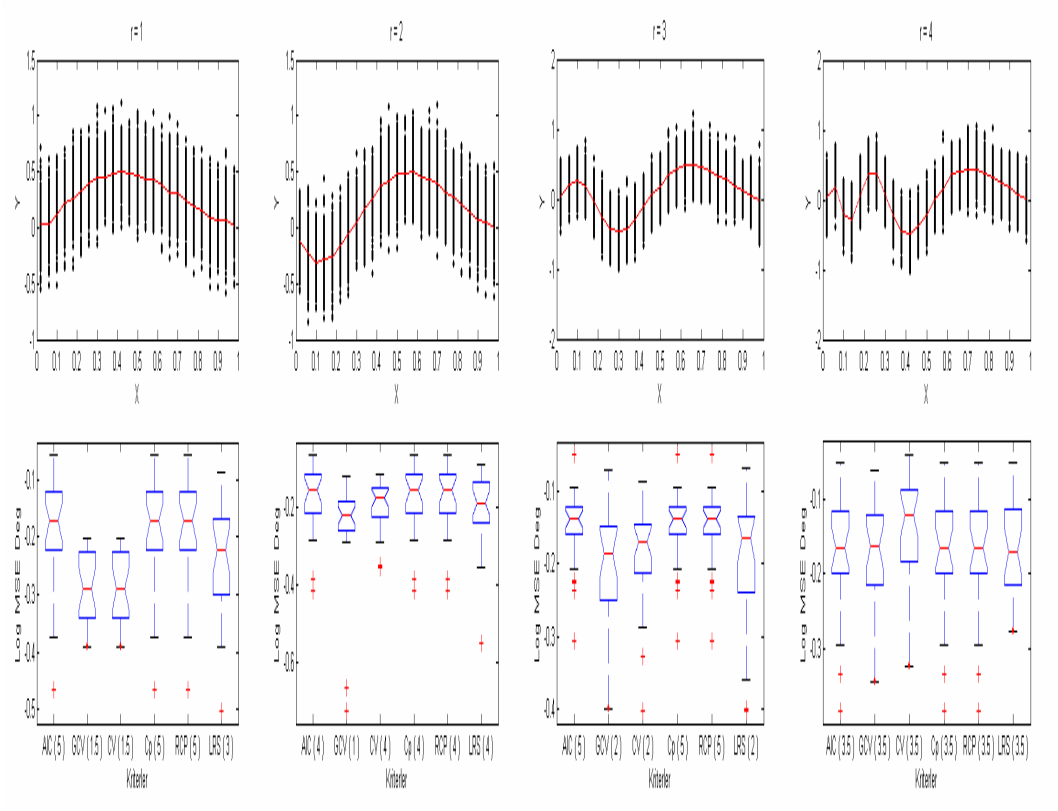
Simülasyonla elde edilen örneklem verileri için toplam 288 faklı sayısal deneyi oluşturan, regresyon fonksiyonların grafikleri ve altı düzeltme parametresi seçim yöntemlerinin $\log_e$MSE değerlerinin kutu grafiklerinden (*boxplot*) bazıları, 5.1-5.16 şekillerinde verilmiştir. Söz konusu şekillerde, üst panelde yer alan grafikler tipik bir benzetim veri dizisiyle gerçek regresyon fonksiyonunu ve alt

panelde yer alan grafikler soldan sağa doğru AIC_c, GCV, CV, Cp, RCP ve LRS kriterlerinin log_eMSE değerlerin kutu grafiklerini göstermektedir. Kutu grafiklerin altında yer alan sayılar ise, söz konusu altı seçim kriterlerinin MSE ölçü değerlerine ilişkin meydanların, Wilcoxon işaretli sıra sayıları testine göre sıralanmalarını göstermektedir. Bunun yanı sıra, her bir faktör ve tekrar edilen her bir örneklem verileri için, yöntemlerin ortalam düzeyinde başarılarını görmek amacıyla, atı seçim yöntemlerinin kutu grafiklerinin altında yer alan sayıların ortalaması alınarak ortalama sıralamalar ede edilmiştir. Söz konusu bu ortalama sıralamalar 5.2-5.7 Tablolarında verilmiştir.

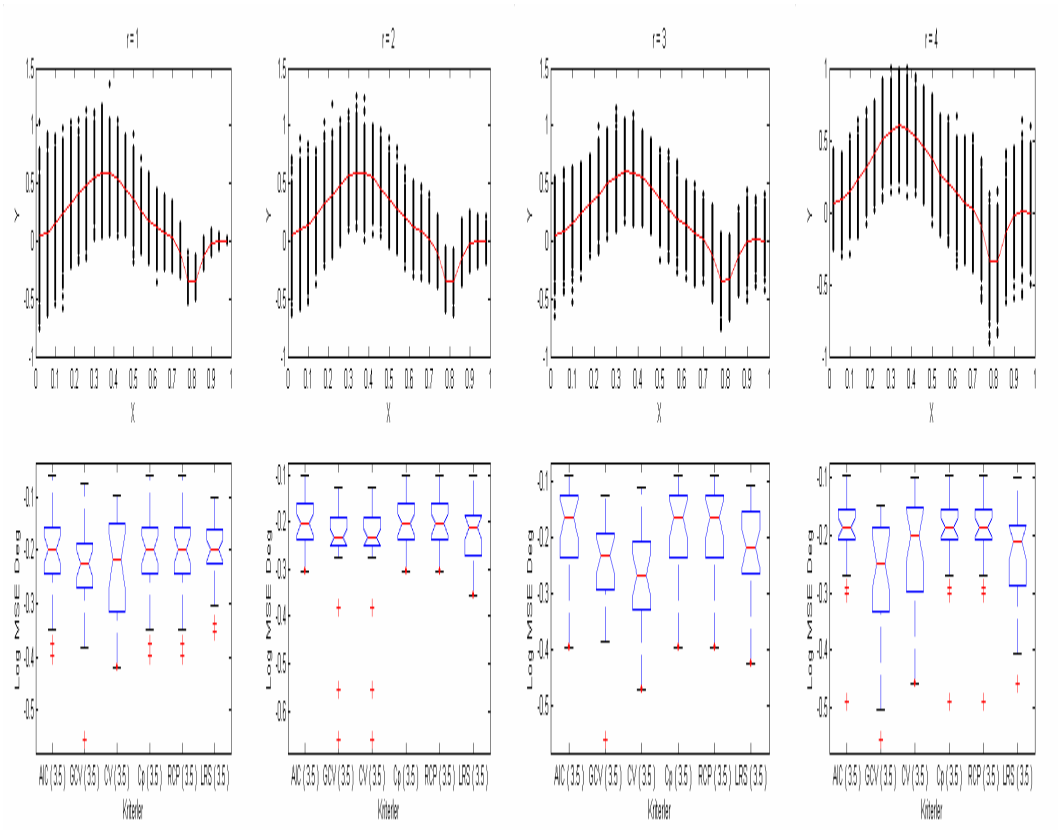
5.1-5.3 Şekilleri incelediğinde, yüksek gürültü düzeyli basit bir regresyon fonksiyonu için GCV kriterinin daha üstün olduğu görülmüştür (bak., Şekil 5.1 ve r = 4). Ancak, kalan iki faktörün yüksek olduğu bir düzeyde, tüm seçim kriterlerinin aynı sıralamayı paylaştığı görülmüştür (bak. Şekil 5.2-5.3 ve r = 4).



Şekil 5.1: n = 25 ve m =100 için gürültü düzeyi faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.2: n = 25 ve m =200 için uzaysal değişim faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.3: n = 25 ve m =500 için varyans faktörüne karşı gelen simülasyon sonuçlarının grafikleri

Tablo 5.2 incelendiğinde; n = 25 birimlik küçük örenklem verileri için gürültü düzeyi, varyans fonksiyonu ve uzaysal değişim faktörlerinin etkileri altında kalan 100 tekrarlı bir deneyde, GCV kriterinin düzgün olarak en iyi olduğu ve tekrar sayısı artarak değişmesi durumunda, altı yöntemin 500 tekrar sonucunda, değişen varyans hatası altında aynı ortalamalı sıralamayı paylaştığı ve toplamda yine GCV kriterinin en iyi yöntem olduğu görülürken, genel olarak toplamda en kötü performansı AIC_c Cp ve RCP kriterlerinin sergilediği görülmüştür.

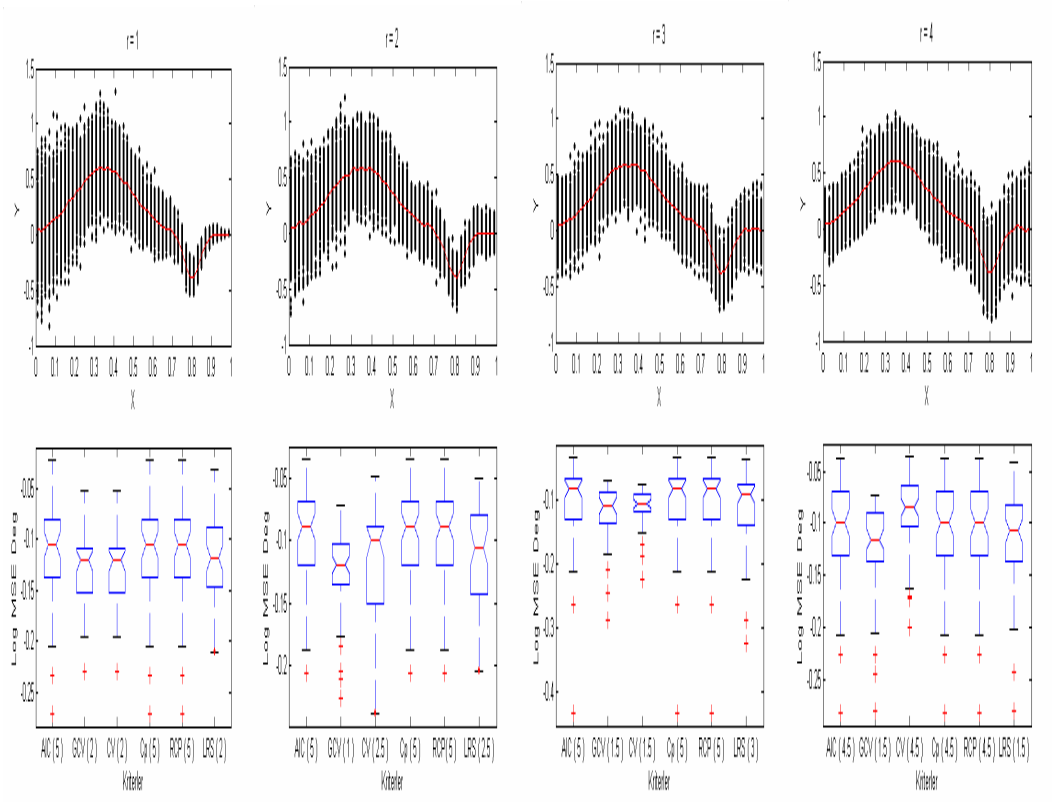
Tablo 5.2: n = 25 hacimlik örneklemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	4,125	4,250	4,125	4,167
GCV	2,375**	2,375**	1,875**	2,208**
CV	4,667	2,750*	3,375*	3,597
Cp	4,125	4,250	4,125	4,167
RCP	4,125	4,250	4,125	4,167
LRS	2,750*	3,125	3,375*	3,083*
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	3,750	4,375	3,875	4,000
GCV	2,250**	2,000**	3,000**	2,417**
CV	4,500	2,750*	3,000**	3,417
Cp	3,750	4,375	3,875	4,000
RCP	3,750	4,375	3,875	4,000
LRS	3,000*	3,125	3,375*	3,167*
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	3,875*	4,500	3,375**	3,917
GCV	3,125**	1,875**	3,375**	2,972**
CV	3,125**	3,000	4,125	3,417
Cp	3,875*	4,500	3,375**	3,917
RCP	3,875*	4,500	3,375**	3,917
LRS	3,125**	2,625*	3,375**	3,042*
Örneklemin Tekrarlanma Sayısı = 500				
AIC_c	3,750	4,000	3,500**	3,750
GCV	2,875**	2,375**	3,500**	2,917**
CV	3,625	3,125*	3,500**	3,417
Cp	3,750	4,000	3,500**	3,750
RCP	3,750	4,000	3,500**	3,750
LRS	3,250*	3,500	3,500**	3,417*

(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.
(*): İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.

Şekil 5.4 incelendiğinde, heterokadastik hata altında ki basit bir regresyon fonksiyonu için, AIC_c, Cp ve RCP kriterlerinin en kötü performasla aynı sıralamayı paylaştıkları görülmüştür. Buna karşılık, değişen bir varyans faktörü etkisi altında kalan örneklem verileri için, GCV kriteri en iyi seçim yöntemi olmuştur (bak. Şekil 5.4 ve r = 2).

Küçük örneklem durumuna benzer olarak, Tablo 5.3 incelendiğinde; gürültü düzeyi, varyans fonksiyonu ve uzaysal değişim faktörü etkisi altında kalan n = 50 birimlik örneklem için tekrar sayılarının artarak değişmesi durumunda da GCV kriterinin düzgün olarak en iyi seçim yöntemi olduğu ortaya çıkmaktadır



Şekil 5.4: n = 50 ve m =350 için varyans faktörüne karşı gelen simülasyon sonuçlarının grafikleri

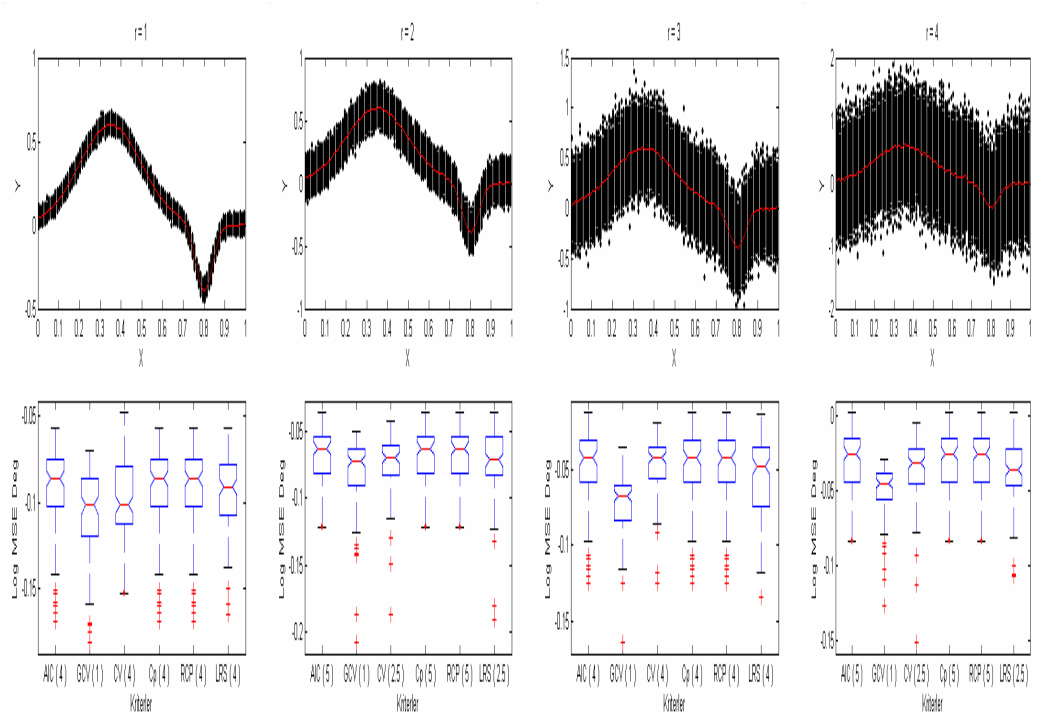
Tablo 5.3: n = 50 hacimlik örneklemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	4,375	4,375	4,875	4,542
GCV	1,875**	1,250**	1,500**	1,542**
CV	2,325*	2,500*	2,000*	2,275*
Cp	4,375	4,375	4,875	4,542
RCP	4,375	4,375	4,875	4,542
LRS	3,750	3,375	2,875	3,333
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	4,000	4,625	4,875	4,000
GCV	2,250*	1,250**	1,500**	2,417**
CV	3,375*	3,375	2,625	3,417
Cp	4,000	4,625	4,875	4,000
RCP	4,000	4,625	4,875	4,000
LRS	3,375*	2,500*	2,250*	3,167*
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	4,000	4,125	4,750	3,917
GCV	2,375**	1,625**	1,250**	2,972**
CV	3,625	3,375*	2,375*	3,417
Cp	4,000	4,125	4,750	3,917
RCP	4,000	4,125	4,750	3,917
LRS	3,000*	3,625	3,125	3,042*
Örneklemin Tekrar Edilme Sayısı = 500				
AIC_c	4,250	4,375	4,750	3,750
GCV	2,375**	1,875**	1,500**	2,917**
CV	2,750*	3,250	2,250*	3,417*
Cp	4,250	4,375	4,750	3,750
RCP	4,250	4,375	4,750	3,750
LRS	3,125	2,750*	3,000	3,417*

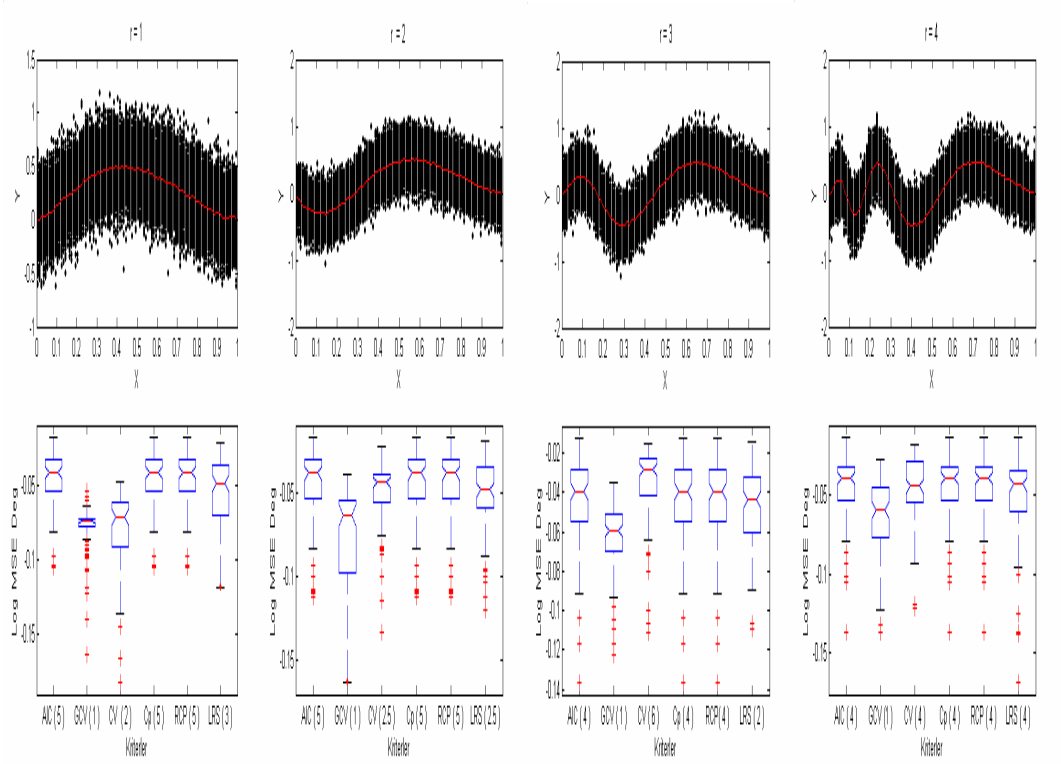
(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.
(*) : İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.

5.5-5.7 Şekilleri incelendiğinde, örneklemdeki tüm gürültü düzeyleri için, Şekil 5.5'te görüldüğü gibi en iyi sıralamayı GCV kriteri almıştır. Buna karşılık, AIC_c, Cp ve RCP kriterlerinin en kötü performansla aynı sıralamayı paylaştıkları

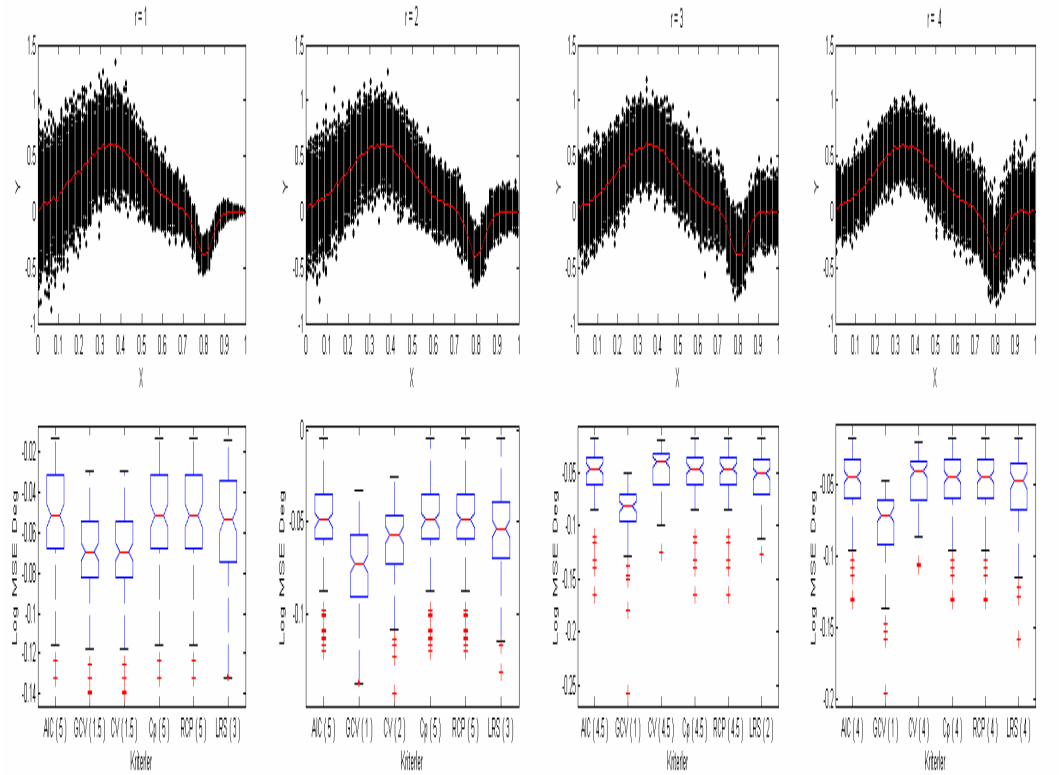
ve benzer bir durum, Şekil 5.6'da görüldüğü gibi, uzaysal değişim sonuçları için de geçerlidir. Diğer yandan, heterokadastik hata altında, Şekil 5.7'de görüldüğü gibi, en iyi seçim yöntemi yine GCV kriteri olurken, yüksek düzeyli değişen varyans hatası altında, AIC_c, CV, Cp RCP ve LRS kriterleri aynı sıralamayı almışlardır (bak. Şekil 5.7 ve r =4). Ancak, varyans faktörünün ilk üç düzeyinde, AIC_c, Cp ve RCP kriterleri, en kötü sıralamayı paylaşmılardır (Şekil 5.7ve r = 1,2,3).



Şekil 5.5: n = 100 ve m =500 için gürültü düzeyi faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.6: n = 100 ve m =350 için uzaysal değişim faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.7: n = 100 ve m =200 için varyans faktörüne karşı gelen simülasyon sonuçlarının grafikleri

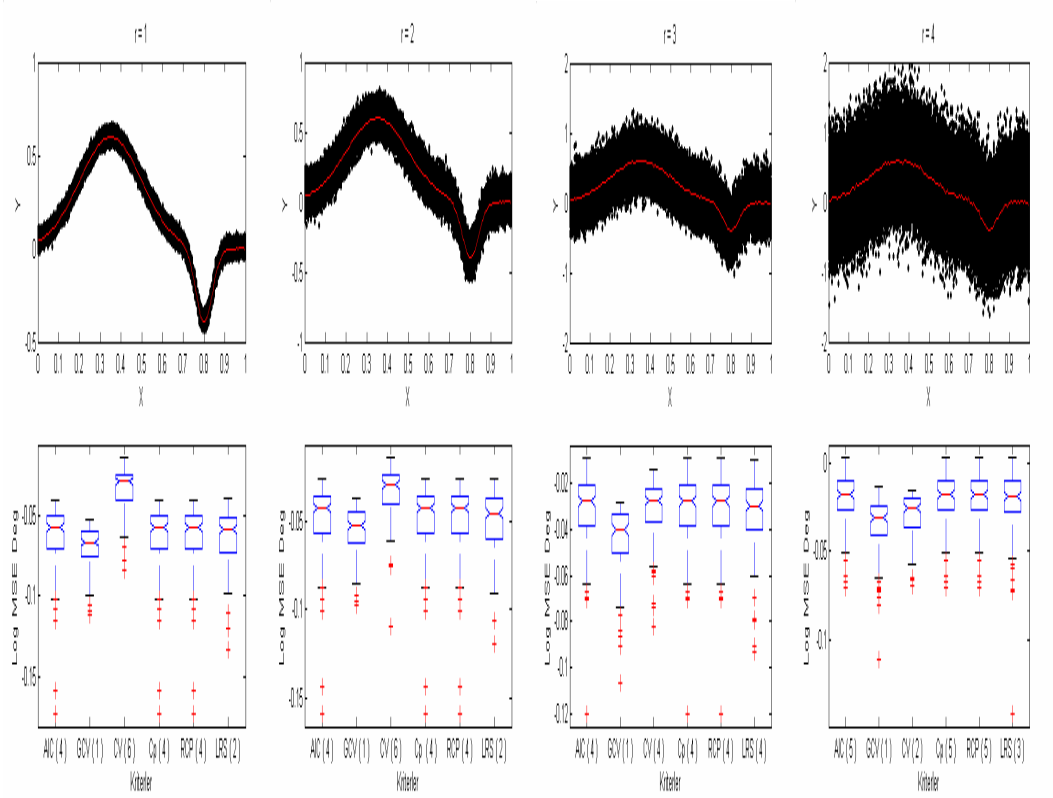
Tablo 5.4 incelendiğinde; n = 50 birimlik örneklemlerin durumuna benzer olarak, gürültü düzeyi, varyans fonksiyonu ve uzaysal değişim faktörlerinin etkileri altında kalan örneklem verileri için, tekrar sayılarının değişmesi durumunda, önceki sonuçlarda olduğu gibi burada da GCV kriterinin düzgün olarak en iyi seçim yöntemi olduğu görülürmüştür. Buna karşılık, en kötü performansla aynı ortalamalı sıralamayı, AIC_c, Cp ve RCP kriterleri paylaşmışlardır.

Tablo 5.4: n =100 hacimlik örnkelemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

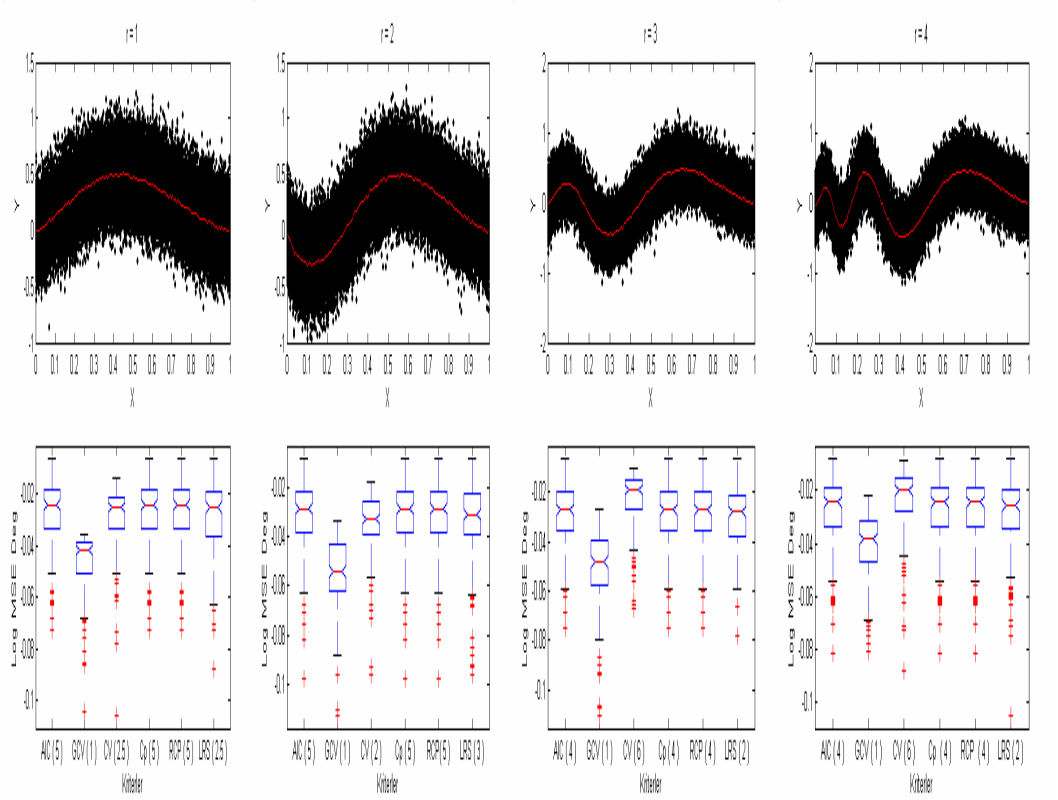
Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	4,750	4,500	4,750	4,667
GCV	1,250**	1,125**	1,250**	1,208**
CV	3,750	3,375	2,375*	3,167
Cp	4,750	4,500	4,750	4,667
RCP	4,750	4,500	4,750	4,667
LRS	3,250*	2,500*	3,125	2,958*
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	4,750	4,750	4,625	4,708
GCV	1,000**	1,250**	1,125**	1,125**
CV	3,250	2,750	2,750*	2,917
Cp	4,750	4,750	4,625	4,708
RCP	4,750	4,750	4,625	4,708
LRS	2,500*	2,375*	3,000	2,625*
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	5,000	4,500	4,750	4,750
GCV	1,375**	1,000**	1,375**	1,250**
CV	1,625*	3,625	2,875	2,708*
Cp	5,000	4,500	4,750	4,750
RCP	5,000	4,500	4,750	4,750
LRS	3,000	2,875*	2,500*	2,972
Örneklemin Tekrar Edilme Sayısı = 500				
AIC_c	4,500	4,750	5,000	4,750
GCV	1,000**	1,375**	1,625**	1,333**
CV	3,250*	2,625*	2,375*	2,750*
Cp	4,500	4,750	5,000	4,750
RCP	4,500	4,750	5,000	4,750
LRS	3,250*	2,750	2,750	2,917

(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.
(*) : İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir

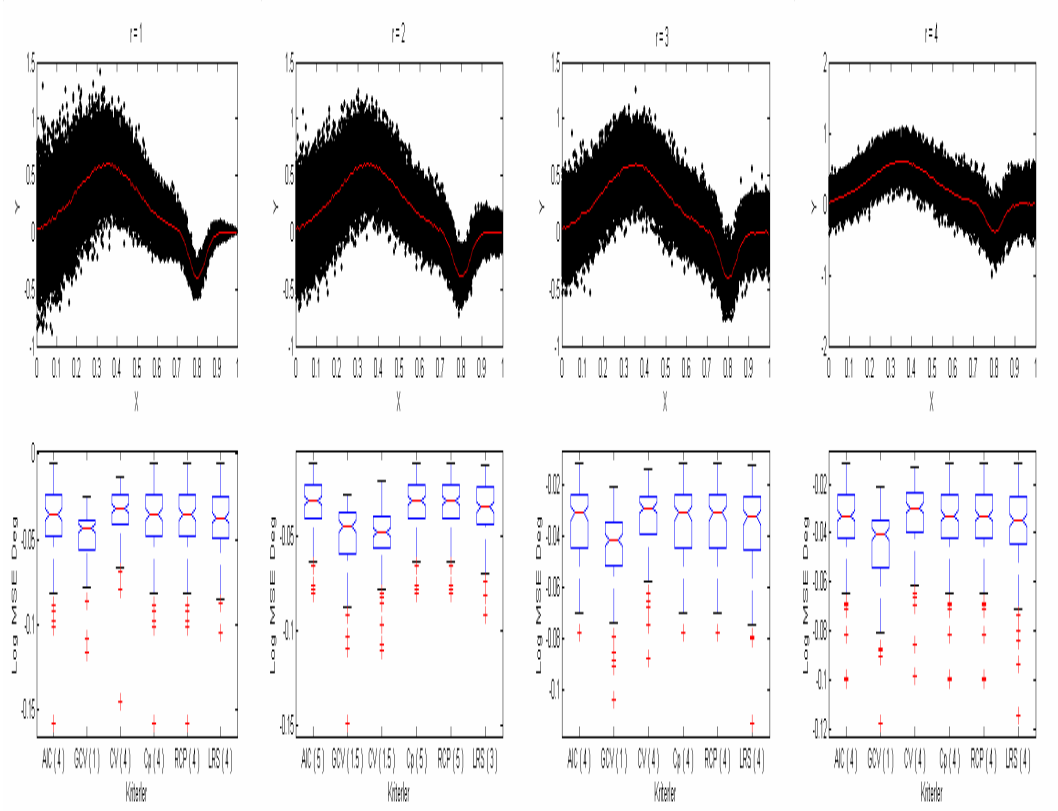
5.8-5.10 Şekilleri incelendiğinde, Şekil 5.8'de görüldüğü gibi tüm gürültü düzeylerinde en iyi sıralamayı GCV kriteri almıştır. Buna karşılık, ilk iki düzeyde CV en kötü performansı gösterirken, AIC_c, Cp ve RCP kriterleri, tüm düzeylerde aynı sıralamayı paylaşmışlardır. Uzaysal değişim faktörünün tüm düzeylerinde, gürültü düzeyinde olduğu gibi, GCV birinci olurken, son iki düzeyde en kötü performansı CV kriteri göstermiştir (bak. Şekil 5.9 ve r =3, 4). Diğer yandan, heterokadastik hata altında, tüm düzeylerde yine GCV kriteri birinci olurken, diğer tüm kriterler aynı sıralamayı paylaşmışlardır (bak. Şekil 5.10 ve r = 1,3 ve 4).



Şekil 5.8: n = 150 ve m =500 için gürültü düzeyi faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.9: n = 150 ve m =350 için uzaysal değişim faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.10: n = 150 ve m =350 için varyans fonksiyonu faktörüne karşı gelen simülasyon sonuçlarının grafikleri

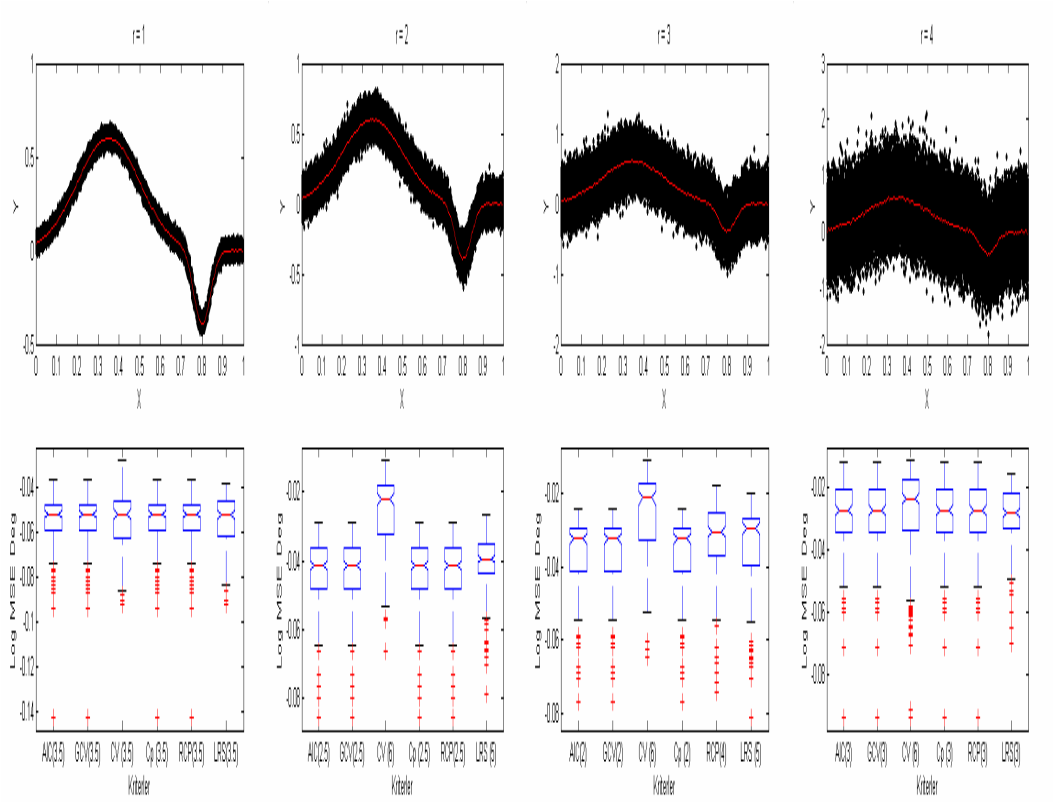
Tüm düzeylere karşı gelen sıralamaların ortalamalarını veren Tablo 5.5 incelendiğinde; gürültü düzeyi, varyans fonksiyonu ve uzaysal değişim faktörlerinin etkileri altında kalan örneklem verileri için değişen tüm tekrarlamalarda, açık bir şekilde düzgün olarak en iyi seçim kriteri GCV ve onun ardından genel ortalamada, ikinci sırayı LRS kriteri alırken, tüm faktör düzeylerinde, en kötü performansla aynı ortalamalı sıralamayı AIC_c, Cp ve RCP kriterlerinin paylaştıkları görülmüştür

Tablo 5.5: n = 150 hacimlik örneklemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

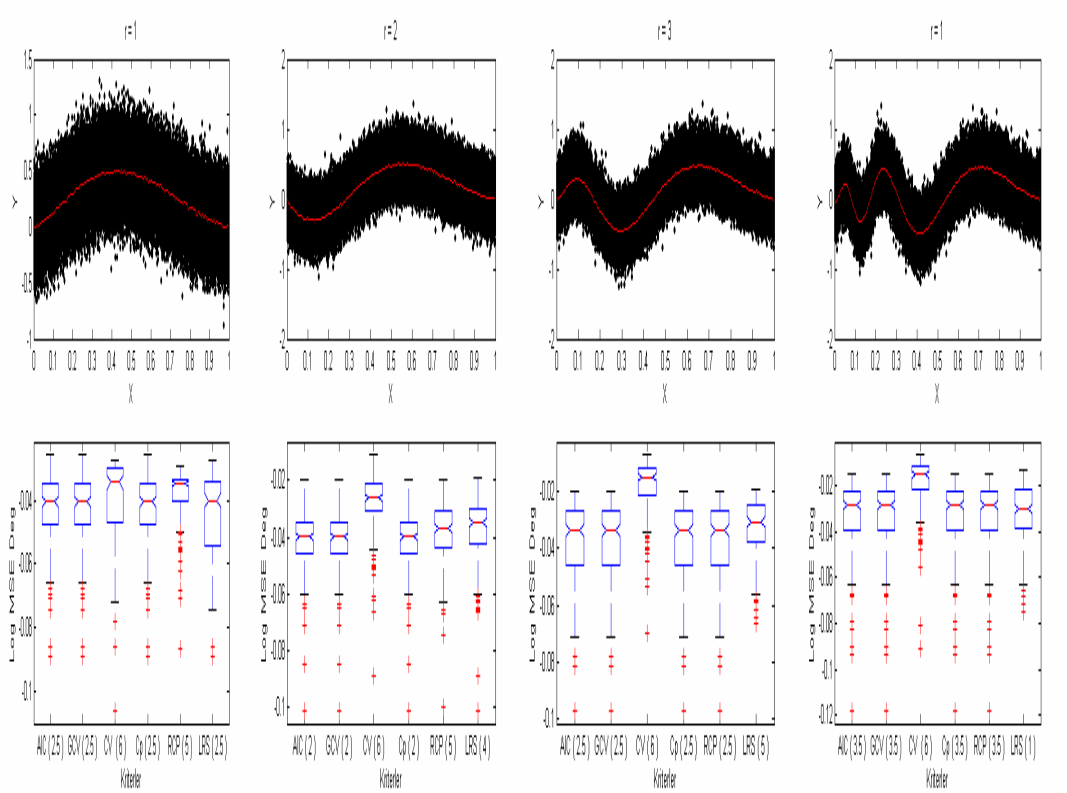
Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	4,750	4,750	4,625	4,708
GCV	1,125**	1,000**	1,375**	1,167**
CV	2,875	4,500	3,625*	3,667
Cp	4,750	4,750	4,625	4,708
RCP	4,750	4,750	4,625	4,708
LRS	2,750*	2,750*	4,125	3,208*
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	4,500	4,750	4,750	4,667
GCV	1,125**	1,125**	1,250**	1,167**
CV	3,000*	4,875	3,000*	3,625
Cp	4,500	4,750	4,750	4,667
RCP	4,500	4,750	4,750	4,667
LRS	3,750	2,250*	3,750	3,250*
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	4,500	4,500	4,250	4,417
GCV	1,125**	1,000**	1,125**	1,083**
CV	3,875	4,125	3,375*	3,792
Cp	4,500	4,500	4,250	4,417
RCP	4,500	4,500	4,250	4,417
LRS	2,500*	2,375*	3,750	2,875*
Örneklemin Tekrar Edilme Sayısı = 500				
AIC_c	4,750	4,500	4,875	4,708
GCV	1,000**	1,250**	1,125**	1,125**
CV	4,500	2,750*	2,625*	3,292
Cp	4,750	4,500	4,875	4,708
RCP	4,750	4,500	4,875	4,708
LRS	2,750*	3,500	2,625*	2,958*

(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri gösterir.
(*) : İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri gösterir.

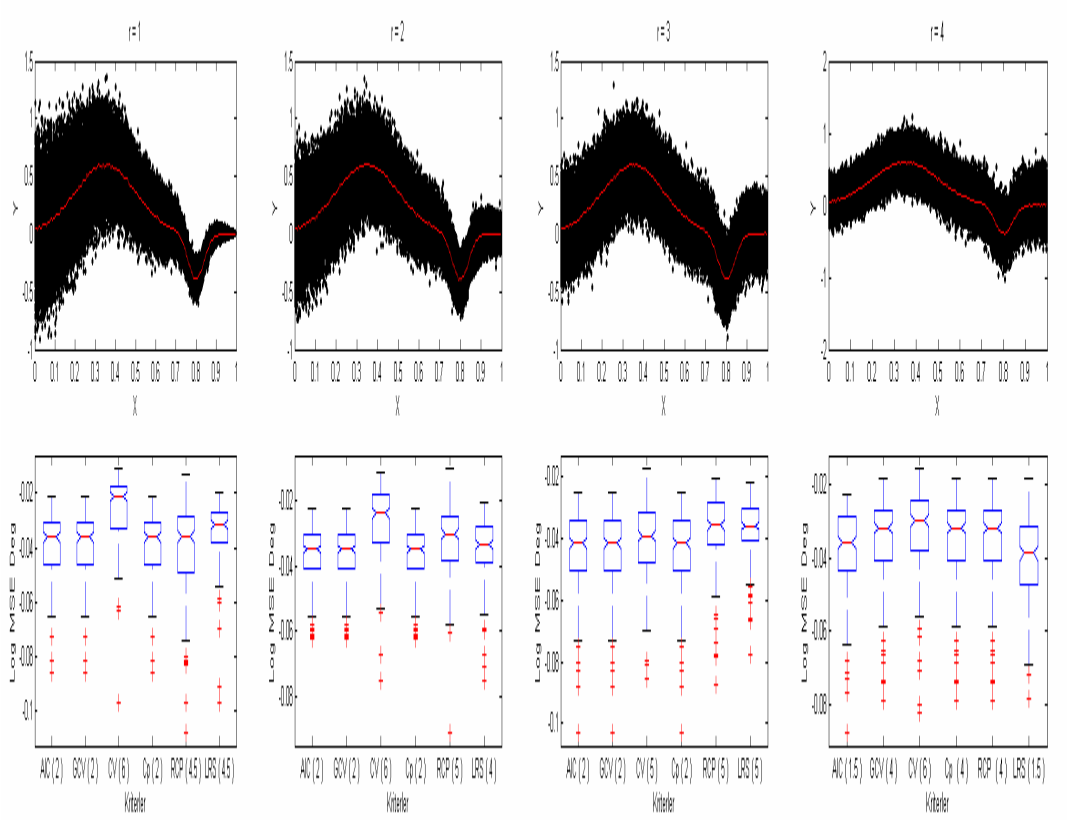
5.11-5.13 Şekilleri incelendiğinde, tüm gürültü düzeylerinde en iyi sıralamayı AIC_c, GCV ve Cp kriterleri, en kötü sıralamayı ise, CV kriteri almıştır (bak. Şekil 5.11 ve r = 2, 3 ve 4). Benzer bir durum, Şekil 5.12'de görüldüğü gibi, uzaysal değişim faktörü için de söylenebilir ancak, yüksek düzeyli bir varyans faktörü için en iyi LRS kriteri olmuştur (bak. Şekil 5.12 ve r = 4). Diğer yandan, heterokadastik hata altında, en iyi yine AIC_c, GCV ve Cp kriterleri olurken, buna karşılık, varyans faktörünün etkilediği tüm düzeylerde en kötü performansı, CV kriteri göstermiştir.



Şekil 5.11: n = 200 ve m =350 için gürültü düzeyi faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.12: n = 200 ve m =500 için uzaysal değişim faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.13: n = 200 ve m =500 için varyans fonksiyonu faktörüne karşı gelen simülasyon sonuçlarının grafikleri

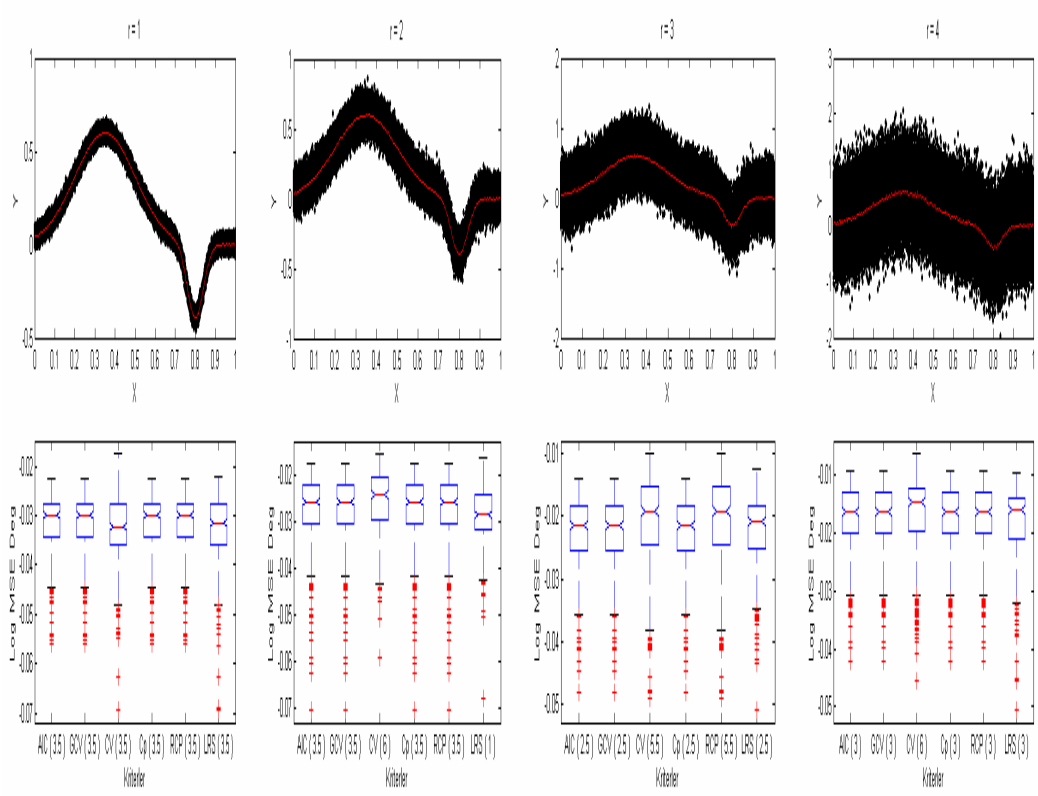
Şu ana kadar en iyi seçim kriteri olan GCV'nin yanırsa, artan örneklem hacmine bağlı olarak, AIC_c ve Cp kriterlerinin de iyi sonuç verdikleri gözlenmiştir. Bu performansları ortalamalar düzeyinde görmek amacıyla, Tablo 5.6 incelendiğinde; gürültü düzeyi, varyans fonksiyonu ve uzaysal değişim faktörlerinin etkileri altında kalan örneklem verileri için tüm tekrarlamalarda, genellikle AIC_c, GCV ve Cp kriterleri en iyi performansla aynı ortalamalı sıralmayı paylaşmışlardır. Toplmada genel olarak, en iyi yöntem AIC_c kriteri olurken, en kötü ortalamalı sıralamayı CV kriterinin aldığı görülmüştür.

Tablo 5.6: n = 200 hacimlik örneklemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

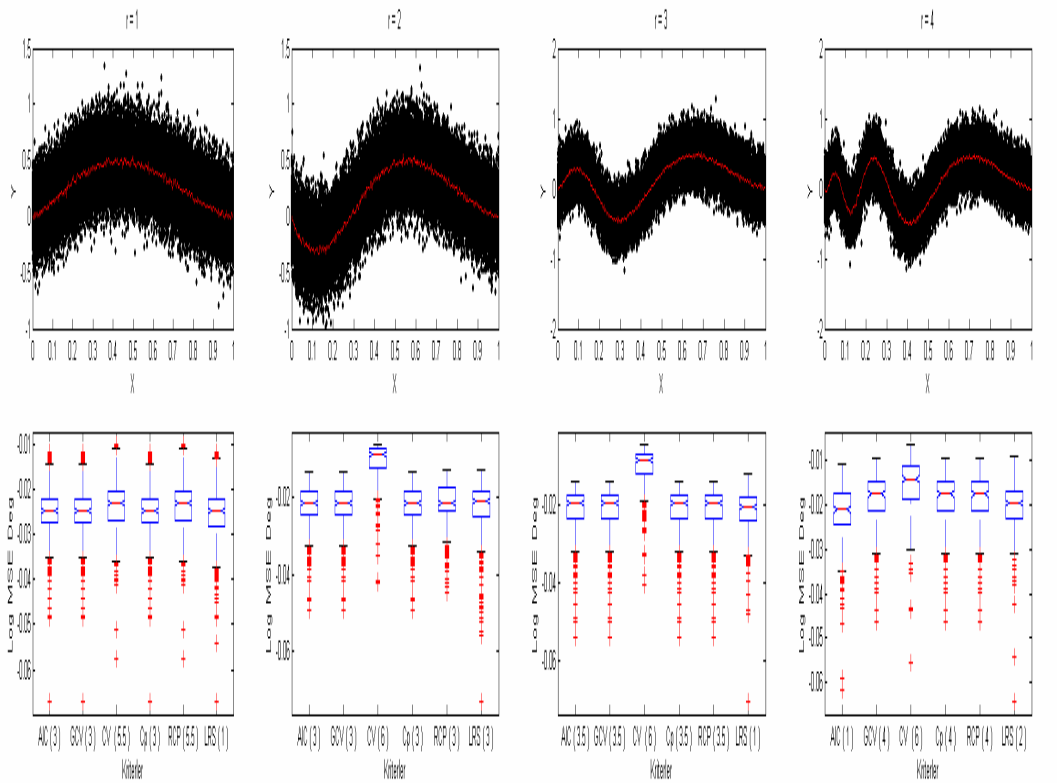
Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	3,375*	2,250**	2,625**	2,750**
GCV	3,375*	2,875*	2,625**	2,958*
CV	4,750	5,250	4,125	4,708
Cp	3,375*	2,875*	2,625**	2,958*
RCP	3,375*	4,375	5,125	4,292
LRS	2,750**	3,375	3,875*	3,333
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	3,000**	2,750**	2,250**	2,667**
GCV	3,000**	2,750**	2,250**	2,667**
CV	4,625	5,125	5,875	5,208
Cp	3,000**	2,750**	2,250**	2,667**
RCP	3,750	4,375	4,000*	4,042
LRS	3,625*	3,250*	4,375	3,750*
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	2,750**	2,875**	2,250**	2,625**
GCV	2,750**	2,875**	3,125	2,912*
CV	5,375	5,375	4,250	5,000
Cp	2,750**	2,875**	3,125	2,912*
RCP	3,250*	3,000*	5,250	3,833
LRS	4,125	4,000	3,000*	3,708
Örneklemin Tekrar Edilme Sayısı = 500				
AIC_c	3,000**	2,625**	1,875**	2,500**
GCV	3,000**	2,625**	2,500*	2,708*
CV	3,875*	6,000	5,750	5,208
Cp	3,000**	2,625**	2,500*	2,708*
RCP	3,875*	4,000	4,625	4,167
LRS	4,250	3,125*	3,750	3,708

(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.
(*) : İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir

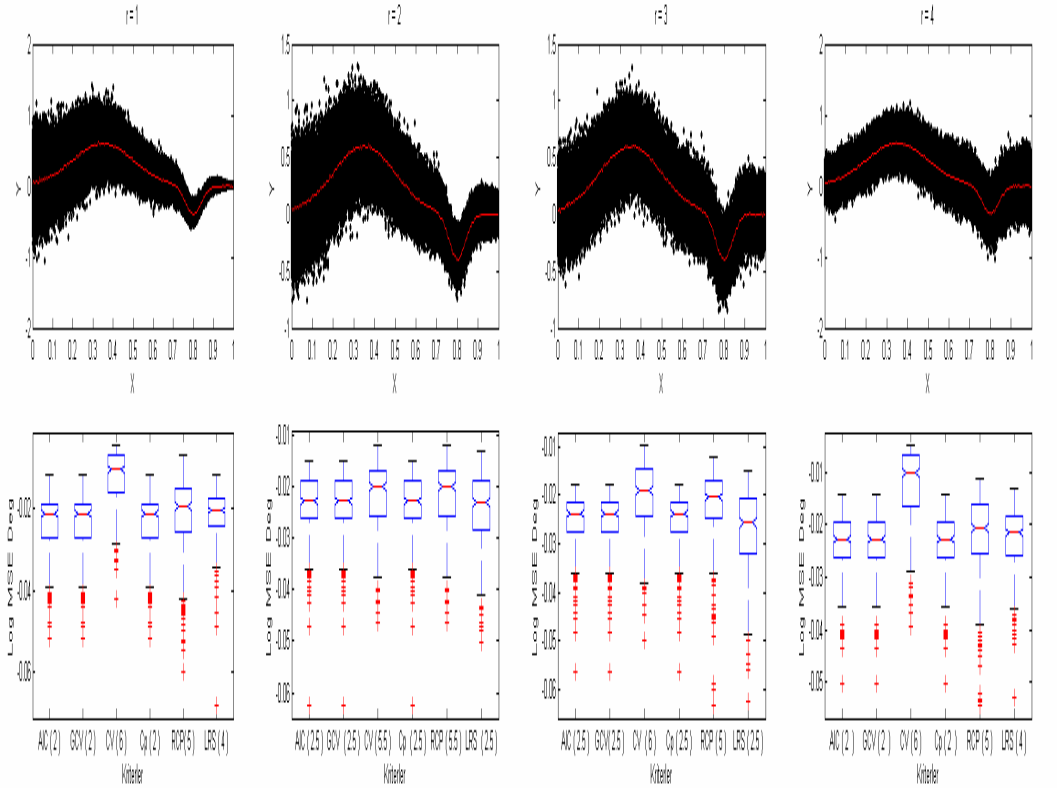
5.14-5.16 Şekilleri incelendiğinde, önceki sonuçlara benzer olarak, tüm gürültü düzeylerinde en iyi sıralamayı yine AIC_c, GCV ve Cp alırken, gürültü düzeyleri arttıkça en kötü performansı CV kriteri almıştır (bak. Şekil 5.14 ve r = 2, 3 ve 4). Uzaysal değişim faktörünün ilk üç düzeyinde, AIC_c, GCV ve Cp kriterleri aynı sıralamayı paylaşmışlardır. Ancak, LRS iki kez en iyi olurken, yüksek bir uzaysal değişim faktörü düzeyinde en iyi AIC_c kriteri olmuştur (bak. Şekil 5.14 ve r = 1, 3 ve 4). Diğer yandan, en kötü performansı yine tüm faktör düzeylerinde CV kriteri sergilemiştir. Son olarak, varyans faktörünün etkisi incelendiğinde, değişen varyans düzeylerinin tümünde, en iyi sıralamayı AIC, GCV ve Cp kriterleri alırken, en kötü sıralamayı CV kriteri almıştır (bak. Şekil 5.16).



Şekil 5.14: n = 350 ve m =350 için gürültü düzeyi faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.15: n = 350 ve m =100 için uzaysal değişim faktörüne karşı gelen simülasyon sonuçlarının grafikleri



Şekil 5.16: n = 350 ve m =350 için varyans fonksiyonu faktörüne karşı gelen simülasyon sonuçlarının grafikleri

Yukarda verilen bilgiler doğrultusunda, şeçim yöntemlerinin gösterdikleri performansları ortalamalar düzeyinde görmek amacıyla Tablo 5.7 incelendiğinde;

Tablo 5.7: n = 350 hacimlik örneklemde altı düzeltme parametresi seçim metotları için ortalaması alınan Wilcoxon testi sıralamaları

Örneklemin Tekrar Edilme Sayısı = 100				
Kriterler	Gürültü Düzeyi	U. Değişim	Varyans Fonk.	Toplam
AIC_c	3,000**	2,625*	2,500**	2,708**
GCV	3,000**	3,375	2,500**	3,308
CV	4,750	5,875	4,250	4,958
Cp	3,000**	3,375	2,500**	3,308
RCP	3,625*	4,000	5,500	4,375
LRS	3,625*	1,750**	3,750	3,041*
Örneklemin Tekrar Edilme Sayısı = 200				
AIC_c	3,000**	2,750**	3,000**	2,917**
GCV	3,000**	2,750**	3,000**	2,917**
CV	4,500	5,125	4,500	4,708
Cp	3,000**	2,750**	3,000**	2,917**
RCP	3,875	3,500*	3,875	3,750*
LRS	3,625*	4,125	3,625*	3,792
Örneklemin Tekrar Edilme Sayısı = 350				
AIC_c	3,125*	2,875**	2,250**	2,750**
GCV	3,125*	2,875**	2,250**	2,750**
CV	5,250	5,250	5,875	5,458
Cp	3,125*	2,875**	2,250**	2,750**
RCP	3,875	4,125	5,125	4,375
LRS	2,500**	3,000*	3,250*	2,917*
Örneklemin Tekrar Edilme Sayısı = 500				
AIC_c	3,000*	2,750**	2,250**	2,667**
GCV	3,000*	2,750**	3,625	3,125
CV	5,750	5,375	3,750	4,958
Cp	3,000*	2,750**	2,875*	2,875*
RCP	3,875	3,875	5,000	4,250
LRS	2,375**	3,500*	3,500	3,125

(**) : En iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.
(*) : İkinci en iyi sıralamayı alan yöntemi ya da paylaşan yöntemleri göstermektedir.

farklı gürültü düzeyleri içeren 350 birimlik bir öreklemin 100 ve 200 kez tekrarlanması durumunda, AIC, GCV ve Cp kriterleri birinci olurken, toplamda ise AIC_c iki kez GCV ve Cp bir kez en iyi ortalamaya sahip sıralamayı almışlardır. Aynı 350 birimlik örneklemin 350 ve 500 kez tekrarlanması durumunda, LRS kriteri en iyi olurken, AIC, GCV ve Cp kriterleri ikinci en iyi ortalamalı sıralamayı paylaşmışlardır. Ancak, toplamda yine AIC_c iki kez, GCV ve Cp bir kez en iyi ortalamaya sahip sıralamayı almışlardır. Uzaysal değişim ve heterokadastik hata altındaki örneklerin simülasyon sonuçlarında ise, büyük çoğunlukla AIC, GCV ve Cp kriterleri en iyi ortalamaya sahip sıralamayı paylaşırken, genelde LRS kriteri ikinci en iyi ortalamalı sıralamayı almıştır. Diğer yandan, önceki simülasyon sonuçlarında olduğu gibi, burada da en kötü ortalamaya sahip bir sıralamayı CV kriterinin aldığı görülmüştür.

Örneklem hacimlerine göre yöntemlerin performansları incelendiğinde, 25-150 hacimlik örneklem aralığında GCV kriteri, gürültü düzeyi, uzaysal değişim ve varyans fonksiyonu faktörlerinin etkisi altında kalan örneklem veriler için düzgün olarak en iyi olurken, ikincilik sıralamasında en iyi LRS kriteri olmuştur (bak. Tablo 5.2-5.5). Buna karşılık AIC_c, Cp ve RCP kriterleri en kötü performansı göstermişlerdir. 200-350 birimlik örneklem verileri incelendiğinde, AIC_c, ve Cp kriterlerinin performanslarının iyileştiği görülmüştür. Bu durumda, söz konusu üç faktör altındaki örneklem verileri için yapılan tekrarların çoğunda, AIC_c, GCV ve Cp kriterleri aynı ortalamalı en iyi sıralamayı paylaşmışlardır. Diğer yandan, genel ortalamanın çoğunda, AIC, kriterinin en iyi olurken, kalanlar GCV, Cp ve LRS şeklinde sıralanır. Ayrıca, gerek faktör düzeylerinde, gerekse genel ortalamada en kötü performansı CV kriteri göstermiştir (bak. Tablo 5.6-5.7).

Örneklem hacimleri artması durumunda simülasyon sonuçlarında nasıl bir değişme olacağını görmek amacıyla 400 ve 500 birimlik örneklemlerin simülasyon sonuçları da incelendi, fakat 200 ve 350 birimlik örneklem sonuçlarına benzer olduğundan bu çalışmada yer verilmedi. Ayrıca bu konuda Mammadov, Yüzer ve Aydın [64] tarafından benzer bir çalışma 4.istatistik kongresinde sunulmuştur.

5.4.3. Simülasyon Sonuçların Oransal Olarak Değerlendirilmesi

Deneysel bir çalışma olarak gerçekleştirilen simülasyonda, her bir faktör düzeyi için farklı büyüklükte (25, 50, 100, 150, 200 ve 350) altı örneklem ve farklı sayılarda (100, 200, 350 ve 500) tekrarlamalar yapmak suretiyle söz konusu seçim yöntemlerinin, örneklem hacimleri ve tekrarlanma sayılarına göre nasıl bir performans izledikleri gözlenmiştir.Simülasyonda toplam 288 sayısal deney yapılmış ve bu deneylerde, söz konusu altı düzeltme parametresi seçim yöntemlerinin MSE medyan değerlerinin Wilcoxon testi sıralanmalarının ortalamalarından elde edilen, 24 ortalama performans Tablosu (6 farklı örneklem ve 4 farklı sayıda tekrarlama için toplam 24 ortalama) hesaplanmıştır. Söz konusu seçim yöntemlerinin ortalamalar düzeyinde birinci ve ikinci en iyi sıralamayı gösteren başarı durumları Tablo 5.8'de özet olarak verilmiştir.

Tablo 5.8'de parantez dışındaki sayılar, yöntemlerin kaç kez birinci olduklarını, parantez içindeki sayılar ise kaç kez ikinci olduklarını göstermektedir. Örneğin, gürültü düzeyi verileri için en iyi tahmin veren λ parametresinin seçiminde, GCV kriteri 21 kez birinci ve 3 kez ikinci olmuştur. Aynı örneklem verileri için RCP kriteri hiç birinci olamazken, sadece 5 kez ikinci olmuştur. Benzer olarak, genel ortalamada GCV kriteri 19 kez birinci ve 3 kez ikinci olurken, RCP kriteri sadece 1 kez ikinci olmuştur. Bu bilgiler doğrultusunda, yapılan simülasyonda çalışmasında altı seçim yönteminin söz konusu en iyi λ düzeltme parametresini seçme başarı oranları ise, Tablo 5.9'da verilmiştir.

Tablo 5.9'da görülen parantez dışındaki sayılar, yöntemlerin birinci olma oranlarını, parantez içindeki sayılar ikinci olma oranlarını göstermektedir. Örneğin,

Tablo 5.8: Yöntemlerin ortalama düzeyinde başarı durumları (birinci ve ikinci olma sayıları)

Kriterler	Gürültü Düzeyi	Uzaysal Değişim.	Varyans Fonk.	Genel Ortalama
AIC_c	5 (3)	8 (1)	10 (-)	8 (-)
GCV	21 (3)	22 (1)	21 (-)	19 (3)
CV	1 (7)	- (7)	2 (8)	- (4)
Cp	5 (3)	7 (1)	7 (2)	3 (4)
RCP	- (5)	- (2)	2 (1)	- (1)
LRS	4 (14)	1 (10)	2 (9)	- (16)

gürültü düzeyi faktörü etkisi altında kalan örneklem verileri için, söz konusu düzeltme parametresinin seçiminde, GCV kriterinin birinci olma oranı %87.5 ve ikinci olma oranı %12.5'tir. Buna karşılık, aynı örneklem verileri için RCP kriterinin birinci olma şansı yokken, fakat %20.8 oranında ikinci olmuştur.

Tablo 5.9: Yöntemlerin ortalama düzeyinde başarı oranları (birinci ve ikinci olma oranları)

Kriterler	Gürültü Düzeyi	Uzaysal Değişim	Varyans Fonk.	Genel Ortalama
AIC_c	0.208 (0.125)	0.333 (0.042)	0.417 (0.000)	0.333 (0.000)
GCV	0.875 (0.125)	0.917 (0.042)	0.875 (0.000)	0.792 (0.125)
CV	0.042 (0.292)	0.000 (0.292)	0.083 (0.333)	0.000 (0.167)
Cp	0.208 (0.125)	0.292 (0.042)	0.292 (0.083)	0.125 (0.167)
RCP	0.000 (0.208)	0.000 (0.083)	0.083 (0.042)	0.000 (0.042)
LRS	0.167 (0.583)	0.042 (0.417)	0.083 (0.375)	0.000 (0.667)

SONUÇ VE ÖNERİLER

Semiparametrik ve parametrik olmayan regresyon modellerinin kestiriminde, splayn düzeltme yöntemi kullanılmış olup, söz konusu yöntem esas itibariyle, cezalı en küçük kareler toplamının minimun problemine dayanır. Böyle bir probleminin çözümünde, bölüm 5.3'de incelenen 6 seçim kriterlerinden her hangi birine göre seçilen bir $\lambda > 0$ düzeltme parametresi ve bu parametreye bağlı bir S_λ düzeltme matrisi gerekir. Söz konusu λ parametresinin seçimi çok önemlidir. Çünkü λ parametresi 0'dan $+\infty$'a değişirken, çözüm interpolasyondan basit bir doğrusal modele değişir. Eğer $\lambda = \infty$ alınırsa, cezalı kareler denklemi sabit eğimli doğrusal regresyon uyumu üretir, buna karşılık $\lambda = 0$ alınırsa tümüyle esnek eğimli bir interpolasyon uyumuna karşı gelir. Bu nedenle istenen çözüm ne doğrusal ne de bir interpolasyon olmalıdır.

Parametrik olmayan ve semiparametrik regresyon modellerinin kestiriminde kullanılan splayn düzeltmenin esasını oluşturan cezalı en küçük kareler yöntemi, böyle bir düzeltme parametresine sahip olması avatajına göre en küçük kareler regresyonundan çok daha iyi sonuç vermektedir. Bölüm 4'te teorik olarak incelenen semiparametrik regresyonu bir uygulama ile desteklemek amacıyla bölüm 4.7'de, evlerin satış fiyatlarına ilişkin bir semiparametrik regresyon modelinin kestirimi, Speckman ve kısmi splayn adı altında iki farklı yaklaşıma göre yapılmıştır. Ayrıca, parametrik olmayan regresyonda iyi bir tahmin sonucu veren düzeltme parametresinin şeçimi için de bir simülasyon çalışması yapılmış olup, özet olarak aşağıdaki sonuçlar elde edilmiştir:

İlk olarak yapılan uygulmada, 1987 yılında Kanada'nın başkenti Ottawa'da satılan 92 müstakil evin satış fiyatı ve evlerin karakteristiklerini gösteren değişkenler arasındaki ilişkiler araştırılmıştır. Bu araştırmada, hem parametrik hemde semiparametrik regresyon analizi yapılmıştır. Elde edilen sonuçlara göre, evlerin satış fiyatlarını AU değişkeni ters yönde etkilerken, kalan tüm değişkenler aynı yönde etkilemiştir. Semiparametrik regresyon modeli ile yapılan parametre tahminlerinin doğrusal regresyon modeli ile yapılan parametre tahminlerinden çok daha üstün olduğu görülmüştür. Diğer bir ifadeyle, semiparametrik regresyon modeli, bağımlı değişkendeki değişimlerin çok önemli bir kısmı açıklarken, parametrik regresyonla kıyaslanmayacak derecede az hata yapmıştır. Ayrıca,

semiparametrik regresyon modelinin parametrelerin aralık kestirimleri, yine parametrik regresyon modelinin aralık kestirimlerine göre oldukça dar bir aralıkta yer almıştır.

Bir diğer uygulama olarak gerçekleştirilen simülasyon çalışmasında, gürültü düzeyi, uzaysal değişim ve varyans fonksiyonu faktörlerinin etkileri altında oluşturulan örneklem verilerinde, söz konusu faktörlerin etkilerini belirleyebilmek için toplam 288 sayısal deney yapılmıştır. Bu sayısal deneyler, her bir faktör düzeyi için farklı sayıda tekrarlama ve faklı büyüklükteki örneklemlerden oluşmaktadır. Söz konusu sayısal deneyleri oluşturan veriler, parametrik olmayan ve semiparametrik regresyonun deneysel uygulamalarında yaygın olarak kullanılan veri seti modellerinden rassal olarak yaratılmıştır.

Simülasyon yoluyla elde edilen 25-150 birimlik örneklemlerde, GCV kriteri, gürültü düzeyi, uzaysal değişim ve varyans fonksiyonu faktörlerinin etkileri altında düzgün olarak birincilik sıralamasında en iyi seçim yöntemi olurken, ikincilik sıralamasında en iyi yöntem LRS kriteri olmuştur. Buna karşılık, aynı hacimlik örneklemlerde, AIC_c, Cp ve RCP kriterleri en kötü performans gösteren yöntemler olmuşlardır. Fakat artan örneklem hacmine bağlı olarak diğer bir ifadeyle, 200 birim ve daha büyük örneklemlerde, bu kriterlerden AIC_c, ve Cp ‘nin performanslarının iyileştiği görülmüştür.

Simülasyon geneline bakıldığında, tüm faktör düzeylerinde ve genel ortalamada, birincilik sıralaması yüksek oranda bir başarı göstergesiyle GCV seçim kriteri, bunu takiben GCV’nin elde ettiği başarıdan çok daha düşük oranlarda AIC_c,ve Cp seçim kriterleri birincilik sıralamasında yer almıştır. Ayrıca, tüm sayısal deneylerin sonucunda, ikincilik sıralamada en yüksek başarı oranı göstergesiyle, en iyi seçim yöntemi LRS kriteri olmuştur. Diğer taraftan, 288 sayısal deneyin büyük çoğunluğunda en kötü performans gösteren yöntemlerin RCP ve CV seçim kriterleri olduğu görülmüştür.

Özet olarak, simülasyon çalışması sonucunda aşağıdaki gözlemler ortaya konmuştur:

- Genel ortalamada AIC_c kriteri 25-150 birimlik örneklemlerin simülasyon sonuçlarında en kötü performansa sahipken, 200-350 birimlik örneklemlerde iyi bir performans göstermiştir.

- Simülasyona genel olarak bakıldığında, faktör düzeyleri ve genel ortalamada en iyi başarıyı GCV kriteri sağlamıştır.
- Yine simülasyon geneline bakıldığında, faktör düzeylerine göre ve genel ortalamada ikincilik sıralamada en iyi başarıyı LRS kriteri sağlamıştır.
- Tüm simülasyon deneylerinde iki klasik metot AIC_c ve Cp kriterleri, tüm faktör düzeylerinde çok yakın sonuçlar vermişlerdir.
- Gürültü düzeyli faktörü için GCV kriterinden sonra AIC_c ve Cp kriterleri, en iyi performansla aynı sıralamayı paylaşmışlardır.
- Heterokadastik hata faktörü altında, GCV en yüksek başarı oranı alırken, CV, RCP ve LRS kriterleri, düşük bir başarı oranı ile aynı sıralamayı paylaşmışlardır.
- Tüm sayısal deneyin büyük çoğunluğunda, RCP ve CV kriterleri en kötü performansa sahip olmuşlardır.

Bu durumda yukarıdaki sonuçlara göre, tavsiyemiz şu şekildedir: Tüm faktör düzeylerinde ve genel ortalamada, birinci sıralamada en yüksek başarı oranına sahip GCV kriteri, 150 birimden büyük olan örneklemlerde GCV'ye ilaveten AIC_c,ve Cp kriterlerini, bunun yanı sıra, 288 sayısal deneye sonucunda, ikinci sıralamada en yüksek başarı oranı gösteren LRS kriterini kullanmayı öneriyoruz.

KAYNAKLAR

[1] HARBRECHT, W., *Determinants of One-Family House Price in the Detroit Area: An Econometric Analysis Based on the Hedonic Price Approach,* Thomas Kick, M.A (Wayne State Uni.), (2002).

[2] HARDLE, W., MÜLER, M., SPERLICH, S. VE WERWATZ, A., *Nonparametric and semiparametric Models,* Springer, New York (2004).

[3] J. VAN HEERDE, H., S.H. LEEFLANG, P. VE R. WITTINK, D., *Semiparametric Analysis to Estimate the Deal Effect Curve,* Journal of Marketing Research, (2001).

[4] HARDLE, W. VE LINTON, O., *Applied Nonparametrics Methods,* Handbook of Econometrics, Edited By R. F. Engle ve D .L Fadden, Elsevier Science B.V, (1994).

[5] HARDLE, W., *Applied Nonparametric Regression,* Cambridge University Press, Cambridge (1991)

[6] FOX, J., *Nonparametric Simple Regression: Smoothing Scatterplots,* Sage Publications, California, USA (2000).

[7] RUST, RONALD , T., *Flexible Regression,* Journal of Marketting Research, **25**, 10-24 (1988).

[8] HARDLE, W., LİANG, H. VE GAO, J., *Partially Linear Models,* Springer, Heidelberg (2000).

[9] YATCHEW, A., *semiparametric Regression for the Applied Econometrician,* Cambridge University Pres (2003).

[10] WALKER, E. VE WRİGHT S. P., *Comparing Curves Using additive Models , Journal of Quality Technology,* **34** (1), 118-129 (2002).

[11] GREEN, P.J. VE SILVERMAN, B.W., *Nonparametric Regression and Generalized Linear Models,* Chapman &Hall, London (1994).

[12] LOADER, C., *Soomthing: Local Regression Principles,* Handbook of Computational Statistics, Ed., J.Gentle, W. Hardle, Y. Mori (2004). http://www.herine.net/stat/papers.html

[13] LOFTSGAARDEN, D.O. VE QUESENBERY, G.P., *A Nonparametric Estimate of a Multivariate Density Function,* Annals of mathematical Statistics, **36**, 1049-1051 (1965).

[14] COVER, T. M VE HART, P. E., *Nearest Neighbour Pattern Clasifiction*, IEEE Transaction on Information Theory, **13**, 12-17 (1967).

[15] SLAMA, R., *Using Nonparametric and Semiparametric Regression in Epidemiology*, England et al., Am. J. Epidemiol, 154 (2001).

[16] NADARYA, E.A., *On Estimating Regression*, Theory Prob. Appl., **10**, 186-190 (1964).

[17] WATSON, G. S., *Smooth Regression Analysis*, Sankhya, Series A, **26**, 359-372 (1964).

[18] WAND, M,. P. VE M. C. JONES, *Kernel Smoothing*, New York: Chapman and Hall (1995).

[19] FAN, J. VE I. GIJBELS, *Local Polynomial Modelling and Additivity in Nonparametric Regression*, Annals of Statistics, **23**, 1896-1920 (1996).

[20] LEE, T. C. M., *Regresion F Smoothing Using the Minimum Description Lenght Principle*, Statistics & Probability Letters, **48**, 71-72 (2000).

[21] WAND, M. P., *A Comparision Regression Spline Smoothing Procedures*, Computational Statistics, **15**, 443-462 (2000).

[22] DE BOORE, C., *A practical Guide to Spline*, New York, Spring Verlag (1978).

[23] MATHEWS, H.,J., *Numerical Methods, for Mathematics*, Science and Engineering, Prentice-Hall İnternational,İnc (1978).

[24] AL-KHAFAJİ, WADİ, AMİR VE TOOLEY R. JHON., *Mumerical Methods in Engineering Practice*, CBS Publishing Japan Ltd. New York (1986).

[25] FAUSETT V.L. VE WHEATLY, O.P., *Applied Numerical Analysis Using MATLAB*, Prentice Hall (1999).

[26] GERALD, F. C. VE WHEATLY, O. P., *Applied Numerical Analysis, Sixth Edition*, Addıson-Wesley (1999).

[27] WAHBA, G., *Spline Models Of Observational Data*, University Of Winconsin At Madison, Pensilvenya (1990).

[28] KOU, S C., *On the Effiency of Selection Criteria in Spline regression*, Probab. Theory Relat. Fields **127**, 153-176 (2003).

[29] EUBANK, L. R., *Spline Regression,* Smoothing and Regression: Approaches, Computation and Application, Edited By Micheal G. Schimek, (2000).

[30] REINCH, C., *Smoothing By Spline Functions,* Numer. Math., **10**, 177-183 (1967).

[31] YATCHEW, A., *Nonparametric Regression Techniques in Economics*, Journal of Economic Literature, **XXXVI**, 669-721 (1988).

[32] KOOP, G. VE POIRIER J.D., *Bayesian Variants of Some Classical Semiparametric regression Techniquues,* Journal of Econometrics, **123**, 259-282 (2004).

[33] GREEN, P. J., JENNISON, C. VE SEHEULT, A., *Analysis of Field Experiments By Least Square Smoothing* , J. Roy. Statis. Soc. B, **47**, 299-315 (1985).

[34] ENGLE, R.F., GRANGER, C.W.J., RICE, C.A. VE WEISS A., *Semiparametric Estimates of the Relation Between Weahter and Electricity Sales,* Journal of Amer. Statist., Assoc., **81**., 310-320 (1986).

[35] SPECKMAN, P., *Kernel Smoothing in Partially Linear Model,* J. Royal Statist., Soc. B., **50**, 413-436 (1988).

[36] ROBINSON, M. P., *Root-N-Consistent Semiparametric Regression,* Econometrica, **56**, **4**, 931-954 (1988).

[37] CHEN, H., *Covergance Rates for Parametric Components in a Partially Linear Models,* Annals of Statistics, **16**, 136-146 (1988).

[38] BREIMAN, L. E., VE FRIEDMAN, J. H., *Estimating Optimal Transformation for Multiple Regression and Correlation (With Discussion),* J. Amer. Statist. Assoc., **80**, 580-619 (1985).

[39] BUJIA, A., HASTIE, T. J., VE TIBSHIRANI, R. J., *Linear Smoother and Additive Models (With Discussion),*Ann. Statist., **17**, 81-89 (1989).

[40] SCHIMEK,G. MICHAEL., *Estimation and İnference in Partially Linear Models with Smoothing Splines,* Journal of Statistical Planning and Inference, **91**, 525-540 (2000).

[41] EUBANK, R.L, KAMBOUR, E.L., KIM, T.C., KIPPLE, K, REESE, S.C. VE SCHIMEK, M., *Estimation in Partiallly Linear Models,* Computational Statistics &Data Analysis, **29**, 27-34 (1998).

[42] RICE, J., *Covergance Rates for Partially Spline Models,* Statis. Prob. Lett., **4**, 203-208 (1986)

[43] GASSER, T., SROKA, L. VE JENNEN-STEINMETZ, C., *Residual Variance Residual Pattern in nonlinear Regression,* Biometrika, **73**, 625-633 (1986).

[44] HASTIE, T. VE TIBSHIRANI, R.J., *Generalized Additive Models,* Chapman & Hall, London (1990).

[45] BOWMAN, A. W. VE AZZALINI, A., *Applied Smoothing Techniques for Data Analysis. The Kernel Approach with S-Plus İllustration,* Claredon Press, Oxford (1997).

[46] RAZ, J., Testing for no *Effect When estimating a Smooth Function By Nonparametric Regression: A Randomization Approach,* J. Amer. Statist. Assoc., **85**, 132-138 (1990)

[47] HONG, Y. VE WHİTE, H., *Consistent Spesifications Testing Via Nonparametric Series Regression,* Econometrica, **63**, 1133-1159 (1995).

[48] RUPPERT D., WAND, M.P. VE CARROLL R.J., *Semiparametric Regression,* Cambridge University Pres (2003).

[49] WAHBA, G., *Smoothing Noisy Data By Spline Function,* Numer. Math., **24**, 383-393 (1975)

[50] CRAVEN, P. VE WAHBA, G., *Smoothing Noisy Data By Spline Function,* Numer. Math., **24**, 383-393 (1979).

[51] EUBANK, R. L., *Nonparametric Regression and Smoothing Spline,* Marcel Dekker İnc., New York (1999).

[52] RİCE, J., *Bandwidth Choice for Nonparametric Regression,* Ann. Statist., **12**, 1215-1230 (1984).

[53] BUCKLEY, M. J., EAGLESON, G. K. VE SILVERMAN, B. W., *The Estimation of Residual Variance in Nonparametric Regression,* Biometrika, **75**, 183-199 (1988).

[54] WAHBA, G. VE WOLD, S., *A Completely Automotic French Curve: Fitting Spline Function By Cross-Validation,* Comunication in Statistics, **4**, 1-17 (1975).

[55] ALLEN, D., *The Relationship Between Variable Selection and Data Augmention and A Method for Prediction,* Technometrics, **16**, 125-127 (1974).

[56] CRAVEN, P. VE WAHBA, G., Smoothing Noisy Data With Spline Functions:Estimating the Correct Degree of Smoothing By the Method of the Generalized Cross-Validation, Numer. Math., **31**, 377-403 (1979).

[57] HURVICH C.M., AND SİMONOFF J.S., AND TASI C.L., *Smoothing Parameter Selection in Nonparametric Regreession Using An İmproved Akaike İnformation Criterion*, J.R. Statist. Soc. B., **60**, 271-293 (1988).

[58] LEE, T. C. M., *Smoothing Parameter Selection For Smoothing Splines: A Simulation Study,* Comput. Statistisc & Data Analysis, **42**, 139-148 (2003).

[59] LEE, T. C. M., *İmproved Smoothing Spline Regression By Combining Estimate Of Different Smoothness*, Statistisc Probabilit Letters (2003).

[60] MALLOWS, C., *Some comments on C_p,* Techometrics, **15**, 661-675 (1973).

[61] KOU, S. C. VE EFRON, B., *Smoothers and the C_p, Generalized Maximum Likelihood, and Extended Exponential Criteria: A Geometric Approach*, JASA, **97**, 766-782 (2002).

[62] LEE, T.M.C. VE SOLO, V., *Bandwidth Selection for Local Linear Regression: A Simulation Study,* Comput. Statist. **14**, 515-532 (1999).

[63] LEE, T. C. M., *A Stabilized Bandwitdth Selection Method for Kernel Smoothing of Periodiagram,* Signal Process, **81**, 419-430 (2001).

[64] MAMMADOV, M. YÜZER, A. F., VE AYDIN, D., *Splayn Düzeltme Regresyonu ve Düzeltme Parametresinin Seçimi,* 4. İatatsitik Kongresi bildiri ve poster özetleri kitabı, Belek- Antalya, 148-149 (2005).

EK-1: Semiparametrik regresyon analizi için algortima

Adım 1: Analizde kullanılan değişkenlere ilişkin veriler girilir.

Adım 2: Parametrik olmayan değişkenin değerleri faklı ve sıralı halde yeniden düzenlenir.

Adım 3: Parametrik olmayan değişkenin sıradan değerleri satırlarda, farklı ve sıralı değerleri sütünlarda gösterilmek suretiyle, bölüm 4.2.1'de tanımlanan N -tekraralanma matrisi elde edilir.

Adım 4: Parametrik olmayan değişkenin farklı ve sıralı değerlerine göre, bölüm 3.2'de tanımlanan Q, R ve $K = QR^{-1}Q^T$ matrisleri elde edilir.

Adım 5: λ düzeltme parametresi için l tane değer girilir.

Adım 6: Her bir l değeri için $S_\lambda = \mathrm{N}(\mathrm{N}^{\mathrm{T}}\mathrm{N} + \lambda K)^{-1}\mathrm{N}^{\mathrm{T}}$ düzeltme matrisi hesaplanarak, bağımlı ve parametrik değişkenlerde dönüşüm sağlanır. Dömüşümü yapılan değişkenlere göre (4.20) yada (4.29) matrisleri elde edilir.

Adım 7: Adım 5'te tanımlanan her bir l değeri için (4.20) veya (4.29) matrisleri, bölüm 5.3'te tanımlanan $\mathrm{AIC_c}$, GCV ve CV seçim kriterlerinde kullanılarak, söz konusu bu kriterlerin minimum değerleri bulunur. Her bir seçim kriterini minimum yapan l değeri λ düzeltme parametresi olarak seçilir ve bu parametreye uygun (4.20) yada (4.29) matrisleri dikkate alınır.

Adım 8: Adım 7'de tanımlanan seçim kriterlerinden herhangi birine göre seçilen λ parametresini kullanılarak, diğer bir deyişle λ_p pilot düzeltme parametresi belirlenerek, (5.11)'de verilen varyans kestiricisi veya (4.30)'da verilen Gasser'in varyans kestiricisi hesaplanır.

Adım 9: Adım 8'de elde edilen sonuçlar yardımıyla, Adım 7'de verilen işlemler bu kez $\mathrm{C_p}$, RCP ve LRS kriterleri için yapılır.

Adım 10: Her bir seçim kriterinin minimum yapan l değerine göre, adım 6'da verilen düzeltme matrisi yardımıyla dönüşümü yapılan bağımlı ve bağımsız değişkenlere göre, semiparametrik modelin (4.18) veya (4.27)'de tanımlanan parametrik katsayıları ve (4.19)'da verilen parametrik olmayan fonksiyonun aldığı değerler vektörü elde edilir.

EK-1: Devam

Adım 11: Adım 7'da tanımlanan (4.20) veya (4.29) matrisleri yardımıyla belirlenen bağımlı değişkeninin tahmini ve gerçek değerleri kullanılarak, semiparametrik modelin R^2 belirlilik katsayısı, $R^2 = \hat{\mathbf{y}}^T \hat{\mathbf{y}} \,/\, \mathbf{y}^{\mathrm{T}} \mathbf{y}$ formülü ile hesaplanır.

Adım 12: Semiparametrik modelin $\hat{\boldsymbol{\beta}}$ parametrik katsayılarının anlamlılığı (4.33) formülü ile verilen, $\mathrm{df} = n - tr\left(S_\lambda\right) - k$ (serbestlik derecesi) ile t-dağılımına sahip bir test istatistiği yardımıyla değerlendirilir. Diğer taraftan yapılacak bir F testini için (4.35)'de tanımlanan formül kullanılır.

Adım 13: Adım 10'da elde edilen $\hat{\boldsymbol{\beta}}$ katsayılar vektörünün $100(1-\alpha)\%$ güven aralığı (4.36)'da verilen formül ile belirlenir.

Adım 14: Parametrik olmayan fonksiyonun aldığı değerler vektörünün $100(1-\alpha)\%$ güven aralığı, (4.38)'de verilen formül yardımıyla belirlenir.

Adım 15: Adım 15'te gerçekleştirilen fonksiyonu biçimsel olarak değerlendirmek için (4.37) formülü ile tanımlanan F-test istatistiği kullanılır.

Adım 16: Parametrik olmayan fonksiyonun aldığı değerler vektörüne ait grafik çizdirilir.

EK-2: Splayn düzeltme regresyonu ve düzetme parametresinin seçimi konusunda yapılan simülasyon çalışması için algoritma

Adım 1: $x=(i-0.5/n)$, $i=1,...,n$ şeklinde tanımlanan parametrik olmayan x değişkeninin n örneklem hacimi ve i'ye bağlı değerleri elde edilir.

Adım 2: Parametrik olmayan değişkenin farklı ve sıralı değerlerine göre, bölüm 3.2'de tanımlanan $(n, n-2)$ *boyutlu* Q, $(n-2, n-2)$ *boyutlu* R ve $K=QR^{-1}Q^{T}$ matrisleri elde edilir.

Adım 3: λ düzeltme parametresi için l tane değer girilir.

Adım 4: Deneyde örneklem oluşturmada kullanılan verileri elde etmek için Tablo 5.1'de genel formda verilen modeller kullanılır.

Adım 5: Her bir l değeri için (3.35)'de verilen $S_{\lambda}=(\mathrm{I}+\lambda K)^{-1}$ düzeltme matrisi değerlendirilip, $\mathbf{f}=(f(x_1),...,f(x_n))$ splayn düzeltme kestiricisine karşı gelen değerler vektörü, (3.34) veya (3.36) förmülleri yardımıyla elde edilir.

Adım 6: Adım 3'te tanımlanan her bir l değeri için $S_{\lambda}=(\mathrm{I}+\lambda K)^{-1}$ matrisi hesaplanır. Bu matris yardımıyla bölüm 5.3'te tanımlanan AIC_c, GCV ve CV seçim kriterlerin minimum değerleri bulunur. Her bir seçim kriterini minimum yapan l değeri, λ düzeltme parametresi olarak seçilir ve bu parametreye uygun $S_{\lambda}=(\mathrm{I}+\lambda K)^{-1}$ düzeltme matrisleri dikkate alınır.

Adım 7: Her bir seçim kriterini minimum yapan l değerine göre Adım 6'da hesaplanan $S_{\lambda}=(\mathrm{I}+\lambda K)^{-1}$ düzeltme matrisleri yardımıyla adım 5'te gerçekleştirilen splayn düzeltme kestiricileri elde edilir.

Adım 8: Adım 6'da tanımlanan seçim kriterlerinden herhangi birine göre seçilen λ parametresini kullanılarak, diğer bir deyişle λ_p pilot düzeltme parametresi belirlenerek, (5.11)'de verilen varyans kestiricisi hesaplanır.

Adım 9: Adım 8'de elde edilen sonuçlar yardımıyla, Adım 6 ve 7'de verilen işlemler bu kez yine bölüm 5.3'te tanımlanan C_p, RCP ve LRS kriterleri için yapılır.

EK-2: Devam

Adım 10: Her bir seçim kriteri için hata kareler ortalaması hesaplanır.

Adım 11: Herbir seçim kriterinin hata kareler ortalamasının logaritması (log MSE) ve medyan değerleri hesaplanır.

Adım 12: MSE medyan değerleri küçükten büyüğe doğru sıralanır.

Adım 13: Herhangi iki seçim metodun MSE değerleri meydanı arasındaki farkın anlamlı olup olmadığı, %5 anlam düzeyinde eşleştirilmiş Wilcoxon işaretli sıra testi ile değerlendirilir.

Adım 14: Tablo 5.1'de verilen her bir faktör için tipik bir benzetim veri dizisiyle gerçek regresyon fonksiyonunun grafiği ve AIC_c, GCV, CV, Cp, RCP ve LRS kriterlerinin $\log_e$MSE değerlerin kutu grafikleri çizdirilir.