

T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

PARAMETRİK OLMAYAN REGRESYONDA
DÜZELTME PARAMETRESİ SEÇİMİ: BİR
SİMÜLASYON ÇALIŞMASI

YÜKSEK LİSANS TEZİ

AKIN FİT

TEMMUZ 2012

MUĞLA

T.C.
MUĞLA SITKI KOÇMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

PARAMETRİK OLMAYAN REGRESYONDA
DÜZELTME PARAMETRESİ SEÇİMİ: BİR
SİMÜLASYON ÇALIŞMASI

YÜKSEK LİSANS TEZİ

AKIN FİT

TEMMUZ 2012
MUĞLA

MUGLA SITKI KOÇMAN ÜNİVERSİTESİ

Fen Bilimleri Enstitüsü

TEZ ONAYI

AKIN FİT tarafından hazırlanan **PARAMETRİK OLMAYAN REGRESYONDA DÜZELTME PARAMETRESİ SEÇİMİ: BİR SİMÜLASYON ÇALIŞMASI** başlıklı tezinin, 27/07/2012 tarihinde aşağıdaki jüri tarafından İstatistik Anabilim Dalı'nda yüksek lisans derecesi için gerekli şartları sağladığı oybirliği/oyçokluğu ile kabul edilmiştir.

TEZ SINAV JURİSİ

Doç. Dr. Dursun AYDIN (**Danışman ve Jüri Başkanı**)

İmza:



İstatistik Anabilim Dalı, Muğla Sıtkı Koçman Üniversitesi, Muğla

Yrd. Doç. Dr. Ummahan ACAR (**Üye**)

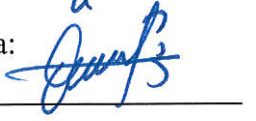
İmza:



Matematik Anabilim Dalı, Muğla Sıtkı Koçman Üniversitesi, Muğla

Yrd. Doç. Dr. Öznur İŞÇİ (**Üye**)

İmza:



İstatistik Ana Bilim Dalı, Muğla Sıtkı Koçman Üniversitesi, Muğla

ANA BİLİM DALI BAŞKANLIĞI ONAYI

Doç. Dr. Dursun AYDIN

İmza:



İstatistik Ana Bilim Dalı Başkanı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

Doç. Dr. Dursun AYDIN

İmza:



Danışman, İstatistik Anabilim Dalı,
Muğla Sıtkı Koçman Üniversitesi, Muğla

Savunma Tarihi: 27/07/2012

Tez alıřmalarım sırasında elde ettiėim ve sunduėum tm sonu, dokman, bilgi ve belgelerin tarafımdan bizzat ve bu tez alıřması kapsamında elde edildiėini; akademik ve bilimsel etik kurallarına uygun olduėunu beyan ederim. Ayrıca, akademik ve bilimsel etik kuralları gereėi bu tez alıřması sırasında elde edilmemiř bařkalarına ait tm orijinal bilgi ve sonulara atıf yapıldıėını da beyan ederim.

Akın Fit

27/07/2012

ÖZET
PARAMETRİK OLMAYAN REGRESYONDA DÜZELTME PARAMETRESİ
SEÇİMİ: BİR SİMÜLASYON ÇALIŞMASI

Akın FİT

Yüksek Lisans Tezi

Fen Bilimleri Enstitüsü

İstatistik Anabilim Dalı

Danışman: Doç. Dr. Dursun AYDIN

Temmuz 2012, 78 sayfa

Bu tezde parametrik olmayan regresyonda düzeltme yöntemleri ele alınmıştır. Parametrik olmayan regresyon modellerin kestiriminde bir uygulama ile Kernel düzeltme kullanılmıştır. Bu yöntemin gerçekleşmesinde 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 parametresi seçtiğini belirlemek amacıyla, TURBO C ortamında yazılan bir program yardımıyla bir simülasyon çalışması yapılmıştır.

Anahtar Kelimeler: Parametrik Olmayan Regresyon, Kernel Düzeltme, Düzeltme Parametresi, Seçim Kriterleri, Düzeltme Yöntemleri

ABSTRACT
**SELECTION OF SMOOTHING PARAMETER IN NONPARAMETRIC
REGRESSION: A SIMULATION STUDY**

Akın FİT

Master of Science (M.Sc.)
Graduate School of Natural and Applied Sciences
Department of Statistics
Supervisor: Doç. Dr. Dursun AYDIN
July 2012, 78 pages

In this thesis, smoothing methods in nonparametric regression taken into account. Kernel smoothing with an application are used for estimation of nonparametric 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 TURBO C environment.

Keywords: Nonparametric Regression, Smoothing Kernel, Smoothing Parameter, Selection Criteria, Smoothing Methods

Sevgili Aileme

ÖNSÖZ

Bu tezin ana hatlarının belirlenmesinde değerli katkılarını esirgemeyen danışmanım Doç. Dr. Dursun AYDIN' a şükranlarımı sunarım.

Ayrıca çalışmamın tüm aşamalarında beni destekleyen değerli anne ve babama minnettar olduğumu belirtmek isterim.

İÇİNDEKİLER

ÖNSÖZ	vii
İÇİNDEKİLER	viii
ÇİZELGELER DİZİNİ.....	x
ŞEKİLLER DİZİNİ.....	xi
SEMBOLLER VE KISALTMALAR DİZİNİ	xiii
1. GİRİŞ.....	1
2. REGRESYON MODELLERİ	3
2.1. Parametrik Regresyon	4
2.2. Yarı Parametrik Regresyon	7
2.3. Parametrik Olmayan Regresyon.....	8
3. PARAMETRİK OLMAYAN REGRESYONDA DÜZELTME KAVRAMI VE DÜZELTME YÖNTEMLERİ	11
3.1. Parametrik Olmayan Regresyonda Düzeltme Kavramı	11
3.1.1. Cezalı en küçük kareler yaklaşımı	13
3.1.1.1. Bir eğrinin pürüzlülüğünü ölçme	13
3.1.1.2. Cezalı en küçük kareler fonksiyonu	14
3.2. Parametrik Olmayan Regresyonda Düzeltme Yöntemleri	15
3.2.1. Splayn düzeltme yöntemi	16
3.2.1.1. Splayn düzeltme tahmin edicisinin elde edilmesi	17
3.2.2. Yerel regresyon düzeltici	19
3.2.3. k -En yakın komşu tahmin edici	20
3.2.4. Kernel (çekirdek) düzeltici	22
3.3. Doğrusal Düzelticilerin İstatistiksel Özellikleri	25
3.3.1. Hata kareler ortalaması	25
3.3.2. Serbestlik derecesi	27
3.4. Hata Varyansının Tahmini	28
3.4.1. Hata kareler yaklaşımı	28
3.4.2. Yerel fark alma yaklaşımı	29
4. KERNEL YÖNTEMİNDE DÜZELTME PARAMETRESİ SEÇİM YÖNTEMLERİ	32
4.1. Çapraz Geçerlilik Ölçütü (CV)	33

4.2. Genelleştirilmiş Çapraz Geçerlilik Ölçütü (<i>GCV</i>)	34
4.3. Mallows'un C_p Ölçütü (C_p)	35
4.4. Geliştirilmiş Akaike Bilgi Ölçütü (AIC_c)	35
4.5. Klasik Pilotları Kullanan Risk Tahmini (<i>RECP</i>)	36
4.6. Sınırlı maksimum olabilirlik ölçütü (<i>REML</i>)	37
4.7. Rice'in T ölçütü (T)	38
4.8. Shibata'nın seçtiği bir model SH ölçütü (SH)	38
5. SİMÜLASYON DENEYİ	38
5.1. Veriler ve Deney Düzenekinin Oluşturulması	39
5.2. Deneysel Değerlendirmeler	40
5.3. Sonuçlar ve Öneriler	71
KAYNAKLAR	73
EKLER	76
Ek A. Kernel Düzeltme ve Kernel Düzeltme Parametresinin Seçimi Konusunda Yapılan Simülasyon Çalışması için Algoritma	76
Ek B. Mevcut X,Y Grafikleri, Kutu Grafikleri ve Wilcoxon Test İstatistikleri Sonuçlarının Elde Edilmesi	77
ÖZGEÇMİŞ	78

ÇİZELGELER DİZİNİ

Çizelge 3.1. Kernel fonksiyonları	23
Çizelge 5.1. Fonksiyonel kalıplar	40
Çizelge 5.2. 1.faktör ve $r=1$ için Wilcoxon test sonuçları	41
Çizelge 5.3. 1.faktör ve $r=6$ için Wilcoxon test sonuçları	42
Çizelge 5.4. $n=50$ 'lik örnekleme 8 düzeltme paramatresi için ortalaması alınan Wilcoxon testi sıralamaları	68
Çizelge 5.5. $n=75$ 'lik örnekleme 8 düzeltme paramatresi için ortalaması alınan Wilcoxon testi sıralamaları	69
Çizelge 5.6. $n=100$ 'lik örnekleme 8 düzeltme paramatresi için ortalaması alınan Wilcoxon testi sıralamaları	70

ŞEKİLLER DİZİNİ

Şekil 3.1. Doğrusal regresyon ve interpolasyon	12
Şekil 5.1. $n=50$ ve $r=1$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği	44
Şekil 5.2. $n=50$ ve $r=2$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği	44
Şekil 5.3. $n=50$ ve $r=2$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği	45
Şekil 5.4. $n=50$ ve $r=4$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	45
Şekil 5.5. $n=50$ ve $r=5$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	46
Şekil 5.6. $n=50$ ve $r=6$ düzeyinde 1. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	46
Şekil 5.7. $n=50$ ve $r=1$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	47
Şekil 5.8. $n=50$ ve $r=2$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	47
Şekil 5.9. $n=50$ ve $r=3$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	48
Şekil 5.10. $n=50$ ve $r=4$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	48
Şekil 5.11. $n=50$ ve $r=5$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	49
Şekil 5.12. $n=50$ ve $r=6$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri .	49
Şekil 5.13. $n=75$ ve $r=1$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	50
Şekil 5.14. $n=75$ ve $r=2$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	51
Şekil 5.15. $n=75$ ve $r=3$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	51
Şekil 5.16. $n=75$ ve $r=4$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	52
Şekil 5.17. $n=75$ ve $r=5$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	52
Şekil 5.18. $n=75$ ve $r=6$ düzeyinde 2. faktöre ilişkin X 'e karşın Y 'nin grafiği.....	53
Şekil 5.19. $n=75$ ve $r=1$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	53
Şekil 5.20. $n=75$ ve $r=2$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	54
Şekil 5.21. $n=75$ ve $r=3$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	54

Şekil 5.22. $n=75$ ve $r=4$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	55
Şekil 5.23. $n=75$ ve $r=5$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	55
Şekil 5.24. $n=75$ ve $r=6$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	56
Şekil 5.25. $n=100$ ve $r=1$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	56
Şekil 5.26. $n=100$ ve $r=2$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	57
Şekil 5.27. $n=100$ ve $r=3$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	57
Şekil 5.28. $n=100$ ve $r=4$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	58
Şekil 5.29. $n=100$ ve $r=5$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	58
Şekil 5.30. $n=100$ ve $r=6$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği	59
Şekil 5.31. $n=100$ ve $r=1$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	59
Şekil 5.32. $n=100$ ve $r=2$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	60
Şekil 5.33. $n=100$ ve $r=3$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	60
Şekil 5.34. $n=100$ ve $r=4$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	61
Şekil 5.35. $n=100$ ve $r=5$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	61
Şekil 5.36. $n=100$ ve $r=6$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	62
Şekil 5.37. $n=100$ ve $r=1$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	62
Şekil 5.38. $n=100$ ve $r=2$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	63
Şekil 5.39. $n=100$ ve $r=3$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	63
Şekil 5.40. $n=100$ ve $r=4$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	64
Şekil 5.41. $n=100$ ve $r=5$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	64
Şekil 5.42. $n=100$ ve $r=6$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği	65

Şekil 5.43. $n=100$ ve $r=1$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	65
Şekil 5.44. $n=100$ ve $r=2$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	66
Şekil 5.45. $n=100$ ve $r=3$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	66
Şekil 5.46. $n=100$ ve $r=4$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	67
Şekil 5.47. $n=100$ ve $r=6$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	67
Şekil 5.48. $n=100$ ve $r=6$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri	68

SEMBOLLER VE KISALTMALAR DİZİNİ

f	Düzgün fonksiyon
f''	f fonksiyonunun 2. türevi
$f \in C^2 \ a, b$	İlk 2 türevi $[a, b]$ kapalı aralığında sürekli olan tüm f fonksiyonlarının kümesi
df	Serbestlik derecesi
λ	Düzeltilme parametresi (bant genişliği)
λ_p	Düzeltilme parametresinin pilot tahmini
S_λ	Parametrik olmayan regresyon için şapka matrisi
W_λ	Kernel düzeltilme için şapka matrisi
AIC_c	Geliştirilmiş Akaike bilgi ölçütü
C_p	Mallow' un C_p ölçütü
$tr(A)$	A matrisinin izi
CV	Çapraz geçerlilik
E(RSS)	Hata kareler toplamının beklenen değeri
EKK	En küçük kareler
EDF	Eşdeğer serbestlik derecesi
GCV	Genelleştirilmiş çapraz geçerlilik
K	Kernel fonksiyonu
k-NN	k -En yakın komşu tahmin edici
MSE	Hata kareler ortalaması
PRESS	Ön tahmin hata kareler toplamı
PS	Pürüzlülük cezası
R	Risk fonksiyonu
RECP	Klasik pilotları kullanan risk tahmini
REML	Sınırlı maksimum olabilirlik ölçütü
RSS	Hata kareler toplamı
SH	Shibata'nın seçtiği bir model SH ölçütü
T	Rice'in T ölçütü

1. GİRİŞ

İstatistik biliminin en önemli konularından birini regresyon analizi oluşturmaktadır. Regresyon analizinin temelinde; gözlenen bir olayın değerlendirilirken, hangi olayların etkisi içinde olduğunun araştırılması yatmaktadır. Bu olaylar bir veya birden çok olacağı gibi dolaylı veya direkt etkileniyor da olabilirler. Regresyon analizi yapılırken, gözlem değerlerinin ve etkilenilen olayların bir matematiksel gösterimle yani bir fonksiyon yardımıyla ifadesi gerekmektedir. Kurulan bu modele regresyon modeli denilmektedir. Doğrusal regresyon analizi bağımlı ve bağımsız değişkenler arasındaki ortalama ilişkinin matematiksel bir fonksiyonla ifade edilmesinde, bağımsız değişkenlerle bağımlı değişkenin doğrusal bir ilişki içerisinde olduğunu varsayar. Regresyon analizi yapılırken kurulan matematiksel modelde yer alan değişkenler bir bağımlı değişken ve bir veya birden çok bağımsız değişkenden oluşmaktadır. Bağımsız değişkenler kurulacak modelde bir değişkenli olarak ele alınırsa, basit doğrusal regresyon modelini, birden fazla bağımsız değişkenli olarak alınırsa, çoklu regresyon modeli konusunu oluşturmaktadır.

Doğrusal regresyon analizi bazı varsayımlara dayanır. Bu varsayımların en önemlisi bağımlı ve bağımsız değişkenler arasındaki ilişkinin şeklinin biliniyor olmasıdır. Varsayımların sağlanmadığı durumlarda yapılan tahminler iyi bir tahmin olma niteliğine sahip olamazlar. Bu durumda daha iyi tahmin yapabilmek amacıyla parametrik regresyondaki doğrusallık varsayımının esnetilmesine olanak sağlayan regresyon yöntemlerine ihtiyaç duyulur. Bu yöntemler parametrik olmayan ve yarı parametrik regresyon yöntemleri olarak bilinen regresyon modelleridir (Aydın, 2005).

Parametrik olmayan regresyon modellerinde, parametrik regresyon modellerine kıyasla her hangi bir matematiksel forma ihtiyaç duyulmadan tahminler gerçekleştirilmektedir. Parametrik olmayan regresyon modellerinde ana kütle regresyon eğrisinin tahmini için regresyon eğrisine ait matematiksel kalıbın herhangi bir form varsayılmadan doğrudan tahmin edilmesi “düzeltme” adı verilen teknikler ile gerçekleştirilmektedir.

Düzeltilme tekniklerinin yanı sıra parametrik olmayan regresyon modelleri için ana kütle regresyon eğrisinin tahmininde büyük önem taşıyan en uygun “Düzeltilme Parametresi”nin seçimi ve en uygun düzeltme tekniğine karar verilmesiyle ilgili ölçütler bulunmaktadır.

Yarı parametrik regresyon modeli için var olan yaklaşımların tümü farklı parametrik olmayan regresyon yöntemlerine bağlıdır. Yarı parametrik regresyon modelleri karmaşık veri kümelerini bizim anlayabileceğimiz biçimde özetleyip, uygulamada verilerin önemsiz detaylarını göz ardı ederken önemli olan özelliklerini muhafaza eder ve böylece sağlam kararlar verilmesini sağlarlar (Ruppert vd., 2003).

Yarı parametrik regresyon modeli, parametrik ve parametrik olmayan regresyon fonksiyonun toplamsal olarak birleşiminden oluşması nedeniyle böyle modellere, kısmi doğrusal regresyon modelleri de denir. Yarı parametrik regresyon, parametrik değişkenlerin etkilerinin sıfır olması ya da bu tür değişkenlerin analizde yer almadığı durumlarda parametrik olmayan regresyon olarak işlem görür. Ayrıca yarı parametrik 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 (Aydın, 2005).

Bu çalışma beş bölümden oluşmaktadır. İlk giriş bölümünde tezin içeriği ve uygulama alanı için yapılan incelemeler gözden geçirilmiştir. Regresyonda, parametrik regresyon, parametrik olmayan regresyon ve yarı parametrik regresyon hakkında temel bilgiler verilmiştir.

İkinci ve üçüncü bölümde parametrik regresyon, parametrik olmayan regresyon ve yarı parametrik regresyon olmak üzere üç farklı regresyon yaklaşımı ile birlikte, regresyonda düzeltme kavramı ve parametrik olmayan regresyonda düzeltme yöntemleri incelenmiştir.

Dördüncü bölümde parametrik olmayan regresyon modellerinde düzeltme parametresinin seçimi için, çapraz geçerlilik ölçütü, genelleştirilmiş çapraz geçerlilik ölçütü, geliştirilmiş Akaike bilgi ölçütü ve Mallows’un C_p ölçütü, pilot tahminlerin

seçimini gerektiren risk tahmin yöntemi, sınırlı ML ölçütü, Rice'ın T yöntemi ve Shibata'nın seçtiği model SH yöntemleri ele alınmıştır.

Beşinci bölümde dördüncü bölümdeki düzeltme parametresi seçim yöntemlerinin hangisinin daha iyi bir sonuç verdiğini belirlemek amacıyla bir simülasyon çalışması yapılmıştır. Simülasyon çalışmasının gerçekleştirilmesinde TURBO C programı kullanılmıştır. Parametrik olmayan düzeltme yöntemlerinden Kernel Düzeltmesi ele alınmıştır. Kernel Düzeltme kestiricisinin kalitesini değerlendirmek için MSE (hata kareler ortalaması) performans ölçü kriteri olarak alınmıştır. Adı geçen performans kriterine göre, düzeltme parametresi seçim yöntemleri başarı sıralamasına tabi tutularak en iyi yöntem ve dolayısıyla en iyi düzeltme parametresi belirlenmiştir.

2. REGRESYON MODELLERİ

Regresyon, mevcut değişkenlerden birinin bağımlı (açıklanan) ve diğerlerinin bağımsız (açıklayıcı) olması durumunda, bağımlı değişkenin (y) bağımsız değişkenlerin ($x_1, x_2, \dots, x_k$) bir fonksiyonu olarak ifade edilmesidir. Bir fonksiyonel bağıntı olan regresyon analizi, bağımsız değişkenlerdeki değişmelerin bağımlı değişkeni hangi yönde ve hangi oranda etkilediğini belirtir. Regresyon analizinin asıl amacı, değişkenler arasındaki fonksiyonel bağıntıyı en iyi şekilde ifade edecek matematiksel denklemi kurmak ve bu denklemi bağımlı değişkenin değerlerinin tahmin edilmesinde ve istatistiksel analizde kullanmaktır. Regresyon analizi ($x_1, x_2, \dots, x_k$) bağımsız değişkenlerine göre (y) bağımlı değişkenin koşullu ortalamasının fonksiyonel bağımlılığını ortaya koyar. Bilinen klasik doğrusal regresyon modeli,

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

biçiminde ifade edilir. Burada, ε_i , 0 (sıfır) ortalamalı, sabit σ^2 varyanslı ve bağımsız değişkenlerden bağımsız bir hata terimidir. Ayrıca, $x_1, x_2, \dots, x_k$ değişkenlerine göre y 'nin koşullu ortalaması, bu değişkenlerin doğrusal bir fonksiyonudur (Aydın, 2005).

Genellikle böyle bir doğrusal model çok faydalı olmasına rağmen, verilerin doğrusal olmaması halinde, bu tür model uygun olmaz. Bazen bu problem, bağımsız değişkenlerde dönüşüm yapılarak atlatılabilir. Bununla birlikte, uygulamada veriler göz önüne alınarak en uygun fonksiyonel şekli tahmin etmek zordur. Ancak parametrik olmayan regresyonla, bağımsız değişkenlere göre bağımlı değişkenin fonksiyonel bağımlılığını belirtmeksizin verilere uygun modeller oluşturulabilir. Buna karşın; 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 yarı parametrik regresyon daha uygun sonuçlar verir (Aydın, 2005).

Bu bölümde bağımsız değişkenlere göre bağımlı değişkeninin fonksiyonel bağıllığını ortaya çıkarmada parametrik regresyon, yarı parametrik regresyon ve parametrik olmayan regresyon olmak üzere üç farklı regresyon gösterilmiştir.

2.1. Parametrik Regresyon

Parametrik regresyon bağımlı ve bağımsız değişken veya bağımsız değişkenler ile bu değişkenler arasındaki ortalama ilişkinin matematiksel bir fonksiyonla ifade edilmesidir. Parametrik regresyon, regresyon fonksiyonunun $(x_1, x_2, \dots, x_k)$ bağımsız değişkenlerinin bir doğrusal fonksiyonu olarak yazılabildiğini varsayar. $E(y|X)$ koşullu beklenen değeri X biliniyorken y 'nin ortalama dağılımının X ile fonksiyonel ilişkisini gösterir. Başka bir ifade ile $X = (x_1, x_2, \dots, x_k)$ bağımsız değişkenlerindeki değişime karşılık y bağımlı değişkeninin ortalama tepkisini ifade eder ve:

$$E(y|X) = X\beta \quad (2.2)$$

biçiminde veya,

$$y = X\beta + \varepsilon \quad (2.3)$$

biçiminde yazılır. Burada $\varepsilon = y - E(y|X)$ ifadesi $E(y|X)$ koşullu beklenen değerinin y 'den sapması olarak tanımlanır (Hardle vd., 2004).

Model (2.3)'de gözlem sayısı n , bağımsız değişken sayısı k olmak üzere; y , $(n \times 1)$ boyutlu bağımlı değişken vektörü, X , $(n \times (k+1))$ boyutlu bağımsız değişkenlerin gözlem değerlerini içeren bir matrisi, β , $(k+1)$ boyutlu bilinmeyen regresyon

katsayıları vektörü, ε ise 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 (2.3) regresyon modeline uyumuna göre, en iyi β değerlerini bulmaktır. Bu bakış açısından, (2.3) 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 (Hardle vd., 2004).

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: Birincisi, 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 yanlış 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, uygulamalarda parametrik tahmin için gerekli olan güçlü sınırlamaları (parametrik model varsayımları) yüklemenin maliyeti de dikkate alınmalıdır (van Heerde vd., 2001).

Parametrik bir modelin en önemli varsayımlarını şu şekilde sıralamak mümkündür;

- Hata terimleri rassaldır ve beklenen değeri 0'dır ($E(u_i) = 0$).

- Hata terimlerinin varyansları gözlemden gözleme değişmemektedir. Hata terimlerine ait varyanslar sabittir ($E(u_i^2) = \sigma_u^2$).
- Hata terimlerinin birbirini takip eden değerleri arasında ilişki yoktur ($KOV(u_i, u_j) = 0, \quad i \neq j$).
- Hata terimleri ile bağımsız değişkenler arasında ilişki olmadığı varsayılmaktadır ($KOV(u_i, x_{ij}) = 0$).

Burada doğrusallık kavramına ilişkin birtakım açıklamalarda bulunmakta fayda vardır. Parametrik yöntemlerle elde edilen regresyon denklemlerinde, ilişkilerin doğrusal varsayılması, bağımsız değişkenlerin önünde bulunan katsayıların doğrusal olması anlamına gelmektedir. Parametrik modellerde yer alan bağımlı değişken ve bağımsız değişkenlerin, doğrusal olması şartı aranmamaktadır. Parametrik regresyon denklemlerine, aşağıda yer alan denklemleri örnek olarak gösterilebilir (Hepsağ, 2007):

$$y_i = \beta_0 + \beta_1 x_i \quad (2.4)$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \quad (2.5)$$

$$y_i = \beta_0 + \beta_1 \sqrt{x_i} \quad (2.6)$$

$$y_i = \beta_0 + \beta_1 \left(\frac{1}{x_i} \right) \quad (2.7)$$

$$y_i = \beta_0 x(x_i^{\beta_1}) x(z_i^{\beta_2}) \quad (2.8)$$

Yukarıda gösterilen tüm denklemler katsayıları itibariyle doğrusaldır. Bu örnek denklemlerden parametrik modellerde varsayılan doğrusallık varsayımının ne olduğu anlaşılmaktadır. Bağımsız değişkenlerin üssel, parabolik, ters ya da kareköklü olarak ifade edilmesi söz konusu parametrik modellerin doğrusallık özelliğini etkilememektedir. Söz konusu doğrusallık ilişkisi, deneye dayalı bir yöntem olan serpilme(dağılma) diyagramları ile gözlemlenebilmektedir. Koordinat sistemi içerisinde; ordinat kısmında yer alan bağımlı değişken değerlerine karşı, apsis

kısmına analizde kullanılması düşünölen bağımsız değışken değeri tek tek yerleřtirilerek söz konusu iki değışkene ait iliřkinin yönü ve boyutu tespit edilmeye çalışılmaktadır.

Parametrik modellere ait tahmin yapılırken sıklıkla kullanılan metot olan En Küçük Kareler Yöntemi (EKK) dir. Bu doğrusal varsayımı dikkate alarak parametre tahminlerini gerçekleştirir. EKK parametrik modellere ait eğilimsiz, tutarlı, etkin ve doğrusal parametre tahmin ediciler elde edilmesinde uygun bir yöntem olarak uygulamalı çalışmalarda kullanılmaktadır. EKK'nin temel amacı,

$x_i y_i$ $\sum_{i=1}^n$ ($i=1, \dots, n$) olarak elde edilmiş gözlemler içerisinde doğrusal bir model geçirerek, gözlemlerin bu doğruya olan uzaklığı arasındaki farkın karesi toplamını minimize etmeye çalışmaktır $\sum \hat{u}_i^2 \Rightarrow \min$ (Hepsağ, 2007).

2.2. Yarı Parametrik Regresyon

Yarı parametrik regresyon modellerinde, bağımlı değışken bir ya da daha fazla açıklayıcı değışkenle doğrusal olarak ilişkili fakat diğer ek değışken ya da değışkenlerle ilişkisinin kolayca parametreleştirilmediğı varsayılır. Değışkenlerin skaler olduğu ve f fonksiyonunun parametrik bir aile içerisinde bulunmadığı bilinen yarı parametrik regresyon modeli;

$$\begin{aligned} y &= E(y|Z, x) + \varepsilon \\ &= Z\beta + f(x) + \varepsilon \end{aligned} \quad (2.9)$$

biçiminde ifade edilir. Model (2.9)'da 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 $E(\varepsilon|Z, x) = 0$ ve $Var(\varepsilon|Z, x) = \sigma_\varepsilon^2$ olduğu başka bir anlatımla, ε hata terimlerinin sıfır ortalamalı ve σ^2 varyansı ile istatistiksel bağımsız ve normal olarak dağıldığı ve $Z = (z_1, \dots, z_k)$ bağımsız değışkenlerinin koşullu ortalaması $X = (x_1, \dots, x_k)$ kestirici değışkenlerinin pürüzsüz bir fonksiyonu olduğu varsayılır (Yatchew, 2003).

Model (2.9)'da, β parametre vektörünü ve $E(y|Z, x) = Z\beta + f(x)$ ortalama vektörünü etkin olarak tahmin etmesinin yanı sıra, uyumun sayısal olarak özetlemesi ve grafiksel olarak görüntülemesine de olanak sağlarlar.

Karşılaşılan bir problemde çok sayıda bağımsız değişkene bağlı olan bir y değişkeni gözlemlendiğini varsayalım. Çoklu doğrusal regresyonda bu bağıllığın doğrusal olduğu varsayılır ve genel doğrusal modellerin bilinen teorileri tahminde kullanılır. Yarı parametrik 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 bağıllığı korumaktır (Aydın, 2005).

2.3. Parametrik Olmayan Regresyon

(x_i, y_i) , $i = 1, 2, \dots, n$ gözlem değerleri verildiğinde, $x = (x_1, x_2, \dots, x_k)$ bağımsız değişkenlerinin ve y bağımlı değişkeninin ölçümleri arasındaki ilişkiyi açıklayan en yaygın yöntem, $E(y|x) = f(x)$ koşullu beklenen fonksiyonunu tahmin etmektir. Bu ilişkiyi açıklayan genel parametrik olmayan regresyon modeli,

$$\begin{aligned} y &= E(y|x) + \varepsilon \\ &= f(x) + \varepsilon \\ &= f(x_1, \dots, x_k) + \varepsilon \end{aligned} \tag{2.10}$$

biçiminde ifade edilir. Eşitlik (2.10)'da f , parametreleri belirginleşmemiş bir fonksiyonel ilişkiyi gösteren bilinmeyen bir fonksiyondur, diğer bir ifadeyle, az sayıda parametre ile basit olarak özetlenemeyen, belirgin bir şekle sahip olmayan ve bağımsız x değişkenlerinin bir fonksiyonudur. Ayrıca (2.10)'da yer alan ε , rassal hata terimi olup $E(\varepsilon|x) = 0$ ve $Var(\varepsilon|x) = \sigma^2(x)$ koşulunu sağlayan bağımsız bir rassal değişkendir (Hardle ve Linton, 1994). Eşitlik (2.10) daki f ikinci mertebeden sürekli türeve sahip olan fonksiyonlar uzayının bir elemanıdır ve $f \in C^2$ a, b eşitliğin düzeltme kısmıdır. f fonksiyonunu tahmin etmek için düzeltme yöntemleri kullanılır. Eşitlik (2.10)'a 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 bağıllığın özel bir türünü gösterip göstermediğini anlatır (Hardle, 1991).

Parametrik olmayan regresyon analizinin amacı parametreleri tahmin etmekten çok, doğrudan f regresyon fonksiyonunu tahmin etmektir. Model (2.10)'un özel bir durumu, yalnızca bir x bağımsız değişkenin yer aldığı 'basit' parametrik olmayan regresyon modeli;

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n \quad (2.11)$$

biçimindedir. 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 (Fox, 2000).

Eşitlik (2.11) da $f \in C^2$ a, b olan bir düzgün fonksiyon, x_i parametrik olmayan bağımsız değişkene ait gözlem değerleri, y_i bağımlı değişkene ait gözlem değerleri, ε_i , 0 ortalamalı ve σ^2 sabit varyanslı, bağımsız özdeş olarak dağılan rassal hata terimleridir.

Parametrik olmayan yaklaşımla bir regresyon ilişkisini belirlemeye çalışmanın 4 avantajı bulunmaktadır. Bunlar,

- Parametrik olmayan tahmin yöntemi bağımlı ve bağımsız değişken arasındaki genel ilişkinin araştırılmasında çok yönlü ve esnek metot sağlamaktadır.
- Herhangi bir fonksiyonel forma ihtiyaç duyan sabit parametrik modeli temel almaksızın gözlemlere ait tahminler yapar.
- Uç değer olarak tabir edilen değerlerin etkisinin analizde yer alması sebebiyle “sahte” gözlemleri bulmak için bir araç niteliği taşımaktadır.
- Bağımsız değişkenler arasında interpolasyon yapılabilmesi ve veri seti içerisinde kayıp gözlemlerin yerine konulabilmesi açısından esnek bir metot oluşturmaktadır (Hepsağ, 2007).

Parametrik olmayan regresyon yönteminin tahmin yaparken kısıtlayıcı varsayımları olmamasına rağmen bazı sakıncaları vardır. Bağımsız değişken sayısı fazla olduğu zaman tahmin yapmak zor olmakta ve elde edilen grafikler karmaşık bir yapıda olmaktadır. Bu durum ‘boyutluluk sorunu’ olarak adlandırılır. Ayrıca parametrik olmayan yöntemle kesikli bağımsız değişkenleri dikkate almak ve bağımsız değişken sayısındaki artışa bağlı olarak y değişkenine ait bireysel etkileri yorumlamak zor olmaktadır.

Parametrik olmayan regresyon aykırı gözlemlerin bulunduğu veri setleri için önemli bir analiz yöntemidir. İstatistiksel çalışmalarda aykırı gözlemlerin etkilerini farklı biçimlerde ele alan güçlü (robust) parametrik yöntemler bulunmaktadır. Bununla birlikte, aykırı gözlemlerden dolayı parametreler bozulduğu için bu güçlü yöntemler bile uygun çözümler üretemeyebilir ve verinin gerçek yapısı modele yansıtılamaz. Bu durumda parametrik olmayan regresyon, ön bilgi sağlamaktadır (Hardle, 1994). Parametrik yaklaşımda çok fazla varsayım yapıldığından dolayı sonuçların güvenilirliği giderek azalmaktadır. Parametrik olmayan yaklaşımda ise hiçbir varsayım yapılmamakta fakat bağımsız değişken sayısının fazla olması durumunda model tahmininin elde edilmesi zor olmaktadır.

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 bağımsız değişken varsa, parametrik olmayan regresyon modelleri yanlış tahminler üretebilir ve onları yorumlamak zor olabilir (Hardle, 2004).

3. PARAMETRİK OLMAYAN REGRESYONDA DÜZELTME KAVRAMI VE DÜZELTME YÖNTEMLERİ

3.1. Parametrik Olmayan Regresyonda Düzeltme Kavramı

Parametrik olmayan regresyon analizinin amacı bilinmeyen yanıt fonksiyonu olan $f(\mathbf{x})$ 'in uygun analizini elde etmektir. Gözlenen hataların azaltılması ile $\mathbf{y}$ 'nin $\mathbf{x}$ 'e göre ortalama bağımlılığının önemli ayrıntılarını vermek yorumu kolaylaştırır. Bu eğri yaklaştırma işlemi genel olarak 'düzeltme (smoothing)' olarak adlandırılır (Hardle, 1991). Düzeltme fikrinin temelinde verileri bir eğriye uydurmak ve daha basit fonksiyonların birleşimi olabilen esnek fonksiyonları kullanmak yatar.

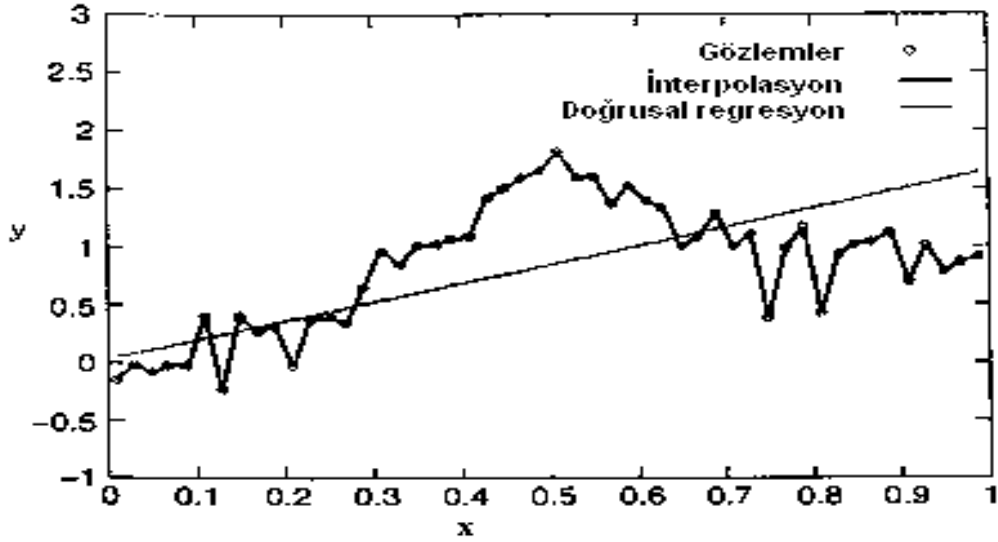
'Düzeltici (smoother)'; $\hat{f}(\mathbf{x}) = \sum_{i=1}^n w_{xi} y_i$ ise $x_1, x_2, \dots, x_k$ olmak üzere bir ya da birden

fazla bağımsız değişkenin fonksiyonu olan $\mathbf{y}$ bağımlı değişkenin sahip olduğu eğilimi ifade etmek için kullanılan bir araçtır ve bağımlı değişkenin kendisinden daha az değişken olan bir eğriyi tahmin etmeyi amaçlamaktadır. Bir düzelticinin en önemli özelliği, değişkenler arasındaki ilişkinin şeklini kesin bir biçimde belirlememesidir (bir doğru gibi) ve bu özelliğinden dolayı parametrik olmayan regresyonda sık kullanılan bir araçtır (Hastie ve Tibshirani, 1990). Parametrik regresyonda elde edilen doğru kesin parametrik bir biçime sahip olduğu için düzeltici değildir. Düzeltme yönteminin sonuç değişkeni yani bir düzeltici tarafından oluşturulan tahmin 'düzgün (smooth)' olarak adlandırılır (Hardle, 1994).

f regresyon fonksiyonunu tahmin edecek en popüler klasik yöntemlerden biri doğrusal regresyon yaklaşımıdır. Kullanılan bu teknik, (x_i, y_i) , $i = 1, \dots, n$, veri dizisi için $f(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$ fonksiyonuna göre;

$$RSS(f) = \sum_{i=1}^n (y_i - f(x_i))^2 \quad (3.1)$$

biçiminde ifade edilen hata (artık) kareler toplamını minimum yaparak, f fonksiyonunu tahmin eder. Denklem (2.9)'daki f fonksiyonu yaklaşık olarak doğrusalsa tahmin için bu yaklaşım etkin olabilir. Fakat f fonksiyonunu doğrusal değilse bu yaklaşım başarısız da olabilir. Böyle bir başarısızlık örneği Şekil 3.1'de verilen örnekte görülmektedir. Söz konusu şekilde görülen doğrusal regresyon uyumu, sadece değişkenler arasında bir ilişkinin varlığını ifade eder (Aydın, 2005).



Şekil 3.1. Doğrusal regresyon ve interpolasyon

Şekil 3.1'de görüldüğü gibi f regresyon fonksiyonunun diğer bir tahmini $(x_i, y_i), i=1, \dots, n$ verilerinin interpolasyonu ile elde edilebilir. 'İnterpolasyon' elde var olan (bilinen) değer noktalarından yola çıkarak, farklı bir yerde ve değeri bilinmeyen bir diğer noktadaki olası değeri tahmin etmeye yarayan yöntemlerin tümüne verilen genel isimdir. En basit tanımı ile var olan sayısal değerleri kullanarak, boş noktalardaki değerlerin tahmin edilmesi olarak açıklanır. Bu tahmin edici için doğrularla birleştirilen gözlemlerde bireysel eğimler ile sabit eğimlilik sağlanır. Bu tahmin edici için hata kareler toplamı $RSS(f)=0$ olur. Doğrusal regresyon uyumu verideki bilginin çok azını kullanırken, bu uyumda verideki bilgi daha fazla içerilir ancak verilerin yararlı bir özetini sağlamada başarısız olur ve daha önemlisi modelin rastgele hata kavramından kaynaklanan ve (2.9) denklemindeki regresyon fonksiyonuna bağlı olabilen verilerdeki özelliklerin veya esas eğilimin doyurucu bir şekilde açıklanmasını gerçekleştirmez. Bu durumda parametre

tahmininde doğrusal regresyon ve interpolasyon uyumlarının esnetilmesini sağlayan daha esnek yaklaşımların kullanılması 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 minimizasyonu dikkate alınmalıdır (Aydın, 2005).

3.1.1. Cezalı en küçük kareler yaklaşımı

En basit şekliyle cezalı en küçük kareler yaklaşımı, regresyon doğrusu boyunca klasik doğrusal regresyondaki model varsayımlarını esneten bir yöntemdir.

Cezalı en küçük kareler yaklaşımının esas amacı hızlı olarak dalgalanan bir eğrinin eğilimini ölçmek ve daha sonra eğri tahmininde sabit eğimli uyumlar ile esnek eğimli uyumlar arasında gerekli uzlaşmayı sağlayacak şekilde tahmin problemini ortaya koymaktır (Green ve Silverman, 1994).

Basit doğrusal regresyon Şekil 3.1’de görülen veriler için uygun değildir. Şekil 3.1’deki gözlem değerlerinden tamamıyla esnek eğimli uyumlar (interpolasyon) ile sabit eğimli uyumlar (bir doğru) arasında bir uzlaşmaya gereksinim olduğu görülmektedir. Bunu sağlamanın bir yolu, modele regresyon fonksiyonun eğimi ile bağlantılı olan ceza fonksiyonunun eklenmesidir (Aydın, 2005).

3.1.1.1. Bir eğrinin pürüzlülüğünü ölçme

$f \in C^2$ a, b olan bir f eğrisinin ne kadar pürüzlü ve dalgalı olduğunu ölçmenin birçok farklı yolu vardır. Bunlardan en çok kullanılanı, iki kez sürekli türevlenebilir f eğrisinin pürüzlülüğünü ölçmek için f fonksiyonunun ikinci türevinin karesinin integralini:

“ $\int_a^b f''(x)^2 dx$ ” almaktır. Bu ölçüme göre sadece doğrusal $f(x)$ fonksiyonları sıfır

pürüzlülüğe sahipken diğer C^2 a, b sınıfındaki tüm fonksiyonlar pozitif bir pürüzlülüğe sahiptir. Pürüzlülüğü hesaplamak için çeşitli yöntemler bulunmaktadır. Özellikle regresyon konusunda, modele bir sabitin veya doğrusal bir fonksiyonun eklenmesi pürüzlülüğün ölçümünü etkilemez. Öyle ki eğer iki fonksiyon sadece bir

sabit veya bir doğrusal fonksiyona göre farklı ise onların pürüzlülüğü benzer olmalıdır. Bu doğal olarak göz önüne alınan eğrinin ikinci türevine bağlı olan bir fonksiyonel pürüzlülük fikrine yol açar (Green ve Silverman, 1994).

3.1.1.2. Cezalı en küçük kareler regresyonu

İkinci mertebeden sürekli türeve sahip herhangi bir f fonksiyonu ve bir $\lambda > 0$ düzeltme parametresi ile cezalı kareler toplamı;

$$\begin{aligned} S(f) &= \sum_{i=1}^n y_i - f(x_i)^2 + \lambda \int_a^b f''(x)^2 dx \\ &= RSS(f) + \lambda J(f) \end{aligned} \quad (3.2)$$

olarak tanımlanır. Eşitlik (3.2)'den elde edilen cezalı en küçük kareler (CEKK) tahmin edici $\hat{f}$ iki kez türevlenebilen tüm f fonksiyonları sınıfında $S(f)$ fonksiyonunu minimum yapan değer olarak tanımlanır. Buradaki λ düzeltme parametresi belirlenmiş bir sabit olup, eğrinin pürüzlülüğünü ölçen $\int_a^b f''(x)^2 dx$ ile verinin uyumunu ölçen $\sum_{i=1}^n y_i - f(x_i)^2$ ifadeleri arasında bir uzlaşma sağlar ve esneklik üzerine konulan önemin bir ölçüsü olarak görülebilir. $S(f)$ 'nin minimize edilmesi uyum iyiliği ve düzeltme arasındaki en iyi uzlaşmayı verecektir. λ büyükse $S(f)$ 'deki 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 yaklaşacaktır ve $\hat{f}$ eğrisi Şekil 3.1'deki gibi bir doğrusal regresyon uyumuna yaklaşır. Eğer λ oldukça küçükse $S(f)$ 'ye ana katkı, hata kareler toplamı (RSS) olur ve $\hat{f}$ tahmin eğrisi veriyi yakından izleyecektir. Limit durumunda ise, λ sıfıra yaklaştığı için, $\hat{f}$ eğrisi Şekil 3.1 de gösterilen esnek eğimli uyum (interpolasyon) eğrisine yaklaşacaktır. Burada önemli olan nokta verileri en iyi temsil eden bir eğri tahmini elde etmek için en uygun λ değerinin seçimidir (Green ve Silverman, 1994).

λ değerinin seçimi ile ilgili olarak düzeltme parametresi seçim yöntemleri 4. bölümde ayrıntılı olarak ele alınmıştır.

3.2. Parametrik Olmayan Regresyonda Düzeltme Yöntemleri

Düzeltme yöntemleri (doğrusal düzelticiler) farklı ölçümler arasındaki fonksiyonel ilişkileri bulmaya yarayan yöntemlerdir. Parametrik regresyon çerçevesinde olduğu gibi verilerin bir ya da daha fazla bağımsız değişken ve bir bağımlı değişken ölçümlerinden oluştuğu varsayılır. Parametrik regresyon yöntemleri bağımsız ve bağımlı değişkenler arasındaki ilişkiyi tanımlamak için fonksiyonel bir şekil (bir doğru gibi) belirlerler. Düzeltme yöntemleri ise uydurulmuş eğrilerin şeklini belirlemek için kendileri veri noktaları sağlayan daha esnek bir yaklaşım sağlarlar (Loader, 2004). Bu bölümde en yaygın düzeltme yöntemlerinden bazıları ele alınacaktır. Ele alacağımız düzelticilerin hepsi,

$$\hat{f}(x) = \sum_{i=1}^n w_{xi} y_i \quad (3.3)$$

yazılışından dolayı doğrusal düzelticilerdir. Burada w_{xi} n vektörü x_i n vektörüne bağlı olarak hesaplanan ağırlık vektörüdür. Eşitlik (3.3)'deki $\hat{f}(x)$ regresyon tahmin edicisi 'düzeltici' olarak adlandırılır. Düzeltme yöntemlerin amacı $f(x)$ ortalama fonksiyonu için parametrik bir model belirlemek değil verileri sağlayacak uygun bir fonksiyonel şekil belirlemektir (Loader, 2004).

3.2.1. Splayn düzeltme yöntemi

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 bir yolu, $\int f''(x)^2 dx$ 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 yöntemi, $f \in C^2[a, b]$ uzayındaki tüm f fonksiyonları arasında;

$$S_\lambda(f) = \sum_{i=1}^n y_i - f(x_i)^2 + \lambda \int_a^b f''(x)^2 dx \quad (3.4)$$

cezalı kareler toplamının minimum yapılması ile elde edilir (Hardle, 2004). Diğer bir ifadeyle, splayn düzeltme kestiricisi olan $\hat{f}$ tahmin eğrisi $C^2[a, b]$ uzayının fonksiyonları arasında $S_\lambda(f)$ cezalı kriterini minimum yapan eğri olarak tanımlanır.

Eşitlik (3.4)'da ilk terim hata kareler toplamını (RSS) ifade etmektedir ve parametrik regresyonda olduğu gibi uyumun verilere yakınlığını ölçer, ikinci terim ise pürüzlülük cezasını gösterip pürüzlüğe bir ceza yükler yani fonksiyondaki eğriliği cezalandırır. İkinci terimde yer alan λ düzeltme parametresi olup 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 (3.4) denklemi sabit eğimli doğrusal regresyon uyumu üretir, buna karşılık $\lambda = 0$ alınırsa tümüyle esnek eğimli bir interpolasyon uyumuna karşı gelir (Eubank, 2000).

Modeli (3.4) şeklinde olan cezalı kareler toplamını minimum yapan f eğrisine 'splayn düzeltme tahmin edicisi' denir ve $x_1, x_2, \dots, x_n$ düğüm noktaları ile bir doğal kübik splayn olarak bilinir. $\hat{f}$ 'nın doğal kübik splayn özelliklerinin pürüzlülük cezasının özel seçiminden kaynaklanan özellikler olduğuna dikkat edilmelidir. Burada bahsedilen splayn düzeltme tahmin edicisi;

$$\hat{f}_\lambda = (I + \lambda K)^{-1} y = S_\lambda y \quad (3.5)$$

biçiminde gösterilir. (3.5)'de splayn düzelticinin y_i 'de bir doğrusal düzeltici olduğunu görebiliriz. Yani $\hat{f}_\lambda(x) = \sum_{i=1}^n w_{\lambda i}(x) y_i$ olacak şekilde $w_{\lambda i}(x)$ ağırlıkları vardır (Hardle, 2004).

Eşitlik (3.5)'de yer alan S_λ matrisine $x = (x_1, x_2, \dots, x_n)$ düğüm noktaları vektörü ile belirlenen bir 'düzeltme matrisi' denir ve doğrusal regresyondaki şapka matrisinin rolünü oynar.

3.2.1.1. Splayn düzeltme tahmin edicisinin elde edilmesi

Teorem 3.1. f ve γ vektörlerinin doğal kübik splayn belirtmesi için gerek ve yeter koşul aşağıdaki gibidir:

$$Q'f = R\gamma \quad (3.6)$$

(3.6) sağlanırsa bu durumda pürüzlülük cezası;

$$\int_a^b f''(x)^2 dx = \gamma' R \gamma = f' K f \quad (3.7)$$

biçiminde yazılabilir (İspat için bkz. (Green ve Silverman, 1994). Teorem 3.1 pürüzlülük cezasının (3.7) biçiminde yazılmasından dolayı splayn düzeltme yöntemi ile tahmin yapmada gerekli bir teoremdir.

Eşitlik (3.2) ile verilen cezalı kareler toplamının minimum yapılması ile elde edilen splayn düzeltme tahmin edicisi $\hat{f}$ 'nin doğal kübik splayn olarak bilinmesi çok önemlidir. Çünkü bu durumda sonsuz boyutlu C^2 a, b düzgün fonksiyonlar kümesinde minimum problemi, (x_i, y_i) gözlem noktaları ve λ düzeltme parametresine bağlı olarak hesaplanan doğal kübik splayn dikkate alınarak $S(f)$ 'nin minimum yapılması problemi ile yer değiştirmiş olur. Bu durumda $\hat{f}$ tahmin fonksiyonu splayn düzeltme tahmin edicisi olarak belirtilebilir. Doğrusal eşitliklerin bir kümesinin çözülmesi ile splayn eğrisinin minimum yapılması aşağıdaki gibidir:

f fonksiyonunun f ve γ vektörleri ve Q ve R matrisleri ile doğal kübik splayn olduğu varsayalım. Bu vektör ve matrislere göre (3.2) cezalı kareler toplamı yeniden yazılabilir. $y = (y_1, \dots, y_n)'$, gözlem değerleri vektörü olsun. x_i düğüm noktalarında $f(x_i)$ değerlerinin vektörü $f = (f_1, \dots, f_n)' = (f(x_1), \dots, f(x_n))'$ olduğundan f 'e göre verilen hata kareler toplamı;

$$RSS = \sum_{i=1}^n y_i - f(x_i)^2 = (\mathbf{y} - \mathbf{f})'(\mathbf{y} - \mathbf{f}) \quad (3.8)$$

biçiminde yazılabilir. Eşitlik (3.7)'dan dolayı $\int f''^2$ pürüzlülük ceza teriminin $\mathbf{f}'\mathbf{K}\mathbf{f}$ olarak ifade edilebileceğini biliyoruz. Buna göre, $\int_a^b f''(\mathbf{x})^2 dx = \mathbf{f}'\mathbf{K}\mathbf{f}$ olduğunu göz önünde bulundurarak, (3.2) ile verilen cezalı kareler toplamı;

$$\begin{aligned} S(f) &= \sum_{i=1}^n y_i - f(x_i)^2 + \lambda \int_a^b f''(\mathbf{x})^2 dx \\ &= (\mathbf{y} - \mathbf{f})'(\mathbf{y} - \mathbf{f}) + \lambda \mathbf{f}'\mathbf{K}\mathbf{f} \\ &= \mathbf{f}'(\mathbf{I} + \lambda\mathbf{K})\mathbf{f} - 2\mathbf{y}'\mathbf{f} + \mathbf{y}'\mathbf{y} \end{aligned} \quad (3.9)$$

biçiminde yeniden yazılır. f splayn düzeltme kestiricisi cezalı en küçük kareler toplamını minimum yapan eğri olarak belirtilmektedir.

$\mathbf{K}$ matrisi yarı-pozitif tanımlı ve $\lambda > 0$ olduğundan, $\lambda\mathbf{K}$ matrisi de yarı-pozitif tanımlı bir matris olacaktır. Dolayısıyla $(\mathbf{I} + \lambda\mathbf{K})$ kesin pozitif tanımlı matristir. Bu yüzden (3.9) kare formu tek bir minimumuma sahiptir. Buna göre, (3.9)'e minimum değer veren $\mathbf{f}$ vektörü, (3.9)'ün $\mathbf{f}$ 'e göre türev fonksiyonuna sıfır değerini veren vektördür:

$$\begin{aligned} S'(\mathbf{f}) &= 2\mathbf{f}'(\mathbf{I} + \lambda\mathbf{K}) - 2\mathbf{y}' = 0 \\ 2\mathbf{f}'(\mathbf{I} + \lambda\mathbf{K}) &= 2\mathbf{y}' \end{aligned} \quad (3.10)$$

Buna göre $\mathbf{f}$ vektörü;

$$\mathbf{f} = (\mathbf{I} + \lambda\mathbf{K})^{-1} \mathbf{y} = \mathbf{S}_\lambda \mathbf{y} \quad (3.11)$$

biçiminde bir tek minimum değeri elde edilir (Green ve Silverman, 1994).

3.2.2. Yerel (local) regresyon düzeltici

Bu yönteme göre bir regresyon fonksiyonu, herhangi bir x noktasının komşuluğunda düşük dereceli bir polinom (doğrusal bir fonksiyon) tarafından çok iyi tahmin edilebilir. Verilen bir x noktası için;

$$y_i = a(x) + b(x)x_i + hata, \quad x - \lambda \leq x_i \leq x + \lambda \quad (3.12)$$

modelini ele alalım. Burada $a(x)$ ve $b(x)$ iki yerel parametredir ve λ yerel komşulukların büyüklüğüdür. Ele alınan gözlemler eşit ağırlıklı ise, doğal olarak bu model aşağıdaki gibi yerel regresyon problemine götürür:

$$\sum_{i=1}^n y_i - a(x) - b(x)x_i \quad (3.13)$$

Model (3.13) ifadesini minimum yapan $a(x)$ ve $b(x)$ bulunarak, bir x noktasından elde edilen regresyon fonksiyonunun tahmini;

$$\hat{f}(x) = \hat{a}(x) + \hat{b}(x)x \quad (3.14)$$

olarak elde edilir (Fan ve Gijbels, 1996). Bağımsız değişkeninin tüm olası değerlerinin kümesindeki noktaların dizisinde bu işlem tekrar edilerek, f regresyon fonksiyonunun parametrik olmayan tahmini elde edilir. Bir yerel ikinci dereceden tahmin,

$$f(x_i) \approx a_0 + a_1(x_i - x) + \frac{a_2}{2}(x_i - x)^2 \quad (3.15)$$

biçiminde tanımlanır (Loader, 2004). p . dereceden genel bir polinom;

$$f(x_i) = \sum_{k=0}^p a_k \frac{(x_i - x)^k}{k!} \quad (3.16)$$

biçimindedir. Yerel tahmin yerel olarak ağırlıklandırılmış en küçük kareler ile uydurulabilir. Yani daha uzak gözlemlere daha düşük ve daha yakın gözlemlere daha yüksek ağırlıklar atayan ağırlıklı regresyon yapılabilir. Ağırlık fonksiyonu ve

düzeltilme parametresi kernel regresyonda olduğu gibi tanımlandığından, bu işlem bazen kernel regresyonu olarak da ifade edilir. 3. dereceden yerel bir polinom (3.16) eşitliğinde $p = 3$ alınır;

$$f(x_i) \approx a_0 + a_1(x_i - \mathbf{x}) + \frac{a_2}{2}(x_i - \mathbf{x})^2 + \frac{a_3}{6}(x_i - \mathbf{x})^3 \quad (3.17)$$

biçiminde yazılıp katsayı tahminleri;

$$\sum_{i=1}^n K\left(\frac{x_i - \mathbf{x}}{\lambda}\right) \left(y_i - \left(a_0 + a_1(x_i - \mathbf{x}) + \frac{a_2}{2}(x_i - \mathbf{x})^2 + \frac{a_3}{6}(x_i - \mathbf{x})^3 \right) \right)^2 \quad (3.18)$$

ifadesini minimum yapacak biçimde seçilir.

3.2.3. k -En yakın komşu tahmin edici (kNN)

Bu komşuluğun genişliği düzeltilme parametresi λ ve K kernel fonksiyonu tarafından yönetilir.

$$y_i = f(x_i) + \varepsilon_i, i = 1, \dots, n \quad (3.19)$$

modelinde $(x_1, y_1), \dots, (x_n, y_n)$ gözlem değerlerini göz önüne alalım. Burada x_i $i=1, \dots, n$ değişkeninin a, b 'de aldığı farklı değerleri gösterir. Ayrıca x_i noktasındaki f 'nin tahmin edicisi x_i 'ye en yakın k tane komşuya ait olan y_i değerlerinin ortalamasıdır. k -en yakın komşu tahmini değişen komşuluklarda ağırlıklı bir ortalama. Bu komşuluk öklid uzaklığındaki x_i 'nin k . en yakın komşular arasında bulunan $\mathbf{x}$ değişkenleri aracılığı ile tanımlanır.

Kuramda, $k - NN$ ağırlık dizisi, Loftsgaarden ve Quesenbery (1965) tarafından tanıtılmış ve sınıflandırma amaçları için Cover ve Hart (1967) tarafından kullanılmıştır. Biçimsel olarak, k -en yakın komşu tahmini;

$$\hat{f}_k(\mathbf{x}) = n^{-1} \sum_{i=1}^n w_{ki}(\mathbf{x}) y_i \quad (3.20)$$

olarak yazılabilir. Burada $w_{ki}(x) \prod_{i=1}^n$ ifadesi;

$$w_{ki}(x) = \begin{cases} 1/k, & i \in J_x \text{ ise} \\ 0, & \text{aksi durumda ise} \end{cases} \quad (3.21)$$

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

biçiminde belirtilen indeks dizisi ile tanımlanan bir ağırlık dizisidir. Burada k bu tahmin edicinin düzeltme parametresidir, tahmin edilen eğrinin pürüzlülüğünün derecesini düzenler ve kernel düzelticilerin düzeltme parametresine (bant genişliğine) benzer bir rol oynar. Bu yöntemle elde edilen f fonksiyonu sürekli. 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 (Slama, 2001).

Ağırlıkların oluşturulması hakkında fikir sahibi olabilmek için aşağıdaki örneği ele alalım. Burada $(x_i, y_i) \prod_{i=1}^5$ gözlemleri $\{(1,5), (7,12), (3,1), (2,0), (5,4)\}$ olsun ve $x = 4$ ve $k = 3$ için $\hat{f}_k(x)$ 'in k -NN tahminini hesaplayalım.

x 'e en yakın k gözlem son üç veri noktasıdır. Böylece $J_x = J_4 = 3, 4, 5$ ve bundan dolayı $w_{k1}(4) = 0$, $w_{k2}(4) = 0$, $w_{k3}(4) = 1/3$, $w_{k4}(4) = 1/3$, $w_{k5}(4) = 1/3$ olup $\hat{f}_k(x)$ 'in k -NN tahmini $k = 3$ ve $x = 4$ için (3.20) ile verilen eşitlikten, $\hat{f}_3(x = 4) = (1 + 0 + 4) / 3 = 5/3$ olarak elde edilir (Hardle, 1994).

Belirlenmiş bir k sabiti için genel olarak; Eğer veriler yoğun bir şekilde dağılmış ise komşuluklar arasındaki uzaklıklar küçük, tam tersi olarak veriler dağınık bir şekilde dağılmış ise komşuluklar arasındaki uzaklıklar büyük olur.

3.2.4. Kernel (çekirdek) düzeltme

Literatürde; Priestley ve Chao Kernel tahmin edicisi, Gasser-Müller Kernel tahmin edicisi, Mack ve Müller Kernel tahmin edicisi, Rosenblatt Kernel yoğunluk tahmin edicisi, Parzen Kernel yoğunluk tahmin edicisi, Ahmad-Lin Kenel tahmin edicisi ve Nadaraya Watson Kernel tahmin edicisi bulunmaktadır. Bu bölümde Kernel Düzeltme Nadaraya Watson tahmin edicisi ele alınmıştır.

Bu yöntemde bir x noktası ortalama fonksiyon $f(x)$ 'in tanım bölgesinde yer almıştır ve bir düzeltme penceresi bu noktanın etrafında belirlenmiştir. En sık kullanılan düzeltme penceresi basitçe $(x - \lambda, x + \lambda)$ aralığı olarak gösterilir. Kernel tahmini düzeltme penceresi içindeki gözlemlerin bir ağırlıklandırılmış ortalamasıdır ve aşağıdaki şekilde gösterilir:

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^n K\left(\frac{x_i - x}{\lambda}\right) y_i}{\sum_{j=1}^n K\left(\frac{x_j - x}{\lambda}\right) y_j} \quad (3.22)$$

Burada λ düzeltme penceresinin yarıçapına karşılık gelen ve bant genişliği olarak da bilinen bir düzeltme parametresi, n gözlem sayısı ve K seçilen kernel fonksiyonudur. Model (3.22) ile verilen kernel düzeltme aşağıdaki biçimde de gösterilebilir (Loader, 2004);

$$\hat{f}_\lambda(x) = \sum_{i=1}^n \left(\frac{K\left(\frac{x_i - x}{\lambda}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{\lambda}\right)} \right) y_i = \sum_{i=1}^n w_{xi} y_i \quad (3.23)$$

Eşitlik (3.23)'de w_{xi} $x - x_i$ uzaklığına bağlı i . gözlem y_i 'ye atanan ağırlık olarak yani i . gözlemin w ağırlığı, $x - x_i$ uzaklığının bir fonksiyonu olarak tanımlanır. Bu yöntemde x 'e bağlı w_{xi} ağırlıkları yardımı ile y_i bağımlı değişkenlerinin bir ağırlıklı ortalaması olarak x noktasında regresyon fonksiyonu tahmin edilir ve w_{xi} katsayıları,

$$w_i(x) = w_{xi} = \frac{K\left(\frac{x_i - x}{\lambda}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{\lambda}\right)} = \frac{K(u)}{\sum_{j=1}^n K(u)} \quad (3.24)$$

biçiminde olup $\sum_{i=1}^n w_{xi} = 1$ dir. Bir doğrusal düzeltici uygun olarak tanımlanmış w_{xi} ağırlıkları için (3.23)'teki gibi gösterilen bir düzelticidir. Bir kernel düzeltme her bir

hedef noktada tahmini üretmek için kernel olarak tanımlanan yerel ağırlıkların açıkça tanımlanmış bir kümesini kullanır. Bir kernel düzeltme hedef noktadan uzağa taşındığından dolayı düzgün bir biçimde azalan ağırlıkları kullanır. Genellikle uzaklık küçükse ağırlık yüksek ve uzaklık büyükse ağırlık düşük olur. Ağırlıklar K tarafından belirtilir ve düzeltme parametresi (bant genişliği) olarak bilinen λ tarafından kontrol edilir. Eşitlik (3.23)'de kernel tahmini bazen 'Nadaraya-Watson' tahmini olarak da adlandırılır (Hardle, 2004). Bu yaklaşım ağırlıkları açıkça belirtmek için K kernel fonksiyonunu kullandığından dolayı bir 'kernel düzeltme' olarak adlandırılır. Kernel fonksiyonları aşağıdaki özelliklere sahip fonksiyonlardır:

- 1) $K(u) \geq 0, \quad \forall u$ için
- 2) $\int_{-\infty}^{+\infty} K(u) du = 1$
- 3) $K(-u) = K(u)$ simetriktir.

Bu özellikler aynı zamanda simetrik bir olasılık yoğunluk fonksiyonunun özellikleridir. Uygulamada kullanılan bazı kernel fonksiyonları Çizelge 3.1'de verilmiştir;

Çizelge 3.1. Kernel fonksiyonları

Fonksiyon adı	Fonksiyon
Epanechnikov	$\frac{3}{4}(1-u^2)I(u \leq 1)$
Üçgensel (Triangular)	$(1- u)I(u \leq 1)$
Düzgün (Uniform)	$\frac{1}{2}I(u \leq 1)$
Normal (Gaussian)	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)I(u < \infty)$
4.dereceden (Biweight,Quartic)	$\frac{15}{16}(1-u^2)^2 I(u \leq 1)$
6. dereceden (Triweight)	$\frac{35}{32}(1-u^2)^3 I(u \leq 1)$
Cosinus	$\frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) I(u \leq 1)$

Çizelge 3.1'de verilen tüm fonksiyonlar yukarıda verilen kernel fonksiyonlarının özelliklerini sağlar. Kernel fonksiyonunun seçimini düzeltme parametresinin seçimine göre daha önemsizdir (Hastie ve Tibshirani, 1990).

Model (3.4)'deki kernel tahmininde λ düzeltme parametresinin rolü önemlidir. Şöyle ki: Büyük λ değerleri için eğri çok yavaş değişir ve düzeltme önemlidir. Bu durumda tahminin varyansı sınırlı fakat tahmin oldukça sapmalıdır. λ değeri küçük olduğunda ise eğri oldukça düzensizdir ve sapmalar sınırlı fakat tahminin varyansı büyüktür. Bu nedenle λ parametresi sapmalar ve tahminin doğruluğu arasında bir uzlaşma sağlar (Hardle, 2004).

Ayrıca model 3.3 ün matris ve vektör formu, $\hat{\mathbf{f}} = \mathbf{W}_\lambda \mathbf{y}$ olarak gösterilir. Burada,

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1' \\ \cdot \\ \mathbf{w}_i' \\ \cdot \\ \mathbf{w}_n' \end{bmatrix} = \begin{bmatrix} w_{11} & \cdot & \cdot & w_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ w_{i1} & \cdot & \cdot & w_{in} \\ \cdot & \cdot & \cdot & \cdot \\ w_{n1} & \cdot & \cdot & w_{nn} \end{bmatrix} \quad (3.25)$$

$$w_{ij} = \frac{K\left(\frac{x_i - x_j}{\lambda}\right)}{\sum_{j=1}^n K\left(\frac{x_i - x_j}{\lambda}\right)} = \frac{K(u)}{\sum_{j=1}^n K(u)} \quad (3.26)$$

şeklindedir. Ayrıca herhangi bir x_i noktasındaki gözlemin (2.11) modelindeki f 'nin tahmini,

$$\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^n w_{ij} y_j = \mathbf{w}_i' \mathbf{y}, \quad i = 1, 2, \dots, n \quad (3.27)$$

olarak yazılabilir. Burada,

$$\hat{f}(x_1) = \mathbf{w}_1' \mathbf{y} = (w_{11}, \dots, w_{1n}) \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \quad (3.28)$$

olur. Yukarıda sözü edilen (3.5) denkleminde splayn düzeltme tahminine benzer olarak (2.11)'de ifade edilen parametrik olmayan regresyonda kernel tahmini,

$$\hat{\mathbf{f}} = \begin{bmatrix} f(x_1) \\ \cdot \\ \cdot \\ \cdot \\ f(x_n) \end{bmatrix}_{n \times 1} = \begin{bmatrix} w_{11} & \cdot & \cdot & \cdot & w_{1n} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ w_{n1} & \cdot & \cdot & \cdot & w_{nn} \end{bmatrix}_{n \times n} \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}_{n \times 1} = \mathbf{W}_\lambda \mathbf{y} \quad (3.29)$$

olarak gösterilir (Aydın, 2007).

3.3. Doğrusal Düzelticilerin İstatistiksel Özellikleri

Düzeltilme parametresi seçimi ve diğer istatistiksel çıkarım yöntemlerini inceleyebilmek için başlangıçta ilk olarak doğrusal düzelticilerin özelliklerinin bilinmesi gerekir. Bu kısımda doğrusal düzelticilerin özelliklerinden bazılarını ele alacağız.

3.3.1. Hata kareler ortalaması (MSE)

Regresyon modelleri için yaygın olarak kullanılan performans kriterlerinden biri hata kareler ortalamasıdır.

Reinsch (1967), σ^2 (varyans) biliniyorsa;

$$\frac{1}{n} \sum_{i=1}^n (f_\lambda(x_i) - y_i)^2 = \sigma^2 \quad (3.30)$$

denklemini 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;

$$MSE(\lambda) = \frac{1}{n} \sum_{i=1}^n (f_{\lambda}(x_i) - f(x_i))^2 = \frac{1}{n} \|(\mathbf{W}_{\lambda})y - f\|^2 \quad (3.31)$$

formülüyle verilen $MSE(\lambda)$ olarak tanımlanır. Wahba (1975), λ parametresinin (3.30) 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 (3.30) denklemi için uygun örneklem hacmini (n) belirlemek çok zor ve bu anlamda sonuç pratik değildir (Craven ve Wahba, 1979).

Amacımız (3.31)'deki MSE değerini minimum yapan λ düzeltme parametresini tahmin etmektir. Eşitlik (3.31)'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, $\mathbf{f} = (f(x_1), \dots, f(x_n))$ ve bu fonksiyonun bir λ parametresine bağlı tahmin değerlerinin vektörü de $\hat{\mathbf{f}}_{\lambda} = (\hat{f}_{\lambda}(x_1), \dots, \hat{f}_{\lambda}(x_n))$ olsun. Buna göre, tahmini hata kareler ortalaması;

$$EMSE(\lambda) = E\left(\frac{1}{n} \|\hat{\mathbf{f}}_{\lambda} - \mathbf{f}\|^2\right) = E\left(\frac{1}{n} \|(\mathbf{W}_{\lambda})y - f\|^2\right) \quad (3.32)$$

ile belirtilen $EMSE(\lambda)$ fonksiyonu ile elde edilir. Mümkün olduğu kadar gerçek f fonksiyonuna yakın olacak $\hat{f}_{\lambda}$ fonksiyonunu tahmin etmek gerekir. (3.32)'deki $EMSE$, tahminler ile gerçek değerler vektörleri arasındaki öklid uzaklığının beklenen değeridir. Model (3.32) ifadesinde bir $E\hat{f}_{\lambda}$ eklenip-çıkartılması ile $EMSE$ aşağıda gösterildiği gibi iki bileşene ayrıştırılabilir:

$$\begin{aligned}
EMSE(\lambda) &= \frac{1}{n} E \left\| (E\hat{f}_\lambda - f) + (\hat{f}_\lambda - E\hat{f}_\lambda) \right\|^2 \\
&= \frac{1}{n} E \left\| (E\hat{f}_\lambda - f) \right\|^2 + \frac{2}{n} E (E\hat{f}_\lambda - f)' (\hat{f}_\lambda - E\hat{f}_\lambda) + \frac{1}{n} E \left\| (\hat{f}_\lambda - E\hat{f}_\lambda) \right\|^2 \\
&= \frac{1}{n} E \left\| (E\hat{f}_\lambda - f) \right\|^2 + \frac{1}{n} E \left\| (\hat{f}_\lambda - E\hat{f}_\lambda) \right\|^2 \quad (3.33) \\
&= \frac{1}{n} \left\| (I - \mathbf{W}_\lambda) f \right\|^2 + \frac{\sigma^2}{n} \text{tr} \mathbf{W}_\lambda^2 \\
&= (\text{Yan}(\text{Bias}))^2 + (\text{Varyans})
\end{aligned}$$

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 (Eubank, 1999).

EMSE ölçümünde, (3.33) modelinde birinci terim yanlılığın karesini ölçerken ikinci terim varyansı ölçmektedir (Hastie ve Tibshirani, 1990). Burada düzeltme parametresinin değerindeki artmalarda $\hat{f}(x)$ tahminini oluşturmak için daha fazla veri kullanıldığı önsezi ile söylenebilir ve bu durumda $\text{var}(\hat{f}(x))$ azalır. Diğer taraftan yerel polinom tahmini küçük aralıklar üzerinde en iyidir. Bundan dolayı λ değeri artar iken yanlılık artışı beklenebilir. Sezgisel olarak büyük komşuluklar düşük varyanslı bir tahmin fakat potansiyel olarak yüksek yanlılık üretecektir ve tam tersi küçük komşuluklar için de söylenebilir. Böylece seçilen düzeltme parametresi (bant genişliği) λ ile yönetilen yanlılık ve varyans arasında bir değiş-tokuş (trade-off) olduğunu söyleriz.

3.3.2. Serbestlik derecesi

Parametrik regresyonda mevcut gözlemlere dayanarak tahmin edilen parametre sayısının k olduğu varsayalım. Bu durumda uygun modelin H şapka matrisinin izi, modelin serbestlik derecesine eşit olup söz konusu modelin serbestlik derecesi k 'ya eşittir. Bu durumda hatanın (gürültü) 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. Hata için eş değer serbestlik derecesi;

$$EDF = tr(I - W_{\lambda}) \quad (3.34)$$

tanımlanmıştır. Eşitlik (3.34)'deki W_{λ} , λ düzeltme parametresi ile kernel düzeltme ile bağlantılı olan parametrik regresyondaki şapka matrisine karşılık gelen bir düzeltici matristir ve $df_{fit} = tr(W_{\lambda})$ modelin serbestlik derecesidir (Aydın, 2005).

3.4. Hata Varyansının Tahmini

Bu çalışmada parametrik olmayan regresyon konusunda birçok istatistiksel yöntemler için ihtiyaç duyulan hata varyansı σ^2 'nin tahmini için, istatistiksel çalışmalarda yaygın olarak kullanılan iki yöntem ele alınmıştır. Bu yöntemlerden ilki parametrik regresyon konusunda olduğu gibi, hata kareler toplamının serbestlik derecesine oranı ile elde edilir. σ^2 'nin tahmini için yaygın olarak kullanılan diğer bir yöntem ise fark almaya dayalı varyans tahminidir.

3.4.1. Hata kareler yaklaşımı

Parametrik regresyon uygulamasında örneklem için hata kareler toplamı;

$$\hat{\sigma}^2(MSE) = \frac{\sum_{i=1}^n (y_i - f(x_i))^2}{df} = \frac{\sum_{i=1}^n \varepsilon_i^2}{n-k} \quad (3.35)$$

Eşitlik (3.35)'deki gibi serbestlik derecesine bölünmesi ile varyansın yansız bir tahmin edicisi elde edilir. Benzer olarak kernel düzeltme konusunda, hata kareler toplamı eşdeğer serbestlik derecesi olarak tanımlanan $tr(I - W_{\lambda})$ ifadesine bölünerek varyansın bir tahmin edicisi elde edilir. λ düzeltme parametresi ile bu tahmin edici aşağıdaki biçimde verilir:

$$\begin{aligned}\hat{\sigma}_\lambda^2 &= \frac{\sum_i (y_i - \hat{f}(x_i))^2}{\text{tr}(I - \mathbf{W}_\lambda)} = \frac{\|(\mathbf{W}_\lambda - I)y\|^2}{\text{tr}(I - \mathbf{W}_\lambda)} \\ &= \frac{RSS}{n - \text{tr}(\mathbf{W}_\lambda)}\end{aligned}\quad (3.36)$$

Burada $\hat{f}, \lambda$ düzeltme parametresi ile hesaplanan kernel düzeltme tahmin edicisidir. Gerçek f regresyon eğrisinin bir düz doğru olduğu özel durumda, tüm λ düzeltme parametreleri için $\hat{\sigma}_\lambda^2$ tahmin edicisi, σ^2 'nin yansız bir tahmin edicisidir (Green ve Silverman, 1994).

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

$$\sigma^2 = \frac{\|(I - \mathbf{W}_\lambda)y\|^2}{n} \quad (3.37)$$

ile verilir (Wahba, 1990).

3.4.2. Yerel fark alma yaklaşımı

Rice (1984), verilerin birinci farklarını kullanarak varyansı;

$$\hat{\sigma}_R^2 = \frac{1}{2}(n-1)^{-1} \sum_{i=2}^n (y_i - y_{i-1})^2 \quad (3.38)$$

formülü ile tahmin etmeyi önermiştir. Bu tahminin 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 analizinde kullanılmaktadır.

Rice (1984), ayrıca verilerin ağırlıklı ikinci farkları üzerine dayalı ikinci bir tahmini önermiştir. Buna göre, f fonksiyonuna bir doğrusal fonksiyonun eklenmesi durumu değiştirmeyecektir. Tahmin, bir polimal değerlerle yerel olarak tahmin edilen f

fonksiyonunun deęerleri ile hata kareler toplamına dayalı olup, bir birini takip eden üçlü noktalara en küçük kareler denklemini uygulayarak elde edilir. Bu tahmin, bir yerel tahmin olarak bir doğru denklemini kullanan, Gasser vd. (1986) tarafından önerilmiştir:

$$\hat{\sigma}_{GSJ}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} c_i^2 \hat{\varepsilon}_i^2 \quad (3.39)$$

Burada;

$$\hat{\varepsilon}_i^2 = \left[y_{i-1} + \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} (x_i - x_{i-1}) - y_i \right]^2, \quad i = 2, \dots, n-1 \quad (3.40)$$

$$c_i^2 = \left[\left(\frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} \right) + \left(\frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} \right) + 1 \right]^2, \quad i = 2, \dots, n-1 \quad (3.41)$$

şeklindedir. Gasser vd. (1986) tarafından önerilen bu varyans kestiricisi, Schimek (2000) ve Eubank vd. (1998) tarafından yapılan çalışmalarda da verildiğı gibi aşağıdaki şekilde de yazılabilir:

$$\hat{\sigma}^2 = \frac{y'A'Ay}{trA'A} \quad (3.42)$$

Buradaki A matrisi girişlerin tümü sıfır fakat i . girişi $a_i c_i$, $(i+1)$. girişi $-c_i$ ve $(i+2)$. girişi $b_i c_i$ olan $(n-2) \times n$ boyutlu bir matristir ve $trA'A = n-2$ 'dir.

4. KERNEL YÖNTEMİNDE DÜZELTME PARAMETRESİ SEÇİM YÖNTEMLERİ

Düzeltilme parametresinin (bant genişliği) 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. Ayrıca literatürde birçok seçim kriteri yöntemleri de mevcuttur. Bu çalışmada ele alınan yöntemler matrissel gösterim temeli altındadır ve yöntemdeki modellerin minimum yapılması esasına dayanır. Yöntemdeki modelleri minimum yapan düzeltme parametresi en iyi düzeltme parametresi olarak ifade edilir.

Düzeltilme parametresi λ 'nın $\hat{f}_\lambda$ 'nin varyansı ve yanlılığı arasında bir uzlaşma sağladığını alt bölüm 3.3.1'den biliyoruz. Bundan dolayı düzeltme parametresi $\lambda > 0$, $\hat{f}_\lambda$ 'nin kalitesi üzerinde çok önemli bir etkiye sahiptir. Böylece parametrik olmayan regresyon konusunda, verilere dayalı olarak elde edilen düzeltme parametresinin seçiminin önemli bir problem olduğunu söyleyebiliriz. Aynı λ değerinin her veri seti ile aynı derecede iyi çalışması beklenemez.

Splayn düzeltmesi yönteminde kullanılan düzeltme parametresi seçim yöntemleri kernel düzeltmesi için genelleştirildi. Burada dikkate aldığımız seçim yöntemleri aşağıda verilmiştir:

- Çapraz Geçerlilik (CV) ölçütü,
- Genelleştirilmiş Çapraz Geçerlilik (GCV) ölçütü,
- Mallows'un C_p ölçütü,
- Klasik Akaike bilgi ölçütünün (AIC) gelişmiş bir biçimi olan AIC_c ölçütü,

- Klasik pilotları Kullanan Risk Tahmini (*RECP*),
- Sınırlı maksimum olabilirlik ölçütü (*REML*),
- Rice'in T ölçütü,
- Shibata'nın seçtiği bir model SH ölçütü,

Söz konusu düzeltme parametresinin seçiminde kullanılan klasik ve risk tahmin metotları, yukarda belirtildiği sırada izleyen alt bölümlerde detaylı olarak belirtilmiştir.

4.1. Çapraz Geçerlilik Ölçütü (CV)

Çapraz Geçerlilik, (x_i, y_i) veri noktalarından birinin çıkarılması ve geri kalan $(n-1)$ veri noktasına bağlı olarak x_i noktasındaki pürüzsüzlüğün yani x_i noktasındaki uyum $\hat{f}_{\lambda}^{-i}(x_i)$ 'in tahmin edilmesidir. $\hat{f}_{\lambda}^{-i}(x_i)$ 'yi tanımlamak için n veri noktasına bağlı doğrusal düzeltmeler ile başlanırsa $n-1$ nokta için karşılık gelen düzeltmeler tanımlanmak zorundadır.

y^{-i} , orijinal y bağımlı değişken vektöründen y_i gözleminin çıkarılması ile elde edilen $n-1$ boyutlu gözlemlerin vektörü olsun. Ayrıca, söz konusu y^{-i} vektörüne uygun düzeltme tahmini $\hat{f}_{\lambda}^{-i}(x_i)$ olsun. Kestirim hatasının çapraz geçerlilik (alışılmış çapraz geçerlilik) ölçütü;

$$CV(\lambda) = n^{-1} \sum_{i=1}^n y_i - \hat{f}_{\lambda}^{-i}(x_i)^2 \quad (4.1)$$

olarak tanımlanır. Burada $\hat{f}_{\lambda}^{-i}(x_i)$, x_i 'deki uyumu gösterir ve i . veri noktasının çıkarılması ile hesaplanır. Çapraz geçerliliğin asıl amacı $CV(\lambda)$ fonksiyonunu minimum yapan λ değerini seçmektir.

$\hat{f} = W_{\lambda} y$ ile verilen doğrusal düzelticiler için çapraz geçerlilik ölçütü aşağıdaki biçimde elde edilir:

$$y_i - \hat{f}_i^{-i}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathbf{W}_{ii}}$$

$$CV(\lambda) = n^{-1} \sum_{i=1}^n (y_i - \hat{f}_i^{-i}(x_i))^2 = n^{-1} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - \mathbf{W}_{ii}} \right\}^2 \quad (4.2)$$

Burada $\mathbf{W}_{ii}$ (3.25)'de tanımlanan matrisin köşegen değerleridir. Böylece λ parametresinin CV tahmini $CV(\lambda)$ fonksiyonunu minimum yapan değerdir (Ruppert vd., 2003).

4.2. Genelleştirilmiş Çapraz Geçerlilik Ölçütü (GCV)

Genelleştirilmiş çapraz geçerlilik ölçütü GCV, çapraz geçerlilik üzerindeki değişikliklerden elde edilir. Doğrusal düzelticiler için GCV (4.2) eşitliğindeki $\mathbf{W}_{ii}$ 'nin ortalama değeri $tr(\mathbf{W}_\lambda)/n$ ile yer değiştirmesi temeline dayanır ve (2.9) ile verilen parametrik olmayan regresyon modeli için genelleştirilmiş çapraz geçerlilik ölçütü;

$$GCV(\lambda) = n^{-1} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - tr(\mathbf{W}_\lambda)/n} \right\}^2$$

$$= \frac{n^{-1} \|(I - \mathbf{W}_\lambda)y\|^2}{1 - n^{-1}tr(\mathbf{W}_\lambda)} \quad (4.3)$$

$$= \frac{n^{-1} \|(I - \mathbf{W}_\lambda)y\|^2}{1 - n^{-1}df_{fit}}$$

olarak elde edilir. Burada $df_{fit} = tr(\mathbf{W}_\lambda)$ olup modelin serbestlik derecesidir. λ parametresinin GCV tahmini $GCV(\lambda)$ fonksiyonunu minimum yapan değerdir (Ruppert vd., 2003).

4.3. Mallows'un C_p Ölçütü

Mallows'un C_p ölçütü splayn düzeltme konusunda yansız risk yöntemi olarak bilinir. σ^2 bilindiğinde λ için bir yansız risk tahmini mevcuttur. Regresyonda böyle bir tahmin Mallows (1973) tarafından önerilmiş ve Craven ve Wahba (1979) tarafından splayn düzeltmeye uygulanmıştır.

Hata kareler toplamının yansız bir tahminini, minimum yapan λ parametresini seçmeyi amaçlayan C_p kriteri;

$$\begin{aligned} C_p(\lambda) &= \frac{1}{n} \|(\mathbf{W}_\lambda - I)y\|^2 + 2\sigma^2 \text{tr}(\mathbf{W}_\lambda) + \sigma^2 \\ &= \frac{1}{n} \|y - \hat{f}_\lambda\|^2 + 2\sigma^2 \text{tr}(\mathbf{W}_\lambda) + \sigma^2 \end{aligned} \quad (4.4)$$

ile verilir. C_p istatistiği, aynı zamanda $E\|\hat{f}_\lambda - f\|^2$ kareli kestirim hatasının da yansız bir tahminidir (Kou ve Efron, 2002).

Craven ve Wahba (1979) tarafından yapılan deneysel sonuçlar göstermiştir ki, (4.4)'de aynı σ^2 kullanıldığında, büyük örneklerde 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 (3.36) ile elde edilir. Böylece $C_p(\lambda)$ kriterini minimum yapan λ seçilmek istendiğinde, model (4.4)'de yer alan σ^2 'yi uygun bir $\hat{\sigma}^2$ kestiricisiyle yer değiştirmek gerekir (Aydın, 2005).

4.4. Geliştirilmiş Akaike Bilgi Ölçütü (AIC_c)

Akaike Bilgi Kriteri (AIC) ilk olarak Kullback –Leibler bilgisinin beklenen değerinin yaklaşık yansız bir kestiricisi olarak parametrik modeller için geliştirilmiştir. Hurvich vd. (1988), 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 vd. (1988), tarafından elde edilen geliştirilmiş AIC_c kriteri, parametrik olmayan düzelticiler için düzeltme parametresinin seçimi için kullanılmıştır. Söz konu geliştirilen AIC_c kriteri;

$$\begin{aligned} AIC_c(\lambda) &= \log(\hat{\sigma}^2) + \frac{1 + \text{tr}(\mathbf{W}_\lambda) / n}{1 - \text{tr}(\mathbf{W}_\lambda) + 2 / n} \\ &= \log(\hat{\sigma}^2) + 1 + \frac{2 \text{tr}(\mathbf{W}_\lambda) + 1}{n - \text{tr}(\mathbf{W}_\lambda) - 2} \end{aligned} \quad (4.5)$$

formülüyle verilir. Eşitlik (4.5)'de de görüldüğü gibi bu kriter $\mathbf{W}_\lambda$ düzeltme matrisinin yalnızca izini kullandığından, düzeltme parametresinin seçimi için uygulaması kolay olan bir kriteridir. (3.37)'de belirtilen varyans fonksiyonu (4.5)'de yerine yazılarak, basit cebirsel işlemlerden sonra, Lee (2003) tarafından da belirtildiği gibi, (4.5)'de tanımlanan AIC_c bilgi kriteri;

$$\begin{aligned} AIC_c(\lambda) &= \log \frac{\sum (y_i - \hat{f}_\lambda(x_i))^2}{n} + 1 + \frac{2 \text{tr}(\mathbf{W}_\lambda) + 1}{n - \text{tr}(\mathbf{W}_\lambda) - 2} \\ &= \log \frac{\|(\mathbf{W}_\lambda - \mathbf{I})\mathbf{y}\|^2}{n} + 1 + \frac{2 \text{tr}(\mathbf{W}_\lambda) + 1}{n - \text{tr}(\mathbf{W}_\lambda) - 2} \end{aligned} \quad (4.6)$$

biçimini alır. Diğer yöntemlerde olduğu gibi, $AIC_c(\lambda)$ kriterini minimum yapan λ değeri düzeltme parametresi olarak seçilir.

4.5. Klasik Pilotları Kullanan Risk Tahmini (*RECP*)

Risk fonksiyonu R , tahmin ile gerçek regresyon fonksiyonu arasındaki uzaklığı ölçer. Risk fonksiyonu aşağıdaki gibi yanlılık ve varyansın toplamı olarak gösterilebilir:

$$R(f, \hat{f}_\lambda) = \frac{1}{n} E \|f - \hat{f}_\lambda\|^2 = \frac{1}{n} \|(\mathbf{W}_\lambda - I)f\|^2 + \sigma^2 \text{tr}(\mathbf{W}_\lambda \mathbf{W}_\lambda') \quad (4.7)$$

Eşitlik (4.7)'de verilen σ^2 ve f 'nin uygun pilot tahminlerinin yerleştirilmesi ile $R(f, \hat{f}_\lambda)$ riskini tahmin etmek ve risk tahmin edicisini minimum yapan λ parametresini seçmek gerekir. Pilot tahminleri seçmek için önerilen bir yöntem; klasik bir yöntem (CV, GCV, AIC_c, C_p) kullanarak bir pilot λ_p değerini seçmek ve bu λ_p değeri ile σ^2 ve f 'nin pilot tahminleri olan $\hat{\sigma}_{\lambda_p}^2$ ve $\hat{f}_{\lambda_p}$ 'nin hesaplanmasıdır (Lee ve Solo, 1999) (Lee, 2001). Böylece parametrik olmayan regresyon modeli için $R(f, \hat{f}_\lambda)$ 'nin tahmin edicisi;

$$R(\hat{f}_{\lambda_p}, \hat{f}_\lambda) = \frac{1}{n} E \|\hat{f}_{\lambda_p} - \hat{f}_\lambda\|^2 = \frac{1}{n} \|(\mathbf{W}_\lambda - I)\hat{f}_{\lambda_p}\|^2 + \sigma_{\lambda_p}^2 \text{tr}(\mathbf{W}_\lambda \mathbf{W}_\lambda') \quad (4.8)$$

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

4.6. Sınırlı Maksimum Olabilirlik Ölçütü (REML)

REML ölçütü λ 'nın optimizasyonu ile bir fonksiyonu olarak davranır. REML tarafından λ tahmin edildiğinde, REML hata varyans tahmini, teorik düzeltme varyans tahmini ile uyuşur. REML ölçütü;

$$\begin{aligned} REML(\lambda) &= \frac{\|(\mathbf{W}_\lambda - I)y\|^2}{\text{tr}(I - \mathbf{W}_\lambda)} \\ &= \frac{nMSE(\lambda)}{n - df_{fit}(\lambda)} \end{aligned} \quad (4.9)$$

biçimindedir. Benzer olarak diğer yöntemlerde olduğu gibi, $REML(\lambda)$ ifadesini minimum yapan λ , düzeltme parametresi olarak seçilir (Aydın ve Memmedli, 2012).

4.7. Rice'ın T Ölçütü

Literatürde diğer kriterler baz alındığında pek rağbet görmeyen bir seçim kriteridir. Model aşağıdaki şekildedir.

$$T(\lambda) = \frac{1}{n} \|(I - \mathbf{W}_\lambda)y\|^2 / (1 - n^{-1} \text{tr}(\mathbf{W}_\lambda)) \quad (4.10)$$

$T(\lambda)$ ifadesini minimum yapan λ , düzeltme parametresi olarak seçilir.

4.8. Shibata'nın Seçtiği Bir Model SH Ölçütü

Bu modelde üstteki T kriteri gibi literatürde pek rağbet görmez. Diğer seçim kriterlerinin yanında kötü sonuçlar vermektedir. Model aşağıdaki şekildedir.

$$SH(\lambda) = \frac{1}{n} \|(I - \mathbf{W}_\lambda)y\|^2 (1 - n^{-1} \text{tr}(\mathbf{W}_\lambda)) \quad (4.11)$$

$SH(\lambda)$ ifadesini minimum yapan λ , düzeltme parametresi olarak seçilir.

5. SİMÜLASYON DENEYİ

Bu bölümde, 4. bölümde ele alınan 8 seçim kriterinden hangisinin daha iyi bir düzeltme parametresi seçtiğini belirlemek amacıyla, bu seçim kriterlerinin karşılaştırılması yapılmıştır. Bunun için, TURBO C programı yardımıyla, her bir faktör düzeyi için farklı büyüklükte 3 örneklem oluşturulmuş ve oluşturulan her bir örneklem için 100 tekrarlamaya yapılmıştır. Seçim kriterine göre $\hat{f}$, kernel düzeltmesi ve kernel fonksiyonlarından da gaussian (normal) kernel fonksiyonu kullanılarak tahmin edilmiştir.

Tekrarlanan her bir örneklem veri dizisi için herhangi bir $\hat{f}$ tahmin eğrisinin kalitesini değerlendirmek için hata kareler ortalaması (*MSE*) kullanılmıştır. Herhangi iki metodun *MSE* değerleri medyanı arasındaki farkın anlamlı olup olmadığı, TURBO C programından veriler alınarak SPSS programı yardımıyla Wilcoxon işaretli sıra sayıları testi ile test edilmiştir.

Böylece, yapılan simülasyon çalışmasında adı geçen kriterler değerlendirilerek, hangi kriterin daha iyi bir tahmin sonucu veren düzeltme parametresi seçtiği belirlenmiştir.

5.1. Veriler ve Deney Düzenekinin Oluşturulması

Deneyde örneklem oluşturmada kullanılan veriler, Çizelge 5.1’de genel biçimde verilen modellerden elde edilmiştir.

Simülasyon deneyi, TURBO C ortamında yazılan bir programla ve SPSS programı yardımıyla gerçekleştirilmiş olup, deneyin planı ve yürütülmesi ise şu şekilde tasarlanmıştır:

Her bir faktör düzeyi için 50, 75 ve 100 olmak üzere üç farklı örneklem oluşturulmuştur. Oluşturulan her bir örneklem 100 kez tekrar edilmiştir. Her bir

faktör düzeyi için tekrar edilen tüm örneklemelerde adı geçen her bir seçim kriterini minimum yapan λ düzeltme parametresi seçilmiştir.

Her bir seçim kriterine göre hesaplanan kernel kestiricileri için *MSE* değerleri hesaplanmıştır. Her hangi iki seçim yönteminin performans ölçüsü olarak dikkate alınan *MSE* değerlerinin farklı olup olmadığı Wilcoxon testi ile test edilmiştir. Çizelge 5.1’de belirtilen 4 faktörün etkisini belirlemek amacıyla faktör düzeyleri $r=1, 2, 3, 4, 5, 6$ olmak üzere, 6 kez değiştirilmiştir.

Çizelge 5.1: Fonksiyonel Kalıplar

Faktör	Genel şekli	Fonksiyonel içeriği
1	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$= \sin(15\pi x_i) + \sigma_r \varepsilon_i$
2	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$= 0.3 \exp -64(x_i - 0.25)^2 + 0.7 \exp -256(x_i - 0.75)^2 + \sigma_r \varepsilon_i$
3	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$= 10 \exp(-10x_i) + \sigma_r \varepsilon_i$
4	$y_{ir} = f(x_i) + \sigma_r \varepsilon_i$	$= \sqrt{x_i(1-x_i)} \sin \left[\frac{2\pi(1+2^{(9-4r)/5})}{x_i + 2^{(9-4r)/5}} \right] + \sigma_r \varepsilon_i$
$\sigma_r = 0.02 + 0.04(r-1)^2, \quad r = 1, 2, 3, 4, 5, 6$ $x_i \sim U(0,1), \varepsilon_i \sim N(0,1) \quad i = 1, \dots, n$		

5.2. DeneySEL Değerlendirmeler

Simülasyon deneyleri sonucunda oluşturulan deneylerin her birinde, 4.bölümde ele alınan 8 seçim kriterinden hangisinin daha iyi bir düzeltme parametresi seçtiğini belirlemek amacıyla, *MSE* değerleri performans ölçü kriteri olarak kullanılmıştır.

Herhangi iki seçim metodun *MSE* değerleri medyanı arasındaki farkın anlamlı olup olmadığı %5 anlam düzeyinde eşleştirilmiş Wilcoxon işaretli sıra testi ile test edilmiştir. Ayrıca seçim metodları Wilcoxon işaretli sıra testine göre şu şekilde sıralanmıştır: Bir metodun *MSE* değeri medyanı anlamlı bir biçimde kalan 7 yöntemde daha az ise, 1 sırası atanacak, bir metodun *MSE* değeri medyanı anlamlı olarak birinden daha büyük fakat altı metottan az ise, 2 sırası atanacak ve benzer biçimde 3-8 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.

Çizelge 5.2’de $n=100$, 3.faktör, $r=1$ ve $r=6$ düzeyi için 100 tekrarlı simülasyona ilişkin wilcoxon testi sonuçları verilmiştir.

Çizelge 5.2: $n=100$, 3.faktör ve $r=1$ için wilcoxon test sonuçları

	Test İstatistiği (c)						
	GCV-CV	CP-CV	AKAIKE-CV	SH-CV	T-CV	ML-CV	RECP-CV
Z	,000 (a)	,000 (a)	,000 (a)	-1,342 (b)	,000 (a)	,000 (a)	,000 (a)
Asymp. Sig. (2-tailed)	1,000	1,000	1,000	,180	1,000	1,000	1,000
a. The sum of negative ranks equals the sum of positive ranks. b. Based on negative ranks. c. Wilcoxon Signed Ranks Test.							
	Test İstatistiği (c)						
	CP-GCV	AKAIKE-GCV	SH-GCV	T-GCV	ML-GCV	RECP-GCV	AKAIKE-E-CP
Z	,000 (a)	,000 (a)	-1,342 (b)	,000 (a)	,000 (a)	,000 (a)	,000 (a)
Asymp. Sig. (2-tailed)	1,000	1,000	,180	1,000	1,000	1,000	1,000
a. The sum of negative ranks equals the sum of positive ranks. b. Based on negative ranks. c. Wilcoxon Signed Ranks Test.							
	Test İstatistiği (c)						
	SH-CP	T-CP	ML-CP	RECP-CP	SH-AKAIKE	T-AKAIKE	ML-AKAIKE
Z	-1,342 (a)	,000 (b)	,000 (b)	,000 (b)	-1,342 (a)	,000 (b)	,000 (b)
Asymp. Sig. (2-tailed)	,180	1,000	1,000	1,000	1,000	1,000	1,000
a. Based on negative ranks. b. The sum of negative ranks equals the sum of positive ranks. c. Wilcoxon Signed Ranks Test.							
	Test İstatistiği (c)						
	RECP-AKAIKE	T-SH	ML-SH	RECP-SH	ML-T	RECP-T	RECP-ML
Z	,000 (a)	-1,342 (b)	-1,342 (b)	-1,342 (b)	,000 (a)	,000 (a)	,000 (a)
Asymp. Sig. (2-tailed)	1,000	,180	,180	,180	1,000	1,000	1,000
a. The sum of negative ranks equals the sum of positive ranks. b. Based on positive ranks. c. Wilcoxon Signed Ranks Test.							

Çizelge 5.3: n=100, 3.faktör ve r=5 için wilcoxon test sonuçları:

	Test İstatistiği (c)						
	GCV-CV	CP-CV	AKAİKE-CV	SH-CV	T-CV	ML-CV	RECP-CV
Z	-1,680 (a)	-5,646 (a)	-8,329 (a)	-8,68 (a)	-8,15 (a)	-8,68 (b)	-8,68 (b)
Asymp. Sig. (2-tailed)	0,093	,000	,000	,000	,000	,000	1,000
a. Based on negative ranks.							
b. Based on positive ranks.							
c. Wilcoxon Signed Ranks Test.							

	Test İstatistiği (c)						
	CP-GCV	AKAİKE-GCV	SH-GCV	T-GCV	ML-GCV	RECP-GCV	AKAİKE-CP
Z	-5,37 (a)	-8,374 (a)	-8,68 (a)	-8,10 (a)	-8,68 (b)	-8,68 (b)	-6,51 (a)
Asymp. Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000
a. Based on negative ranks.							
b. Based on positive ranks.							
c. Wilcoxon Signed Ranks Test.							

	Test İstatistiği (c)						
	SH-CP	T-CP	ML-CP	RECP-CP	SH-AKAİKE	T-AKAİKE	ML-AKAİKE
Z	-8,682 (a)	-6,093 (a)	-8,682 (b)	-8,68 (b)	-8,68 (a)	-2,36 (b)	-8,68 (b)
Asymp. Sig. (2-tailed)	,000	,000	,000	,000	,000	,018	,000
a. Based on negative ranks.							
b. Based on positive ranks.							
d. Wilcoxon Signed Ranks Test.							

	Test İstatistiği (b)						
	RECP-AKAİKE	T-SH	ML-SH	RECP-SH	ML-T	RECP-T	RECP-ML
Z	-8,68 (a)	-8,68 (a)	-8,68 (a)	-8,68(a)	-8,68 (a)	-8,68 (a)	-8,68 (a)
Asymp. Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000
a. Based on positive ranks.							
b. Wilcoxon Signed Ranks Test.							

Çizelge 5.3'deki wilcoxon test istatistikleri sadece n=100, 3.faktör r=1 ve 3.faktör r=5 için test sonuçlarını göstermektedir. Her bir faktör ve her bir faktör düzeyi için bu ikili karşılaştırmalar yapılmıştır. Bir faktör ve bir faktör düzeyi için 28 ikili karşılaştırma mevcuttur. Bir faktör için 6 faktör düzeyi ele alındığından dolayı, bir faktörde $28 \times 6 = 168$ ikili karşılaştırma yapılmıştır. Toplamda da 4 faktörümüz olduğu için $4 \times 168 = 672$ ikili karşılaştırma yapılmıştır. Bu ikili karşılaştırmalar aşağıdaki gibi verilen hipotez testi ile gerçekleştirilir.

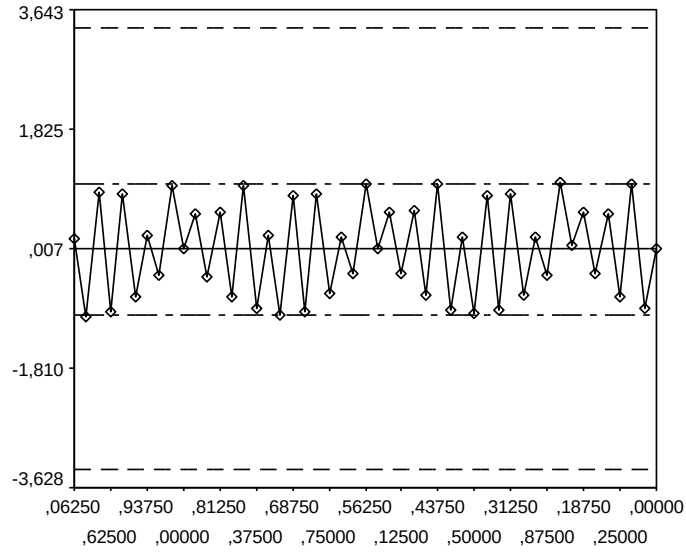
$$\begin{aligned} H_0 : M &= 0 \\ H_1 : M &\neq 0 \end{aligned} \tag{5.1}$$

Eşitlik 5.1 deki şekliyle hipotezler kurulur. Wilcoxon test istatistikleri sonuçlarındaki en alt satır “p” değeridir. bu “p” değerlerine bakarak model 5.1 deki şekliyle kurulan hipotezler sonuçlandırılır.

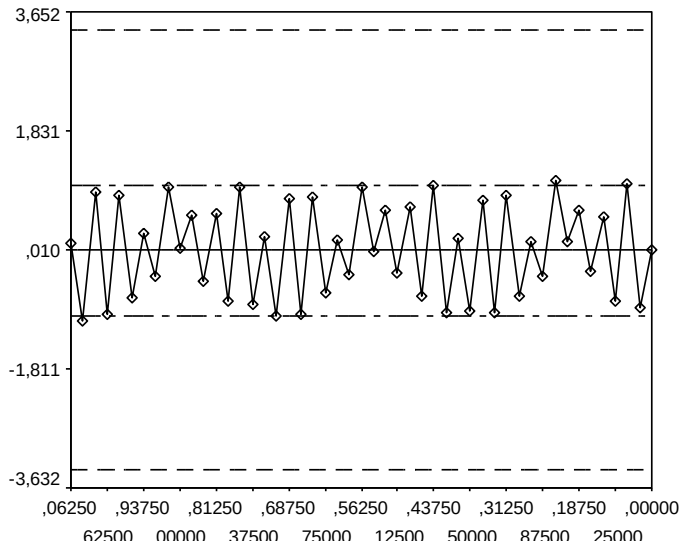
Eğer $p < (\alpha = 0.05)$ ise hipotez red edilir ve anlamlıdır şeklinde ifade edilir. $p \geq (\alpha = 0.05)$ ise hipotez red edilemez ve anlamlı değildir şeklinde ifade edilir. Burada hipotezin anlamlı olması demek ön test ile son test arasında yani ele aldığımız herhangi bir ikili arasında medyana göre ve $\alpha = 0.05$ anlamlılık düzeyinde fark olduğunun göstergesidir. Anlamlı olmaması durumunda ise medyana göre ve $\alpha = 0.05$ anlamlılık düzeyinde aralarında fark olmadığının göstergesidir.

Test istatistik sonuçlarından öncede belirttiğimiz gibi herhangi iki seçim metodun *MSE* değerleri medyanyı arasındaki farkın anlamlı olup olmadığı %5 anlam düzeyinde eşleştirilmiş Wilcoxon işaretli sıra testi ile test edilmiştir. Bir metodun *MSE* değeri medyanyı anlamlı bir biçimde kalan 7 yöntemde daha az ise, 1 sırası atanacak, bir metodun *MSE* değeri medyanyı anlamlı olarak birinden daha büyük fakat altı metottan az ise, 2 sırası atanacak ve benzer biçimde 3-8 sıraları atanacak. Farklı medyan değerlerine sahip fakat anlamlı olmayan metotlar aynı ortalama sırayı paylaşacaklardır.

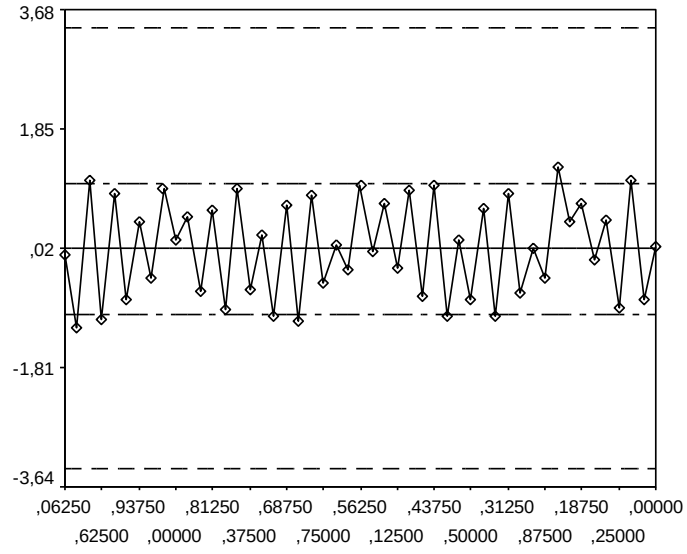
Buradan hareketle çizelge 5.2’deki 3.faktörü $r=1$ düzeyi için değerlendirilecek olursa, 28 adet ikili karşılaştırma için 28 adet hipotez kurulur ve baktığımızda her biri $p \geq \alpha$ olduğundan, burada 8 kriterin her birine 4.5 değeri atanır. 3.faktör $r=5$ düzeyi için değerlendirdiğimizde ise *CV-GCV* ikilisinin anlamlı olmadığı fakat kalan 27 ikilinin anlamlı çıktığı görülmüştür. Bundan sonraki adımda *MSE* değerlerine bakılır ve üstte de ifade edildiği gibi sıra atamaları yapılır. Normalde 1 den 8 e kadar atama yapılabilinirdi. Fakat iki değer anlamlı olmadığından aynı değerde atama yapılır. Yani *CV* ’ye ne atanırsa *GCV* kriterine de aynı değer atanır. Dolayısıyla 3.faktör ve $r=5$ düzeyi için 1 den 7 ye kadar atama yapılacaktır.



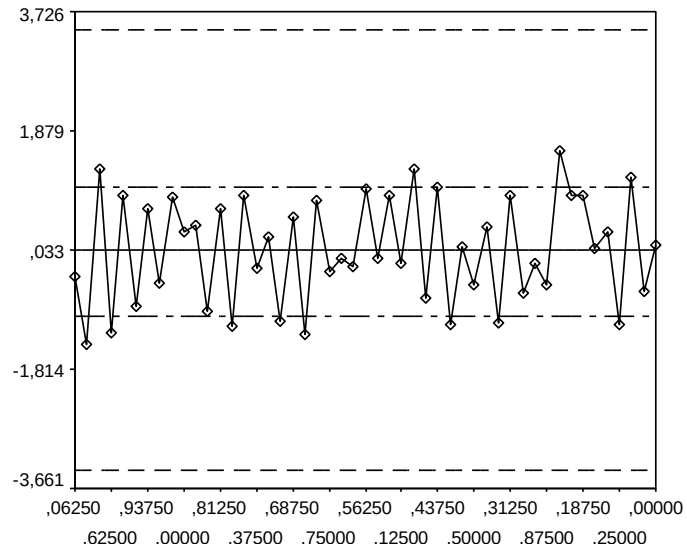
Şekil 5.1. $n=50$ ve $r=1$ düzeyinde 1. faktöre ilişkin X'e karşın Y'nin grafiği



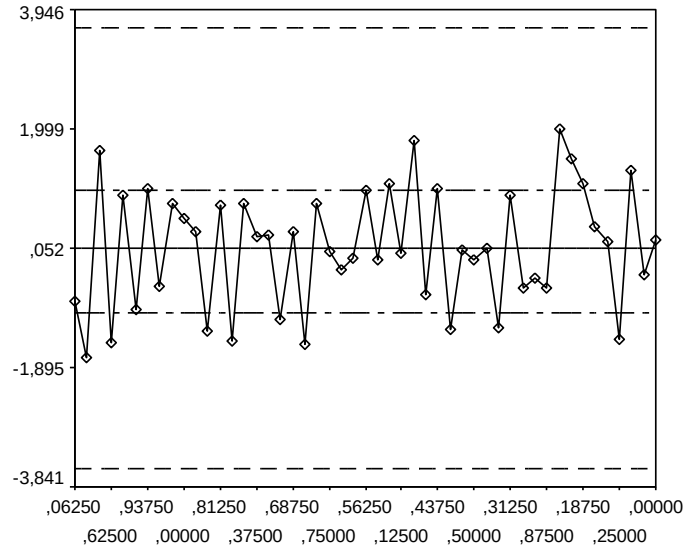
Şekil 5.2. $n=50$ ve $r=2$ düzeyinde 1. faktöre ilişkin X'e karşın Y'nin grafiği



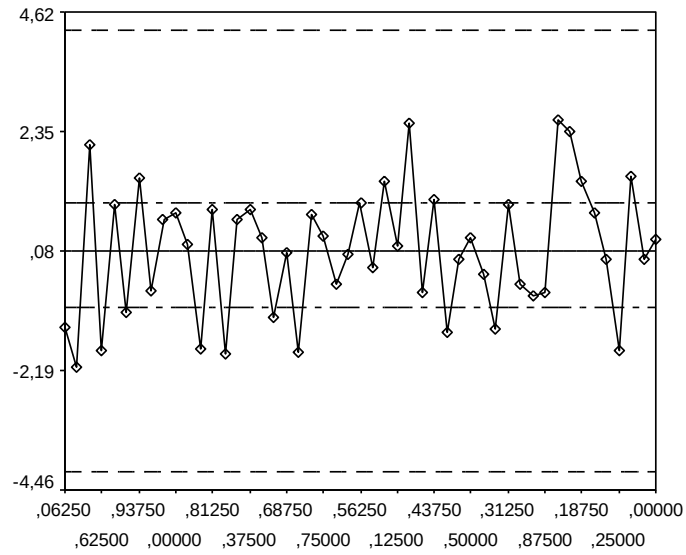
Şekil 5.3. $n=50$ ve $r=3$ düzeyinde 1.faktöre ilişkin X'e karşın Y'nin grafiği



Şekil 5.4. $n=50$ ve $r=4$ düzeyinde 1. faktöre ilişkin X'e karşın Y'nin grafiği

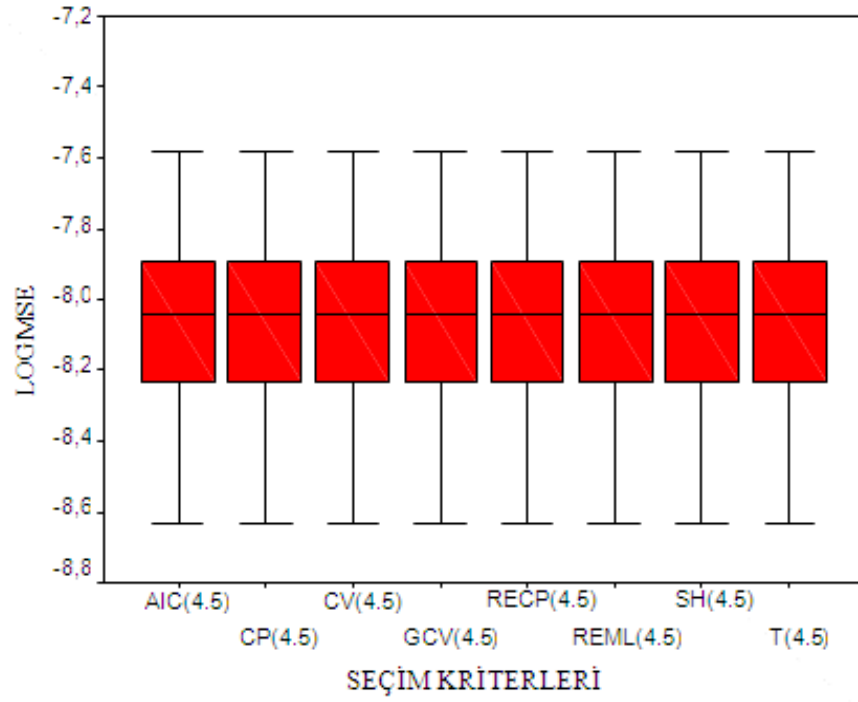


Şekil 5.5. $n=50$ ve $r=5$ düzeyinde 1. faktöre ilişkin X'e karşın Y'nin grafiği

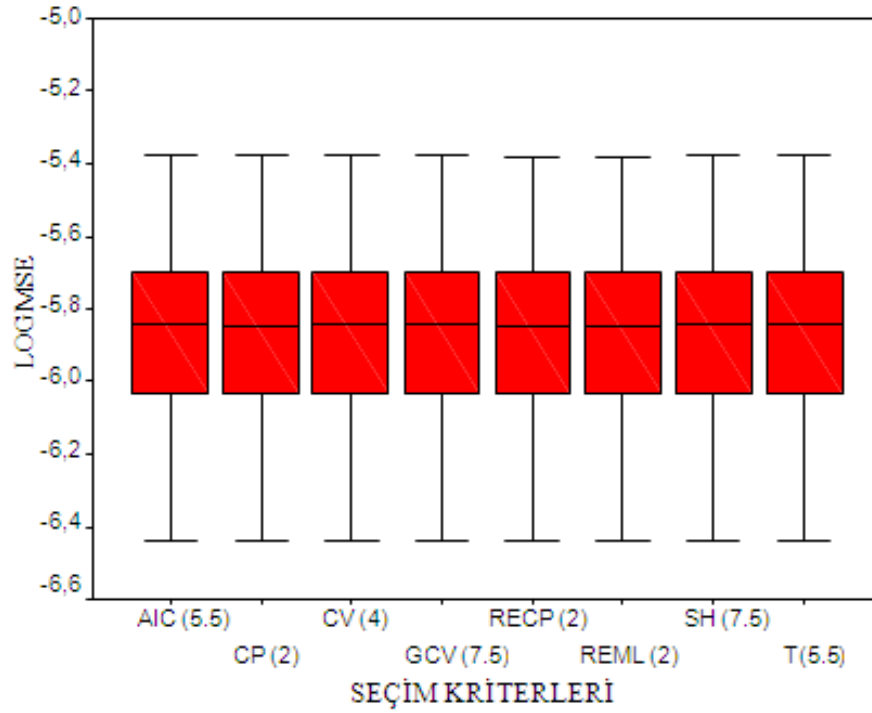


Şekil 5.6. $n=50$ ve $r=6$ düzeyinde 1. faktöre ilişkin X'e karşın Y'nin grafiği

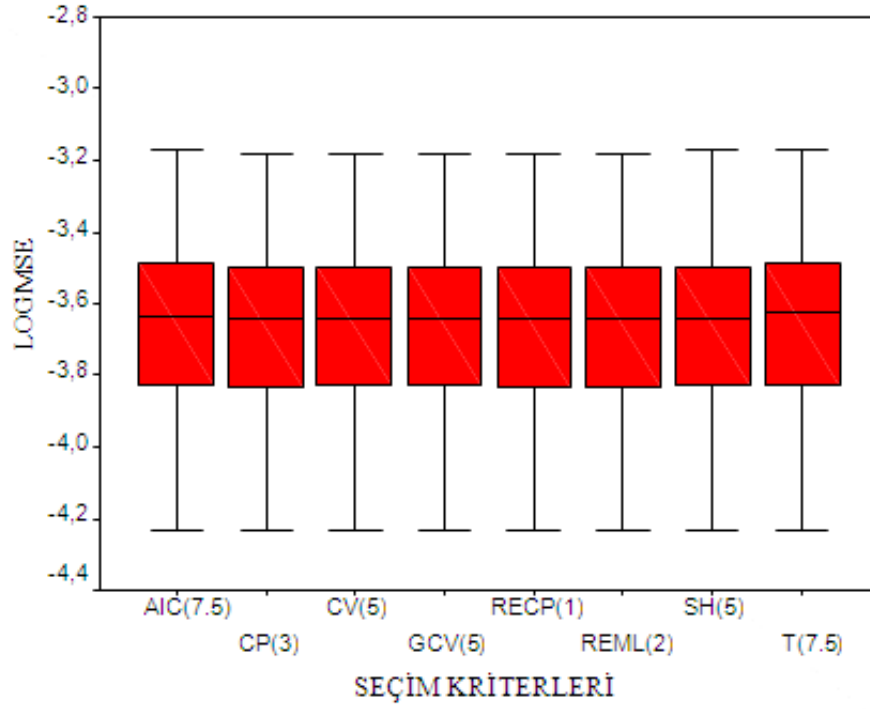
Üstteki 6 şekil $n=50$ ve sırasıyla $r=1,2,3,4,5,6$ düzeyi için 1.faktöre ilişkin X'e karşın Y'nin grafikleridir. Buradaki 6 şekilden herhangi birini ele aldığımızda bize faktörün şekli konusunda bilgi verir. Ayrıca düzeyler ise, düzeyin artışına bağlı olarak Y değerlerinin artışı net olarak göstermektedir.



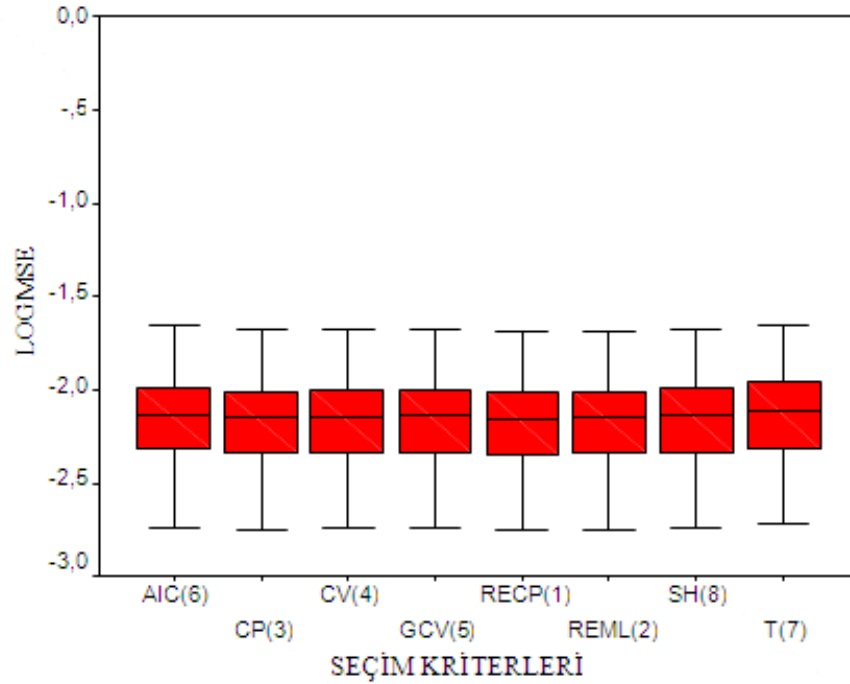
Şekil 5.7. $n=50$ ve $r=1$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



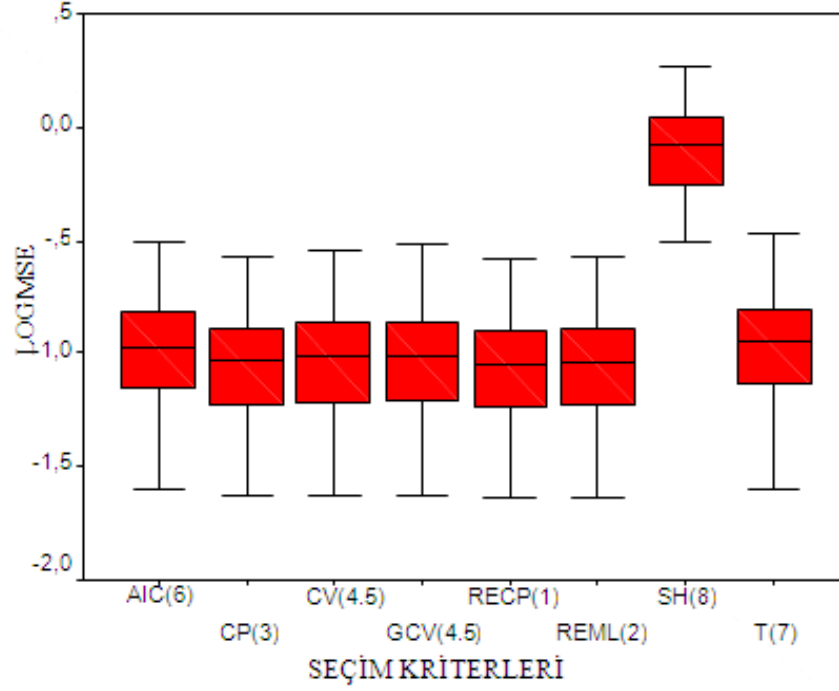
Şekil 5.8. $n=50$ ve $r=2$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



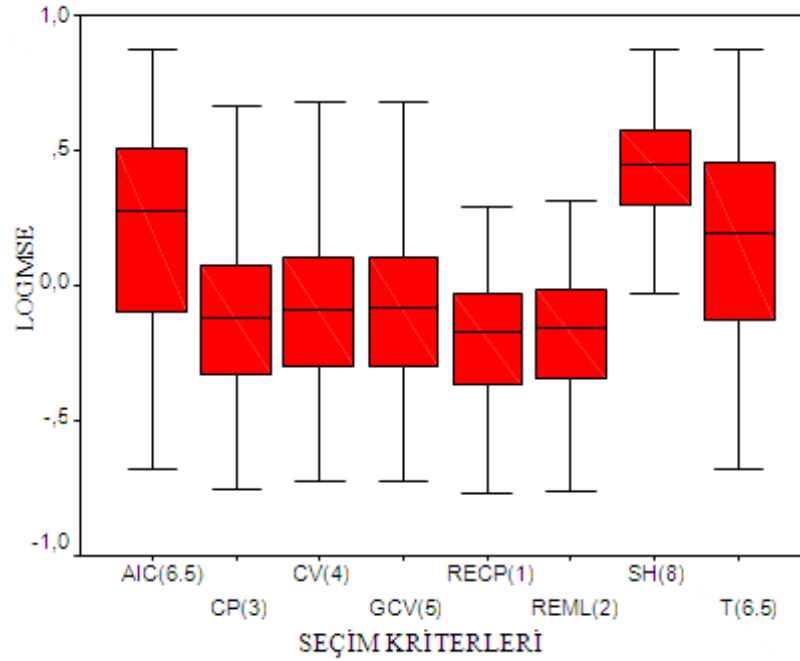
Şekil 5.9. $n=50$ ve $r=3$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



Şekil 5.10. $n=50$ ve $r=4$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



Şekil 5.11. $n=50$ ve $r=5$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



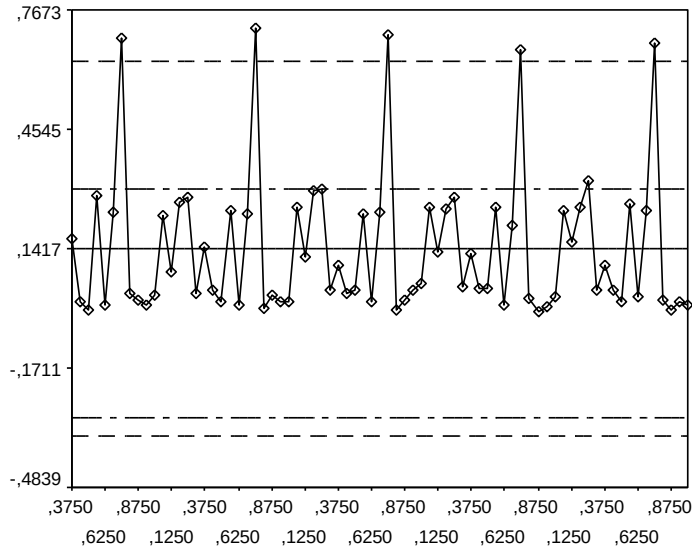
Şekil 5.12. $n=50$ ve $r=6$ düzeyinde 1. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri

Şekil 5.7'den şekil 5.12'ye kadar olan üstteki grafikler, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerinin $n=50$, 1. faktör ve $r=1,2,3,4,5,6$ düzeyi için kutu grafikleridir.

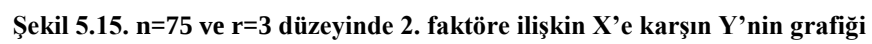
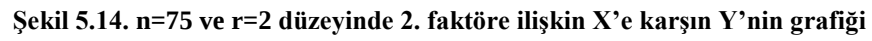
Şekil 5.7'deki kutu grafiğini değerlendirecek olursak, görüldüğü gibi tüm seçim kriterleri aynı sıralamayı almıştır. Yani $r=1$ düzeyi için herhangi bir seçim yöntemi diğerinden daha üstün değildir.

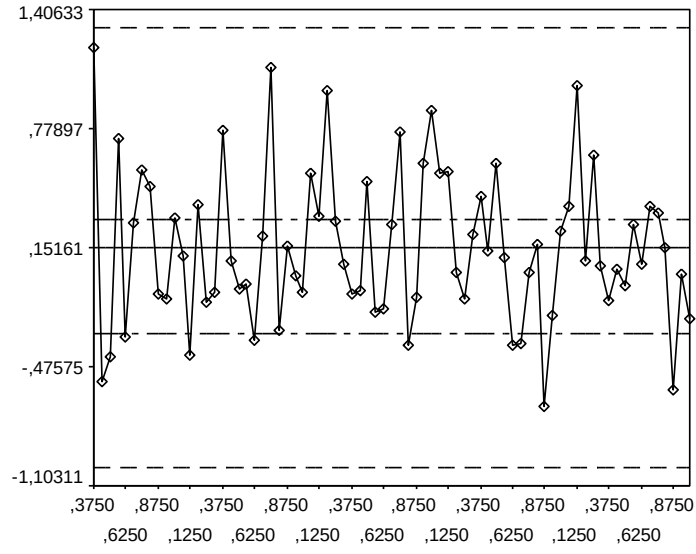
Şekil 5.12'deki kutu grafiğini değerlendirdiğimizde ise, $r=6$ düzeyi için *T-AKAİKE* ikilisinin birbirine üstünlük kurmadığını diğer seçim kriterlerinin ise birbirine karşı üstünlük kurduğunu ve farklı sıra atamalarını aldığı görülmektedir.

Sıralamanın yapılışı ise, Wilcoxon'a bağıntılı şekildedir. Her bir seçim kriterinde MSE değerlerinin medyanı %5 anlamlılık düzeyinde, anlamlı bir şekilde kalan 7 yöntemden az ise 1 sırası atanır. Bir seçim kriteri, MSE değerleri medyanı anlamlı olarak birinden daha büyük fakat kalan altı yöntemden daha az ise 2 sırası atanır, benzer şekilde 3-8 atanır. Farklı medyan değerlerine sahip fakat anlamlı olmayan metotlar aynı ortalamalı sırayı paylaşırlar. $r=6$ düzeyi için kutu grafiğine baktığımızda, bu şekilde sıra atamaları yapılmıştır. Kutu grafiğinde seçim kriterlerinin bitişiğindeki parantez içinde hangi sıra atamasına sahip olduğu görülmektedir.

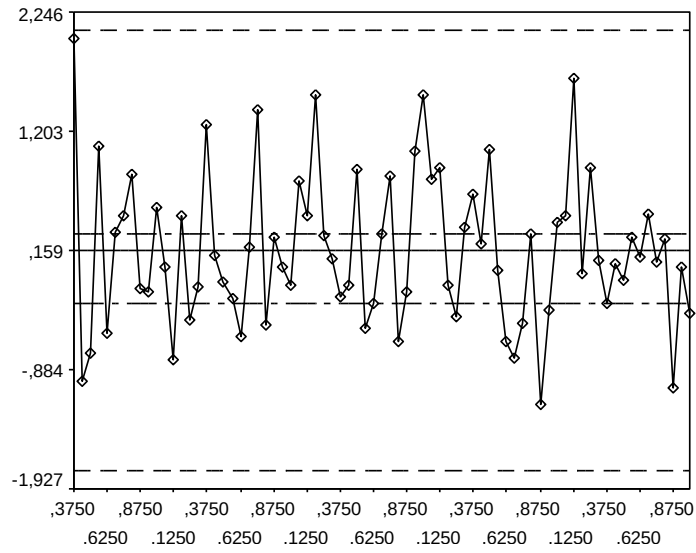


Şekil 5.13. $n=75$ ve $r=1$ düzeyinde 2. faktöre ilişkin X'e karşın Y'nin grafiği

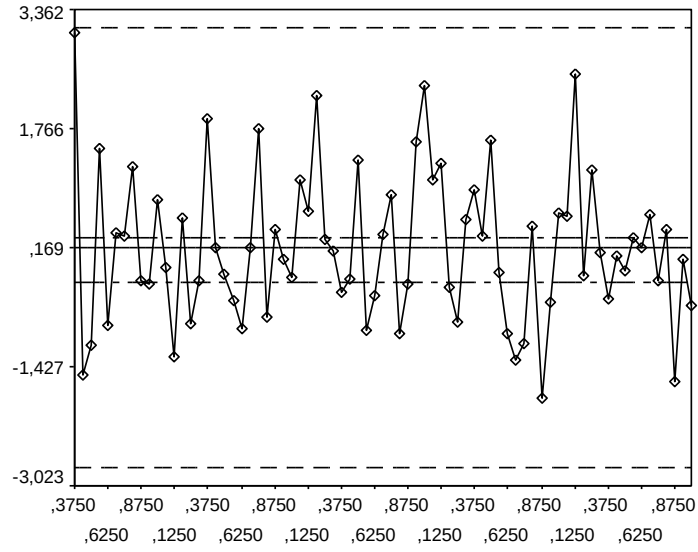




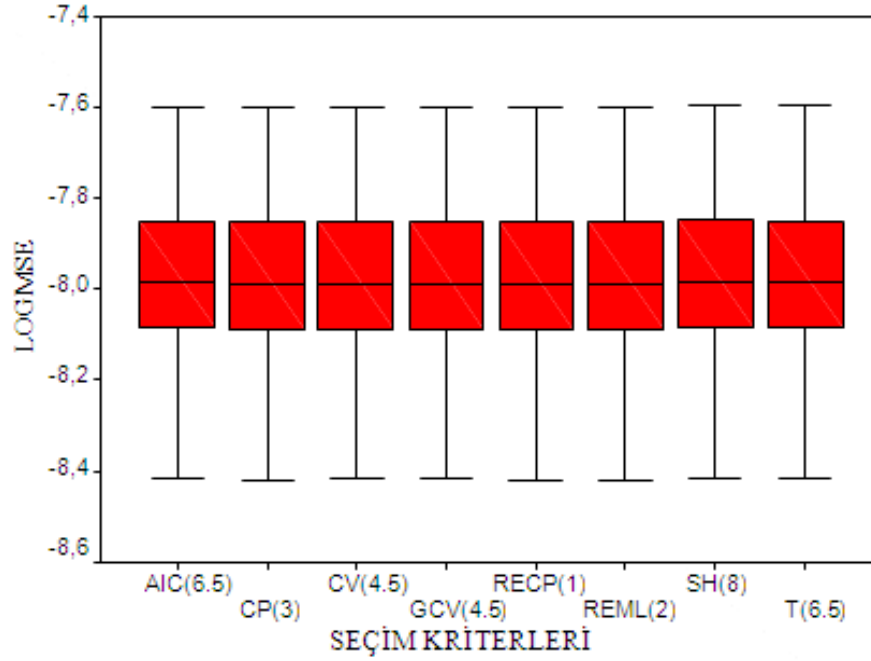
Şekil 5.16. $n=75$ ve $r=4$ düzeyinde 2. faktöre ilişkin X'e karşın Y'nin grafiği



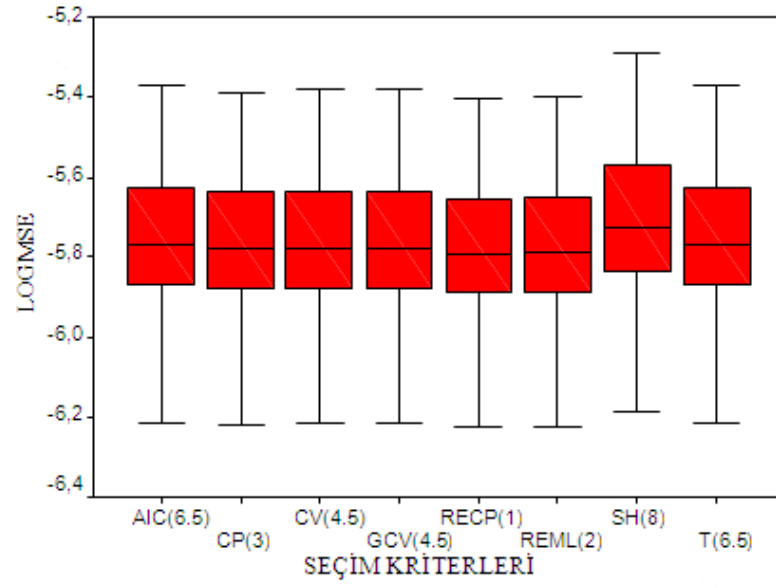
Şekil 5.17. $n=75$ ve $r=5$ düzeyinde 2. faktöre ilişkin X'e karşın Y'nin grafiği



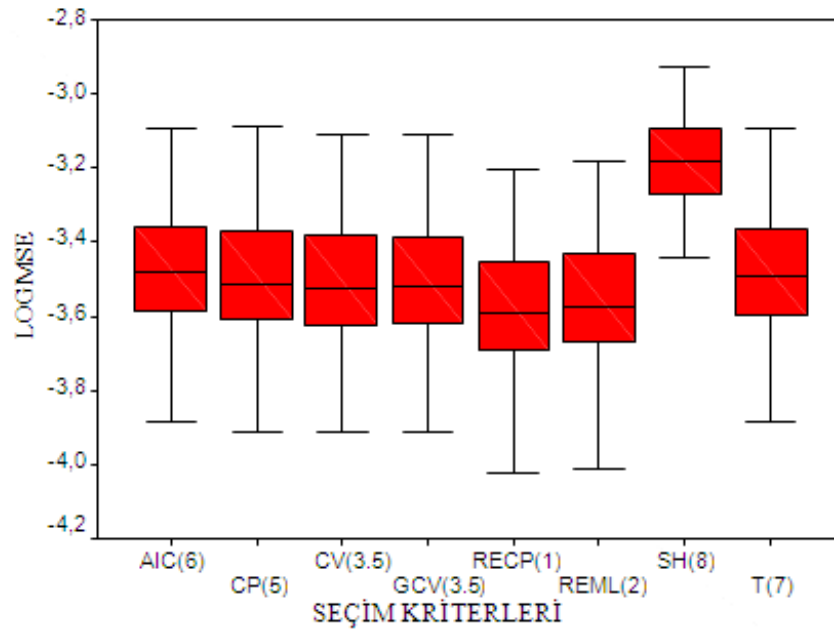
Şekil 5.18. $n=75$ ve $r=6$ düzeyinde 2. faktöre ilişkin X'e karşın Y'nin grafiği



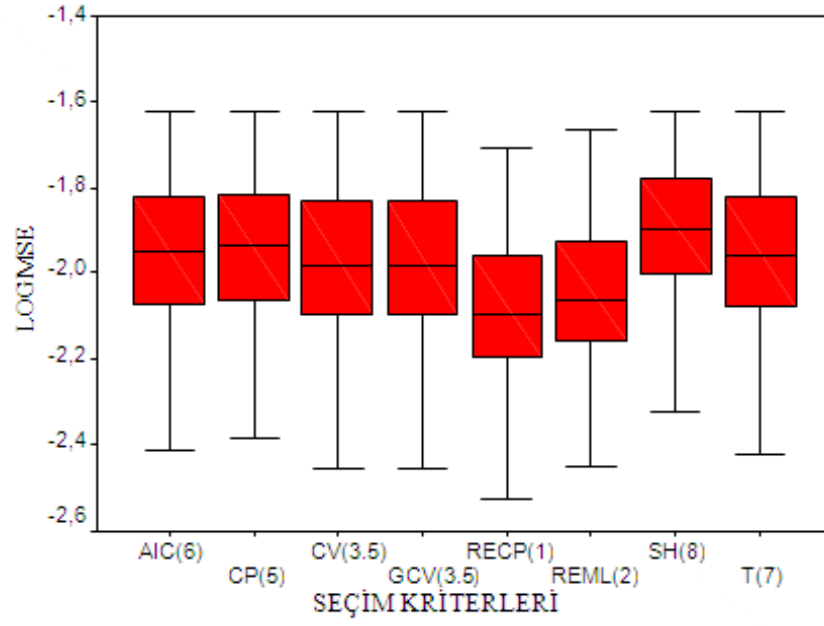
Şekil 5.19. $n=75$ ve $r=1$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



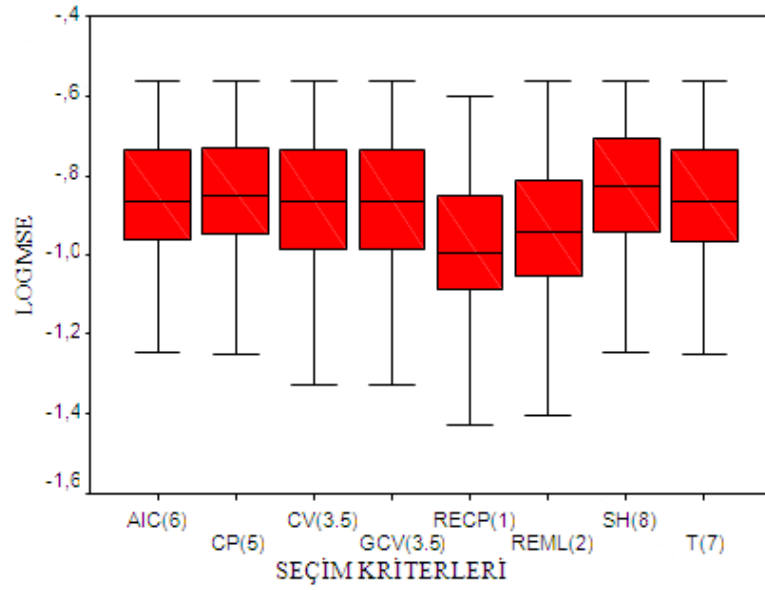
Şekil 5.20. n=75 ve r=2 düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



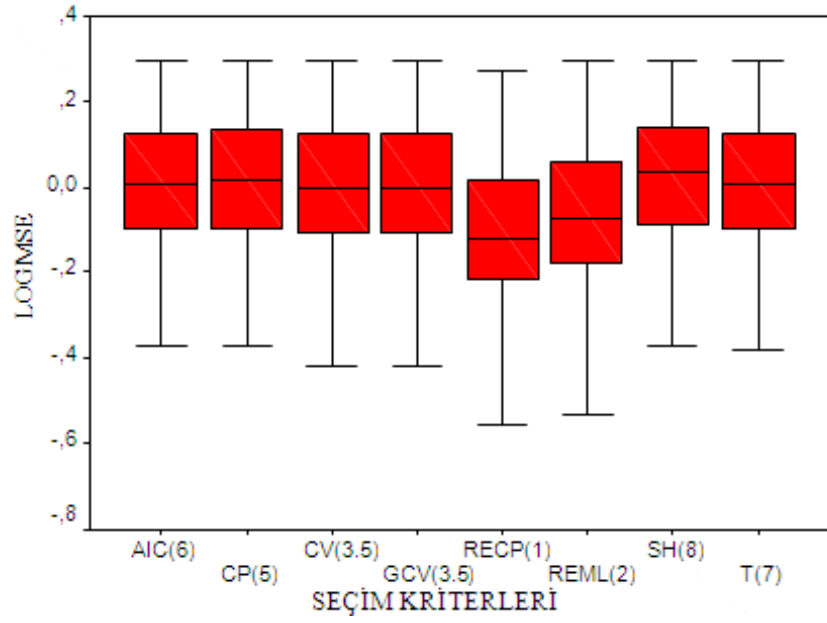
Şekil 5.21. n=75 ve r=3 düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



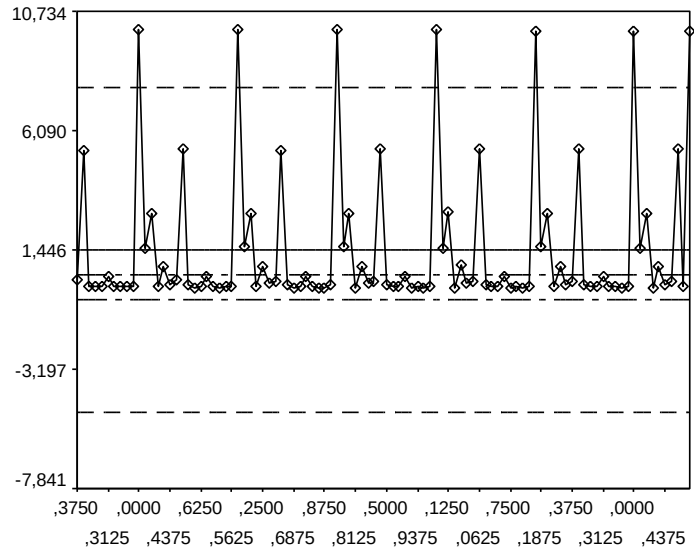
Şekil 5.22. n=75 ve r=4 düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



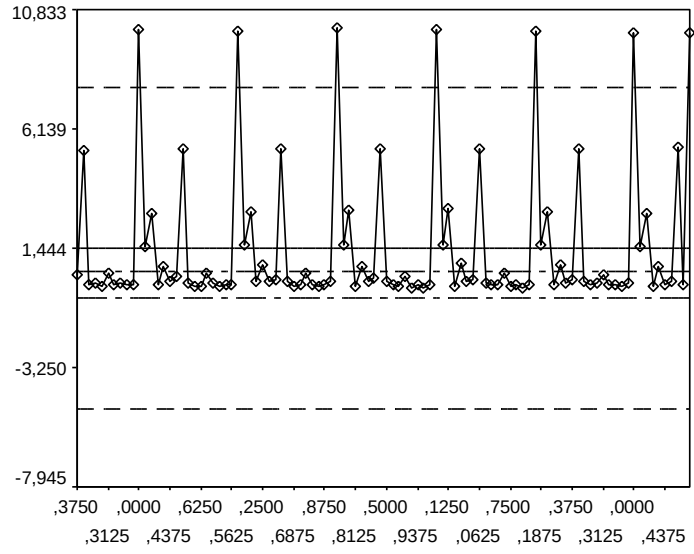
Şekil 5.23. n=75 ve r=5 düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



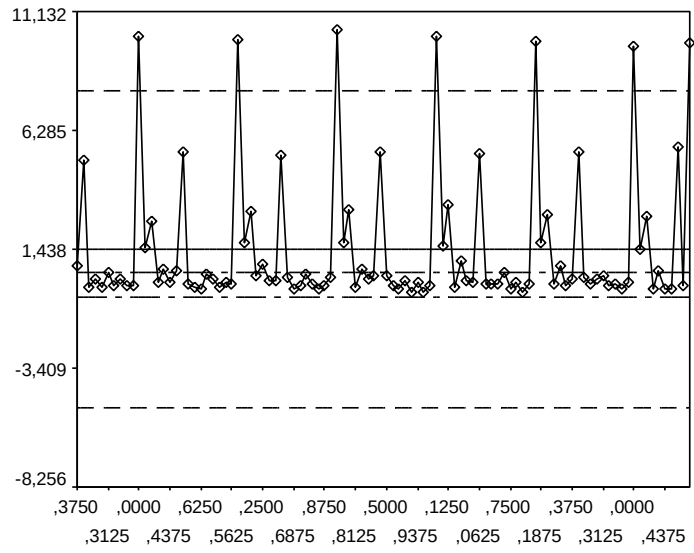
Şekil 5.24. $n=75$ ve $r=6$ düzeyinde 2. faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



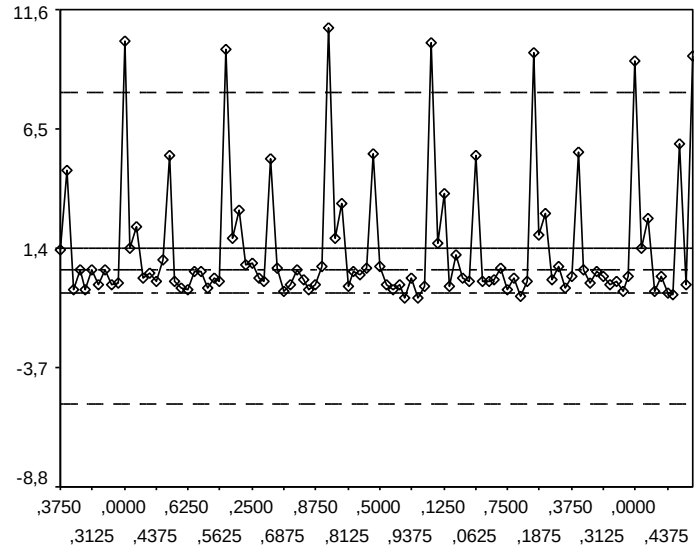
Şekil 5.25. $n=100$ ve $r=1$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



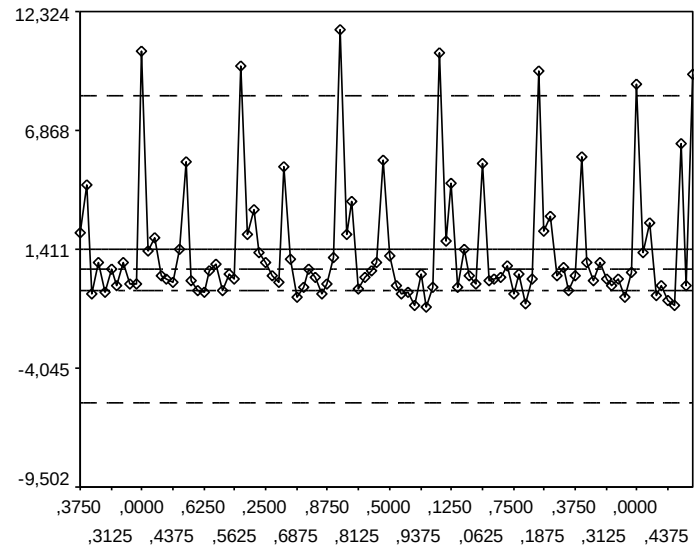
Şekil 5.26. $n=100$ ve $r=2$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



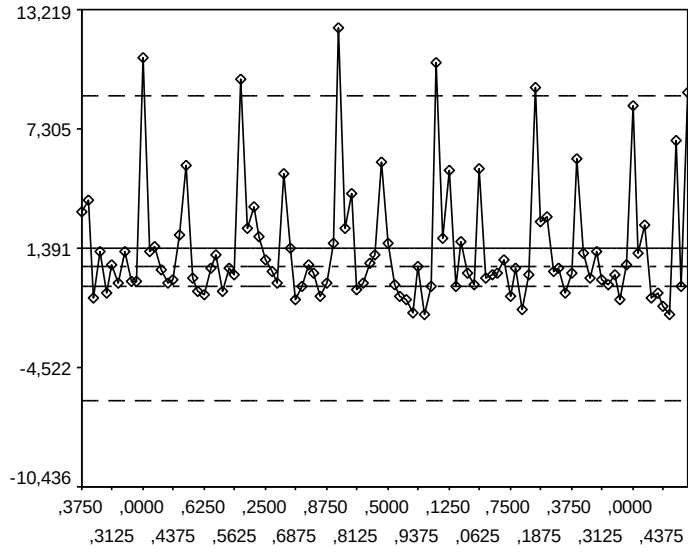
Şekil 5.27. $n=100$ ve $r=3$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



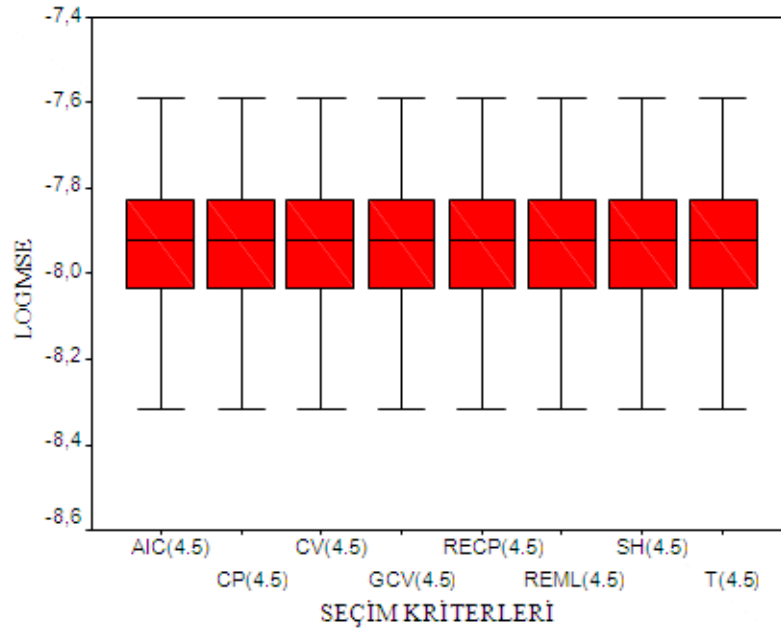
Şekil 5.28. $n=100$ ve $r=4$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



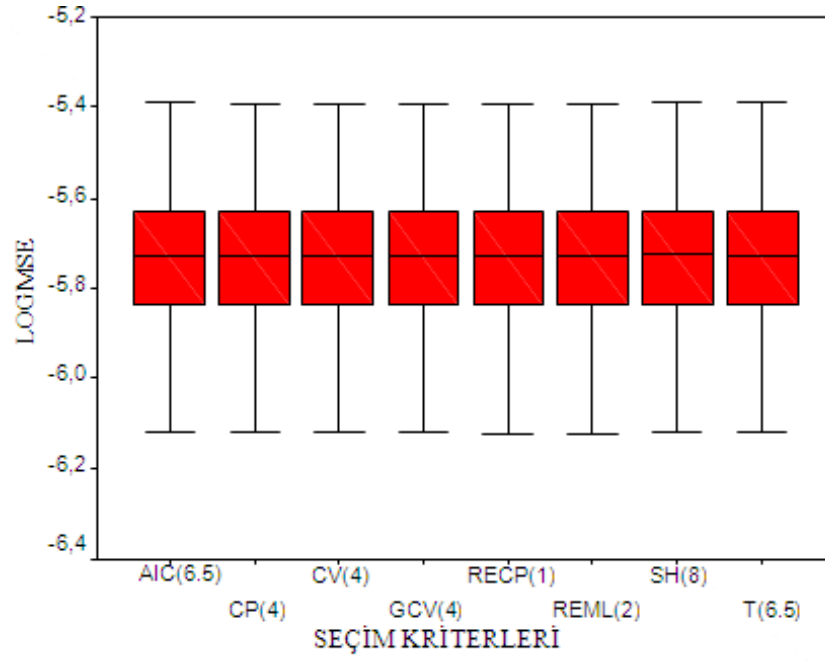
Şekil 5.29. $n=100$ ve $r=5$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



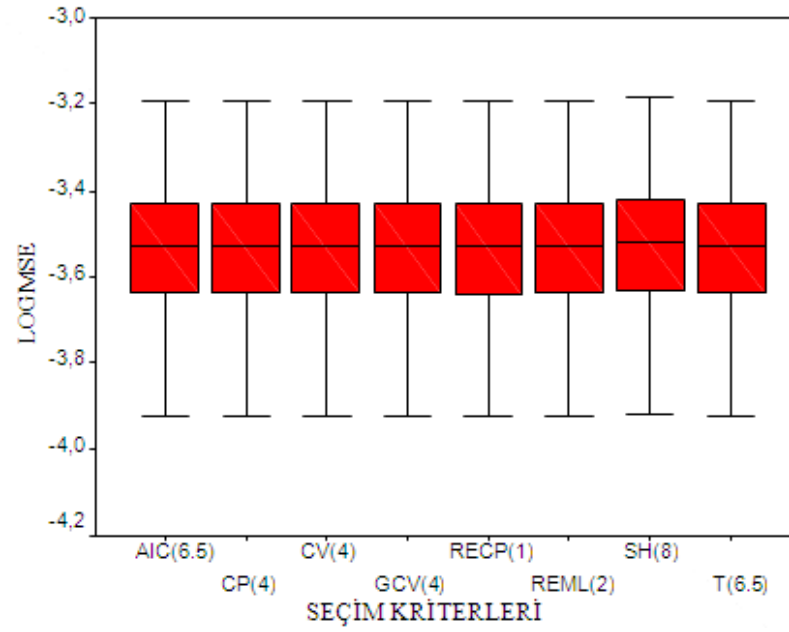
Şekil 5.30. $n=100$ ve $r=6$ düzeyinde 3.faktöre ilişkin X'e karşın Y'nin grafiği



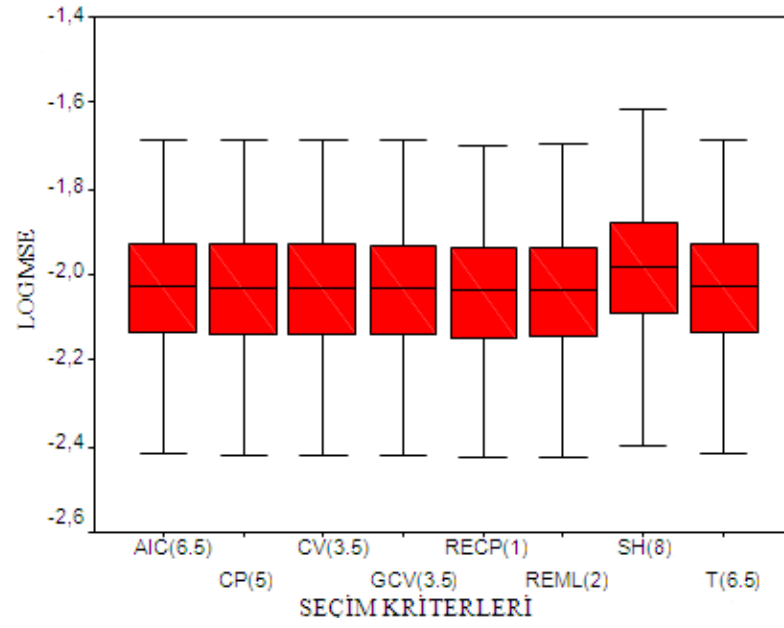
Şekil 5.31. $n=100$ ve $r=1$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



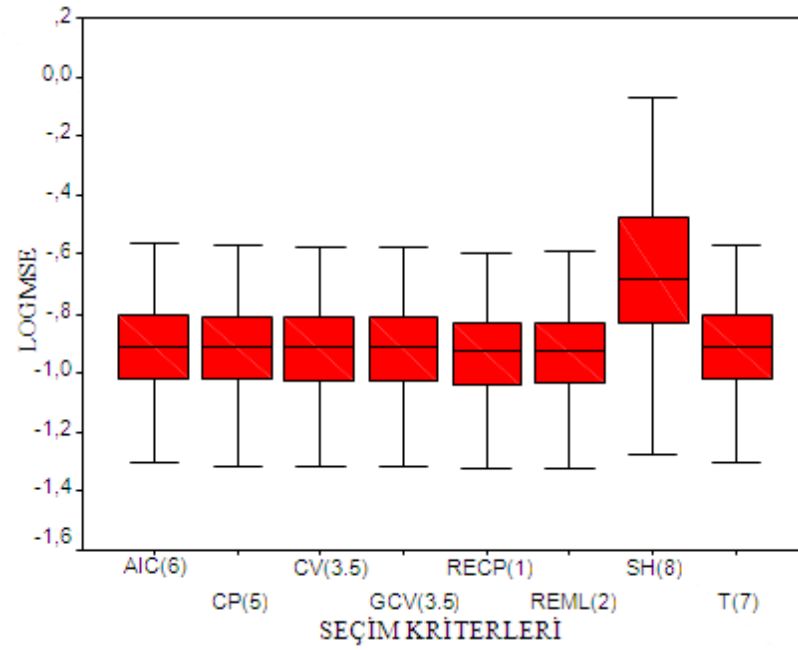
Şekil 5.32. $n=100$ ve $r=2$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



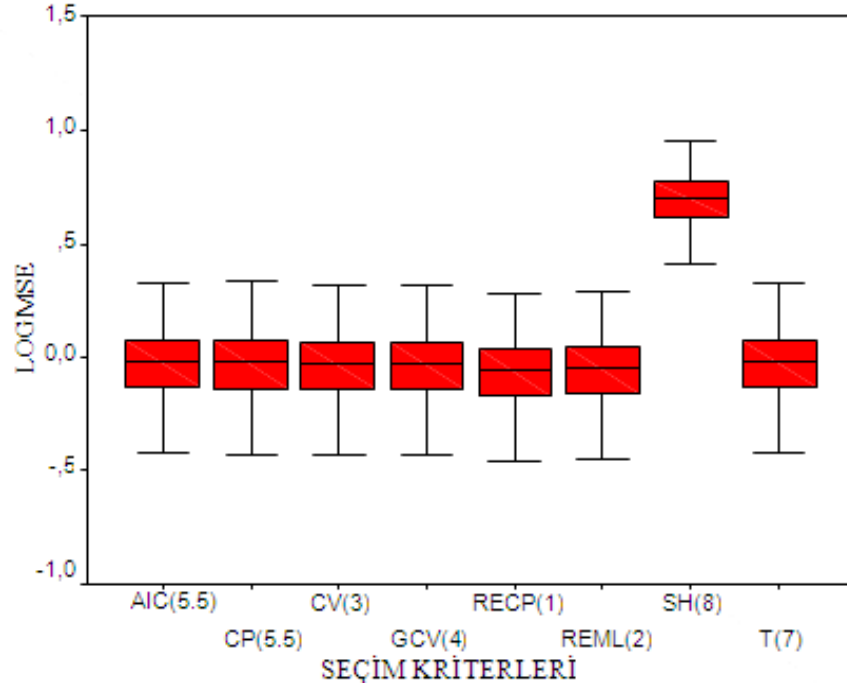
Şekil 5.33. $n=100$ ve $r=3$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



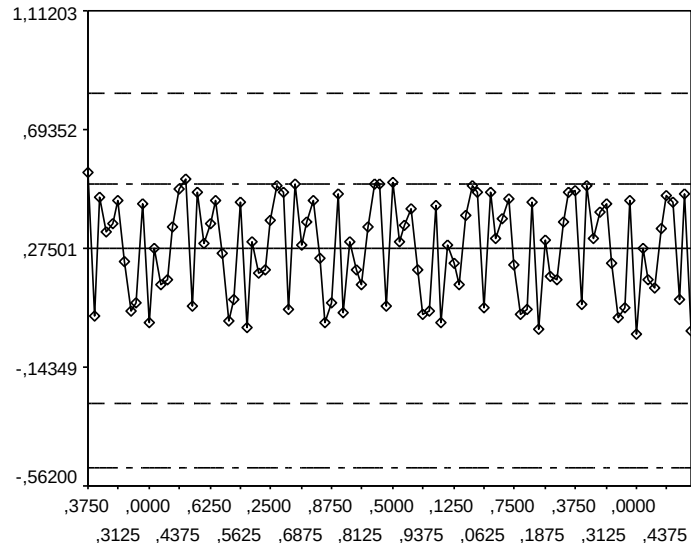
Şekil 5.34. $n=100$ ve $r=4$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



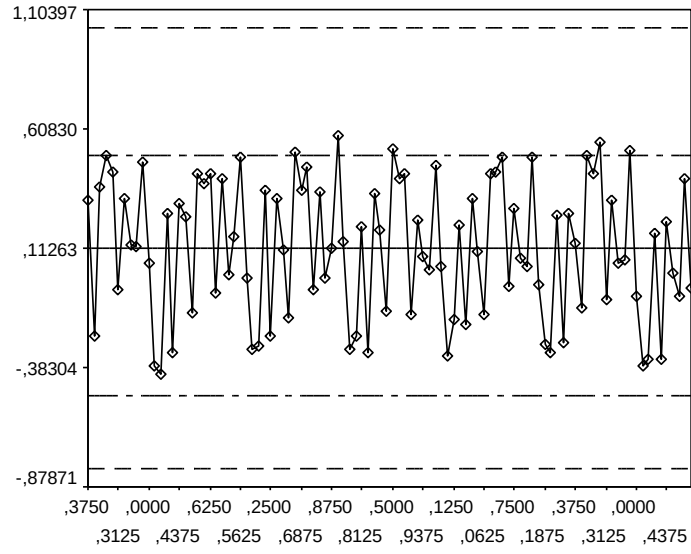
Şekil 5.35. $n=100$ ve $r=5$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



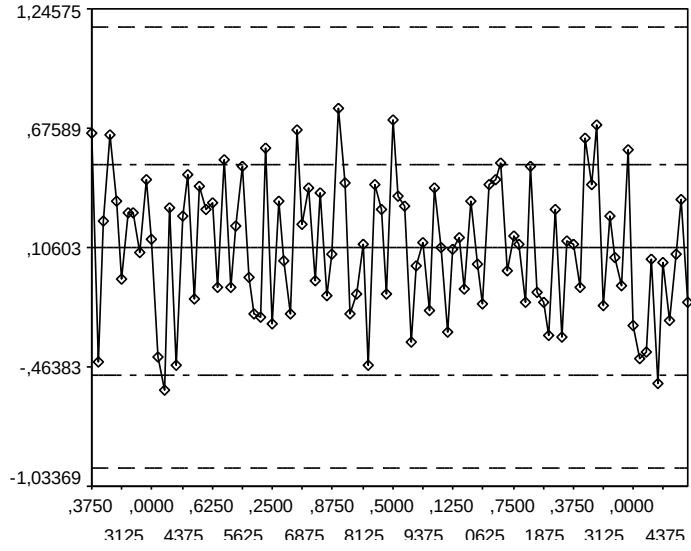
Şekil 5.36. $n=100$ ve $r=6$ düzeyinde 3.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



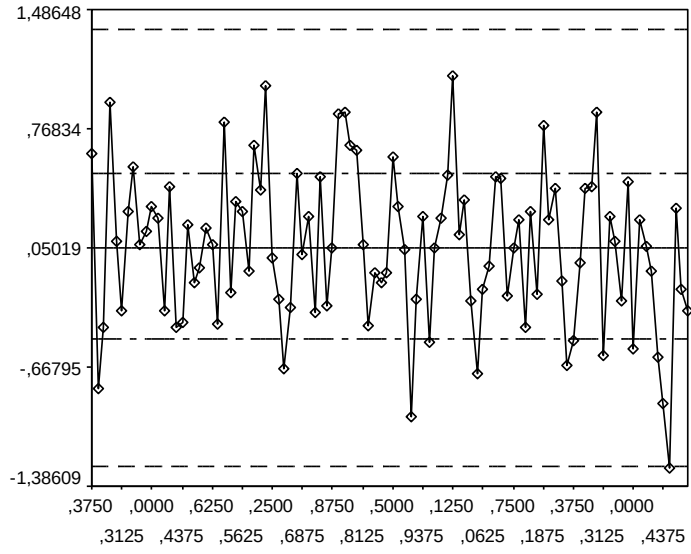
Şekil 5.37. $n=100$ ve $r=1$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



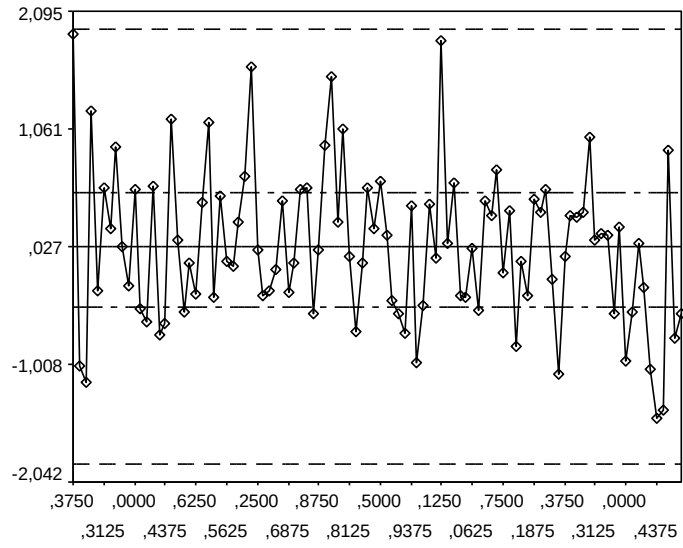
Şekil 5.38. $n=100$ ve $r=2$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



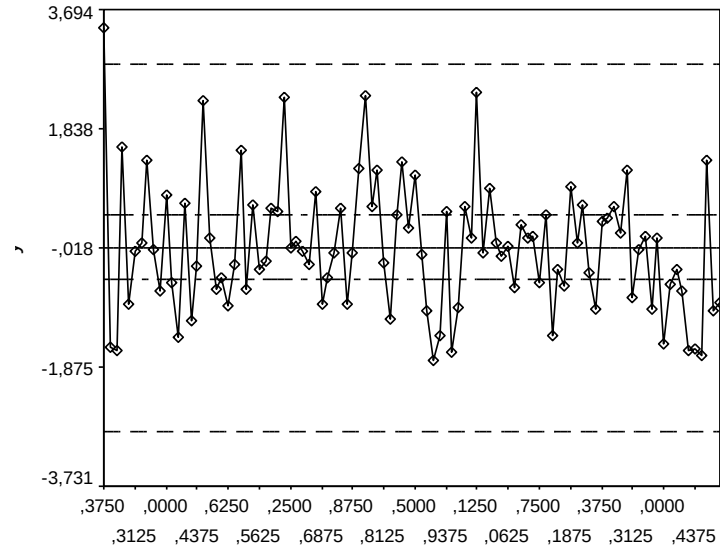
Şekil 5.39. $n=100$ ve $r=3$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



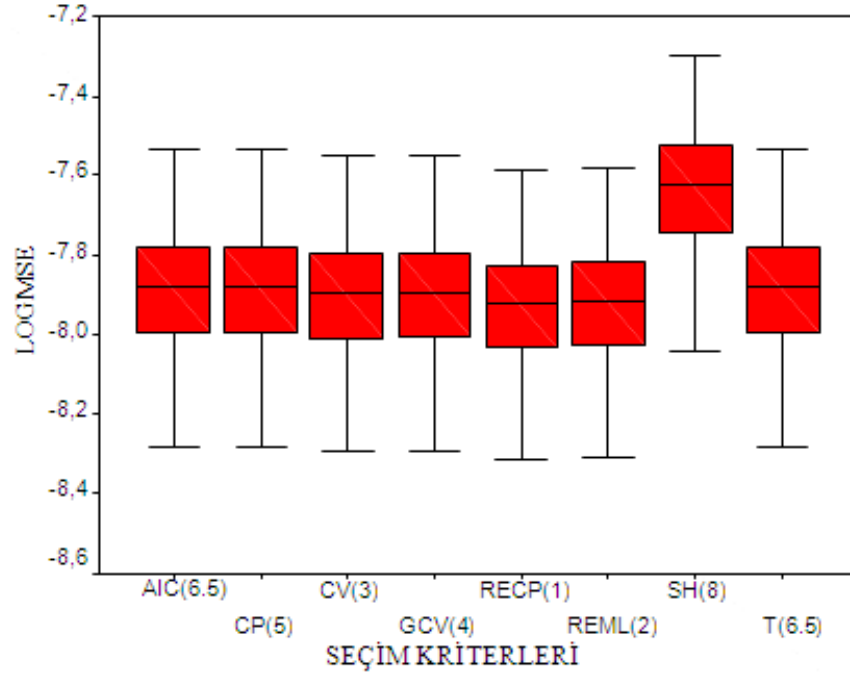
Şekil 5.40. $n=100$ ve $r=4$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



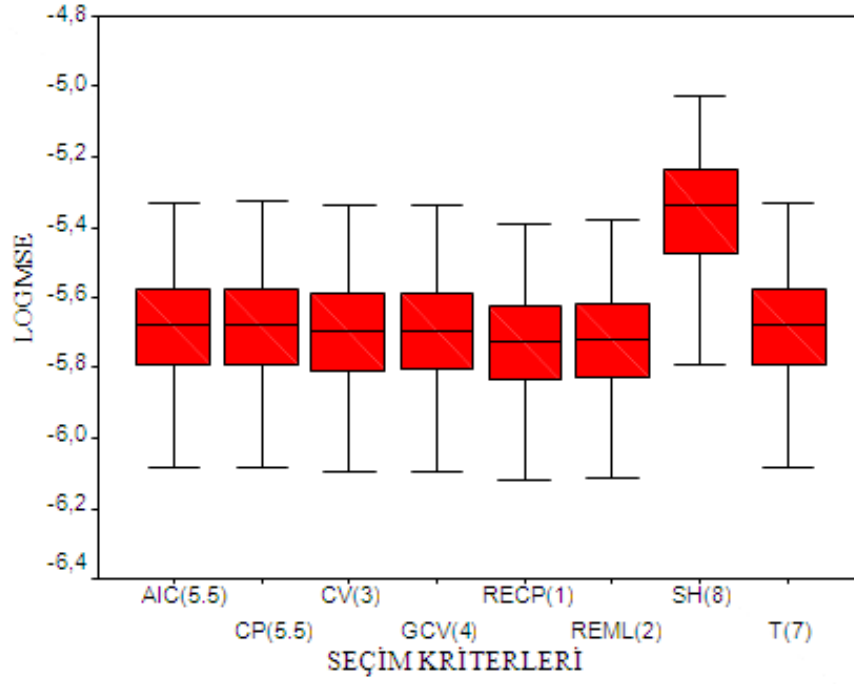
Şekil 5.41. $n=100$ ve $r=5$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



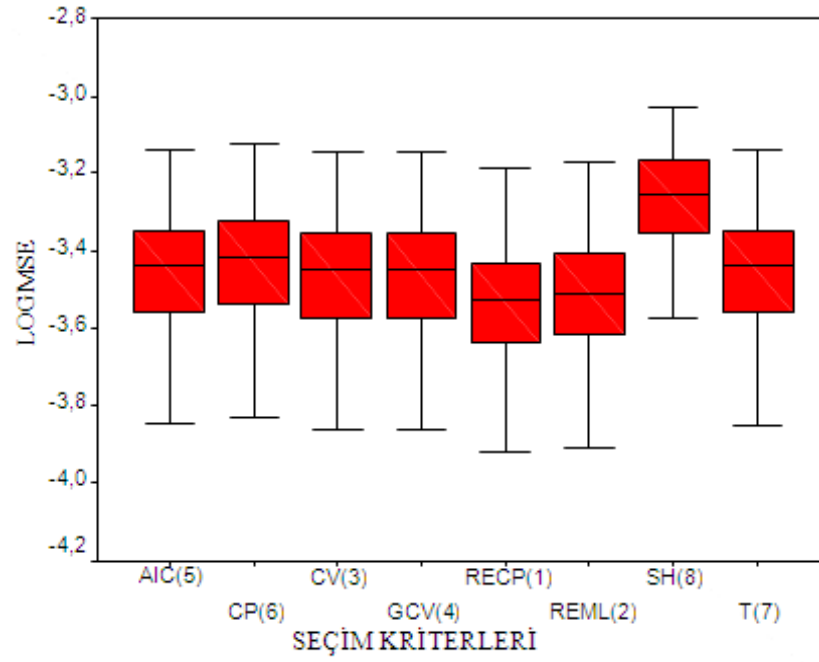
Şekil 5.42. $n=100$ ve $r=6$ düzeyinde 4. faktöre ilişkin X'e karşın Y'nin grafiği



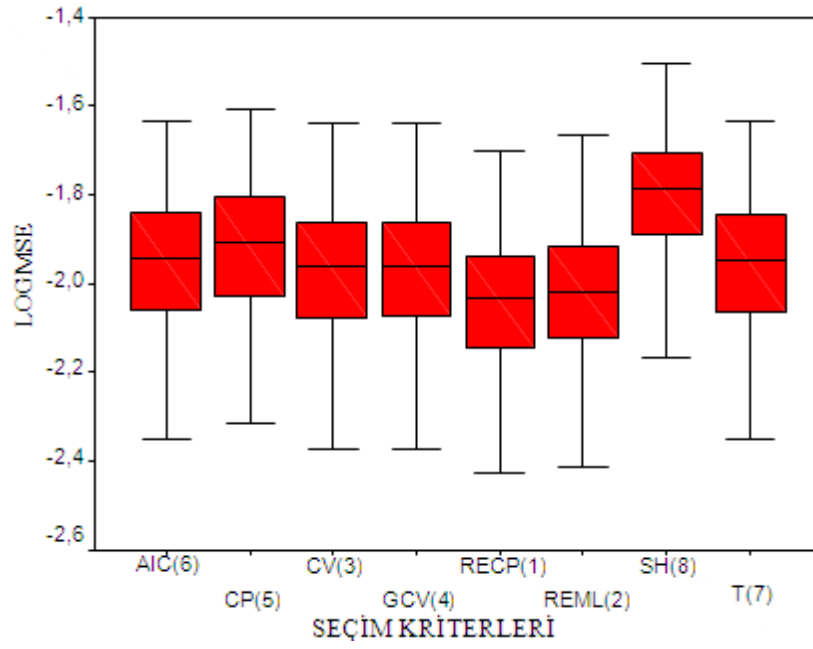
Şekil 5.43. $n=100$ ve $r=1$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



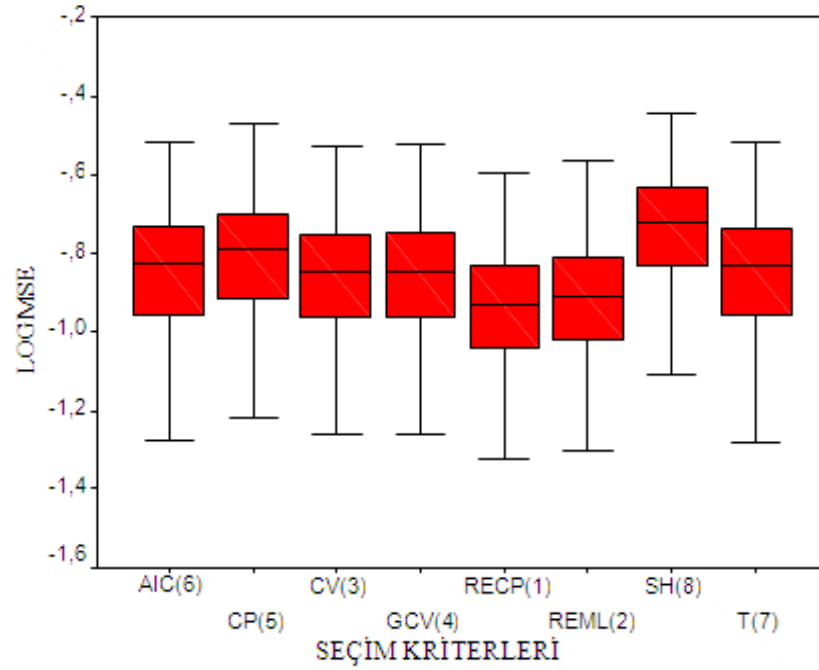
Şekil 5.44. $n=100$ ve $r=2$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



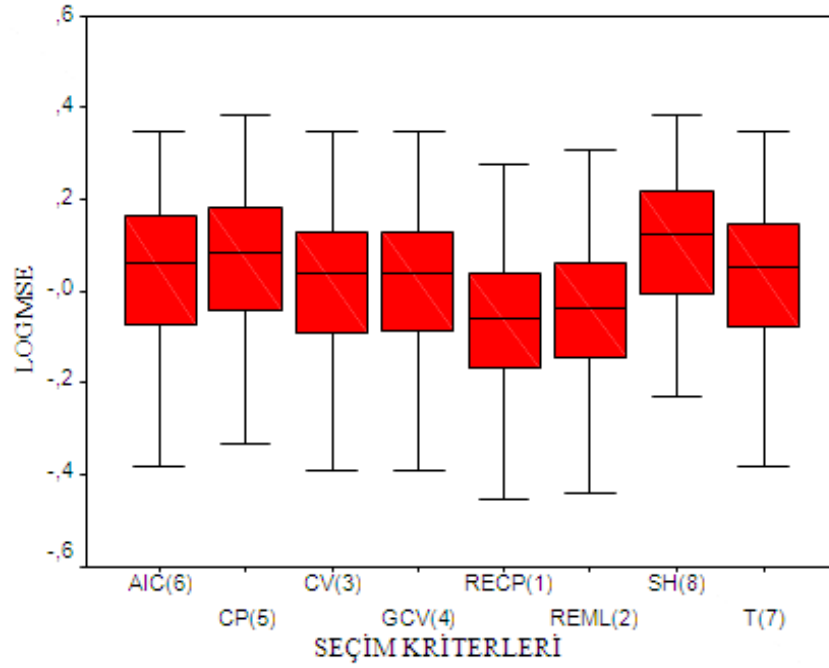
Şekil 5.45. $n=100$ ve $r=3$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



Şekil 5.46. n=100 ve r=4 düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



Şekil 5.47. n=100 ve r=5 düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri



Şekil 5.48. $n=100$ ve $r=6$ düzeyinde 4.faktöre ilişkin, soldan sağa AIC, CP, CV, GCV, RECP, REML, SH ve T için logMSE değerlerinin kutu grafikleri

Simülasyon genelinde tek bir örneklem için 672 wilcoxon testi yapılmıştır. Bu diğer ele aldığımız iki farklı örneklem için de uygulanmıştır ve aşağıdaki sonuçlar elde edilmiştir.

Aşağıdaki çizelge 5.4, 5.5 ve 5.6’da $n=50, 75$ ve 100 ile 8 düzeltme parametresi ve 4 farklı faktör altında $r=1,2,3,4,5,6$ düzeylerinin ortalaması alınan Wilcoxon testi sıralamaları bulunmaktadır.

Mesela CV seçim kriterine ve 1.faktöre karşılık gelen değer, 1.faktör için $r=1,2,3,4,5,6$ düzeylerinin Wilcoxon testi sonucuna göre ve MSE Medyan değerlerine göre sıra atamalarının ortalamasıdır. Diğer seçim kriterlerine karşı gelen değerlerde benzer şekilde elde edilmiştir.

Çizelge 5.4. $n=50$ lik örneklemde 8 düzeltme parametresi için ortalaması alınan Wilcoxon testi sıralamaları

Kriterler	1.Faktör	2.Faktör	3.Faktör	4.Faktör	Ortalama
CV	4.3333	3.4166	4.25	3.3333	3.83
GCV	5.25	4.25	4.5833	4.4166	4.62
C_p	3.0833	4.3333	3	4.25	3.67

Çizelge 5.4. (devam)

Kriterler	1.Faktör	2.Faktör	3.Faktör	4.Faktör	Ortalama
AIC_c	6	6.75	6.1666	6.8333	6.44
SH	6.8333	6.75	6.6666	8	7.06
T	6.3333	6.5	7.3333	6.1666	6.58
$REML$	2.4166	2.25	2.3333	2	2.25
$RECP$	1.75	1.4166	1.6666	1	1.46

Çizelge 5.4'ü incelediğimizde; n=50 birimlik örneklem verisi için 4 farklı faktör etkisi altında ve 100 tekrarlı bir deneyde, $RECP$ kriterinin en iyi yöntem olduğu görülmüştür. $REML$ kriterinin de en iyi 2. yöntem olduğu görülmektedir.

Çizelgenin sağ tarafındaki değerler 4 faktörün genel ortalamasını göstermektedir. Toplam ortalama da CV , GCV ve C_p kriterinin yakın sonuçlar verdiği ve en kötü performans olarak da AIC_c , SH ve T kriterlerinin sergilediği görülmüştür.

Çizelge 5.5. n=75 lik örneklemde 8 düzeltme parametresi için ortalaması alınan Wilcoxon testi sıralamaları

Kriterler	1.Faktör	2.Faktör	3.Faktör	4.Faktör	Ortalama
CV	3.75	4.5	4.1666	3.3333	3.94
GCV	3.75	4.5	4.1666	3.8333	4.06
C_p	4.75	3	3.5	5.5	4.19
AIC_c	5.75	6.5	6	5.3333	5.89
SH	7.4166	8	7.4166	8	7,71
T	5.75	6.5	6.3333	7	6.39
$REML$	2.75	2	2.8333	2	2.39
$RECP$	2.0833	1	1.5833	1	1.42

Çizelge 5.5'i incelediğimizde; n=75 birimlik örneklem verisi için 4 farklı faktör etkisi altında ve 100 tekrarlı bir deneyde, $RECP$ kriterinin en iyi yöntem olduğu görülmüştür. $REML$ kriterinin de en iyi 2. yöntem olduğu görülmektedir.

Toplam ortalama da CV , GCV , C_p ve AIC_c kriterinin yakın sonuçlar verdiği ve en kötü performans olarak da SH ve T kriterlerinin sergilediği görülmüştür.

Çizelge 5.6. n=100 lük örnekleme 8 düzeltme parametresi için ortalaması alınan Wilcoxon testi sıralamaları

Kriterler	1.Faktör	2.Faktör	3.Faktör	4.Faktör	Ortalama
<i>CV</i>	3.75	3.1666	3.75	3	3.42
<i>GCV</i>	3.9166	4	3.9166	4	3.96
<i>C_p</i>	4.9166	5.5	4.6666	5.25	5.08
<i>AIC_c</i>	5.5833	5.5	5.9166	5.83	5.71
<i>SH</i>	7.4166	8	7.4166	8	7.71
<i>T</i>	5.5833	6.8333	6.3333	6.9166	6.42
<i>REML</i>	2.75	2	2.4166	2	2.29
<i>RECP</i>	2.0833	1	1.5833	1	1.42

Çizelge 5.6'yı incelediğimizde ise; n=100 birimlik örneklem verisi için 4 farklı faktör etkisi altında ve 100 tekrarlı bir deneyde, *RECP* kriterinin tekrar en iyi yöntem olduğu görülmüştür. *REML* kriterinin de en iyi 2. yöntem olduğu görülmektedir. Toplam ortalama *CV* ile *GCV* nin ve sonrada *C_p* ile *AIC_c* kriterlerinin yakın sonuçlar verdiği görülmektedir. En kötü performans olarak da *SH* ve *T* kriterlerinin sergilediği görülmüştür.

Simülasyonla elde edilen örneklem verileri için, regresyon fonksiyonların grafikleri ve sekiz düzeltme parametresi seçim yöntemlerinin $\log_e MSE$ değerlerinin kutu grafiklerinden (box plot) bazıları, 5.1-5.48 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*, *C_p*, *CV*, *GCV*, *RECP*, *REML*, *SH* ve *T* kriterlerinin $\log_e MSE$ değerlerin kutu grafiklerini göstermektedir.

Kutu grafikleri incelediğinde, yüksek düzeyli bir regresyon fonksiyonu için *RECP* ve *REML* kriterlerinin az bir farkla da olsa daha üstün olduğu, *SH* kriterinin de bariz bir şekilde kötü performans sergilediği görülmüştür. Ancak, gürültü düzeyinin yüksek olmadığı durumlarda yani faktör düzeyinin 1,2 gibi olduğu durumlarda tüm seçim kriterlerinin çoğu zaman aynı sıralamayı paylaştığı görülmüştür.

5.3. Sonuçlar ve Öneriler

Parametrik olmayan regresyon modellerinin kestiriminde, Nadara Watson kernel düzeltme yöntemi kullanılmış olup, söz konusu yöntem esas itibarıyla, cezalı en küçük kareler toplamının minimum problemine dayanan kısmı ele alınmıştır. Böyle bir probleminin çözümünde, bölüm 4’de incelenen 8 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’den $+\infty$ ’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 regresyonda iyi bir tahmin sonucu veren düzeltme parametresinin seçimi için bir simülasyon çalışması yapılmış olup, özet olarak aşağıdaki sonuçlar elde edilmiştir:

Gerçekleştirilen simülasyon çalışmasında, 4 farklı fonksiyonel faktörlerinin etkileri altında oluşturulan örneklem verilerinde, söz konusu faktörlerin etkilerini belirleyebilmek için 6 faktör düzeyi her bir faktörde ele alınmış ve farklı büyüklükte örneklem için de uygulanmıştır. Toplamda $4 \times 6 \times 3 = 72$ deney yapılmıştır. Söz konusu veriler çizelge 5.1 deki şekliyle 50, 75 ve 100 örneklem altında ve 100 tekrarlı rassal bir şekilde yaratılmıştır. Tüm düzeltme parametresi seçim kriterlerinde en iyi λ düzeltme parametresi bulunmuş ve seçim kriterlerini karşılaştırma yapabilmek için bu en iyi λ düzeltme parametresine karşı gelen *MSE* değerleri TURBO C programıyla hesaplanmıştır. Daha sonra ise *MSE* değerlerinin %5 anlamlılık düzeyinde karşılaştırmaları için, TURBO C programıyla elde edilen tüm *MSE* değerleri SPSS programına aktarılmış ve gereken analizler yapılmıştır.

Özet olarak, simülasyon çalışması sonucunda aşağıdaki gözlemler ortaya konmuştur;

- Genel ortalama *RECP* kriteri, 50, 75 ve 100 birimlik örneklemelerde en iyi performansı göstermiştir.
- Yine simülasyon geneline bakıldığında, faktör düzeylerine göre ve genel ortalama ikincilik sıralamasında en iyi başarıyı *REML* kriteri sağlamıştır.

- Tüm simülasyon deneylerinde iki metot CV ve GCV kriterleri, tüm faktör düzeylerinde çok yakın sonuçlar vermişlerdir. Küçük örneklem için C_p kriteri de bu iki kritere yakın sonuç vermiştir. Fakat örneklem arttıkça C_p kriteri bu iki kriterden ayrılmaya başlamış ve AIC_c kriteriyle yakın sonuçlar vermeye başlamıştır.
- Tüm sayısal deneyin büyük çoğunluğunda, SH ve T kriterleri en kötü performansa sahip olmuşlardır.
- Box plot grafikleri detaylı incelendiğinde, en kötü performansa sahip olan SH ve T dışındaki kalan 6 kriterin özellikle küçük faktör düzeyinde birbirine çok yakın sonuçlar verdiği, faktör düzeyi arttıkça ise azar azar farkla birbirlerinden ayrıştığı gözlemlenmiştir.

Bu durumda yukarıdaki sonuçlara göre, tavsiyemiz şu şekildedir: Tüm faktör düzeylerinde ve genel ortalama, birinci sıralamada en yüksek başarı $RECP$ kriterini, ikinci sıralamada $REML$ kriterini kullanmayı öneriyoruz. Bunun yanı sıra sırasıyla, CV , GCV , C_p ve AIC_c kriterlerinin de ilk iki sırayı paylaşan $RECP$ ve $REML$ kriterlerinden çok az farkla ayrıldığı da göz ardı edilmemelidir.

KAYNAKLAR

- Aydın, D. (2005) *Semiparametrik regresyon modellemede splayn düzeltme yaklaşımı ile tahmin ve çıkarsamalar*, Doktora Tezi, Anadolu Üniversitesi, Eskişehir, 128s.
- Aydın, D. (2007) A Comparison of the nonparametric regression models using smoothing spline and kernel regression, *Int J of Math, Phys and Eng Sci*, 36: 253-257.
- Aydın, D. ve Memmedli, M. (2012) Optimum smoothing parameter selection for penalized least squares in form of linear mixed effect models, *Optim: a J of Math Program and Oper Res*, 61(4): 459-476.
- Cover, T.M. ve Hart, P.E. (1967) Nearest neighbor pattern clasifiction, *IEEE Trans on Inf Theory*, 13: 12-17.
- Craven, P. ve Wahba, G. (1979) Smoothing noisy data by spline function, *Numer Math*, 31: 377-403.
- Eubank, R.L., Kambour, E.L., Kim, J.T., Kliple, K., Reese, C.S. ve Schimek, M. (1998) Estimation in partially linear models, *Comput Stat & Data Anal*, 29: 27-34.
- Eubank, R.L. (1999) *Nonparametric Regression and Spline Smoothing*, 2.Baskı, Marcel Dekker, New York, 338s.
- Eubank, R.L. (2000) Spline Regression, 1-16, Schimek, M.G. (editör), *Smoothing and Regression: Aproaches, Computation and Application*, John Wiley & Sons, USA, 581s.
- Fan, J. ve Gijbels, I. (1996) Local polynomial modelling and additivity in nonparametric regression, *Ann Statist*, 23: 1896-1920.
- Fox, J. (2000) *Nonparametric simple regression: smoothing scatterplots*, Sage publications, California, 87s.

- Gasser, T., Spoka, L. ve Jennen-Steinmetz, C. (1986) Residual variance and residual pattern in nonlinear regression, *Biom*, 73: 625-633.
- Green, P.J. ve Silverman, B.W. (1994) *Nonparametric regression and generalized linear models*, Chapman & Hall, London, 182s.
- Hardle, W. (1991) *Applied nonparametrics methods*, Cambridge University Press, Cambridge, 337s.
- Hardle, W. ve Linton, O. (1994) Applied nonparametrics methods, *Handb of Econom*, 4: 2295-2339.
- Hardle, W., Muller, M., Sperlich, S. ve Werwatz, A. (2004) *Nonparametric and semiparametric models*, Springer, New York, 305s.
- Hastie, T. ve Tibshirani, R. (1990) *Generalized additive models*, Chapman & Hall, London, 335s.
- Hepsağ, A. (2007) *Parametrik olmayan regresyon ve tahmin yöntemleri: imkb’de işlem gören hisse senetlerine ait piyasa riskinin tahmini üzerine bir uygulama*, Yüksek Lisans Tezi, İstanbul Üniversitesi, İstanbul, 126s.
- Hurvich, C.M., Simonoff, J.S. ve Tsai, C.L. (1998) Smoothing parameter selection in nonparametric regression using an improved akaike information criterion, *J R Statist Soc: B*, 60: 271-293.
- Kou, S.C. ve Efron, B. (2002) Smoothers and the C_p , generalized maximum likelihood, and extended exponential criteria, *Jasa*, 97: 766-782.
- Lee, T.C.M. ve Solo, V. (1999) Bandwidth selection for local linear regression: a simulation study, *Comput Statist*, 14: 515-532.
- Lee, T.C.M. (2001) A Stabilized bandwidth selection method for kernel smoothing of periodogram, *Signal Process*, 81: 419-430.
- Lee, T.C.M. (2003) Improved smoothing spline regression by combining estimates of different smoothness, *Stat & Probab Lett*, 67: 133-140.

- Lee, T.C.M. (2003) Smoothing parameter selection for smoothing splines: a simulation study, *Comput Stat & Data Anal*, 42: 139-148.
- Loader, C. (2004) Smoothing: local regression techniques, *CASE*, Berlin, 28s.
- Loftsgaarden, D.O. ve Quesenbery, G.P. (1965) A Nonparametric estimate of a multivariate density function, *Ann of Math Statist*, 36: 1049-1051.
- Mallows, C. (1973) Some comments on C_p , *Techometrics*, 15: 661-675.
- Reinch, C. (1967) Smoothing by spline functions, *Numer Math*, 10: 177-183.
- Rice, J. (1984) Bandwidth choice for nonparametric regression, *Ann Statist*, 12: 1215-1230.
- Ruppert, D., Wand, M.P. ve Carroll, R.J. (2003) *Semiparametrik regression*, Cambridge University Press, Cambridge, 382s.
- Schimek, M.G. (2000) Estimation and inference in partially linear models with smoothing splines, *J of stat plan and inference*, 91: 525-540.
- Slama, R., Werwatz, A., Boutou, O., Ducot, B., Spira, A. ve Hardle, W. (2003) Does male age affect the risk of spontaneous abortion? An aproach using semiparametric regression, *Am J Epidemiol*, 157(9): 815-824.
- Van Heerde, H.J., Leeftang, P.S.H. ve Wittink, D.R. (1999) *Semiparametric analysis to estimate the deal effect curve*, University of Groningen, Groningen, 39s.
- Wahba, G. (1975) Smoothing noisy data by spline function, *Numer Math*, 24: 383-393.
- Wahba, G. (1990) *Spline models for observational data*, University of Wisconsin at Madison, Pensilvanya, 169s.
- Yatchew, A. (2003) *Semiparametric regression for the applied econometrician*, Cambridge University Press, Cambridge, 215s.

Ek A. Kernel düzeltme ve kernel düzeltme parametresinin seçimi konusunda yapılan simülasyon çalışması için algoritma

Adım 1. $x_i \sim U(0,1)$, $i = 1, \dots, n$ olarak elde edilir.

Adım 2. $\varepsilon_i \sim N(0,1)$, $i = 1, \dots, n$ şeklinde n hacmine bağlı olarak elde edilir.

Adım 3. Çizelge 5.1'deki faktörler $r=1,2,3,4,5,6$ düzeyleri ve n örneklem hacmine bağlı olarak belirlenir.

Adım 4. λ değerleri denene yoluyla optimum aralığı belirlenmiştir. $\lambda \geq 0.001$ 'den başlatılır ve bu adıma her geri döndüğünde ifade 0.001 arttırılır.

Adım 5. Denklem 3.26 yardımıyla **W** matrisi elde edilir.

Adım 6. $\hat{f} = W_{\lambda} y$ olarak hesaplanır.

Adım 7. $RSS = \sum_{i=1}^n \hat{y}_i - f_i(x_i)^2$ hesaplanır.

Adım 8. 4. bölümde ifade edilen Kernel düzeltmesinde, düzeltme parametresi seçim yöntemleri, uygun denklemler yardımıyla hesaplanır.

Adım 9. adım 4'e geri dönülür ve farklı λ değerleri için n örneklem hacmine bağlı olarak işlem tekrar edilir.

Adım 10. adım 8'deki seçim kriterlerinin her birinin minimum değeri hesaplanır. bu minimum değere karşı gelen λ belirlenir.

Adım 11. adım 10'daki minimum λ değerine karşı gelen MSE değerleri hesaplanır.

Adım 12. adım 2'ye dönülür ve bu işlem 100 kez tekrar edilir.

Ek B. Mevcut X,Y grafikleri, kutu grafikleri ve Wilcoxon test istatistikleri sonuçlarının elde edilmesi

Adım 1. Ek A’da bulunan ilk üç adım aynen uygulanır.

Adım 2. Çizelge 5.1 de belirtilen formüller yardımıyla n örneklem hacmine bağlı olarak veriler elde edilir.

Adım 3. Bu veriler SPSS programı kullanılarak, X,Y grafikleri, kutu grafikleri ve Wilcoxon test istatistik sonuçları elde edilir.

ÖZGEÇMİŞ

Kişisel Bilgiler

Ad Soyad : Akın Fit
Uyruk : T.C.
Doğum Yeri ve Tarihi : Muğla 1/10/1985
Medeni Hali : Bekar
Telefon : 0 536 386 6172
E-posta : statistics48@hotmail.com

Eğitim

Alınan Derece	Aldığı Kurum/Üniversite	Mezuniyet Yılı
Lise	Turgut Reis Lisesi	2003
Lisans	Ankara Üniversitesi	2007

Yabancı Dil

İngilizce	Başlangıç	Orta	İleri
Yazma		X	
Konuşma		X	
Anlama		X	
Okuma		X	

Bilimsel Faaliyetler

-

Yayınlar

-

Hobiler

- 1.Futbol oynamak ve seyretmek.
- 2.Satranç oynamak.
- 3.Yüzme