T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÜLKELERİN KARBONDİOKSİT EMİSYONLARININ MODELLENMESİNDE PARAMETRİK VE PARAMETRİK OLMAYAN REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRMASI

Murat SALMAN

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Dr. Öğr. Üyesi Elif TUNA

Ağustos, 2020

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÜLKELERİN KARBONDİOKSİT EMİSYONLARININ MODELLENMESİNDE PARAMETRİK VE PARAMETRİK OLMAYAN REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRMASI

Murat SALMAN tarafından hazırlanan tez çalışması 24.08.2020 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Dr. Öğr. Üyesi Elif TUNA
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Dr. Öğr. Üyesi Elif TUNA, Danışman
Yıldız Teknik Üniversitesi _____

Doç. Dr. Atıf Ahmet EVREN, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Nilgün YILDIZ, Üye
Marmara Üniversitesi _____

Danışmanım Dr. Öğr. Üyesi Elif TUNA sorumluluğunda tarafımca hazırlanan Ülkelerin Karbondioksit Emisyonlarının Modellenmesinde Parametrik ve Parametrik Olmayan Regresyon Yöntemlerinin Karşılaştırması başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Murat SALMAN

İmza

Aileme

TEŞEKKÜR

Üniversite eğitim hayatımız boyunca gerek Lisans gerekse Yüksek Lisans yapan biz İstatistik bölümü öğrencilerine yardımlarını esirgemeyen her zaman anlayışlı ve yardımsever olan değerli hocam Dr. Öğr. Üyesi Elif TUNA' ya saygı ve sevgilerimi sunarım. Bununla birlikte söylemem gerekir ki hayatımın her aşamasında her daim desteklerini çok yakından ve samimiyetle hissettiğim sevgili aileme gösterdikleri anlayış ve sonsuz güven için çok teşekkür ederim.

Murat SALMAN

İÇİNDEKİLER

SİMGE LİSTESİ viii

KISALTMA LİSTESİ ix

ŞEKİL LİSTESİ x

TABLO LİSTESİ xi

ÖZET xii

ABSTRACT xiv

1 GİRİŞ 1

1.1 Literatür Özeti 1
1.2 Tezin Amacı 4
1.3 Hipotez 4

2 PARAMETRİK OLMAYAN REGRESYON MODELLERİ 5

2.1 Regresyon Analizi 5
2.2 Parametrik Model Varsayımları 6
2.3 Parametrik Olmayan Regresyon Modelleri 7
2.4 Parametrik ve Parametrik Olmayan Regresyon Modelleri Hakkında 8
2.4.1 Parametrik Olmayan Yöntemlerin Avantajları 8
2.4.2 Parametrik Olmayan Yöntemlerin Dezavantajları 8
2.4.3 Avantajlar ve Dezavantajlar Genel Bir Değerlendirme 9
2.5 Parametrik ve Parametrik Olmayan Modellerin Karşılaştırılması 9
2.5.1 Değişkenler Açısından Karşılaştırma 9
2.5.2 Varsayımlar Açısından Karşılaştırma 10
2.5.3 Fonksiyonel Şekil Açısından Karşılaştırma 10
2.5.4 Veriye Güvenme Açısından Karşılaştırma 10
2.5.5 Kullanım Seçimi Açısından Karşılaştırma 11

2.5.6 Test Yapılması Açısından Karşılaştırma 11

3 DÜZGÜNLEŞTİRME KAVRAMI 12

3.1 Düzgünleştirme Kavramı 12
3.2 Düzgünleştirme Parametresi 13
3.3 Düzgünleştirme Parametresinin Seçimi 14
3.4 Düzgünleştirme Parametresinin Seçiminde Ortalama Hata Kare 14
3.5 Düzgünleştirme Parametresi Seçim Yöntemleri 15
3.6 Penalizing Fonksiyon Seçicisi 17
3.6.1 Plug-in Seçicisi 18
3.6.2 Akaike Bilgi Kriteri Seçicisi 20
3.6.3 Düzeltilmiş Akaike Kriteri Seçicisi 20
3.6.4 Fisher Bilgi ile Akaike Bilgi Kriteri 20

4 KERNEL REGRESYON MODELİ VE TAHMİNCİLERİ 22

4.1 Kernel Regresyon Modeli 22
4.2 Nadaraya-Watson Tahmincisi 23
4.3 En Yakın Komşu Tahmincisi 25
4.4 Kernel Regresyon Teorisi 26
4.5 Belirsizlik ve Güven Aralığı 27

5 UYGULAMA 30

5.1 Parametrik Olmayan Regresyon Analizinin Uygulanması 30
5.1.1 Kernel Regresyon Yöntemi (N-W Tahmincisi) 30
5.2 Parametrik Regresyon Analizi (EKK Yöntemi) 33
5.2.1 Parametrik Regresyon Analizi ile Tahmin 33
5.2.2 Parametrik Regresyonun Varsayımlarının İncelenmesi 34
5.3 Analiz Sonuçlarının Değerlendirilmesi 37

6 SONUÇ VE ÖNERİLER 38

KAYNAKÇA 39

TEZDEN ÜRETILMIŞ YAYINLAR 42

SİMGE LİSTESİ

s	Aralık
I	Birim matris
λ	Ceza parametresi
CV(μ̂)	Çapraz doğrulama skoru
H	Düzgünleştirme matrisi
h	Düzgünleştirme parametresi
$\hat{m}^3(x)$	En yakın komşu tahmincisi
k	En yakın komşu sayısı
$I(.)$	Gösterge fonksiyonu
n	Gözlem sayısı
x, x_0	Hedef gözlem değeri
$d_k(x)$	Hedef gözlem değerinin k. en yakın komşu değerine uzaklığı
m	Hedef gözlem değerine en yakın gözlem sabit oranı
W_i	Kernel ağırlığı
$K(.)$	Kernel fonksiyonu
$\hat{m}(x_0)$	Kernel regresyon tahmincisi
K	Kesim noktası
m(.)	Parametrik olmayan regresyon fonksiyonu
$m''(x)$	Parametrik olmayan regresyon fonksiyonunun eğimindeki değişim oranı
$f(.)$	Olasılık yoğunluk fonksiyonu
$\hat{f}$ (.)	Olasılık yoğunluk fonksiyonu tahmincisi
h_{opt}	Optimal düzgünleştirme parametresi
$W_{ni}(x)$	Regresyon eğrisi ağırlıklar toplamı tahmini
$\hat{h}_T$	Rice'ın düzgünleştirme parametre seçicisi
$\hat{h}_S$	Shibita'nın model seçicisi
S	Şapka matrisi veya düzgünleştirme matrisi

KISALTMA LİSTESİ

AIC	Akaike Bilgi Kriteri
AIC_c	Düzeltilmiş Akaike Bilgi Kriteri
AMISE	Yaklaşık Ortalama Birleştirilmiş Hata Karesi
CV	Çapraz Geçerlilik Seçicisi
EKC	Çevresel Kuznets Eğrisi
EKK	En Küçük Kareler
EKKY	En Küçük Kareler Yöntemi
GSYİH	Gayri Safi Yurt İçi Hasıla
ISE	Birleştirilmiş Hata Kare
KDE	Kernel Yoğunluk Tahmincisi
K-NN	En Yakın Komşu Tahmincisi
KOV	Kovaryans
LCE	Yerel Sabit Tahmincisi
LCSV	En Küçük Kareler Çapraz Geçerlilik Seçicisi
LCV	Benzerlik Çapraz Geçerlilik
MISE	Ortalama Birleştirilmiş Hata Kare
ML	En Çok Olabilirlik Yöntemi
MSE	Ortalama Hata Kare
N-W	Nadaraya-Watson Tahminleyeni

ŞEKİL LİSTESİ

Şekil 5.1 Ülkelerin Kişi Başı $CO2$ Emisyonu ve Gsyih Değerleri [32].....32
Şekil 5.2 $CO2$ Emisyonu ile Gsyih Arasındaki Kernel Düzgünleştirmesi.....33
Şekil 5.3 Hata Kalıntılar Grafiği.....35

TABLO LİSTESİ

Tablo 5.1 Gsyih ile Kurulan Kernel Regresyon Sonuçları.....31
Tablo 5.2 Parametrik Regresyon Analiz Sonuçları.....34
Tablo 5.3 Spearman Korelasyon Katsayısı Sonuçları.....35
Tablo 5.4 Parametrik ve Parametrik Olmayan Regresyon Analiz Sonuçları.....37

ÖZET

Ülkelerin Karbondioksit Emisyonlarının Modellenmesinde Parametrik ve Parametrik Olmayan Regresyon Yöntemlerinin Karşılaştırması

Murat SALMAN

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Danışman: Dr. Öğr. Üyesi Elif TUNA

Gayri safi milli hasıla (Gsyih) bir ülkenin en önemli gelişmişlik göstergelerinden biridir. Kişi başına milli gelirin miktarı bir çok etken sayesinde artış gösterebilir. Gsyih'ın yüksek olan ülkeler genel olarak gelişmişlik seviyesi yüksek olan ülkelerdir. Üretime dayalı gelişmişlik seviyesi ülkelerin sanayi ve teknolojik alanda ileri seviyede olmalarına bağlı olarak değişmektedir. Teknolojik gelişme ve sanayi yatırımları ile birlikte Karbondioksit salınımı miktarındaki artış günümüzde dünyanın ekolojik dengesi ve insan sağlığını etkileyen en büyük problemlerden birisi olmuştur. Ülkeler bunu önlemek adına çeşitli protokoller imzalayarak sanayi üretimindeki karbondioksit salınımının azaltılması için önlemler almaya çalışmaktadırlar. Bu çalışmada amaç kişi başına düşen milli gelir artışı ile kişi başında düşen karbondioksit seviyesi arasındaki artışın yönünü belirleyerek aralarındaki ilişkiyi daha net olarak görebilmeyi sağlamaktır. Bu kapsam dahilinde 2014 yılına ait Open bank data verileri kullanılıp bağımlı değişken olarak Karbondioksit emisyonu bağımsız değişken Gsyih seçilmiştir. Öncelikle parametrik regresyon modeli uygulanarak verilerin normal dağılıma uygun olup olmadığı hata

kareleri ortalaması ve elde edilen grafiksel çıktılar ile incelenmiş, bulunan sonuçlar üzerinden parametrik olmayan regresyon yöntemleri uygulanarak daha iyi R^2 ve ortalama hata kare sonuçları alınmış ve grafiklerle desteklenmiştir.

Anahtar Kelimeler: Karbondioksit, emisyon, Gsyih

ABSTRACT

Comparison of Parametric and Nonparametric Regression Methods in Modeling Carbon Dioxide Emissions of Countries

Murat SALMAN

Department of Statistics

Master of Science Thesis

Advisor: Dr. Lecturer Elif TUNA

Gross national product (Gdp) is one of the most important development indicators of a country. The amount of national income per person can increase due to many factors. Countries with high Gdp are generally those with a high level of development. The level of development based on production varies depending on high level of countries in industry and technology. With the technological development and industrial investment, the increase in the amount of carbon dioxide emissions has become one of the biggest problems affecting the world's ecological balance and human health today. Countries try to take measures to reduce carbon dioxide emissions in industrial production by signing various protocols to prevent this. The aim of this study is to determine the direction between the increase in per capita income and the level of carbon dioxide per capita and to see the relationship between them more clearly. Within this scope Open bank data for 2014 was used and the independent variable of Carbon dioxide emission and Gdp was chosen as the dependent variable. First of all, the parametric regression

model was applied and whether the data was suitable for normal distribution was examined with the mean of the error squares and the graphical outputs obtained. By applying non-parametric regression methods, better R^2 and mean error square results were obtained and supported with graphics.

Keywords: Carbon dioxide, emission, Gdp

1
GİRİŞ

1.1 Literatür Özeti

Ortadoğu ve Kuzey Afrika ülkelerinde ekonomik büyüme ve Karbondioksit (CO_2) emisyonu arasındaki ilişkiyi incelemişlerdir. Birçok alanda ülkeler arasında benzer düzenlemeler olması nedeniyle aralarında korelasyon olduğunu doğrulamışlardır. Bölgesel düzeyde kişi başına gelir ile kişi başına karbondioksit salınımı arasında ters U şekline sahip hipotezin geçerli olduğu ancak ülke düzeyinde (Ürdün hariç) olmadığı sonucuna varmışlardır. Nedensellik tahmini için panel data düzeltme modeli kullandılar ve GSYİH'dan CO_2 emisyonuna gerçekten nedensellik olduğunu buldular, ancak kısa vadede her ülke için ekonomik büyüme düzeyinde güvenilirdi. Son olarak, Tunus dışındaki tüm ülkeler için uzun vadede değişkenlerin ortak hareketini doğruladılar [1].

Toda-Yomato (TY) prosedürünü kullanarak Çin örneğini 1960-1970 yılları arasındaki verilerle incelediler. TY prosedürünü kullanarak gerçek GSYİH'dan enerji tüketimine ve uzun süre enerji tüketiminden CO_2 emisyonuna tek yönlü bir nedensellik buldular. Bununla birlikte, GSYİH ve CO_2 emisyonları arasında bir bağlantı bulamadılar [2].

Jalil ve Mahmud (2009), Çin örneğinde Çevresel Kuznets Eğrisi (EKC) hipotezinin geçerliliğini ilk inceleyen kişiler arasındaydı. Çevresel bozulmayı için kişi başına CO_2 emisyonu ve ekonomik büyüme için kişi başına GSYİH ile otoregresif dağıtılmış gecikme çerçevesi kullandılar. Bulgular EKC hipotezini destekledi. Nedensellik çıkarımıyla ilgili olarak, GSYİH'dan CO_2 emisyonuna kadar tek yönlü standart doğrusal nedensellik buldular [3].

ABD örneğini ve enerji tüketimi ile gelirin CO_2 emisyonu üzerindeki etkisini incelemiştirler. Test öncesi önyargıyı önlemek ve nedenselliği test etmek için Toda-Yomamoto prosedürünü kullandılar. Sonuçlar, CO_2 emisyonu ve gelir arasında nedensellik kanıtı göstermemekle birlikte, enerji tüketiminden CO_2 emisyonuna

kadar tek yönlü bir nedensellik göstermiştir. Bu nedenle, çevre politikalarının CO_2 emisyon fiyatlarındaki düşüş için enerji tüketimini azaltmaya odaklanması gerekmektedir [4].

Aynı ülke için gerçek GSYİH'ya paralel olarak CO_2emisyonu ve nükleer enerji tüketimi arasındaki ilişkiyi incelemişlerdir. Ana odak, nükleer ve / veya yenilenebilir enerji tüketiminin CO_2 emisyonunu azaltmaya yardımcı olabileceği hipotezini kanıtlamaktı. Aynı zamanda, TY modelini kullanarak GSYİH ve CO_2 emisyonları arasındaki nedenselliği test ettiler ve iki yönlü bir nedensellik buldular [5].

Kore'de emisyonların ve ekonomik büyümenin birbirine olan bağımlılığını incelenmiştir. Zaman serisinin doğrusal olmayışını saptamak için düzgün geçişli otoregresif bir model (Star) kullandılar. Bulgular, CO_2 emisyonu ile ekonomik büyüme arasında iki yönlü nedensellik olduğunu göstermektedir. Doğrusal Granger nedensellik testleri uygulandığında yazarlar nedensellik bulamadılar [6].

Türkiye için yapılan başka bir çalışmada CO_2 emisyonu, enerji tüketimi ve gelir arasındaki nedensel ilişkileri incelenmiştir. Veriler 1960-2005 dönemine aittir. Dış ticareti de yaklaşımına dahil etmiştir. Mal üretimi sırasında diğer ülkelerde oluşan kirliliğin aynı çerçeveye eklenmesi gerektiğini savunmuştur. İlk olarak EKC hipotezini reddetti ve daha sonra ARDL eşbütünleşme prosedürü ve Granger nedensellik testi ile devam etti. Testler, CO_2 emisyonları ile kısa vadede ve uzun vadede ki gelir arasında iki yönlü bir nedensellik olduğunu göstermektedir [7].

CO_2 emisyonu, enerji tüketimi, ekonomik büyüme ve ticaret arasındaki ilişkiyi araştırmak araştırılmıştır. ARDL ve Gregory-Hansen eşbütünleşme testlerini kullandılar ve verileri için EKC hipotezini doğrulamayı başardılar. Araştırmalarının bir sonraki adımında bir Granger nedensellik testi uyguladılar ve gerçek GSYİH'nın hem kısa hem de uzun vadede CO_2 emisyonuna neden olduğunu buldular [8].

Rusya için, çıktının emisyon üzerindeki etkisini bulmak ve EKC hipotezini desteklemek veya reddetmek için eşbütünleşme testleri uygulamıştır. Sonuçlar, gerçek çıktıdan kaynaklanan emisyonlar üzerinde olumsuz bir etki göstererek hipotezin reddine yol açtı. Araştırmanın nedensellik ile ilgili bölümünde CO_2

emisyonu, GSYİH ve enerji kullanımı arasında çift yönlü bir Granger nedensellik olduğunu buldular [9].

Hindistan örneğinde, enerji tüketimi, CO_2 emisyonu ve gelir arasındaki nedenselliği incelenmiştir. Ölçeklendirmeyi önlemek için birleştirilmiş veri kullanmışlardır. Veriyi 1971 ile 2006 yılları arasında sınırlandırıp Toda Yamamoto yapısını kullanıldı. Sonuçlar, enerji tüketimi ile gelir arasında nedensellik ve enerji tüketimi ile CO_2 emisyonu arasında iki yönlü bir nedensellik olduğunu göstermiştir. En önemlisi, kısa vadede emisyon ve gelir arasında her iki yönde de nedensellik göstermemiştir. Örneklemin dışındaki bir dönemde nedenselliğin gücünü ve değişkenlerin yeniliklere nasıl tepki verdiğini ölçmek için Genelleştirilmiş Etki-Tepki İşlevi (GIRF) kullanmışlardır. Kısa vadede ortaya çıkan nedensellik sonuçlarının aksine, gelir ve CO_2 emisyonları arasında bir etkileşim gözlemlediler [10].

19 ülkenin finansal gelişimini ve enerji tüketimi ile CO_2 emisyonu arasındaki nedensel ilişkilere odaklanılmıştır. Pedroni eşbütünleşme testini kullanarak elde edilen bulgular ile CO_2 emisyonu ve finansal gelişme arasında uzun soluklu bir ilişkiye işaret etmiştir. Granger nedensellik testi, kişi başına GSYİH ile CO_2 emisyonu arasında iki yönlü bir ilişki olduğunu göstermiştir [11].

Fransa verisini aynı kapsamda incelenmiştir. Bulgular, ekonomik büyümeden emisyon artışına tek yönlü bir nedenselliğe işaret etmiştir ve Ekc hipotezini desteklendi [12].

Fei Li ve arkadaşları 22 yıl boyunca Çin'in 30 eyaletinde çalışmıştır. Pedroni'nin panel eşbütünleşme testini kullanmışlardır ve kişi başına GSYİH'da %1'lik bir artışın enerji tüketiminde neredeyse % 0.50 ve CO_2 emisyonunda % 0.41 - 0.43 oranında bir artışa yol açtığını buldular [13].

Wang ve arkadaşları panel veri analizini ve değişkenlerine GSYİH'nın karesini ekleyerek önceki çalışmalardan kendilerini ayırmışlardır. Büyüme ve emisyonlar arasında U şeklinde bir ilişki bulduklarından EKC hipotezini reddettiler. Uzun vadede, sonuçları ekonomik büyüme enerji tüketiminden CO_2 emisyonuna ve CO_2

emisyonu ile ekonomik büyümeden enerji tüketimine bir nedensellik olduğunu gösterdi [14].

İran'daki ekonomik büyüme, emisyonlar ve fosil yakıt tüketimi arasındaki nedensel ilişkilere odaklanılmıştır. Sonuçlar, GSYİH'dan CO_2 emisyonlarına nedensellik olduğunu gösterdi. Ayrıca, GSYİH'dan enerji tüketimine nedensellik olduğunu gösterdiler. Toplam yakıt tüketim oranı CO_2 emisyonlarını etkilemedi [15].

Türkiye örneğinde ekonomik büyüme, enerji türketimi, istihdam oranı ve CO_2 emisyonları arasındaki nedensel ilişkiye odaklanılmıştır. Yaklaşımları Ardl modelini kullanarak eşbütünlemeyi test etmek ve hata düzeltme tabanlı Granger nedensellik modellerine sahip değişkenler arasındaki nedenselliğin yönünü bulmaktı. Bulguları, kişi başına düşen GSYİH ile kişi başına düşen CO_2 emisyonu arasında nedensellik kanıtı olmadığını göstermektedir. [16].

Ardl sınır testi yaklaşımını ve Hindistan örneğinde Johansen prosedürü kullanılmıştır. Enerji tedarikini, yatırımını ve istihdamını çok değişkenli yapıya dahil etti. CO_2 ve GSYİH arasında uzun vadeli bir nedensellik kuramadı. Kısa vadede çalışma, ikisi arasında iki yönlü bir ilişki olduğunu gösterdi. [17].

1.2 Tezin Amacı

Tez çalışmasının amacı, parametrik olmayan regresyon modeli kullanarak veri setinden elde edilen grafikler ver değerlerle ülkelerin kişi başına CO_2 salınımları ile GSYİH arasındaki ilişkinin yönünü, kullanılan yöntem ile daha iyi bir şekilde açıklamayı hedeflemektedir.

1.3 Hipotez

CO_2 emisyonlarının çevresel bozulma üzerindeki olumsuz etkisi daha önce yapılan bilimsel çalışmalarda ortaya konulmuştur. Bu çalışmada GSYİH değişkeni ile CO_2 emisyonu arasındaki ilişkinin parametrik olmayan regresyon yöntemleri ile parametrik regresyon yöntemlerinden daha iyi sonuçlar vereceği savunulmaktadır. Ortalama Hata Kriteri ve R^2 ile farkları ortaya koyulmuş ve grafikler ile desteklenmiştir.

2
PARAMETRİK OLMAYAN REGRESYON MODELLERİ

2.1 Regresyon Analizi

Regresyon analizi herhangi bir değişkenin (bağımlı değişken) bir veya birden fazla değişkenle (bağımsız – açıklayıcı) arasındaki ilişkinin matematik bir fonksiyon şeklinde yazılmasıdır. Bu fonksiyona regresyon denklemi adı verilmektedir [18].

Regresyon analizi ile ilgili ilk çalışmalar Francis Galton tarafından 19. Yüzyılın sonlarında yapılmıştır. Galton çalışmasında anne babaların boyu ile çocuklarının boyları arasındaki ilişkiyi incelemiş ve kısa boylu anne babaların çocuklarının kısa, uzun boylu anne babaların çocuklarının boylarının uzun olmasına rağmen, çocuklarının boylarının anakütle boy ortalamasına doğru yaklaşma eğiliminde olduğunu görmüştür. Bu eğilimi "ortaya doğru çekilme = regression to mediocrity" olarak adlandırmıştır. Galton'un çalışmaları bugün değişkenler arasındaki istatistik ilişkilerini inceleyen "Regresyon Analizi (Regression Analysis)"nin başlangıcı olmuştur [19].

Geniş bir yelpazede regresyon analizi yapabilmek birçok konuda bize yardımcı olur. Regresyon analizini aşağıda bahsedilen konularda kullanabilirsiniz:

- Birden fazla bağımsız değişkeni modellemede
- Sürekli ve kategorik değişkenleri dahil etmede
- Eğriliği modellemek için polinom terimleri kullanmada
- Bir bağımsız değişkenin etkisinin başka bir değişkenin değerine bağlı olup olmadığını belirlemek için etkileşim terimlerini değerlendirmede

Örneğin, aşağıdaki konular hakkında araştırma yapmak istiyorsanız regresyon analizini kullanmalısınız.

- Sosyo-ekonomik statü ve ırk eğitimdeki başarıyı etkiler mi?
- Eğitim ve IQ kazançları etkiler mi?
- Egzersiz alışkanlıkları ve diyet kiloyu etkiler mi?

- Kahve ve sigara içmek mortalite riskiyle ilişkili midir?
- Belirli bir egzersiz müdahalesinin, diğer fiziksel aktivitelerden farklı bir etkisi olan kemik yoğunluğu üzerinde bir etkisi var mıdır?

Regresyon analizi her zaman uygun bir çözüm vermeyebilir, analizden verimli sonuçlar almak için modelin çok karmaşık veya çok fazla sayıda değişken kullanıldığı durumlarda grafiklerden yardım almak anlamlı olacaktır. İki değişken arasındaki ilişkiyi inceleme noktasında saçılım grafiklerinden yardım alınabilir.

2.2 Parametrik Model Varsayımları

Parametrik bir modele ait varsayımlar aşağıdaki şekilde sıralayabiliriz:

- Hata terimleri rassaldır ve beklenen değeri 0'dır.

$$E(u_i) = 0. \tag{2.1}$$

- Hata terimlerine ait varyanslar sabit ve gözlemden gözleme değişmemektedir.

$$E(u_i^2) = \sigma_u^2. \tag{2.2}$$

- Birbirini takip eden hata terimleri arasında bir ilişki yoktur,

$$KOV\ (u_i, u_j) = 0\ i \neq j. \tag{2.3}$$

- Bağımsız değişkenler ve hata terimleri arasında bir ilişki olmadığı varsayılmaktadır. $KOV\ (u_i, X_j) = 0.$ (2.4)

Burada konuya açıklık getirmek için birkaç açıklama yapmak gerekmektedir. Parametrik yöntemler kullanarak elde ettiğimiz regresyon denklem ilişkilerinin doğrusal olduğunun varsayılması, bağımsız değişkenlerimizin önündeki katsayıların da doğrusal olduğu anlamını taşımaktadır. Parametrik modelimizde yer verdiğimiz bağımlı ve bağımsız değişkenlerimizin doğrusal olma koşuluna bakılmamaktadır. Aşağıda yer alan denklemler parametrik regresyon denklemlerine örnek olarak verilebilir.

$$Y_i = \beta_0 + \beta_1 X_i \tag{2.5}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 \tag{2.6}$$

$$Y_i = \beta_0 + \beta_1 \sqrt{X_i} \tag{2.7}$$

$$Y_i = \beta_0 + \beta_1 \left(\frac{1}{X_i}\right) \tag{2.8}$$

$$Y_i = \beta_0 . X_i^{\beta_1} . Z_i^{\beta_1} \quad (2.9)$$

Ortaya çıkan tüm koşullar katsayıları açısından doğrusaldır. Parametrik modellerde doğrusallık varsayımının ne olduğu bu test koşullarından yakalanır. Bağımsız faktörlerin üssel, parabolik, kareköklü ya da ters olarak ifade edilmesi bu parametrik modellerin doğrusallığını etkilemez. Doğrusallık ilişkisi bir test stratejisidir ve dağılım hakkında fikir sahibi olabilmek için serpilme (dağılım) grafikleri izlenebilir. Grafik üzerinde x ve y eksenlerinde bağımlı ve bağımsız değişkenlere ait belirlenen noktalar üzerinden değişkenlere ait ilişkinin doğrultusu ve büyüklüğü bulunmaya çalışılmaktadır. Parametrik modeller ile ilgili tahmin yapılırken en çok kullanılan yöntemlerden biri En Küçük Kareler Yöntemi (EKKY) varsayımı üzerinden parametre tahminleri yapılır. EKKY parametrik modellere ait yansız, güvenilir ve etkinliği yüksek olan ve doğrusal değişkenler tahmin edilmesinde etkili bir yöntem olarak uygulamalarda kullanılmaktadır. EKKY hata kareler toplamının en küçük yapılmasını amaçlayan bir istatistiksel yöntemdir.

2.3 Parametrik Olmayan Regresyon Modelleri

Regresyon analizinin amacı bağımsız değişken ya da değişkenlerin bağımlı değişkendeki değişimleri ne şekilde açıklanabildiğini matematiksel bir model ile ifade etmektir. Amaç ilişkiyi açıklayan en iyi modeli bulmak olduğundan, parametrik veya parametrik olmayan yöntemlerden en iyi sonucu veren tercih edilir [20]. Parametrik yöntemlerin kullanılabilmesi için dağılımın normal olması ve varsayımları sağlaması gerekmektedir. Bu varsayımların gerçekleşmeme durumunda parametrik olmayan yaklaşımdan yararlanılabilir. Temelde parametrik testler uygulanmadan isteğe bağlı olarak parametrik regresyon modeli oluşturmadan önce parametrik olmayan tahmin yapılabilir. Bu modeli oluşturmak isteyenin amacına uygun olmak koşuluyla kendi seçimlerine bağlı olmaktadır.

N tane (X_i, Y_i) veri çifti gözlenmiş ise regresyon eğrisi, m regresyon fonksiyonu ve ε hata terimi olmak üzere,

$$Y_i = m(X_i) + \varepsilon_i, \; i=1,2,3,...,n \quad (2.10)$$

şeklinde tanımlanır [21].

m(x)'in önceden belirlenmiş şekline ait herhangi bir bilgi elimizde yok ise parametrik olmayan regresyon modelinden söz edilebilir.

2.4 Parametrik ve Parametrik Olmayan Regresyon Modelleri Hakkında

Regresyon analizinde temel amaç bilinmeyen m(x) fonksiyonu için bir bakış açısı geliştirmektir. Burada bakılması gereken iki farklı yöntem olan parametrik ve parametrik olmayan yöntem vardır. En çok kullanılan yöntem parametrik olandır. Bu yaklaşımda fonksiyonun şekli önceden belirlenmiş fonksiyonel bir şekle sahiptir. Parametrik modeller arasında, En Küçük Kareler (EKK) ve En Çok Olabilirlik Yöntemi (Maximum Likelihood-ML) bilinmeyen parametrelerin tahmininde kullanılan en yaygın yöntemlerdendir. EKK yönteminde anakütle regresyonu denkleminin parametreleri hakkında çıkarım yapılırken varsayımların sağlandığı varsayılır. Parametrik olmayan yaklaşım ise regresyonun şeklinin esnek bir yapıda olduğunu söyler.

2.4.1 Parametrik Olmayan Yöntemlerin Avantajları

Parametrik olmayan yöntemlerde hesaplamalar kolayca tahmin edilebilir. Verilerin nicel olarak ölçülmesi gerekmez ancak nitel bir formatta olabilir, sıralı bir yapıda da olabilir. Parametrik istatistiksel çıkarımlar kadar kısıtlayıcı varsayımları yoktur. Parametrik olmayan yöntemler parametrik bir yaklaşımla sorunlu olabilecek beklenmedik, dışsal gözlemlerle uğraşmak için yararlı olabilir. Parametrik olmayan yöntemler verinin elde edildiği anakütlenin hangi dağılımdan geldiği hakkında güçlü varsayımlar gerektirmez. Parametrik olmayan yöntemleri çoğunlukla anlamak kolaydır. Bu teknikler parametrik tekniklerde olduğu gibi birimlerin büyüklüklerinden daha çok onların sıralarıyla ilgilendikleri için normal dağılım varsayımı gerektiren analizlerin uygulanamadığı birçok durumda kullanılabilir.

2.4.2 Parametrik Olmayan Yöntemlerin Dezavantajları

Parametrik olmayan yöntemler, daha geleneksel yaklaşımlarla karşılaştırıldığında güçsüz olabilir. Örnek boyutu küçükse veya karşılık gelen parametrik yöntem (örneğin verilerin normalliği) için varsayımlar geçerliyse bu özellikle endişe

vericidir. Parametrik olmayan yöntemler, etkilerin tahmini yerine hipotez testlerine yöneliktir. Parametrik olmayan tahminler ve ilişki güven aralıkları elde etmek çoğu zaman mümkündür, ancak bu genellikle basit değildir.

2.4.3 Avantajlar ve Dezavantajlar Genel Bir Değerlendirme

Birçok kişi, parametrik veya parametrik olmayan testlerin kullanılması arasındaki kararın verilerimizin normal olarak dağılıp dağılmamasına bağlı olarak değiştiğine inanır. Küçük bir veri kümeniz varsa, dağılım belirleyici bir faktör olabilir. Ancak çoğu durumda, bu sorun aşağıdakilerden dolayı kritik değildir:

- Parametrik analizler birçok veri kümesi için normal olmayan dağılımları analiz edebilir. Parametrik olmayan analizler, karşılanması daha zor olabilecek başka sağlam varsayımlara sahiptir.
- Ortalama daha iyi bir ölçüm ise ve yeterince büyük bir örneklem boyutuna sahipse, parametrik test genellikle daha iyi, daha güçlü bir seçimdir.
- Medyan daha iyi bir ölçüm ise, örneklerimizin boyutuna bakılmaksızın parametrik olmayan bir testi düşünebiliriz.
- Örneklem büyüklüğümüz küçük ise parametrik olmayan bir test kullanmak zorunda kalabiliriz.

2.5 Parametrik ve Parametrik Olmayan Modellerin Karşılaştırılması

Parametrik ve parametrik olmayan modellerin kullanım seçimi, veriye güvenme, değişkenler gibi birçok farklı yönden karşılaştırması yapılmaktadır. Regresyon analizi, regresyon fonksiyonu hakkında istatistiki çıkarımlarda bulunmak için kullanılan yöntemleri içermektedir [22].

2.5.1 Değişkenler Açısından Karşılaştırma

Her iki modelde bağımlı ve bağımsız değişkenler bulunmaktadır ve bunlar arasındaki ilişki incelenmektedir. Parametrik olmayan regresyon analizinin parametrik regresyon analizinden ayrım noktası ilişkinin parametrik olmayan fonksiyonlarla açıklanmasıdır. Bu noktada verinin kendisi aslında fonksiyonun nasıl görüneceğini söylemektedir.

2.5.2 Varsayımlar Açısından Karşılaştırma

Parametrik bir model parametrelerden gelen verilerin tüm bilgilerini kullanır. Modelin mevcut durumundan ileriye dönük bir tahmin yapabilmemizi sağlar ve bunlar tamamiyle bir parameter ile belirlenir. Tahmini yapılan modeler varsayımların istenildiği gibi olması durumunda kolaylıkla yorumlanabilirler. Varsayımların geçerli olmaması durumunda yapacağımız tahminler yanıltıcı olabilir ve regresyon ilişkisi hakkında eksik ve yanıltıcı bilgiler verir. Parametrik olmayan yaklaşımda kurulan hipotez bir parametreye dayalı olmaksızın ele alınır.

2.5.3 Fonksiyonel Şekil Açısından Karşılaştırma

Parametrik bir model parametrelerdeki veriler hakkında ki tüm bilgileri tutar. Modelin mevcut durumundan gelecekteki bir verinin değerini tahmin etmek için bilinmesi gereken tek şey parametrelerdir. Örnek olarak bir değişkenli doğrusal regresyon modelinde iki parametre vardır. Bu iki parametreyi bilmek yeni bir değer tahmin etmemizi sağlayacaktır. Parametrik regresyon analizinde veriler normal dağılmalıdır ve nicel yapıda olmalıdır (Likert, Aralıklı, Oransal). Öte yandan bir parametrik olmayan model verinin daha göze batmayan ince ayrıntılarını yakalar ve bunlar hakkındabilgi verir. Tahmin edebilmek için mevcut modele bağlı veri kümesinden daha fazla bilgi geçmesine olanak verir. Parametrik olmayan modellerde ise bağımlı ve bağımsız değişkenler arasında fonksiyonel ilişki bilinmemektedir. Parametrik olmayan yöntem anakütleye yönelik tahminlerde bulunmaz. Yapılan testler popülasyonun parametrelerini kıyaslamaya girmeksizin karar verilmesini sağlar.

2.5.4 Veriye Güvenme Açısından Karşılaştırma

Parametrik olmayan tahminciler veriden daha derin çıkarımlar yapabilmesi ve mevcut modele ait veri kümesinden daha fazla bilgi geçirmesi yönünden veriye daha fazla güvenmektedir. Parametrik olmayan regresyon aykırı gözlemlerin (outliers) bulunduğu veri setleri için önemli bir analizdir. Literatürde aykırı gözlemlerin etkilerini farklı biçimlerde ele alan güçlü (robust) parametrik yöntemler vardır. Bununla birlikte aykırı gözlemlerden dolayı parametreler bozulduğu için bu güçlü yöntemler ile uygun çözümler üretemeyebilir ve verinin gerçek yapısı modele

yansıtılamaz. Bu durumda parametrik olmayan regresyon, x doğrultusunda belirli bir parametrik model olmaksızın ön bilgi sağlar [23].

2.5.5 Kullanım Seçimi Açısından Karşılaştırma

Eğer regresyon fonksiyonu için uygun bir parametrik biçim elde edilemezse parametrik olmayan tahmincilerin kullanılması daha avantajlı olacaktır.

2.5.6 Test Yapılması Açısından Karşılaştırma

İstatistik kullanan çoğu insanın parametrik analizlere parametrik olmayan analizlerden daha aşina olduğunu söylemek daha doğru olabilir. Parametrik olmayan testlere dağıtımsız testler de denir, çünkü verilerinizin belirli bir dağılımı izlediğini varsaymazlar. Verileriniz parametrik testin varsayımlarını özellikle de normal dağılan veriler hakkındaki varsayımı karşılamadığında parametrik olmayan testler kullanmanız gerektiğini duymuş olabilirsiniz.

3
DÜZGÜNLEŞTİRME KAVRAMI

3.1 Düzgünleştirme Kavramı

Parametrik olmayan regresyon modellerinde, regresyon eğrisine ait matematiksel kalıbın herhangi bir form varsayılmadan doğrudan tahmin edilmesi "düzgünleştirme" adı verilen teknikler ile gerçekleştirilir. Yerel ortalama yaklaşımı, düzgünleştirme teknikleri içerisinde uygulamada yer alan en basit yaklaşımdır. Düzgünleştirme yaklaşımında regresyon eğrisinin tahmini için hesaplanan $W_{ni}(X)$ ağırlıkları toplamının diğer ifadeyle $l(X_i = x)$ gösterge fonksiyonunun 1'e eşit olması, bu yaklaşımın temelini oluşturmaktadır. Hesaplanan ağırlıklar, regresyon eğrisine ait fonksiyonun belirlenmesinde kullanılmaktadır. Gösterge fonksiyonu tarafından yaratılan yoğunluk fonksiyonu 1'e entegre olan bir özelliğe sahiptir, ancak bu durum elde edilen regresyon eğrisinin düzgün olmaması gibi bir dezavantajı da beraberinde getirmektedir. Düzgünleştirmenin ana teması lokal olarak ortalama alma fikrinden gelmektedir [24]. Parametrik olmayan regresyon bir dizi regresyon fonksiyonunun fonksiyonel formunun verilerden hesaplanması ve herhangi bir doğrusallık varsayımı, düzgün regresyon fonksiyonunun çok daha zayıf bir varsayımı ile değiştirilir; bu nedenle, bağımlı ve bağımsız değişkenler arasındaki ilişkinin yapısı göz önüne alındığında, incelenen değişkenler arasındaki ilişki hakkında önceden bilgi sahibi olunmadığında veya parametrik regresyon kullanan modellemenin çok zor olduğu durumlarda bunu kullanmak uygundur. Bu özellik parametrik olmayan regresyonu çok esnek kılar [25]. Parametrik olmayan bir regresyon belirli bir model varsaymaz. Bu durumda model çok geneldir ve aşağıdaki gibi temsil edilir:

$$E(Y|X=x)=m(x) \tag{3.1}$$

burada m(x) bilinmeyen düzgünleştirilmiş bir fonksiyondur ve "y" ile "x" arasındaki ilişkinin fonksiyonel biçimini ifade eder. Amaç verilerden m(x)'in fonksiyonel biçimini tahmin etmektir [26]. Bu tahmin, parametrik olmayan bir tahmin yöntemi ile elde edilir.

m(x) ve m̃(x) tahmin edildikten sonra, her x_i değeri için tahmini $E(y|x_i)$ elde edilir; ve bu bilgi bir anlamda bir dağılım diyagramı tarafından sağlanan bilgilere eşdeğerdir. Bir düzgünleştiricinin en önemli özelliği, değişkenler arasındaki ilişkinin biçimini kesin bir biçimde varsaymamasıdır ve bu özelliğinden dolayı parametrik olmayan regresyonda sık kullanılan bir araçtır [27].

Parametrik regresyonda ise doğru keskin bir parametrik biçime sahiptir ve bu nedenle düzgünleştirici değildir. Bildiğimiz en basit düzgünleştirici hareketli ortalamadır. Düzgünleştirici tarafından tahmin edilen eğriye "düzgün" (Smooth) denir. Tek bir bağımsız değişkenin olduğu durumlarda ise "serpilme diyagramı düzgünleştirmesi" (Scatter Plot Smoothing) olarak adlandırılmaktadır. Serpilme diyagramındaki noktalar, herhangi bir olasılıksal model olmaksızın düzlem üzerindeki noktaların birikimi olarak kabul edilmektedir [28].

3.2 Düzgünleştirme Parametresi

Parametrik olmayan regresyon modeli tahmininde kullanılan yöntemlerden biri de "K" Kernel fonksiyonudur. Düzgünleştirmenin doğru bir şekilde yapılabilmesi uygun fonksiyonun ve düzgünleştirme parametresinin seçimi ile ilgilidir.

$$\int_{-\infty}^{+\infty} K(u)du = 1 \tag{3.2}$$

Kernel düzgünleştiricilerden yararlanılarak, regresyon eğrisine ait "düzgün" tahminlerin gerçekleştirilmesi bu tahmincilere ait ağırlıkların tanımlanmasına ve hesaplanmasına bağlıdır. Kernel düzgünleştiricilerinde tek boyutlu bir değişken için ağırlıklar şu şekilde tanımlanır.

$$W_{ni} = \frac{K_{h_n}(x_i - x)}{\hat{f}(x)} \tag{3.3}$$

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(x_i - x) \tag{3.4}$$

$$K_{h_n} = \frac{1}{h_n} \mathrm{K}(\mathrm{u} / h_n) \tag{3.5}$$

3.3 Düzgünleştirme Parametresinin Seçimi

Parametrik olmayan regresyonlarda parametre seçimi önem teşkil etmektedir. Yapılan testlerin işe yarar sonuçlar vermesi bu seçime diğer bir ifade ile bant genişliğine (Bandwith) bağlıdır. Düzgünleştirme parametresinin (Smoothing parameter) seçilmesi tahminin varyansı ve sapma miktarı arasındaki denge durumu hakkında bilgi verir. H değeri eğrinin bant genişliğini gösterir. En uygun bant seçimi daha önce de vurguladığımız gibi düzgünleştirme parametresi için çok önemlidir. Bant genişliği (h) değerine göre, dağılımın eğrisi değişkenlik gösterebilir. Dağılımın seçimi (Gauss, Quartic, Uniform) uygun ise fonksiyon daha anlaşılır ve daha iyi dağılım hakkında yorum yapılabilir. Bu seçimin en iyi şekilde yapılmasının temel amacı Ortalama Hata Karesini (MSE) azaltmaktır.

3.4 Düzgünleştirme Parametresinin Seçiminde Ortalama Hata Kare

Düzgünleştirme parametresinin seçiminde çeşitli yöntemler kullanılabilir. MSE standart parametrik istatistiksel yöntemlerden aşina olduğumuz gibi ana kütle regresyon eğrisi üzerinde bulunan tahmini noktalarımız ile gerçek parametre değerleri arasındaki yakınlığı ölçmek için kullandığımız bir yöntemdir. Bu kriterlere dayanan en yaygın yöntemler Birleştirilmiş Hata Kare (ISE) ve Birleştirilmiş Ortalama Hata Kare (MISE)'dir. Temel düzeyde ortalama hata kare aşağıdaki gibi ifade edilebilir.

$$\mathrm{MSE}(\hat{m}) = \mathrm{E}(\hat{m}(\mathrm{x}) - \mathrm{m}(\mathrm{x}))^2 \tag{3.6}$$

Bu ifadeyi varyans'a sapmanın karesi olarak da gösterebiliriz. ISE tek bir noktada fonksiyonu tahmin etmek yerine eğrinin bütününde bir tahminleme amaçlanmaktadır.

$$\mathrm{ISE}\hat{m}(.;\mathrm{h}) = \int\{\hat{m}(x;h) - \mathrm{m}(\mathrm{x})\}^2 \tag{3.7}$$

$$\mathrm{MISE}(\hat{f}) = \int E\left\{\hat{f}(x) - f(x)\right\}^2 dx \tag{3.8}$$

$$= \int \mathrm{MSE}_{\mathrm{x}}(\hat{\mathrm{f}})\mathrm{dx}$$

$$= \int\{\mathrm{E}\hat{\mathrm{f}}(\mathrm{x}) - \mathrm{f}(\mathrm{x})^2\,\mathrm{dx} + \int \mathrm{var}\,\hat{\mathrm{f}}(\mathrm{x})\mathrm{dx}$$

MISE'nin ardında yatan temel düşünce yoğunluk fonksiyonunun tek bir x noktasında tahminini elde etmektense daha geniş bir aralıkta bu tahmini elde etmektir. Bu ifade'den de görülebileceği gibi h değerinin çok küçük seçilmesi halinde tahmincinin sapması küçük fakat varyansı büyük olacaktır. Diğer bir deyişle h değerinin küçük seçilmesi halinde çok az sayıda gözlem önemli düzeyde ağırlığa sahip olacak ve elde edilen yoğunluk tahmini kaba bir görüntüye sahip olacaktır. Buna karşılık çok büyük bir tanımlama kullanıldığında çok sayıda gözlem önemli düzeyde ağırlığa sahip olur ve yoğunluğu aşırı derecede düzleştirir (örneğin çift modlu bir yoğunluğun tek modlu tanımlanmasına neden olabilir). Böyle bir durumda yoğunluğu hesaplanan x değerine çok yakın olmayan gözlemler ait olmadıklarından tahmin sürecine ek bir sapmanın katılması söz konusu olur. Dolayısıyla varyans ile sapma arasında bir alışveriş söz konusudur. Yüksek bir h değeri (daha fazla gözlem içermesi sonucu) yoğunluk tahmininin varyansının azalmasına ancak daha fazla sapmaya neden olmaktadır [29].

3.5 Düzgünleştirme Parametresi Seçim Yöntemleri

Parametrik olmayan regresyonda düzgünleştirme parametresinin belirlenmesi için çeşitli yöntemler geliştirilmiştir. Düzgünleştirme yöntemleri, farklı ölçümler arasında işlevsel ilişkiler bulmaya çalışır. Standart regresyonda olduğu gibi, verilerin bağımlı değişkeni ve bir veya birden fazla tahminci değişkenin ölçümlerinden oluştuğu varsayılmaktadır. Standart regresyon teknikleri, tahminci ve bağımlı değişkeni arasındaki ilişkiyi tanımlamak için fonksiyonel bir form belirtir. Düzgünleştirme yöntemleri daha esnek bir yaklaşımla veri noktalarının uydurulmuş eğrinin şeklini belirlemelerini sağlar.

Çapraz Doğrulama Yöntemi (CV), regresyonun öngörü hatasının tahminini sağlar ve bu da hata ölçütünün bir yaklaşımıdır. Yöntemin çerçevesi açıktır. Verilerdeki her gözlem için yöntem, modelleme işleminden çıkarılan gözlem ile çekirdek regresyonu kullanarak tahmin hatasını değerlendirir. En uygun bant genişliği, tüm gözlemlerden tahmin hatalarının karelerinin toplamını en aza indirecek şekilde seçilir. CV, Akaike Bilgi Kriteri ve Shibata'nın ölçütünün örnek olduğu, veriye dayalı yöntemler adı verilen daha büyük bir bant genişliği seçim yöntemleri sınıfından biridir. Çapraz doğrulama tahmin problemine odaklıdır. $x = x_0$ değerinde yeni bir

gözlem yapılırsa ve bağımlı Y_0 , $\hat{Y} = \hat{\mu}(x_0)$ ile tahmin edilirse, tahmin hatası (3.9)' de verilen ifade şeklinde olur:

$$E((Y_0 - \hat{Y}_0)^2) \tag{3.9}$$

CV yöntemi bu miktarı tahmin etmek için kullanılabilir. Buna karşılık, her gözlem (x_i, Y_i) veri kümesinden çıkarılır ve kalan n-1 gözlem düzgünleştirilerek tahmin edilir. Bu CV skorunu verir:

$$CV(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{\mu}_{-i}(x_i))^2 \tag{3.10}$$

Burada $\hat{\mu}_{-i}(\cdot)$, tek veri noktası (x_i, Y_i) veri kümesinden çıkarıldığında düzeltilmiş tahmini belirtir; tahmini hesaplamak için yalnızca kalan n -1 veri noktaları kullanılır. Şeklen her bir CV (leave-one-out) regresyonun hesaplanmasında $\hat{\mu}_{-i}(\cdot)$ sayısal olacaktır ve bu nedenle de ilk bakışta CV skorunun hesaplanması oldukça maliyetli görünmektedir. Ancak hemen hemen tüm yaygın doğrusal düzgünleştiriciler için kullanılan dikkate değer bir sadeleştirme vardır:

$$\hat{\mu}_{-i}(x_i) = \frac{\hat{\mu}(x_0) - l_i(x_i)\, Y_i}{1 - l_i(x_i)}. \tag{3.11}$$

Bu basitleştirme ile CV kriteri:

$$CV(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}(x_i))^2}{1 - l_i(x_i)^2}. \tag{3.12}$$

Genelleştirilmiş çapraz doğrulama (GCV), $l_i(x_i)$ etki değerlerinin her biri ortalama v_1/n ile değiştirilir. GCV, düzgünleştirme parametresini seçmek için geleneksel bir yöntem olan CV'nin modifiye edilmiş bir şeklidir. Ve böylece:

$$GCV(\hat{\mu}) = n \frac{\sum_{i=1}^{n} (Y_i - \hat{\mu}(x_i))^2}{(n - v_1)^2} \tag{3.13}$$

şekline gelir. Öte yandan ISE kriteri düzgünleştirme parametresinin seçimi için çapraz geçerlilik metodunda kullanılmaktadır. Üzerinde konuştuğumuz bu yönteme "En Küçük Kareler Çapraz Geçerliliği" (Least Squares Cross Validation-LSCV) de denilmektedir. Bu yöntemi :

$$ISE(h) = \int (\hat{m}(x) - m(x))^2 dx \tag{3.14}$$

$$\mathrm{ISE}(h)_1 = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} K \circ \mathrm{K}\left(\frac{x_i - x_j}{h}\right) - \frac{1}{2n}\sum_{i=1}^{n}\hat{m}_{-i}(x_i) \qquad (3.15)$$

$\mathrm{ISE(h)}_{\mathbf{1}}$ değerinin minimum yapılmasıyla hesaplanmaktadır.

Başka bir çapraz geçerlilik metodu "benzerlik çapraz geçerlilik" (Likelihood Cross Validation-LCV) metodudur. Bu yöntemdeki ana düşünce benzerlik fonksiyonunu maksimum yapan düzgünleştirme parametresini seçmektir. Genel yazım şekli aşağıda formülde gösterildiği gibidir:

$$\mathrm{L(h)} = \prod_{i=1}^{n}\hat{m}_h(x_j) \qquad (3.16)$$

$$\mathrm{L(h)} = \prod_{j=1}^{n}\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - X_j}{h}\right) \qquad (3.17)$$

$$= \prod_{j=1}^{n}\frac{1}{nh}\left[K(0) + \prod_{i\neq j} K\left(\frac{X_i - X_j}{h}\right)\right]$$

3.6 Penalizing Fonksiyon Seçicisi

$\mathrm{H} \downarrow 0$'ın enterpolasyona, yani $\hat{\mathrm{m}}_{\mathrm{h}}(\mathrm{x_j}) \rightarrow \mathrm{Y_j}$'e yol açtığı açıktır, böylece minimize edilecek fonksiyon, yani p(h), keyfi olarak küçük olabilir. Öte yandan, bu kesinlikle büyük bir $\hat{\mathrm{m}}_{\mathrm{h}}$ varyansına neden olacaktır, böyle bir kriter fonksiyonun yanlı ve varyansı dengelemeyeceğini gösterir. Sonuç olarak düzeltilmiş MSE, h'yi (en azından yaklaşık olarak) makul bir anlamda çok küçük seçerken cezalandırır. Bu nedenle (3.18) eşitliği yardımıyla hesaplanır:

$$\mathrm{G(h)} = \frac{1}{n}\sum_{\mathrm{j}=1}^{\mathrm{n}}(\mathrm{Y_j} - \hat{\mu}_h(x_i))^2 \;\Xi\left(\frac{1}{\mathrm{n}}\mathrm{W}_{h,j}(\mathrm{X}_j)\right)\mathrm{w}(\mathrm{X}_j) \qquad (3.18)$$

Burada $\mathrm{w}(\mathrm{X}_j) = 1_{[X_{l+1}, X_{n-1}]}$ sınırın yakınında kırpma yapmak için kullanılır. $\Xi(.)$ birinci dereceden Taylor genişlemesi ile cezalandırıcı bir işlevi vardır:

$$\Xi(\mathrm{u}) = 1 + 2\mathrm{u} + \mathrm{O}(u^2), \mathrm{u} \rightarrow 0. \qquad (3.19)$$

Bant genişliği h ne kadar küçük seçilirse, $\mathrm{W}_{h,j}(\mathrm{X}_j)$ ve cezalandırma faktörü $\Xi\,\frac{1}{\mathrm{n}}\,\mathrm{W}_{h,j}(\mathrm{X}_j)$ artar. G'nin birinci dereceden Taylor açılımını yapar ve düşük dereceli terimlerini göz ardı edersek, h' dan bağımsız olarak, üstel açılımın bir derecesine bağlı veya bir rotasyonuna kadar, kabaca MSE(h)'e eşit olduğunu göstermek kolay olur. Aşağıda açılımı sağlayan bir dizi cezalandırma fonksiyonu için öneri sunulmuştur. Bunlardan bazıları şu şekildedir.

Shibata'nın model seçicisi $\hat{\mathrm{h}}_{\mathrm{S}} = \mathrm{argmin}\ \mathrm{G_S(h)}$

ile $\Xi_S(u) = 1 + 2u.$ (3.20)

Rice'ın düzgünleştirme parametresi seçicisi $\hat{h}_T = \text{argmin } G_T(h)$

ile $\Xi_T(u) = (1 - 2u)^{-1}.$ (3.21)

3.6.1 Plug-in Seçicisi

CV ve düzeltilmiş MSE yöntemlerinin aksine, plug-in yöntemleri, optimum bant genişliğini hesaplamak için yerel doğrusal tahmin edicinin $\hat{m}h(x)$ MISE veya Asimtotik MISE (AMISE) değerini en aza indirmeye çalışır. Koşullu ağırlıklı AMISE (3.22) eşitliği yardımıyla hesaplanır:

$$AMISE(\hat{m}_h(x)|X_1, , \dots, Xn) \tag{3.22}$$

$$= \frac{1}{nh} \|K\|_2^2 \int_S \sigma^2(x)dx \frac{h^4}{4} \mu_2{}^2(K) \int_S m''(x))^2 f(x)dx,$$

Burada birinci toplam hesaplanan ortalama entegre yaklaşık varyans ve ikinci toplam hesaplanan ortalama yanlı entegre karedir. Gösterim şu şekildedir: f(x), X'in yoğunluğunu gösterir:

$$\|K\|_2^2 = \int K(u)^2 \, du, \; \mu_l(K) = \int u^l K(u)du \tag{3.23}$$

ve f yoğunluğunu S = [a, b] ⊂ R ile belirtir. Makul bir bant genişliği bulmak varyansı dengelemek anlamına gelir ve dolayısıyla, h'a göre minimize edilerek optimal bir bant genişliği elde edilir. Açıkcası, MISE esas olarak bilinmeyen fonksiyonlardan ve parametrelerden oluşur. Sonuç olarak, seçim yöntemlerinin ana zorluğu uygun ikameler veya tahminler bulmaktır. Ağırlığın en aza indirilmesi h, aşağıda verilen formülde AMISE optimal bant genişliğine (h_{AMISE}) yol açar:

$$h_{AMISE} = \left(\frac{\|K\|_2^2 \int_S \sigma^2(x)dx}{\mu_2^2 . \int_S m^{"}(x))^2 \, f(x)dx \, .n} \right)^{1/5}, \tag{3.24}$$

burada S = [a, b] ⊂ R, n boyutunda X örneklemine destektir. Açıkca, (3.24) de verilen denklemdeki bilinmeyen iki ölçü, uygun tahminlerle değiştirilmesi gereken $\int_S \sigma^2$ ve $\int_S m''(x))^2 f(x)dx$'dir. Kuartik kernel kullanılarak aşağıdaki formül elde edilir:

$$h_{AMISE} = \left(\frac{35 \cdot \int_S \sigma^2(x)dx}{\theta_{22} \cdot n} \right)^{1/5} , \quad \theta_{rs} \int_S m^{(r)}(x) m^{(s)}(x) f(x) dx, \qquad (3.25)$$

burada m^lm'nin l. Türevini belirtir. Optimum bant genişliğinin elde edilmesinde üç farklı strateji tanımlanmıştır. Simülasyon çalışmalarındaki performans nedeniyle, pratik olarak bant genişliği seçici h_{rot} kullanılmıştır. Bu yöntem çok açıktır, çünkü bilinmeyen miktarların yerini parametrik OLS tahmin edicileri alır. Uyumlu ek yöntemi, çapraz geçerlilik ve cezalandırma fonksiyonu yöntemleri kadar etkin sonuçlar vermesine ragmen, ilk adımda bant aralığının seçiminde bir belirsizliğe sahiptir[31]. Plug-in (eklenti) fikri, bilinmeyen miktarları temel olarak üç farklı stratejiyle değiştirmektir: Rule-of-thumb (pratik) bant genişliği seçici h_{rot}, bilinmeyen miktarlar, parametrik EKK tahmin edicileri ile değiştirilir. Direct plug-in (doğrudan eklenti) bant genişliği seçici, bilinmeyen miktarları, parametrik olmayan ikinci tahminci için 'önceki bant genişliklerini' seçmemiz gereken parametrik olmayan tahminlerle değiştirir. İkinci aşamada bu bant genişliklerinin hesaplanması için parametrik bir tahmin kullanıyoruz. Bootsrap tabanlı bant genişliği seçimi h_{SB} ve h_{WB}, bilinmeyen ifade, ön yükleme yöntemleri ile tahmin edilir. Düzgünleştirilmiş bootstrap(h_{SB}) durumunda, yine bilinmeyen ifadeler tahmni edilirken wild bootstrap yöntemi h_{WB} doğrudan $\hat{m}_h$ MISE'yi tahmin eder ve h değerine göre en aza indirir. Her iki yöntemde bir önceki bant genişliği gerektirir. Önceden bant genişliği gerektirmeyen, ancak sayısal olarak kapalı denklemleri çözmeye çalışan bir bant genişliği seçici de vardır. Bu prosedür, Kernel yoğunluk tahmininde denklemi çözme yaklaşımını izler. Bununla birlikte, bu bant genişliği seçicinin sonuçları doğrudan eklenti (direct plug-in) yaklaşımının sonuçlarından eşit oranda daha iyi değildir, ancak çok daha büyük bir hesaplama çabası gerektirir ve bu nedenle uygulamada oldukça çekici değildir. Direct plug-in yaklaşımında bilinmeyenin yerini parametrik olmayan tahminler alır. Daha sonra θ_{22} parametrik olmayan tahmincisi için bir bant genişliği g gereklidir. Bariz bir aday olarak, σ^2'nin parametrik olmayan tahmincisinin AMISE'yi en aza indiren bant genişliği g_{AMSE} dir. Ayrıca, σ^2'nin parametrik olmayan tahmincisi için bir önceki bant genişliği λ_{AMSE} belirlenmelidir.

Önceki bant genişlikleri bir parametrik EKK-block-fit ile hesaplanır. Parametrik olmayan bir tahminci θ_{22} (g_{AMSE}) şu şekilde tanımlanabilir:

$$\theta_{22}(g) = n^{-1} \sum_{i=1}^{n} [\hat{m}^{(2)}(X_i)]^2, \tag{3.26}$$

Direct plug-in bant genişliği $h_{d\,pi}$:

$$h_{DPI} = \left[35 \frac{\sigma^2(\hat{\lambda}_{AMSE})(b-a)}{\hat{\theta}^{.05}{}_{22}(\hat{g}_{AMSE})}\right]^{1/5}. \tag{3.27}$$

3.6.2 Akaike Bilgi Kriteri Seçicisi

Akaike Bilgi Kriteri (AIC) veri üreten model ile uygun bir aday model arasında beklenen Kullback Leibler bilgisini tahmin etmek için geliştirilmiştir. AIC model seçiminde yaygın olarak kullanılmaktadır, çünkü uygun bir modelin beklenen Kullback-Leibler bilgisinin bir tahminidir. AIC yanlı bir tahmin edicidir.

$$AIC = n(log\hat{g}\,\sigma^2 + 1) + 2p, \tag{3.28}$$

burada p ve $\hat{\sigma}^2$, alt kümeler modelinin sırasıyla parametre sayısı ve varyansıdır. AIC, denklemi en aza indiren alt kümelerin seçimini içerir. AIC yaklaşık olarak yansız bir tahminidir. AIC beklenen entropi, beklenen K-L bilgisi veya beklenen çapraz entropinin örnek bir tahminidir.

3.6.3 Düzeltilmiş Akaike Kriteri Seçicisi

AIC'in Hurvich ve Tsai tarafında düzeltilmiş bir versiyonu:

$$AIC_c = AIC + \frac{2(p+1)(p+2)}{n-p-2}. \tag{3.29}$$

Sugiura tarafından başlangıçta doğrusal modelleri için önerilen bir yöntem olan AIC_c, hem regresyon hem de zaman serilerinde yaklaşık olarak etikilidir. Doğrusal regresyon için AIC_c, aday model ailesinin gerçek modeli içerdiğini varsayarak tarafsızdır. AIC_c, AIC'den daha iyi yanlılık özelliklerine sahiptir.

3.6.4 Fisher Bilgi ile Akaike Bilgi Kriteri

Bu bölümde, değişken seçimi için kullanılan AIC için sapmalı bir düzetlme önerilmektedir. Kullback-Leibler bilgisi yerine gerçek ve yaklaşık modeler arasındaki tutarsızlığın ölçüsü olarak değerlendirebiliriz. Fisher bilgisi:

$$i_n(\theta) = nE_0\left[\frac{\partial \log -g(y;\theta)}{\partial\theta}\right], \tag{3.30}$$

burada $g(y;\theta)$ yaklaşık model altında olabilirlik fonksiyonu belirtir ve

$$g_{0,\sigma^2}(y) = \frac{1}{2\pi\sigma^2}\exp\left[-\sum_{i=1}^{n}\frac{(y_i-h(\theta))^2}{2\sigma^2}\right]. \tag{3.31}$$

$h(\theta)$ ile ilgili olabilirlik fonksiyonunun türevini alarak aşağıdaki şekile gelir,

$$i_n(\theta) = \frac{n}{2}E_F\left[\frac{(\mu+\varepsilon-h(\theta))'(\mu+\varepsilon-h(\theta))}{\sigma^2}\right]. \tag{3.32}$$

$E_F(i_n(\theta))$, gerçek bir F modeli için, yaklaşan benzeyen ailenin kalitesini değerlendirmek için makul bir kriterdir. Tabii ki, $E_F(i_n(\theta))$ bilinmiyor, ancak tahmin edilebilir. Benzeyen ailenin gerçek modeli içerdiğini varsayıyoruz. Güçlü bir varsayım olmasına rağmen AIC'in türetilmesinde de kullanılabilir. Böylece $E(i_n(\theta))$'i aşağıdaki gibi yazabiliriz:

$$E(i_n(\theta)) = \frac{n}{\hat{\sigma}^2}\left\{\frac{np}{n-p-2}+2\right\}. \tag{3.33}$$

AIC, $E(i_n(\theta))$'de yaklaşık yansız bir tahmin elde etmek üzere tasarlanmıştır.

$$AICF = \frac{n}{\hat{\sigma}^2}\left\{\frac{np}{n-p-2}+2\right\} \tag{3.34}$$

$E(i_n(\theta))$'nin yaklaşık yansız bir tahmincisidir.

4
KERNEL REGRESYON MODELİ VE TAHMİNCİLERİ

4.1 Kernel Regresyon Modeli

X_0 noktası verildiğinde, $m(x_0)$ değeri ile ilgilendiğimizi varsayalım. Bu değeri tahmin etmek için basit bir yöntem olarak değerlendirebiliriz. $m(x_0)$ düzgün olduğunda, bir $X_i \approx x_0$ gözlemi $m(X_i) \approx m(x_0)$ anlamına gelnekte olacaktır. Böylece, bağımlı değişken $Y_i = m(X_i) + \varepsilon_i \approx m(x_0) + \varepsilon_i$ şeklinde elde edebiliriz. Bu gözlemi kullanarak, gürültü $\varepsilon_i$azaltmak için örnek ortalamasını kullanabiliriz. Bu nedenle $m(x_0)$ tahmincisi, eşdeğeri X_0'a yakın olan cevapların ortalamasını almaktır.

Daha somut hale getirmek için h > 0 bir eşik değeri olsun. (4.1) eşitliği şeklinde yazılır:

$$\hat{m}_{loc}(x_0) = \frac{\sum_{i=1}^{n} Y_i I(|X_i - x_0| \leq h)}{\sum_{i=1}^{n} Y_i I(|X_i - x_0| \leq h)} = \sum_{i=1}^{n} \frac{I(|X_i - x_0| \leq h)}{\sum_{\ell=1}^{n} I(|X_\ell - x_0| \leq h)} . Y_i = \sum_{i=1}^{n} W_i(x_0) Y_i, \qquad (4.1)$$

Burada

$$W_i(x_0) = \frac{I(|X_i - x_0| \leq h)}{\sum_{\ell=1}^{n} I(|X_\ell - x_0| \leq h)} \qquad (4.2)$$

her gözlem için ağırlıktır. Tüm i = 1, …., n; değerleri için ele alırsak $\sum_{i=1}^{n} W_i(x_0) = 1$ ve $W_i(x_0)>0$'ın $W_i(x_0)$'lerin gerçekten ağırlık olduğunu söyleyebiliriz. (4.2)'deki denklem, yerel ortalamalamalarını tahmin edicinin ağırlıklı ortlama tahmincisi olarak da yazılabileceğini gösterir, böylece i. ağırlık $W_i(x_0)$, Y_i bağımlı değişkeninin $\hat{m}_{loc}(x_0)$ değerine katkısını belirler.

Yerel ortalama tahminciyi oluştururken, komşu noktalardaki dert eşiklere h mesafesi içinde olanlara eşit ağırlık verilir, ancak h eşiğinin dışındakilere tamamen göz ardı edilir. Bu sert eşik, sürekli olmayan tahminciye yol açar. Sorundan kaçınmak için, ağırlıkların başka bir yapısını düşünüyoruz. İdeal olarak, X_0'a yakın olan gözlemlere daha fazla ağırlık vermek ve "düzgün" bir ağırlığa sahip olmasını istiyoruz.

Gauss fonksiyonu:

$G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ uygun bir aday gibi görünmektedir. Şimdi tahminci oluşturmak için Gauss işlevini kullanıyoruz. Önce ağırlığı yapılandırıyoruz:

$$W_i{}^G(x_0) = \frac{G\left(\frac{x_0 - X_i}{h}\right)}{\sum_{\ell=1}^{n} G\left(\frac{x_0 - X_\ell}{h}\right)} \tag{4.3}$$

h > 0 miktarı, yerel ortalamadaki eşik ile benzer miktardır, ancak şimdi Gaussian'ın düzgünleştirme bant genişliği olarak işlev görmektedir. Ağırlığı oluşturduktan sonra yeni tahmincimiz:

$$\hat{m}_G(x_0) = W_i{}^G(x_0) Y_i = \sum_{i=1}^{n} \frac{G\left(\frac{x_0 - X_i}{h}\right)}{\sum_{\ell=1}^{n} G\left(\frac{x_0 - X_\ell}{h}\right)} Y_i = \frac{\sum_{i=1}^{n} Y_i G\left(\frac{x_0 - X_i}{h}\right)}{\sum_{\ell=1}^{n} G\left(\frac{x_0 - X_\ell}{h}\right)}. \tag{4.4}$$

Bu yeni tahminci, yerel ortalamadan daha düzgünce değişen ve istediğimiz gibi düzgün bir ağırlığa sahiptir.

(4.1) ve (4.4) denklemlerinden hareketle, bu yerel tahmincilerin hepsinin benzer bir formda olduğu görülebilir:

$$\hat{m}_h(x_0) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x_0 - X_i}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{x_0 - X_\ell}{h}\right)} = \sum_{i=1}^{n} W_i{}^K(x_0) Y_i, \quad W_i{}^K(x_0) = \frac{K\left(\frac{x_0 - X_i}{h}\right)}{\sum_{\ell=1}^{n} K\left(\frac{x_0 - X_\ell}{h}\right)}, \tag{4.5}$$

burada K bir fonksiyondur. K bir Gauss olduğunda, tahminci (4.4) elde ederiz. K [-1, 1] üzerinde bir örnek olduğunda yerel ortalamayı elde ederiz (4.1). (4.5) denklemindeki tahminciye Kernel regresyon tahmincisi veya Nadaraya-Watson Tahmincisi (N-W) denir. K fonksiyonu, Kernel Yoğunluk Tahmincisi (KDE)'deki Kernel fonksiyonu ile benzer bir rol oynar ve bu nedenle Kernel fonksiyonu olarak da adlandırılır. Ve h > 0 miktarı, KDE'deki düzgünleştirme bant genişliğine benzer de denir.

4.2 Nadaraya-Watson Tahmincisi

Veriler (y_i, X_i) olsun, y_i gerçek değerli ve X_i bir q- vektördür ve hepsinin sürekli olarak f(y,x) bir ortak yoğunluk ile dağıldığını varsayalım. $f(y|x) = f(y,x)/f(x)$, y_i verilen X_i'nin koşullu yoğunluğu olsun, burada $f(x) = \int f(y,x)dy$, X_i'nin marjinal yoğunluğudur. X_i için y_i 'nin regresyon fonksiyonu eşitlik (4.6)'daki gibidir:

$$g(x) = E(y_i|X_i = x). \tag{4.6}$$

Bu parametrik olmayan bir şekilde g'deki minimum varsayımlarla tahmin edilmek istenmektedir. X_i'nin x'e eşit olduğu çok sayıda gözlemimiz olsaydı, bu gözlemler için y_i'lerin ortalama değerlerini alabilirdik. Ancak X_i sürekli dağılıma uyduğundan, aynı değere eşit birden fazla gözlem gözlemleyemeyeceğiz.

Çözüm, x'in bir komşusunu düşünmek ve X_i'nin x de pozitif bir yoğunluğa sahip olması durumunda, bu komşulukta bulunan bir takım gözlemleri izlememiz gerekir ve bu sayının örneklem büyüklüğü ile artmakta olduğunu bilmemiz gerekiyor. Regresyon fonksiyonu g(x) sürekli ise, bu komşulukta makul bir sabit olmalıdır (yeterince küçük ise), bu nedenle bu gözlemler için ortalama y_i değerlerini alınmalıdır. Yapılması gereken açık olarak, komşuluktaki gözlemlerin sayısına karşı komşuluk üzerindeki g(x) cinsinden değişimi yapacak komşuluğun büyüklüğünü belirtmektir. x_i'nin herhangi bir komşuluğunda çok sayıda X_i gözlemlenecektir. h > 0'dan büyük olduğu bazı bant genişlikleri için x'in komşuluğunun $x \pm h$ olmasına izin verilir. Böylece basit bir parametrik olmayan g(x) tahmincisi, gözlemler için y_i'lerin ortalama değeri (4.7) formülünde X_i bu komşulukta olacak şekilde elde edilir, yani:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} 1(|X_i - x| \leq h) y_i}{\sum_{i=1}^{n} Y1(|X_i - x| \leq h)} \tag{4.7}$$

$$= \frac{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right)}$$

burada k(u) düzgün kerneldir.

Genel olarak, kernel regresyon tahmincisi bu formu alır, burada k(u) bir kernel fonksiyonudur. N-W veya yerel sabit tahmincisi (Local Constant Estimator) olarak da bilinir.

q > 1 olduğunda tahminci:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} K(H^{-1}(X_i - x)) y_i}{\sum_{i=1}^{n} K(H^{-1}(X_i - x))} \tag{4.8}$$

burada K(u) çok değişkenli bir kernel fonksiyonudur.

Alternatif olarak regresyon fonksiyonu aşağıdaki gibi yazılabilir:

$$g(x) = \frac{\int y f(y,x) dy}{f(x)} \tag{4.9}$$

burada $f(x) = \int f(y,x)\,dy$ X_i'nin marjinal yoğunluğudur. Şimdi yoğunluk fonksiyonları parametrik olmayan tahminlerle değiştirirsek,

$$\hat{f}(y,x) = \frac{1}{n|H|h_y}\sum_{i=1}^{n} K(H^{-1}(X_i - x))k\left(\frac{y_i - y}{h_y}\right) \tag{4.10}$$

burada h_y, y doğrultusunda düzgünleştirme için bir bant genişliğidir. Yani,

$$\hat{f}(x) = \int \hat{f}(y,x)\,dy \tag{4.11}$$

$$= \frac{1}{n|H|h_y}\sum_{i=1}^{n} K(H^{-1}(X_i - x)) \int k\left(\frac{y_i - y}{h_y}\right) dy$$

$$= \frac{1}{n|H|}\sum_{i=1}^{n} K(H^{-1}(X_i - x))$$

ve

$$\int y\hat{f}(y,x)\,dy = \frac{1}{n|H|h_y}\sum_{i=1}^{n} K(H^{-1}(X_i - x)) \int yk\left(\frac{y_i - y}{h_y}\right) dy \tag{4.12}$$

$$= \frac{1}{n|H|}\sum_{i=1}^{n} K(H^{-1}(X_i - x))\, y_i$$

Böylelikle oranı alırsak (4.13) eşitliği elde edilir:

$$\hat{g}(x) = \frac{\frac{1}{n|H|}\sum_{i=1}^{n} K(H^{-1}(X_i - x))y_i}{\frac{1}{n|H|}\sum_{i=1}^{n} K(H^{-1}(X_i - x))} \tag{4.13}$$

$$= \frac{\sum_{i=1}^{n} K(H^{-1}(X_i - x))y_i}{\sum_{i=1}^{n} K(H^{-1}(X_i - x))}$$

yine N-W elde edildi ve h_y bant genişliği kayboldu. Tahmin edici, $\hat{f}$ (x) < 0 olacak şekilde x değerleri için eksik tanımlanmıştır. Bu X_i dağılımının kuyruklarında meydana gelebilir. Yüksek dereceli kerneller $\hat{f}$ (x) < 0 verebildiğinden, birçok yazar regresyon için yalnızca ikinci dereceden kernelin kullanılmasını önermektedir.

4.3 En Yakın Komşu Tahmincisi

En Yakın Komşu Yöntemi (k-NN), ayrıca Cover ve Hart (1967), istatistiksel ayrımcılık alanındaki en basit ve en sezgisel tekniklerden birini temsil eder. Kullanılan değişkenlere göre, yeni gözleme en yakın öğrenme kümesinden gözlem sınıfına yeni bir gözlem yerleştirildiği parametrik olmayan bir yöntemdir. Bu benzerliğin belirlenmesi mesafe ölçütlerine dayanmaktadır. k ≥ 1 tamsayısını düzeltirsek

$$\hat{m}(x) = \frac{1}{k}\sum_{i\in N_k(x)} Y_i, \tag{4.14}$$

burada $N_k(x)$, $X_1, \ldots, X_n$'ler ile x arasındaki en yakın k indekslerini içerir. Bu hiç de kötü bir tahmin edici değildir ve birçok uygulamada, basitliğinden dolayı kullanılmaktadır. Komşuların k sayısını değiştirerek, tahmini fonksiyon $\hat{m}$'de geniş bir esneklik aralığı elde edebiliriz, küçük k büyük k'a göre daha esnek olduğu söylenebilir. Ancak sınırlamaları vardır, uygun $\hat{m}$ fonksiyonu, özellikle küçük veya ortalama k için her zaman düzgünleştirilmiş görünür. Bu aşağıdaki gibi bir yazım için olanak sağlar:

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x)\, Y_i \tag{4.15}$$

burada ağırlıklar $w_i(x)$, $i = 1, \ldots, n$ şu şekilde tanımlanır:

$$w_i(x) = \begin{cases} 1/k, & X_i\ x'e\ en\ yakın\ noktalardan\ biri\ ise \\ 0, & değil\ ise \end{cases} \tag{4.16}$$

Kernel tahmincilerde aralık $[x_0 - h, x_0 + h]$ ve k-NN yönteminde aralık ise $[x_0 - d_k(x_0), x_0 + d_k(x_0)]$ şeklinde olmaktadır. Burada k yaklaşık olarak $2nd_k(x)\hat{m}(x)$ olmalıdır [32].

4.4 Kernel Regresyon Teorisi

$\hat{m}_h$ tahmincisinin bazı istatistiksel özelliklerinin inceleyeceğiz. Türev ayrıntıları üzerinde durmayacağız.

Yanlılık, x noktasında Kernel Regresyonunun yanlılığı (4.17) eşitliği şeklinde olur:

$$bias(\hat{m}_h(x)) = \frac{h^2}{2}\mu K(m''(x) + 2\frac{m'(x)p'(x)}{p(x)} + o(h^2), \tag{4.17}$$

Burada $p(x)$, $X_1, \ldots, X_n$ ve $\int x^2 K(x)dx$ değişkenlerinin olasılık yoğunluk fonksiyonudur ve KDE' de olduğu gibi kernel fonksiyonunda ki aynı sabittir.

Yanlılık iki bileşene sahiptir: eğrilik bileşeni $m''(x)$ ve tasarım bileşeni $\frac{m'(x)p'(x)}{p(x)}$. Eğrilik bileşeni, KDE'deki bileşene benzer; regresyon fonksiyonu çok fazla kıvrıldığında, Kernel Düzgünleştirme yapıyı düzleştirerek baz yanlılıkları ortaya çıkaracaktır. Tasarım eğilimi olarak da bilinen ikinci bileşen, KDE'deki eğilime kıyasla yeni bir bileşendir. Bu bileşen, değişken p(x) yoğunluğuna bağlıdır. Bazı

çalışmalarda, p(x) yoğunluğunun tasarım olarak da adlandırılması için ortak değişkenlerin değerlerini seçebilmekteyiz (bu yüzden tasarım yanlılığı olarak da bilinir).

Varyans tahmincisi,

$$Var(\hat{m}_h(x)) = \frac{\sigma^2 \cdot \sigma^2{}_K}{p(x)} \cdot \frac{1}{nh} + o\left(\frac{1}{nh}\right), \tag{4.18}$$

burada $\sigma^2 = Var(\varepsilon_i)$, regresyon modelinin hatasıdır ve $\sigma^2{}_K = \int K^2(x)dx$ Kernel Fonksiyon Sabitidir (KDE de olduğu gibi). Bu ifade bize olası varyans kaynakları hakkında bilgi verir. İlk olarak, σ^2 arttığında varyans artar. Bu anlamlıdır çünkü σ^2 gürültü seviyesini belirtir. Gürültü seviyesi büyük olduğunda, tahmin hatasının artmasını bekliyoruz. İkincisi, ortak değişken p(x)'in yoğunluğu varyans ile test orantılıdır. Bu da anlamlıdır, çünkü p(x) büyük olduğunda, x çevresinde daha fazla veri bulunur ve bu da ortalamadan aldığımız numunenin boyutunu artırır. Son olarak, yakınsama oranı KDE ile aynı olan O $\left(\frac{1}{nh}\right)$ dır.

MSE ve MISE: Sapma ve varyans ifadesini kullanarak, x noktasındaki OHK

$$MSE\left(\hat{m}_h(x)\right) = \frac{h^4}{4}\mu^2{}_K \left(m''(x) + 2\frac{m'(x)p'(x)}{p(x)}\right)^2 \tag{4.19}$$

$$+ \frac{\sigma^2 \cdot \sigma^2{}_K}{p(x)} \cdot \frac{1}{nh} + o(h^4) + o\left(\frac{1}{nh}\right)$$

ve MISE:

$$MISE(\hat{m}_h) = \frac{h^4}{4}\mu^2{}_K \int \left(m''(x) + 2\frac{m'(x)p'(x)}{p(x)}\right)^2 dx \tag{4.20}$$

$$+ \frac{\sigma^2 \cdot \sigma^2{}_K}{nh} \int \frac{1}{p(x)} dx + o(h^4) + o\left(\frac{1}{nh}\right).$$

(4.6) denkleminde ana bileşenleri optimize ederek düzgünleştirme bant genişliğinin optimum değerini elde ederiz:

$$h_{opt} = C^* \cdot n^{-1/5}, \tag{4.21}$$

burada C^* p ve K'ya bağlı bir sabittir.

4.5 Belirsizlik ve Güven Aralığı

Tahminci kalitesini nasıl değerlendiririz $\hat{m}_h$?

Bunu testi yapmak için Bootstrap'I kullanabiliriz. Bu durumda, Empirical Bootstrap, Residual Bootstrap ve Wild Bootstrap kullanılabilir. Ancak her birinin biraz farklı varsayımlara dayandığı unutulmamalıdır. $(X_1{}^*, Y_2{}^*),\cdots,(X_n{}^*, Y_n{}^*)$ bootstrap örneği olsun. Bootstrap örneği denklem (4.5)'e uygulanarak, $\hat{m}_h{}^*$ olarak gösterilen bir bootstrap kernel regresyon elde ederiz. Bootstrap prosedürü B kez tekrar edilirse:

$$\hat{m}_h{}^{*(1)},\cdots,\hat{m}_h{}^{*(B)}, \tag{4.22}$$

B bootstrap Kernel Regresyon Tahmincisi'dir. Böylece $\hat{m}_h(x)$ varyansını örnek varyansı ile tahmin edebiliriz:

$$\widehat{Var}_B\left(\hat{m}_h(x)\right) = \frac{1}{B-1}\sum_{\ell=1}^{B}\left(\hat{m}_h{}^{*(\ell)} - \hat{m}^*{}_{h,B}(x)\right), \bar{\hat{m}}^*{}_{h,B}(x) = \frac{1}{B}\sum_{\ell=1}^{B}\left(\hat{m}_h{}^{*(\ell)}(x)\right). \tag{4.23}$$

Benzer şekilde, OHK'I tahmin edebiliriz. Bununla birlikte, belirsizliği tahmin etmek için bootstrap kullanırken, çok dikkatli olmak zorundayız çünkü h çok küçük veya çok büyük olduğunda, Bootstrap tahmini hedefini yakınlaştırmayabilir.

$h = O\left(n^{-1/5}\right)$ seçtiğimizde, varyansın bootstrap tahmini tutarlıdır, ancak OHK'nın bootstrap tahmini tutarlı olamayabilir. Başlıca neden, Bootstrap'ın varyansı tahmin etmesi yanlılıktan daha kolaydır. Bu nedenle, h değerini bu şekilde seçtiğimizde, hem yanlılık hem de varyans OHK'ye çok katkıda bulunur, bu yüzden yanlılığı görmezden gelemeyiz. Bununla birlikte, Bootstrap yanlılığı tutarlı bir şekilde tahmin edemez bu sebepten dolayı OHK tahmini tutarlı değildir.

Güven aralığı, m(x) güven aralığı oluşturmak için kernel regresyonunun aşağıdaki özelliklerini kullanacağız:

$$\sqrt{nh}(\hat{m}_h(x) - E(\hat{m}_h(x))) \xrightarrow{D} N\left(0, \frac{\sigma^2 \cdot \sigma^2{}_K}{p(x)}\right) \tag{4.24}$$

$$\frac{\hat{m}_h(x) - E(\hat{m}_h(x)}{Var(\hat{m}_h(x)} \xrightarrow{D} N(0,1). \tag{4.25}$$

Varyans üç şeye bağlıdır: σ^2, $\sigma^2{}_K$ ve $p(x)$. $\sigma^2{}_K$ miktarı biliniyor çünkü Kernel Fonksiyonun sadece bir özelliğidir. Ortak değişkenlerin p(x) yoğunluğu bir KDE kullanılarak tahmin edilebilir, yani bilinmeyen kalıntı σ^2 gürültü seviyesi. Bu durumda artıkları (residuals) kullanarak tahmin edebiliriz. Kalıntıları hatırlarsak:

$$e_i = Y_i - \hat{Y}_{\hat{i}} = Y_i - \hat{m}_h(X_i) \tag{4.26}$$

$\hat{m}_h \approx m$ olduğunda, artık gürültü e_i'ye yaklaşır. $\sigma^2 = Var(\varepsilon_1)$ miktarı, tahmin etmek için artıkların örnek varyansını kullanabiliriz (artıkların ortalamasının 0 olduğuna dikkat edin).

$$\sigma^2 = \frac{1}{n - 2v + \tilde{v}} \sum_{i=1}^{n} e_i^2 \tag{4.27}$$

burada v, $\tilde{v}$ serbestlik dereceleridir. Böylece a, $(1 - \alpha)$ güven aralığı kullanılarak:

$$\hat{m}_h(x) \pm z_{1-\alpha/2} \frac{\hat{\sigma} \cdot \sigma_K}{\sqrt{\hat{p}_n(x)}}, \tag{4.28}$$

elde edilir, burada $\hat{p}_n(x)$ ortak değişkenlerin KDE'sidir.

5
UYGULAMA

5.1 Parametrik Olmayan Regresyon Analizinin Uygulanması

Parametrik olmayan regresyon yöntemleri ile tahmin yapabilmek için Kernel regresyon (N-W Tahmincisi) kullanılmış olup, değişkenlerin istatistik olarak anlamlı olup olmadıkları test edilip sonuçlar kendi aralarında kontrol edilmiştir. Karşılaştırmalarda R^2 ve modellerin standart hatası kullanılmıştır.

5.1.1 Kernel Regresyon Yöntemi (N-W Tahmincisi)

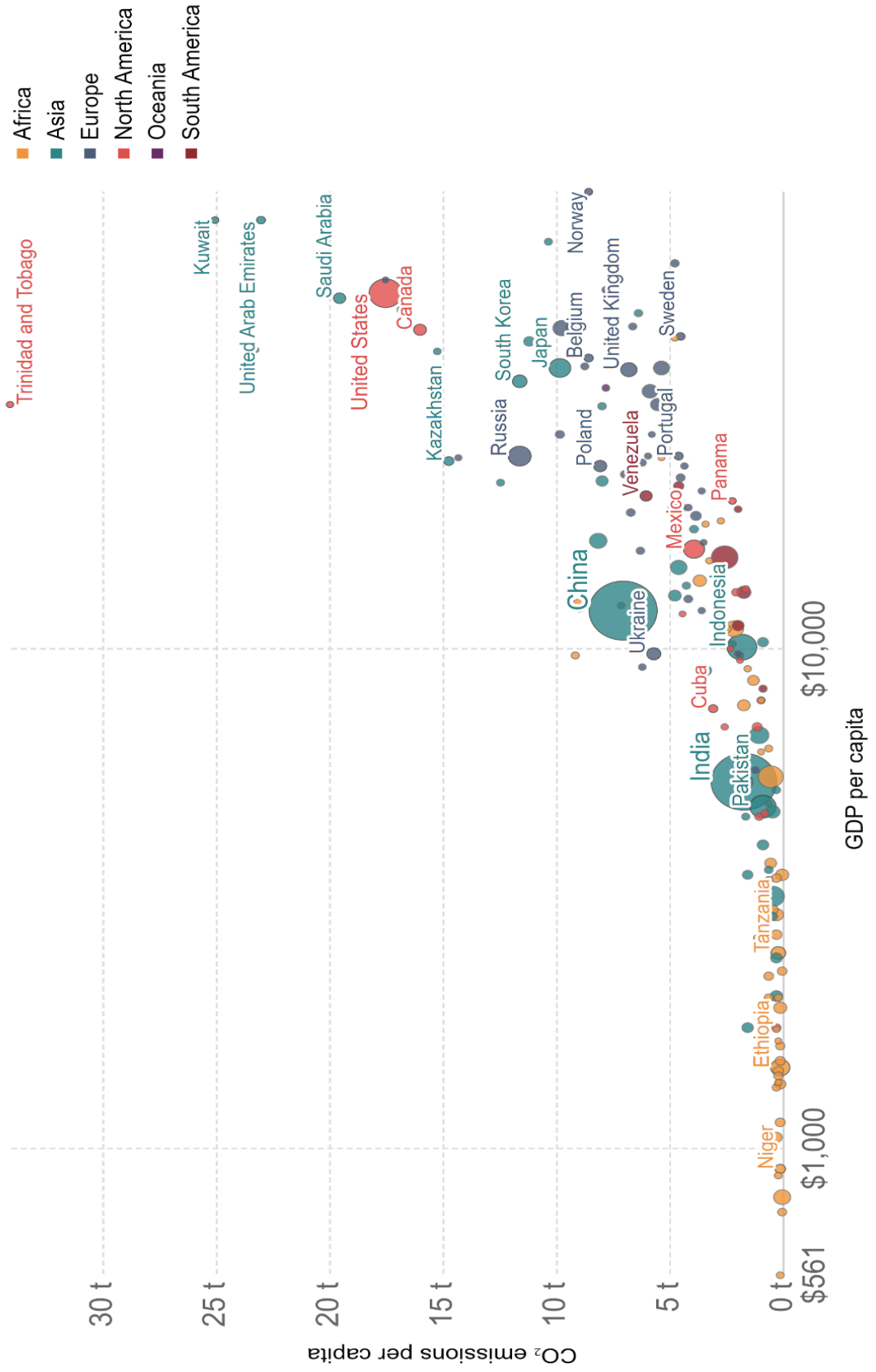
Kernel regresyon yönteminde çapraz geçerlilik yöntemi ile bant genişliği belirlenmiştir. Çapraz geçerlilik yöntemi güçlü (robust) ve yaklaşık olarak optimal bir yöntemdir [33]. Uygulamalarda çok kullanıldığı bilinen Gauss Kernel ağırlık fonksiyonu olarak seçilmiştir. Tek GSYİH ile kurulan modelde analize ait sonuçlar Tablo 5.1'de gösterilmiştir. Ayrıca 194 ülkeye ait CO_2 emisyonu ve GSYİH değerleri görsel olarak Şekil 5.1'de gösterilmiştir.

Tablo 5.1 Gsyih ile kurulan Kernel regresyon sonuçları

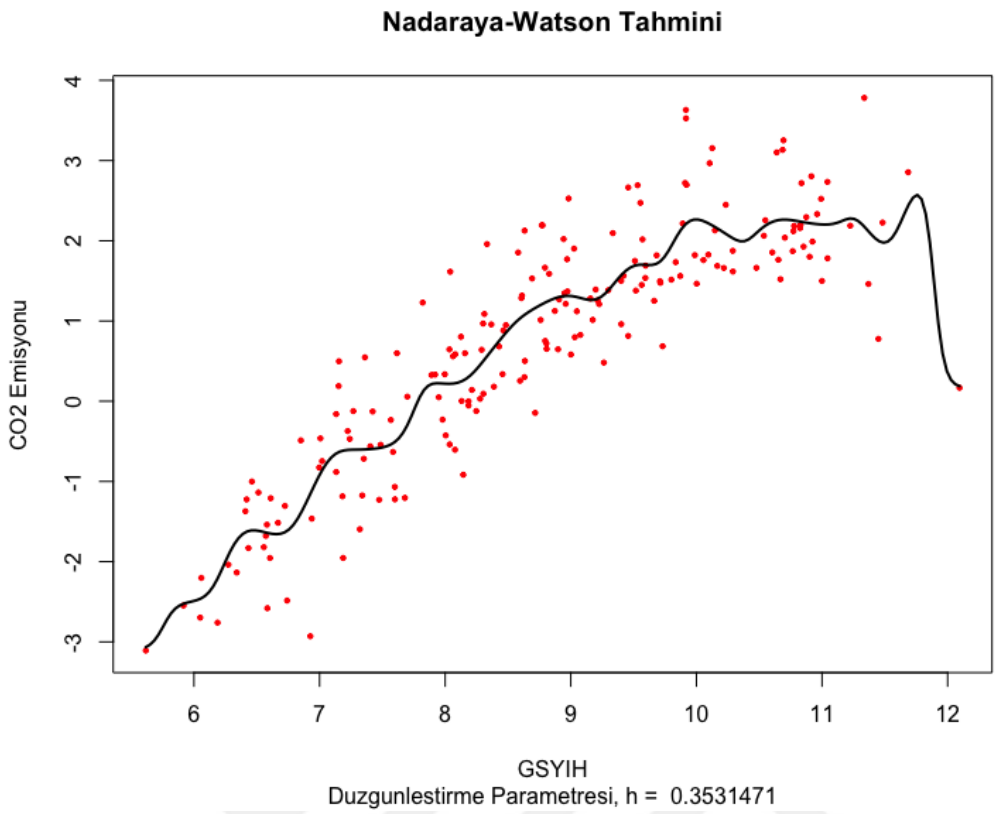
Bağımlı Değişken	CO_2 emisyonu
Bağımsız Değişken	Gayri Safi Yurt İçi Hasıla (GSYİH)
Gözlem Sayısı	194
Modelin Standart Hatası	0.632856
Band Genişliği	0.3531471
R^2	0.819619
p Değeri	2,22e-16***

***0.001 düzeyinde istatistik olarak anlamlıdır.

Oluşturmuş olduğumuz tek bağımsız değişkeni olan modelin standart hatası 0.632856 ve modelin uyum iyiliğini ölçen R^2 değeri 0.819619 olarak bulunmuştur. Çapraz geçerlilik yöntemi ile buluna bant genişliği 0.3531471'dir. GSYİH'nın yer aldığı modelin anlamlılığı test edildiğinde p değeri 2,22e-16 bulunduğundan H_0 hipotezi reddedilmekte ve GSYİH'nın 0,001 düzeyinde istatistiksel olarak anlamlı olduğu söylenebilir. Şekil 5.2'de Nadaraya-Watson tahmini için CO_2 emisyonu ile GSYİH arasında yapılan düzgünleştirme verilmiştir.



Şekil 5.1 Ülkelerin kişi başı CO_2 emisyonu ve Gsyih değerleri [32]



Şekil 5.2 CO_2 emisyonu ile Gsyih arasındaki Kernel düzgünleştirmesi

5.2 Parametrik Regresyon Analizi (EKK Yöntemi)

CO_2 emisyonu ile GSYİH Arasındaki ilişki parametrik olmayan regresyon analizi ile incelenmiş daha sonra bu ilişki EKK yöntemi ile parametrik regresyon ile incelenmiştir. Her iki regresyon analizinden elde edilen sonuçlar karşılaştırılmıştır.

5.2.1 Parametrik Regresyon Analizi ile Tahmin

Öncelikle regresyon denklemi oluşturularak, kısmı regresyon katsayıları tahmin edilmiş ve elde edilen regresyonun tahminlere uygulanabilmesi için varsayımları gerçekleştirip gerçekleştirmediği incelenmiştir. Parametrik regresyon analiziyle ortaya çıkan sonuçlar Tablo 5.2 ve Tablo 5.3 de gösterilmiştir.

y = CO_2 emisyonu

x_1 = Gayri safi yurt içi hasıla (Gsyih)

Değişkenler olmak üzere, parametrik regresyon denklemi

$$y = a_0 + b_1 x_1$$

olarak kullanılmaktadır. Parametrik regresyon denklemi aşağıdaki gibidir:

$$y = -6.98748 + 0.88045 x_1$$

Tablo 5.2 Parametrik regresyon analiz sonuçları

Bağımlı Değişken	CO_2 emisyonu
Bağımsız Değişken	GSYİH
Gözlem Sayısı	194
Tahminin Standart Hatası	0.7602
Durbin Watson İstatistiği	2.0072
R	0.8599
R^2	0.7395
F Değeri	545.1
p Değeri	2,2e-16

*0.001 düzeyinde istatistik olarak anlamlıdır.

Analiz sonucunda elde ettiğimiz modelin geçerliliğinin belirlenmesi adına F testi yapılması gereklidir. Analiz sonucunda elde ettiğimiz F değeri olan 545.1; sonucunu $F_{\alpha,df_1,df_2} = F_{0,01;1,192} = 6{,}635$ ile karşılaştırdığımızda, H_0 reddedilmektedir GSYİH'nın anakütlede CO_2 emisyonunu etkilediği anlaşılmaktadır.

5.2.2 Parametrik Regresyonun Varsayımlarının İncelenmesi

Regresyon modelimizin tahminlerde kullanılabilmesi için modelin varyasımları sağlaması gerekmektedir.

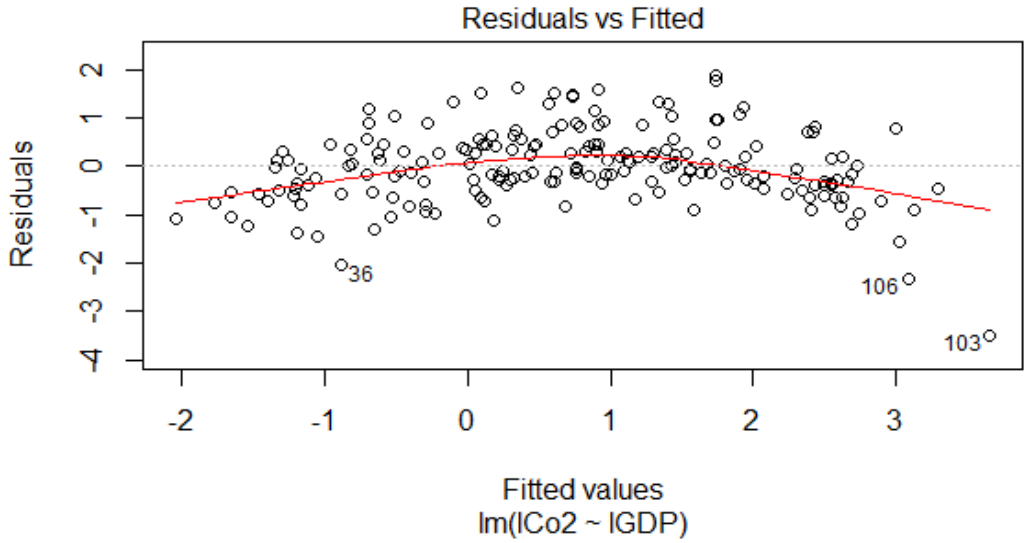
Çoklu Doğrusal Bağlantı: Bağımsız değişkenler arasındaki ilişki, çoklu regresyon analizi yapılması gereken durumlarda incelenir. Parametrik regresyon analizinde adım adım regresyon analizi uygulanması ve analizde tek bir bağımsız değişken bulunmasında dolayı çoklu doğrusal bağlantı sorunu söz konusu olmamaktadır.

Değişen Varyanslılık: Değişen varyanslılığın araştırılmasında elde edilen regresyon denklemindeki hata terimlerinin mutlak değeri ile bağımsız değişkenlerin Spearman Sıra Korelasyonu sonucu bulunarak araştırılmıştır. Tablo 5.3 de sonuçlar verilmiştir.

Tablo 5.3 Spearman korelasyon katsayısı sonuçları

	Hata Terimi	GSYİH
Hata Terimi	1	0.8713908
GSYİH	0.8713908	1

Tabloya bakıldığında hata terimleri ile bağımsız değişken arasında anlamlı bir ilişki olduğu gözlemlenmektedir. Sonuçlara bakılarak değişen varyans durumunun olduğunu ve buna bağlı olarak varyansın sabitlik varyasımına uymadığı söylenebilir.



Şekil 5.3 Hata kalıntılar grafiği

Değişen varyans, artıklarda belirgin bir koni şekli üretir. Temel olarak değişen varyans, tahmin değerleri arttıkça artıkların varyansının da artmasıdır. Şekil 5.3 incelendiğinde artıklar (residuals) grafiğinden değişen varyans olduğu kolayca görülebilmektedir.

Hata Terimlerinin Normal Dağılması: Hata terimlerinin normal dağılım varsayımını sağlayıp sağlamadığını kontrol etmek amacıyla analizde Kolmogorov Smirnov testi ile incelenmiştir. Elde edilen sonuç değeri 0.22474 tablo değeri 0.11 değerinden büyük olması nedeniyle %1 anlamlılık düzeyinde hataların normal dağıldığını öngören H_0 hipotezini reddediyoruz.

Hata Terimlerinin Otokorelasyonu: Otokorelasyonun hata terimlerinin üzerindeki varlığının incelenmesi amacıyla Durbin-Watson (D-W) testinden yararlanılmıştır.

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} \left(e_i\right)^2} \tag{5.1}$$

Şeklinde gösterilmektedir. Verilen formüle göre hesaplanan değer 2.0072'dir. %1 anlamlılık seviyesinde k = 1 bağımsız değişken sayısına göre tablo değeri 1.664'dür. Hesap değerimiz tablo değerinden büyük olduğundan dolayı H_0 kabul edilerek otokorelasyon olmadığı söylenebilir.

Varsayımlardan elde edilen sonuçlar, parametrik regresyondan elde edilen sonuçların tahminde kullanılamayacağı yönündedir. Parametrik regresyon analizinde elde ettiğimiz değerlere bakarsak, Spearman Sıra Korelasyon testinden elde edilen 0.8713908 sonucu hata terimleri ile bağımsız değişkenler arasında sabit varyans varsayımının olmadığını göstermektedir. Analizden elde ettiğimiz sonuçlarla kurulan modelde çoklu korelasyon katsayısı 0.8599 olarak bulunmuştur. Bu sonuca göre bağımsız değişkenimiz GSYİH, CO_2 emisyonunu %85.9 oranında etkilemektedir. Bu oran yüksek bir orandır ama CO_2 emisyonunu etkileyen başka değişkenlerde olabilir.

5.3 Analiz Sonuçlarının Değerlendirilmesi

CO_2 emisyonu ile GSYİH arasındaki bulunan ilişkiyi inceleyebilmek amacıyla önce parametrik olmayan bir yöntem olan Nadaraya-Watson Kernel Regresyon yöntemi daha sonrada EKK yöntemi ile parametrik regresyon analizi yapılmıştır. Analizlerden elde edilen sonuçlar incelendiğinde CO_2 emisyonu ile GSYİH arasında anlamlı bir ilişki bulunmuştur. Analizler için bir özet Tablo 5.4 aşağıdaki gibidir.

Tablo 5.4 Parametrik ve parametrik olmayan regresyon analiz sonuçları

Yöntem	N	R^2	Ortalama Tahmin Hatası
Kernel Regresyon (N-W Tahmincisi)	194	0,8196	0,6328
Parametrik Regresyon (EKK Yöntemi)	194	0,7395	0,7602

R^2 değerlerine bakıldığında sırasıyla %81 ve % 73 bulunmuştur, standart hatalar ise %63 ve %76'dır. Sonuçlara bakarak değerlendirme yapacak olursak en iyi sonucu veren Kernel Regresyon Nadaraya-Watson tahmincisidir. Bunun nedeni parametrik olmayan regresyon yöntemi ile, modelde bulunan değişkenler arasındaki ilişki önceden belirlenmiş bir fonksiyonla sınırlandırılmayıp daha esnek bir çerçevede yorumlanması dolasıyla daha etkilidir denilebilir. Şekil 5.2 incelendiğinde bu fark gayet iyi şekilde ortaya konulduğu görülmektedir. Parametrik olmayan regresyon analizinde çıkan R^2 değerinin yüksek çıkması elimizdeki bağımsız değişkenin bağımlı değişkeni çok iyi derecede açıklamasının bir ifadesidir. Parametrik regresyonun varsayımları sağlamaması ayrıca model tahmininin parametrik olmayan regresyon üzerinde yapılması gerektiği yönündeki bulguları desteklemektedir.

6
SONUÇ VE ÖNERİLER

Ekonomi ile yaşam kalitemiz arasında bir ilişki olduğu uzun zamandan beri araştırmacılar tarafından araştırılan bir konudur. Ancak kişi başına düşen GSYİH meydana çıkan sera gazlarının havaya karışımını, özellikle de karbondioksit gazının salınımını nasıl etkilediği konusunda görüşler vardır. Dünya bankası veri tabanından alından veriler incelendiğinde karbondioksit emisyonu ve GSYİH arasında pozitif korelasyon olduğu anlaşılabilir, yani bir ülkenin GSYİH'sı arttıkça karbondioksit salınımı da artar. Çoğunlukla malzeme üretiminde ve hizmetinde GSYİH'da artışa sebebiyet veren durumlar karbondioksit salınımına neden olmaktadır. Çoğu ürünün üretiminde fosil yakıtlar kullanılması bu miktarın artışında önemli bir pay sahibidir. Analizlerden elde edilen korelasyon katsayısının yüksekliği de bu iki değişkenin birbiri ile bağlantılı olduğunu açıkca göstermektedir. Bu çalışmada sırasıyla parametrik ve parametrik olmayan EKK ve Kernel regresyon yöntemleri karşılaştırılmış veriyi açıklaması bakımından parametrik olmayan Kernel regresyon yönteminin daha iyi sonuçlar verdiği görülmüştür. Tablo 5.1'de görüldüğü gibi tahminin standart hatası 0.63 ve R^2 değeri 0.81 dir. Tablo 5.2'da EKK yöntemiyle modelin standart hatası 0.76 ve R^2 değeri 0.73 bulunmuştur. Bu sonuçlara bakıldığında Kernel regresyon yöntemi ile bulunan tahminin standart hatasının daha küçük olması ve R^2 değerleri bakımından yöntemin açıklayıcılığının daha yüksek bulunması Kernel regresyon yönteminin kullanılması gerektiğini ortaya koymuştur. Ancak CO_2 emisyonunu etkileyen sanayileşme, şehirleşme, enerji tüketimi gibi değişkenlerde olduğu unutulmamalıdır. Bununla birlikte tekil olarak belirli ülkelerin incelendiği çalışmalarda görülecektir ki üretimde dünya üzerinde önemli paya sahip ülkeler hem GSYİH hem de CO_2 emisyonunu miktarları açısından üst sıralardadırlar. CO_2 emisyonunu düşürmek amacıyla yenilenebilir enerjiye ve çevrenin korunmasına yönelik yapılan yatırımlarla birlikte gelişmekte olan ülkelere yapılacak olan yatırımlar ile birlikte CO_2 emisyonunu miktarının düşürülmesi sağlanabilir.

KAYNAKÇA

[1] Arouri, M.E.H., A.B. Youssef, H. M'henni and C. Rault (2012). Energy consumption, economic growth and CO2 emissions in Middle East and North African countries. Energy Policy 45, 342-349.

[2] Zhang, X.P. and X.M. Cheng (2009). Energy consumption, carbon emissions, and economic growth in China. Ecological Economics 68, 2706-2712.

[3] Jalil, A. and S.F. Mahmud (2009). Environmental Kuznets Curve for CO2 emissions. A cointegration analysis for China. Energy Policy 37, 5167-5172.

[4] Soytas, U., S. Ramazan and E.T. Ewing (2006). Energy consumption, income, and carbon emissions in the United States. Ecological Economics 62, 482-489.

[5] Menyah, K. and Y. Wolfe-Rufael (2010). CO2 emissions, nuclear energy, renewable energy and economic growth in the US. Energy Policy 38, 2911-2915.

[6] Kim, S., K.Lee and K. Nam (2010). The relationship between CO2 emissions and economic growth: The case of Korea with nonlinear evidence. Energy Policy 38, 5938-5946.

[7] Halicioglu, F. (2009). An econometric study of CO2 emissions, energy consumption, income and foreign trade in Turkey. Energy Policy 37, 1156-1164.

[8] Shahbaz, M., H.H. Lean and M.S. Shabbir (2012). Environmental Kuznets Curve hypothesis in Pakistan: Cointegration and Granger causality. Renewable and Sustainable Energy Reviews 16, 2947-2953.

[9] Pao, H.T., H.C. Yu and Y.H. Yang (2011). Modeling the CO2 emissions, energy use, and economic growth in Russia. Energy 36, 5094-5100.

[10] Alam, M.J., I.A. Begum, J. Buysse, S. Rahman and G.V. Huylenbroeck (2011). Dynamic modeling of causal relationship between energy consumption, CO2

emissions and economic growth in India. Renewable and Sustainable Energy Reviews 15, 3243-3251.

[11] Al-mulali, U. C.N.B. Che Sab (2012). The impact of energy consumption and CO2 emission on the economic and financial development in 19 selected countries. Renewable and Sustainable Energy Reviews 16, 4365-4369.

[12] Ang, J.B. (2007). CO2 emissions, energy consumption and output in France. Energy Policy 35, 4772-4778.

[13] Li, F., S. Dong, X. Li, Q. Liang and W. Yang (2010). Energy consumption-economic growth relationship and CO2 emissions in China. Energy Policy 39, 568-574.

[14] Wang, S.S., D.Q. Zhou, P. Zhou and Q.W. Wang (2011). CO2 emissions, energy consumption and economic growth in China: A panel data analysis. Energy Policy 39, 4870-4875.

[15] Lotfalipour, M.R., M.A. Falahi and M. Ashena (2010). Economic growth, CO2 emissions, and fossil fuels consumption in Iran. Energy 35, 5115-5120.

[16] Ozturk, I. and A. Acaravci (2010). CO2 emissions, energy consumption and economic growth in Turkey. Renewable and Sustainable Energy Reviews 14, 3220-3225.

[17] Ghosh, S. (2010). Examining carbon emissions economic growth nexus for India: A multivariate cointegration approach. Energy Policy 38, 3008-3014.

[18] Neyran Orhunbilge, Uygulamalı Regresyon ve Korelasyon Analizi, 2. Baskı, İstanbul, 2000, s.12.

[19] Francis Galton, “Regression Towards Mediocrity in Hereditary Stature”, Journal of Anthropological Institute of Great Britain and Ireland, Vol. 15, 1886, pp.246–263

[20] Çağlayan, E. (2012) Nonparametrik regresyon modelleri. İstanbul: Derin Yayınları

[21] Hardle, W. (1990). Applied nonparametric regression (No. 19). Cambridge University Press

[22] Randall L. Eubank, Nonparametric Regression and Spline Smoothing, Second Edition, 1990, s.2

[23] Hardle, W. (1990). Applied nonparametric regression (No. 19). Cambridge University Press s. 11

[24] Adrian Pagan, Aman Ullah, Nonparametric Econometrics, Cambridge University Press, 1999, s. 9.

[25] Eubank, R. L. (1999). Nonparametric regression and spline smoothing. New York. CRC Press.

[26] Keele, L. J. (2008). Semiparametric regression for social sciences. UK: John Wiley Sons

[27] T.Hastie and R.Tibshirani, Generalized Additive Models, Chapman-Hall, 1990, s.9

[28] Ruppert, M.P.Wand, R.J.Carroll, Semiparametric Regression, Cambridge University Press, 2003, s.57

[29] Adrian Bowman, Peter Hall, Tania Prvan, "Bandwith Selecton for the Smoothing of Distribution Functions", Biometrika, Vol.85, No.4, (De. 1998), s.799-808

[30] "Applied Nonparametric Methods" Cowles Foundation Discussion Paper 1069, 1994, s.1-23.

[31] Myoug-Jae Lee, 1990, "methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models", Springer-Verlag, New York, s:41.

[32] https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

TEZDEN ÜRETILMIŞ YAYINLAR

İletişim Bilgisi: mrtslmn1081@gmail.com

Konferans Bildirileri

1. M. Salman ve E. Tuna, "Nonparametrik Yöntemler ve Bir Uygulaması", 7. Uluslararası Bilimsel Araştırmalar Kongresi, 13-14 Mart 2020, Ankara.