T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

SEMİPARAMETRİK REGRESYON VE BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Seda BAĞDATLI

Danışman: Prof.Dr. Münevver TURANLI

İstanbul – 2010

T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

İSTATİSTİK ANABİLİM DALI

SEMİPARAMETRİK REGRESYON VE BİR UYGULAMA

YÜKSEK LİSANS TEZİ

Seda BAĞDATLI

Danışman: Prof.Dr. Münevver TURANLI

İstanbul – 2010

T.C.
İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

ONAY SAYFASI

Yüksek Lisans Öğrencisi …….'ın "……..." konulu tez çalışması jürimiz tarafından ……. Yüksek Lisans tezi olarak oybirliği / oyçokluğu ile başarılı bulunmuştur.

İmza

Tez Danışman : …….

Jüri Üyesi : …….

Jüri Üyesi : …….

ONAYLI

Yukarıdaki jüri kararı Enstitü Yönetim Kurulunun ..... / ..... / ..... tarih ve …… kararı ile onaylanmıştır.

Prof.Dr.Sezgin Alsan
Müdür T.

Hazırlamış olduğum tez özgün bir çalışma olup YÖK ve İTİCÜ Lisansüstü Yönetmeliklerine uygun olarak hazırlanmıştır. Ayrıca, bu çalışmayı yaparken bilimsel etik kurallarına tamamıyla uyduğumu; yararlandığım tüm kaynakları gösterdiğimi ve hiçbir kaynaktan yaptığım ayrıntılı alıntı olmadığını beyan ederim. Bu tezin ihtiva ettiği tüm hususlar şahsi görüşüm olup İstanbul Ticaret Üniversitesinin resmi görüşünü yansıtmamaktadır.

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Klasik (parametrik) regresyon teknikleri, bağımlı değişkenin bağımsız değişkenlerle doğrusal bir ilişki içerisinde olduğunu ve ilişkinin şeklinin biliniyor olduğunu varsayar. Bu varsayımların sağlanamaması durumunda ise parametre tahminleri güvenilir olmamaktadır. İlişkinin şeklinin bilinmediği ya da bilinen parametrik matamatiksel kalıplara uymadığı durumlarda parametrik olmayan regresyon teknikleri kullanılmaktadır. Ancak bu teknikler birden fazla bağımsız değişken olma durumunda çok boyutluluğun yarattığı sıkıntı nedeniyle özellikle yorumlama aşamasında zorluklara neden olmaktadır. Birden fazla bağımsız değişken söz konusu olduğunda, bağımsız değişkenlerin bazıları bağımlı değişkenle doğrusal ilişki içerisinde bulunabilirken, bazıları doğrusal olmayan ilişki içerisinde bulunabirler. Bu tür ilişkilerin modellenebilmesi için, parametrik ve parametrik olmayan regresyon fonksiyonunun toplamsal olarak birleşiminden oluşan semiparametrik regresyon modellerinden yararlanılmaktadır.

Bu çalışmanın amacı, son yıllarda uygulama alanında sıkça kullanılmaya başlanan semiparametrik regresyon analizini inceleyerek gerçek veriler üzerinde uygulanabilirliğini göstermektir. Çalışmanın konusu olarak; son yıllarda konut sektöründeki gelişmelerle birlikte hızla yaygınlaşan "site" yerleşiminin finansal boyutu semiparametrik regresyon modelleri ile incelenmiştir. Semiparametrik regresyon modellerinin tahmini, splayn düzeltme tekniği ile R ortamında yazılan bir program yardımıyla gerçekleştirilmiştir.

Anahtar Kelimeler: Parametrik olmayan regresyon, düzeltme teknikleri, splaynlar, toplamsal modeller, semiparametrik regresyon, otomatik düzeltme.

ABSTRACT

Classical (parametric) regression techniques are based on the assumption that the independent variable is correlated linearly with the dependent variables and the pattern of this relation is known. When such assumption cannot be verified, parameter estimations fail to be reliable. In cases where the way of correlation is not known or it does not comply with the known parametric mathematical patterns, nonparametric regression techniques are to be applied. One shortcoming concerning this procedure emerges particularly in the interpretation process due to problems brought about by multidimensional aspect of the existence of more than one independent variable. Whenever confronted with a case that includes more than one independent variable, some of the independent variables correlate linearly with the dependent variable; at other times some of the independent variables might correlate nonlinearly. In order to establish a modeling for such relations, semiparametric regression models, comprising the aggregate of parametric and nonparametric regression function, are utilized.

The purpose of the study is to examine the semiparametric regression analysis, which has started to be utilized frequently in recent years in practice, and to demonstrate its feasibility on actual data. The scope of the study, the financial aspect of highly widespread "housing estate (group of buildings)" as a result of developments in housing industry recently, has been examined through semiparametric regression models. The estimation of semi-parametric regression models has been conducted via software made on R medium by spline smoothing technique.

Key words: Nonparametric regression, smoothing techniques, splines, additive models, semiparametric regression, automatic smoothing.

1. GİRİŞ

Son yıllarda bilgisayar teknolojisinin gelişmesiyle, bilinen istatistiksel tekniklerin eksikliklerini gidermek amacıyla bir çok teknik geliştirilmiştir. Bu istatistiksel tekniklerin en önemlilerinden biri de semiparametrik regresyon analizidir. Bilindiği üzere, Klasik (parametrik) regresyon teknikleri, bağımlı değişkenin bağımsız değişkenlerle doğrusal bir ilişki içerisinde olduğunu ve ilişkinin şeklinin biliniyor olduğunu varsayar. Bu varsayımların sağlanamaması durumunda parametre tahminleri güvenilir olmamaktadır. Değişkenler arasındaki ilişkilerin doğrusal olmadığı ve şeklinin bilinen fonksiyonlardan herhangi birine uymadığı durumlarla karşılaşılabilmektedir. Parametrik olmayan regresyon analizi bu durumda alternatif bir çözüm yöntemi olarak sunulmaktadır. Parametrik olmayan regresyon analizi birden fazla bağımsız değişken olduğunda karmaşıklığa neden olmaktadır. Ayrıca bazı durumlarda, bağımsız değişkenlerin bazıları bağımlı değişkenle doğrusal ilişki içerisinde bulunabilirken, bazıları doğrusal olmayan ilişki içerisinde bulunabirler. Bu tür bir ilişkinin modellenebilmesi için, parametrik ve parametrik olmayan regresyon fonksiyonunun toplamsal olarak birleşiminden oluşan semiparametrik regresyon modelleri geliştirilmiştir.

Semiparametrik regresyon modeli ile ilgili ilk çalışmalar, 1985 yılında Green ve Yandell tarafından semiparametrik genelleştirilmiş doğrusal modeller isimli makale ile, 1986 yılında Engle, Granger, Rice ve Weiss tarafından günlük ortalama hava sıcaklığı ile elektrik satışları arasında ilişkiyi splayn düzeltme yöntemi kullandığı semiparametrik regresyon modeli ile tanıtmışlardır. Daha sonra, 1988 yılında Speckman kısmi doğrusal modellerin uygulanması, yine 1988 yılında Robinson semiparametrik regresyon modellerinin kestirimi ile 80'li yılllarda semiparametrik regresyon modelleri tanıtılmıştır. Bu yıllardan günümüze kadar semiparametrik regresyon, genelleştirilmiş semiparametrik regresyon ve bir çok semiparametrik yöntem geliştirilmiştir.

2000'li yıllarda yapılan çalışmalarda semiparametrik regresyon yöntemlerinin bir çok alanda kullanıldığı ve geliştirildiği gözlemlenmektedir. (Bu bağlamda bu çalışmayı destekleyen bir çok çalışma bulunmaktadır). 2001 yılında Lın ve Carrol tarafından kümelenmiş verilerde, 2003 yılında Kim, Cohen, Carroll tarafından eşleşmiş vaka kontrol çalışmalarında, 2006 yılında Gronniger tarafından vucüt-kitle indeksi ve ölüm arasındaki ilişkilerin incelenmesinde, 2009 yılında Delis ve Papanikolaou tarafından banka verimliliğini etkileyen unsurların belirlenmesinde ve burada yer verilemeyen bir çok çalışmada semiparametrik modeller uygulanmıştır.

Bu çalışmanın amacı, semiparametrik regresyon modellerinin kullanım alanlarını ve getirdiği çözümleri tanıtmak, bu modellerin parametrelerine ilişkin tahmin ve çıkarımlar yapmaktır.

Çalışma altı bölümden oluşmaktadır. Birinci bölüm giriş bölümü olarak ele alınmış ve bu bölümde çalışmanın amacı, konusu, semiparametrik regresyon analizinin tarihsel gelişimi ve yapılan son çalışmalar anlatılmıştır.

İkinci bölümde; klasik regresyon analizi, doğrusal olmayan ilişki kavramı, doğrusal olmayan ilişkileri doğrusallaştırma yöntemi olarak ele alınan üstel dönüştürme yöntemleri, parametrik olmayan regresyon analizi, regresyonda düzeltme kavramı kısaca açıklanmaya çalışılmıştır. Bu açıklamalardan sonra, semiparametrik regresyon modellerinin temelini oluşturan düzeltme teknikleri incelenmiştir. Düzeltme tekniklerinden splayn düzeltme kavramı, geniş olarak ele alınmış ve açıklanan bütün düzeltme tekniklerinin karşılaştırılması yapılmıştır.

Üçüncü bölümde, düzeltme tekniğine karar verildikten sonra düzeltme parametresinin değerinin belirlenmesi aşaması ve araştırmacının düzgünleştirici tahmini aşamasında hiçbir karar vermesini gerektirmeyen otomatik düzeltme tekniği açıklanmıştır. Lokal polinomiyal regresyon modelleri ve splaynlar için otomatik düzeltme tekniklerinden en çok kullanılanları (çapraz geçerlilik, genelliştirilmiş çapraz geçerlilik) anlatılmış ve karşılaştırmalar yapılmıştır.

Dördüncü bölümde, toplamsal modeller ve genelleştirilmiş toplamsal modeller açıklanmış, daha sonra semiparametrik regresyon modeli, bu modellerin tahmini için en sık kullanılan backfitting algoritması ele alınmıştır. Son aşamada ise, semiparametrik regresyon modellerinde çıkarım yöntemleri incelenmiştir.

Beşinci bölümde ise uygulama olarak, site içerisindeki dairelerin satış fiyatını etkileyen özellikler ve fiyat ile bu özelliklerin arasındaki ilişkiler semiparametrik regresyon analizi ile incelenmiştir. Bu çalışmada ve daireler ile ilgili yapılan uygulamada tüm değerlendirmeler R ortamında yazılan bir programla gerçekleştirilmiştir. Söz konusu uygulamada oluşturulan semiparametrik modellerin performansları karşılaştırmalı bir biçimde incelenmiştir.

Sonuç bölümünde ise uygulamadan elde edilen sonuçlar değerlendirilmiştir.

2. REGRESYON TÜRLERİNE GENEL BİR YAKLAŞIM VE DOĞRUSAL OLMAYAN İLİŞKİ KAVRAMI

Regresyon analizi, en geniş tanımıyla, değişkenler arasındaki ilişki ve bağıntıların araştırılmasıdır. Belirli amaç ve varsayımlar altında bağımsız değişken(ler)in bağımlı değişkene nasıl bağlanacağını araştırma sürecidir. Amaçlar ise en genel biçimde, bağımlı değişkendeki değişimi açıklama, herhangi bir gözleme karşılık gelen ortalama y değerini bulma, noktalara en iyi eğriyi uydurma olarak sıralanabilir (Erar, 2006: 7).

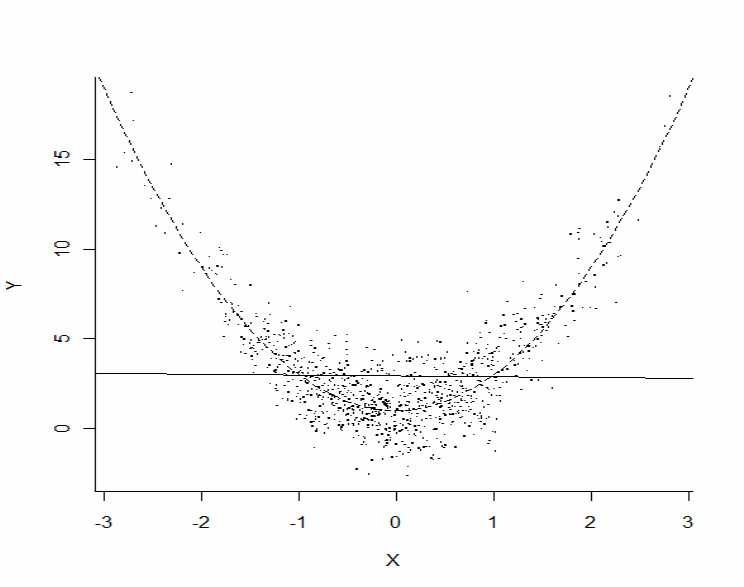
Bilinen klasik doğrusal regresyon modeli,

$$y_i = B_0 + B_1 X_{i1} + B_2 X_{i2} + ... + B_k X_{ik} + \varepsilon_i, \qquad i = 1,...,n \qquad (2.1)$$

şeklinde ifade edilir. Burada, ε_i, sıfır ortalamalı, sabit varyanslı ve açıklayıcı değişkenlerden bağımsız bir hata terimidir. Ayrıca, $X_1,\ldots, X_k$ değişkenlerine göre y'nin koşullu ortalaması, $X_1,\cdots,X_k$ değişkenlerinin bir doğrusal fonksiyonudur. Genellikle böyle bir doğrusal model çok yararlı olmasına karşın, bağımsız değişkenlerin bağımlı değişkenle doğrusal olmayan bir ilişki *(nonlinearity)* içerisinde olmaları durumunda, bu tür bir model uygun değildir. Böyle bir problem ile karşılaşıldığında ilk olarak bağımsız değişkenlerde dönüşüm yapmak düşünülmektedir.

Klasik (parametrik) regresyon modelinde, modelin geçerli sonuçlar verebilmesi için, bir diğer deyişle en iyi β değerini bulabilmek için regresyon fonksiyonun şeklini doğru belirlemek gerekmektedir. Bu amaçla öncelikle, bağımsız değişkenlerle bağımlı değişken arasında nasıl bir ilişki olduğu ortaya çıkarılmalıdır. Doğrusal olmayan ilişkiler analiz edilmediği takdirde, değişkenler arasında ilişki olmadığı sonucuna varılabilir. Şekil 2.1'de X bağımsız ve Y bağımlı değişkenleri arasında basit doğrusal bir regresyon modeli kurulmuş ve düz çizgi ile belirtilen tahmini regresyon doğrusu elde edilmiştir. Basit doğrusal regresyon sonuçlarına göre X ve Y değişkenleri arasında ilişki olmadığı sonucuna varılır. Ancak modelin sağ tarafına karesel terim eklendiğinde Şekil 2.1'de kesikli çizgi ile gösterilen tahmini regresyon doğrusu elde edilmiştir. Bu sonuçlara göre ise X ve Y değişkenleri arasında

güçlü doğrusal olmayan bir ilişki olduğu sonucuna varılmıştır. Doğrusal olmayan ilişkilerin göz ardı edilmesi, yanlış sonuçlara yol açaçaktır (Keele, 2008: 4-5).



Şekil 2. 1: X ve Y Değişkenleri için Tahmini Regresyon Doğruları Grafiği

2.1. Üstel Dönüştürme Yöntemleri

Bazı durumlarda doğrusal olmayan ilişkiler, değişkenlere bazı dönüşümler uygulayarak doğrusallaştırılabilmektedir. Bu modeller, doğrusallaştırılmış modeller olarak adlandırılır. Genel olarak kullanılan fonksiyonel şekiller, doğrusal, yarı-logaritmik ve logaritmik-doğrusal modellerdir. Doğrusal ve doğrusallaştırılmış modeller için belirli varsayımlar, bağımsız değişkenler ile bağımlı değişken arasında açık ve kesin, deterministik bir ilişkiyi belirtir (Aydın, 2005: 6).

Bir çok üstel dönüştürme yöntemi bulunmaktadır. Örneğin, pozitif bir X değişkenini ele alır ve X^{λ} dönüşümünü gerçekleştirmek istersek λ'nın değerine bağlı olarak dönüştürme yöntemi farklılık gösterir (Weisberg, 2005: 8). Tablo 2.1'de dönüştürme yöntemleri gösterilmektedir.

Tablo 2. 1: Üstel Dönüştürme Yöntemleri

λ	**Dönüştürme Yöntemi**
0	Logaritmik Dönüşüm
1	Dönüşüm Yapılmaması
2	Kuadratik (Karesel) Dönüşüm
3	Kübik Dönüşüm
1/2	Karekök Dönüşümü
1/3	Küpkök Dönüşümü

Üstel dönüştürme yöntemleri doğrusal olmayan ilişkileri modellemek için çok kullanışlı modellerdir. Örneğin, herhangi bir ülkede başbakanlık seçimleri yapılırken yaşın artması tecrübenin artması anlamına geldiğinden yaşça büyük olan insanların seçilme olasılığı gençlere göre daha fazladır. Bu durumda yaşın, seçilme durumu ile ilişkisinin doğrusal olduğu söylenebilir. Ancak, belirli bir yaştan sonra yaş artışı bu kişinin oy almasını ters yönde etkileyecektir. Böyle bir durumda ilişkiyi doğrusal olarak modellemek doğru olmayacaktır. Bu tip bir modelde, yaş değişkenine uygun dönüştürme yöntemi uygulanır ve modelin sağ tarafına hem doğal değişken hem de dönüştürme işlemi uygulanmış değişken eklenir. Eğer dönüştürme işlemi uygulanmış değişken istatistiksel olarak anlamlı ise, ilişkinin doğrusal olmadığına karar verilir.

Üstel dönüştürme yöntemleri çok kullanışlı olmalarına rağmen bazı zorlukları bulunmaktır. Bunlardan en önemlileri, hangi üstel dönüştürme yönteminin seçileceği ve hangi yöntemin modelin fonksiyonel şeklini daha iyi belirleyeceğidir. Bu sorunun en uygun çözümü, bütün dönüştürme yöntemlerini uygulayıp hata kareler toplamını minimum yapan λ değerinin seçilmesidir. Ancak burada da modeli yorumlama aşamasında sorunlar yaşanabilmektedir (Berk, 2006: 9).

Üstel dönüştürme yöntemleri sadece pozitif değişkenlere uygulanabilmektedir. Eğer değişkende negatif değerler ya da sıfır var ise değişkenin bütün değerlerine sabit bir sayı ekleyerek dönüştürme yöntemlerinin uygulanması gerekmektedir (Weisberg, 2005: 11) .

Sonuç olarak, üstel dönüştürme yöntemleri, çok kullanılmasına karşın, bahsedilen bu gibi durumlarda kullanışsız hale gelmektedir. Ancak doğrusal olmayan ilişkilerin modele katılmaması daha büyük sorunlara yol açmaktadır. Bu durumda çözümlenmesi çok zor olan ve bilgisayar programlama dillerine ihtiyaç duyan bazı yöntemler geliştirilmiştir (Hastie, Tibshirani ve Friedman, 2003: 11). Bu bağlamda bu tekniklere alternatif olarak parametrik olmayan ve semiparametrik teknikler geliştirilmiştir.

2.2. Parametrik Olmayan Regresyon Analizi

Parametrik olmayan regresyon analizi bağımlı değişken ile bağımsız değişken arasındaki ilişkinin fonksiyonel şekli hakkında kesinlik olmaması durumunda ortaya çıkmaktadır. Parametrik olmayan yaklaşımlar ilişkinin fonksiyonel şeklinin belirlenmesinde kolaylık sağlamalarının yanında, düşünülemeyen fonksiyonel şekillerin ortaya çıkarılmasında da yardımcı olmaktadır (Çağlayan, 2002: 8).

Genel parametrik olmayan regresyon modeli,

$$y = E(y/x) + \varepsilon \qquad (2.2)$$

$$y = f(x) + \varepsilon = f(x_1,...x_p) + \varepsilon$$

biçiminde ifade edilir. (2.2) modelinde belirtilen f, parametreleri belirginleşmemiş bir fonksiyonel ilişkiyi gösteren bilinmeyen bir fonksiyondur. Model (2.2)'de 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, 1991: 4-5).

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

$$y_i = f(x_i) + \varepsilon_i \ , \ i = 1,...,n \qquad (2.3)$$

biçimindedir. Bu model bir x değişkenine karşı y değişkenin aldığı değerlerin dağılımının grafiksel olarak gösterilmesinde kullanıldığından, genellikle ''dağılım grafiğini düzeltme ya da düzgünleştirme'' (scatterplot smoothing) olarak adlandırılır (Aydın, 2005: 9).

Parametrik olmayan regresyon modelinin amaçları; iki değişken arasında genel bir ilişkiyi açıklayan bir model uydurmak, sabit bir parametrik modeli referans almaksızın yapılacak gözlemlerin bir kestirimini vermek, izole edilen noktaların etkisini bulmak için bir araç sağlamak ve komşu x değerleri arasında interpolasyon veya kayıp değerleri yerine getiren esnek bir metot oluşturmaktır.

2.3. Regresyonda Düzeltme Kavramı

Regresyon analizinin amacı, bilinmeyen regresyon fonksiyonu için uygun bir tahmin elde etmektir. Gözlemsel hataları azaltarak y'nin x'e göre ortalama bağımlılığının önemli ayrıntılarını vermek, yorumu kolaylaştırır. Böyle bir eğri yaklaştırma (tahmin) işlemi genel olarak "düzeltme (smoothing)" olarak adlandırılır (Hardle, 1991: 18).

Kullanılan bu teknik, $RSS(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2$ şeklinde ifade edilen artık (hata) kareler toplamını minimum yaparak, f fonksiyonunu tahmin eder (Aydın, 2005: 12-13).

2.4. Parametrik Olmayan Regresyonda Düzeltme Teknikleri

Düzeltme teknikleri genel olarak üç gruba ayrılmaktadırlar (Keele, 2006: 2).

i. Kernel Düzeltmesi

ii. Lokal Polinomiyal Regresyon

iii. Splayn Düzeltme Tekniği

Tüm bu tekniklerde, bir düzeltme düzeyinin belirtilmesi gerekir. Bu düzey düzeltme parametresi veya düğüm sayılarının bir fonksiyonudur. Uyum eğrisi; $\hat{f}(x) = S_\lambda y$ şeklindeki vektörle belirlenir. S_λ, pozitif bir düzeltme düzeyi ve x değişkenine bağlı olan, fakat y değişkenine bağlı olmayan $(n \times n)$ tipinde bir düzeltme ya da şapka matrisi olarak bilinir. Bu tekniklerin temel amacı, ortalama fonksiyonu için parametrik bir şekil belirlemek değil, verilere en iyi uyumu sağlayacak bir fonksiyonel şekil belirlemektir (Aydın, 2005: 15-16).

2.4.1. Kernel Düzeltmesi

$\{x_i, y_i\}$ $i = 1,...,n$ gözlem değerleri olmak üzere parametrik olmayan regresyonun temel düşüncesi, ham verilerin ağırlıklı ortalamasını kullanarak f regresyon fonksiyonunu tahmin etmektir. Söz konusu ağırlıklar x_i noktalarında oluşan X uzayındaki uzaklıkların azalan bir fonksiyonudur. x_i noktasındaki kestirim için y_j gözlemiyle ilşkili bu tür bir ağırlıklandırma şeması Nadarya (1964) ve Watson(1964) tarafından önerilmiştir:

$$w_{ij} = K\left(\frac{x_{i-}x_j}{h}\right) / \sum_{j=1}^{n} K\left(\frac{x_{i-}x_j}{h}\right) = \frac{K(u)}{\sum K(u)} \qquad (2.4)$$

(2.4) eşitliğinde n gözlemlerin sayısını K ise seçilen ve kernel olarak bilinen, sınırlı ve integrali 1'e eşit olan simetrik bir fonksiyunu, u ağırlıkları hesaplamak için kullanılan parametreyi ve h değeri ise bant genişliği veya düzeltme parametresini ifade etmektedir.

Uygulamada kullanılan farklı tipte kernel fonksiyonları vardır. Ancak kernel fonksiyonun seçimi bant genişliğinin seçiminden daha az bir öneme sahiptir. Bu fonksiyonlar sıfıra göre simetrik ve negatif olmayan değerler almaktadırlar. Ayrıca, ikinci mertebeden türevlenebilirlerdir (Fox, 2008: 477-478).

Herhangi bir x_i noktasındaki f regresyon fonksiyonunun kernel tahmini,

$$\hat{y}_i = \hat{f}(x_i) = \sum_{j=1}^{n} w_{ij} y_j \qquad (2.5)$$

olarak ifade edilir. (2.5) denklemi, matris formunda yeniden yazılarak aşağıdaki biçimde de ifade edilebilir.

$$\hat{f} = Wy \qquad (2.6)$$

Bu ifadedeki W matrisine kernel şapka matrisi veya kernel düzeltme matrisi adı verilir (Aydın, 2007: 730-731).

Kernel fonksiyonlarında h parametresinin seçimi önemli bir rol oynar. Büyük h değerleri için eğri çok yavaş değişir ve düzeltme önemlidir. Bu durumda, tahminin varyansı sınırlı, fakat tahmin oldukça sapmalıdır. Küçük h değerleri için eğri düzensizdir. Bu durumda, sapmalar sınırlı fakat tahminin varyansı büyüktür. Bu yüzden h parametresi sapmalar ve tahminin doğruluğu arasında bir denge sağlar (Keele, 2008: 23).

2.4.2. Lokal Polinomiyal Regresyon

Lokal polinomiyal regresyon yönteminde kernel fonksiyonları kullanılmaktadır ve özel noktalarda p. dereceden polinomların tahmini ile lokal olarak elde edilen tahminlerden yoğunluk fonksiyonu tahmin edilebilir (Çağlayan, 2002: 57).

Lokal polinomiyal regresyon tahmini için izlenecek adımlar şöyle özetlenebilir (Fox, 1997: 418-421). Bir x_0 değerine yakın gözlemlere en büyük ağırlıklar verilir. Bu ağırlıklar sayesinde $|x - x_0|$ büyürken simetriklik ve düzgünlük bozulur. x_0 değerine düzgünleşme ya da düzeltme parametresi h'dan daha uzak gözlemlerin ağırlığı sıfırdır. h sabit olabilir veya x_0 değerinin en yakın komşusunun sabit oranı ölçüsünde düzeltilir. Bu oran aralık (span) adını alır ve s olarak ifade edilir. Aralık sıfır ile bir arasında ve bire eşit olabilir. Lokal tahminde kullanılan gözlemlerin sayısı $m = [n.s]$ olacaktır. Burada kullanılan köşeli parantezler, çıkan sonucunen yakın bir tamsayıya yuvarlanacağını, n ise toplam gözlem sayısını ifade etmektedir. s değeri genellikle başlangıç olarak 0.50 alınır ve deneme yanılma yöntemi ile arttırılır veya azaltılır (Keele, 2008: 29).

Kernel ağırlıkları belirlendikten sonra p. dereceden polinomiyal regresyon ağırlıklı en küçük kareler yöntemi ile,

$$\frac{y_i}{w_i} = \alpha + \beta_1 \frac{x_i}{w_i} + \beta_2 \frac{x_i^2}{w_i} + \ldots + \beta_p \frac{x_i^p}{w_i} + \frac{\varepsilon_i}{p_i} \qquad (2.7)$$

(2.7) modeli, $\sum_{i=1}^{n} w_i^2 \varepsilon_i^2$ ifadesini minimize edicek şekilde tahmin edilir. Bu aşamadan sonra (2.7) modelinden tahmin edilen artıklara dayanan, güçlü (robust) ağırlıklar hesaplanır. Büyük artıklara sahip gözlemlere küçük ağırlıklar, küçük artıklara sahip gözlemlere ise büyük ağırlıklar verilir.

Bu güçlü artıklar,

$$\delta_{k=} B(e_k / 6s) \qquad (2.8)$$

(2.8) eşitliğindeki gibi ifade edilir. (2.8) eşitliğindeki e_k lokal artıkları, s $|e_k|$'nin medyan değerini, B ise iki ağırlıklı kernel fonksiyonunu ifade eder. Bu işlemler en uygun regresyon doğrusu belirlenene kadar tekrarlanır.

Lokal polinomiyal regresyon tahmini için uygulanan adımlar, ağırlıklar olmadan gerçekleşirse LOESS[1], ağırlıklar ile anlatılan adımlarla gerçekleşirse LOWESS[2] adını alır (Cleveland, 1993: 41).

Lokal polinomiyal regresyon modelinde aralık değerinin (span) seçimi en önemli aşamadır. Bu değer olması gerektiğinden çok büyük ise gerçek yoğunluk aşırı düzgünleştirilmiş (oversmoothing) olabileceğinden tahmin sapmalı , çok küçük ise tahmin büyük varyanslıdır (Dinardo ve Tobias, 2001: 28). Bu durumda tahmin edilecek güven bantlarını olumsuz etkilenecektir. Dolayısıyla sapma-varyans dengesini sağlayacak en iyi aralık değeri seçilmelidir. Parametrik olmayan regresyon tahmin eğrisi çok pürüzlü ise aralık değeri arttırılır. En iyi aralık değerinin belirlenmesi için otomatik seçim vb. yöntemler bulunmaktadır.

Lokal regresyon modeli ile tahmin yapılacaksa farklı polinom dereceleri, "LOWESS" ile tahmin yapılacaksa farklı ağırlıklar seçilebilir. Farklı polinom dereceleri ve farklı ağırlıklar tahminde önemsiz sayılabilecek değişikliklere neden olurlar. Bu durumda, polinomun derecesini mümkün oldukça azaltmak en uygun yoldur. Çünkü polinomun derecesi arttıkça modelleki katsayılar artacak, bu durum da varyansın artmasına neden olacaktır (Cleveland,

[1] LOESS: Local Regression.
[2] LOWESS: Local Weighed Regression.

1993: 51). Farklı ağırlık seçimine genel olarak rastlanmamaktadır. Bilgisayar programlarının çoğunda üç ağırlıklı kernel kullanılmaktadır.

Lokal Polimomiyal regresyon modellerinde çıkarım yapmak için güven aralıklarının parametrik olmayan karşılığı olan güven bantları hesaplanmalıdır. Güven bantlarının hesaplanması için serbestlik derecesinin belirlenmesine gerek vardır. Bu yaklaşım hat matrisi (H) yaklaşımı olarak adlandırılır. Parametrik olmayan regresyon modellerinde H matrisi, eşitlik (2.9) ile gösterilen S matrisine eşit olmaktadır.

$$S = X(X'X)^{-1}X' \tag{2.9}$$

Buradan (2.10) formülü ile hata terimlerinin varyansının tahmini hesaplanır. Bu formülde $tr(S)$, parametre sayısını göstermek üzere $n - tr(S)$ artıklar için serbestlik derecesini ifade etmektedir.

$$S_e^2 = \frac{\sum e_i^2}{n - tr(S)} \tag{2.10}$$

Bu aşamadan sonra (2.11) formülü ile Y_i değerleri için güven aralığı tahmini yapılır.

$$\hat{y} \pm 2S_e^2\sqrt{s_{ij}} \tag{2.11}$$

Güven bantları tahmin edilen regresyon doğrusuna yakınsa tahmin güvenilir olmaktadır (Keele, 2008: 41).

Lokal polinomiyal regresyon modellerinin parametrik regresyon modellerinden üstün olup olmadığı F testi yardımı ile karşılaştırılabilir.

2.4.3. Splayn Düzeltme Tekniği

Splayn (spline), bir dizi veri noktalarına polinomiyal bir eğri uydurma ya da bu noktalar arasından pürüzsüz olarak geçen ve bir çok parçadan oluşan esnek bir eğri olarak adlandırılır. Splayn fonksiyonlar da bu fikrin uygulaması olan matematiksel araçlardır. Bu fonksiyonların temel düşüncesi, tanımlanan aralığı bağımsız değişkenler yardımıyla alt aralıklara bölerek, her bir alt aralıkta farklı bir polinomiyal fonksiyon ile bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellendirerek, istenilen mertebeden türevi olan sürekli bir fonksiyon elde etmektir. Splaynlar parametrik olmayan fonksiyonu tahmin eden yöntemlerden biri olarak sunulur ve genellikle parametrik olmayan regresyon ortamında ele alınırlar.

Splaynlar parametrik bir fonksiyon şeklinde değilken, çoğu durumlarda polinomiyal gösterime sahip olan temel fonksiyonların bir birleşimi olarak yazılabilir ve böylece bir anlamda parametrik olurlar (Aydın, 2005: 25).

Splayn düzeltme tekniği, bu çalışmanın konusu olan semiparametrik regreyon yöntemlerinde en sık kullanılan tekniktir.

Splayn düzeltme tekniğine geçmeden önce splayn çeşitleri hakkında genel bilgiler verilecek ardından bu teknik ayrıntılarıyla incelenecektir.

Öncelikle basit regresyon splaynları ele alınacaktır. Splaynlar en genel haliyle regresyon doğrusunun yönünü değiştiren ve bir çok kukla değişkenden oluşan regresyon modelleridir. Şekil 2.2'de düz çizgi X ve Y değişkenleri için tahmin edilen regresyon doğrusunu göstermektedir. Ancak, görüldüğü üzere X ve Y değişkenleri arasında doğrusal olmayan bir ilişki söz konusudur. Splaynların altında yatan mantık, böyle bir ilişki için iki ayrı regresyon doğrusu tahmin etmektir. Birinci regresyon doğrusu iki değişken arasındaki yaklaşık negatif ilişkiyi, ikinci regresyon doğrusu ise yukarıya dönüşü temsil edecektir. Splayn modellerini tahmin edebilmek için, doğrusal regresyon doğrularının birleşeceği noktayı belirlemek gerekir bu noktalara düğüm (knot) adı verilir. Şekil 2.2'de düğüm yeri ve sayısı bellidir. Tek düğüm burada yeterli olacaktır.

c_1, tek düğüm değeri olarak splayn regresyon modeli (2.12) modelindeki gibi oluşturulur.

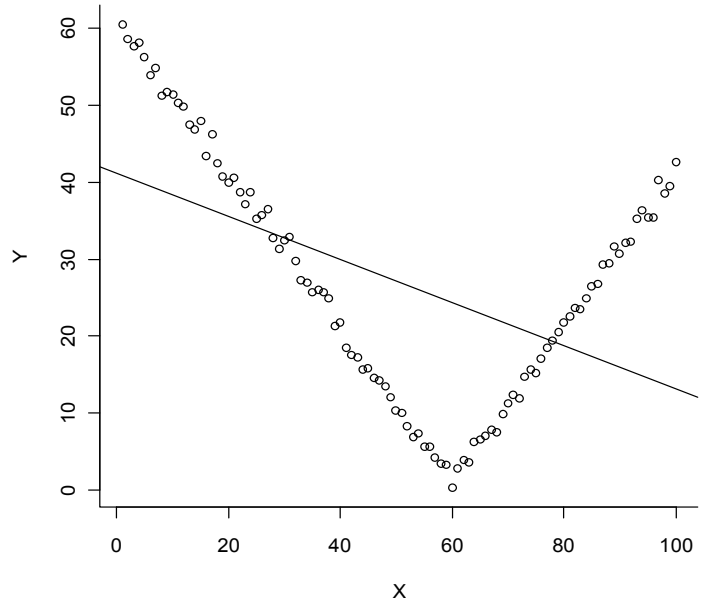
$$y = \alpha + \beta_1 x + \beta_2 (x)_+ + \varepsilon \qquad (2.12)$$

$(x)_+$, fonksiyonu ise (2.13) eşitliğindeki gibi belirlenir.

$$(x)_+ = \begin{cases} x & \text{x} > \text{c}_1 \ \text{ise} \\ 0 & \text{x} \le \text{c}_1 \ \text{ise} \end{cases} \qquad (2.13)$$

Parça fonksiyonlar doğrusal ise (2.12)'deki eşitlik, (2.14) eşitliğindeki ayrı ancak birleştirilmiş iki regresyon doğrusuna dönüşür.

$$y = \begin{cases} \alpha + \beta_1 x & \text{x} \le \text{c}_1 \ \text{ise} \\ \alpha + \beta_1 x + \beta_2 (x - c_1) & \text{x} > \text{c}_1 \ \text{ise} \end{cases} \qquad (2.14)$$



Şekil 2. 2: X **ve** Y **Değişkenleri İçin Tahmin Edilmiş Regresyon Doğrusu**

İki ayrı regresyon regresyon doğrusunu tahmin edebilmek için bir takım baz fonksiyonlara ihtiyaç vardır. Splayn tahmininde baz fonksiyonlar büyük öneme sahiptir. Şekil (2.2)'de görüldüğü gibi tahmin edilecek iki parça fonkiyon olduğu durumda iki baz fonksiyonu belirlemek gerekir. Bu baz fonksiyonlarından bir tanesi düğüm noktasının sağ tarafı için bir diğeri de sol tarafı için belirlenir. Baz fonksiyonu eklemek model matrisine ek bir bağımsız değişken eklemek anlamına gelmektedir (Keele, 2008: 53). Baz fonksiyonunu uygulamak için düğüm noktasının neresi olacağına karar verilmedir. (Şekil 2.2)'de ki gibi bir durumda düğüm noktası tam dönme noktasıdır ($X = 60$). Ancak, birçok durumda bundan çok daha karmaşık doğrusal olmayan ilişkiler ile karşılabilir. Bu durumda daha fazla düğüm noktası ve farklı baz fonksiyonları kullanılmalıdır. Baz fonksiyonunda yapılan değişiklik X ve Y değişkenleri arasındaki tahmin edilen uyumu çok fazla etkilemezken, hesaplamada ve yorumlamada çok büyük kolaylık sağlar (Ruppert, Wand ve Carroll, 2003: 69).

Yukarıda açıklandığı gibi, düğümler arasında eğrisellik söz konusu ise parça regresyon doğruları polinomiyal olarak belirlenebilir. Polinomiyal regresyon fonksiyonlarının düğüm noktalarında birinci türevlerinin tanımlı olduğu, bu durumda splayn tahmininin keskin dönüşlere sahip olmayacağını gösterir (Keele, 2003: 55).

Polinomiyal regresyon fonksiyonlarından, kübik splayn baz fonksiyonları, verideki iniş çıkışlara iyi uyum sağladığından, karesel baz fonksiyonlarına göre daha çok tercih

edilmektedir. İki düğüm noktasına (c_1, c_2) ve kübik baz fonksiyonuna sahip splayn modeli (2.15)'de görüldüğü gibidir.

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - c_1)_+^3 + \beta_5 (x - c_2)_+^3 + \varepsilon \qquad (2.15)$$

Splayn tahmini için oluşturulan modeldeki parametreler, düğüm noktalarının sayıları tarafından belirlenir. Örneğin; k adet düğüm noktası olduğunda kübik splayn fonksiyonu, $k + 4$ adet regresyon katsayısı (sabit katsayı dahil olmak üzere) içermektedir.

İkinci olarak doğal splaynlar incelenecektir. Doğal kübik splaynlar, her bir düğüm noktasının arasına parça fonksiyonlar oluştururlar. Bu durumda, ilk düğümden önce ve son düğümden sonra değişkende bir düşüş gerçekleşirse bunun için herhangi bir parça fonksiyon uydurulmamakta yani tahmin edilemektedir. Doğal kübik splaynlar, belli bir gözlem aralığının uç noktalarında f fonksiyonunun ikinci mertebeden türevlerinin sıfır olması koşulunu sağlar (Aydın, 2005: 32). Özetle doğal kübik splaynlar, x başlangıç noktasından ilk düğüm noktasına kadar olan aralıkta ve son düğüm noktası ile x sınır değeri aralığında doğrusal $f(x)$ fonksiyonları uydurur. Böylelikle x değeri için uç noktalardaki ani değişmeler kontrol altına alınmış olmaktadır.

Kübik splanlar (doğal yada diğerleri) için oluşturulan model matrisinin kolonları, x'deki dönüştürme işlemlerinden sonra yüksek korelasyonlu olmaktadırlar (özellikle çok fazla düğüm noktası kullanıldığında). Bu durum çoklu doğrusal bağıntıya neden olmaktadır. Çoklu doğrusal bağıntı durumunda model matrisi tekil matrise dönüşür ve tahmin süreci gerçekleşemez (Ruppert, Wand ve Carroll, 2003; 70).

Çoklu doğrusal bağıntı problemi B-splayn fonksiyonları ile çözümlenmektedir. Buradaki temel düşünce parça fonksiyonları yeniden ölçeklendirmektir. k adet düğüm noktası içeren kübik B-splayn fonksiyonu (2.16) modelinde gösterildiği gibidir. (2.17) ve (2.18) modelleri ise baz fonksiyonu göstermektedir. (2.17) modeli incelendiğinde baz fonksiyonunda ki yeniden ölçeklendirme açıkça görülmektedir. B- splayn baz fonksiyonlarında yeniden ölçeklendirme, baz fonksiyonun model matrisindeki çoklu doğrusal bağıntı problemini azaltmaktadır (Eiler, Paul, Brian ve Marx, 1996: 100).

$$f(x) = \sum_{i=1}^{k} B_i^2(x) \beta_i \qquad (2.16)$$

$$B_i^2(x) = \frac{x - c_i}{c_{i+2+1}} B_i^{2-1}(x) + \frac{c_{i+2+1} - x}{c_{i+2+1}} B_{i+1}^{2-1}(x) \qquad i = 1,...k \qquad (2.17)$$

$$B_i^{-1}(x) = \begin{cases} 1 & \text{c}_\text{i} \le x < c_{i+1} \quad \text{ise} \\ 0 & \text{d.d} \end{cases} \qquad (2.18)$$

Bütün splayn modellerinde, düğüm noktalarının nereye yerleştirileceğinden daha çok kaç adet düğüm noktasının kullanılacağı önem taşımaktadır (Stone, 1986: 312). Düğüm noktaları özellikle bilgisayar programlarında kartil ve kantiller kullanılarak otomatik yollarla yerleştirilmektedir. Eğer veri setinde şekil (2.2)'deki gibi çok açık durumlar var ise düğüm noktalarını otamatik yolla yerleştirmek sağlıklı olmayacaktır. Önemli olan nokta, kaç adet düğüm noktasının seçileceğidir. Bu durumun nedeni ise, düğüm noktalarının sayısının parça fonksiyonların sayısını etkilemesidir. Düğüm sayısı arttıkça esneklik artmaktadır. Gereğinden az düğüm sayısının kullanılması durumunda çok düz, düşük varyanslı ancak yanlı olabilecek model uydurma (tahmin) gerçekleşirken, gereğinden fazla düğüm sayısı kullanılması ise veriye tam uygun, yansız ancak yüksek varyanslı bir model tahmin edilmesini sağlayacaktır. Bu durum ise "fazla kapsama (overfitting)" problemini ortaya çıkaracaktır (Keele, 2008: 60).

Düğüm noktası sayısının belirlenmesinin başlıca iki yolu bulunmaktadır. Birinci yol bir başlangıç sayısı vermek ve deneme yanılma yöntemi ile devam etmektir. Örnek büyüklüğü 100'den büyük ise başlangıç noktası olarak beş düğüm sayısı, 30'dan küçük ise 3 düğüm sayısı olarak alınabilmektedir. Çok pürüzlü bir görünüm söz konusu ise düğüm sayısı arttırılırken, çok düz bir görünümde düğüm noktası azaltılır (Keele, 2008: 60). İkinci yol ise, Akaike Bilgi Kriterini (AIC)[3] kullanmaktır (Eiler, Paul, Brian ve Marx,1996: 100). Çünkü, düğüm sayısı eklemek modele ek bir parametre eklemek demektir. Düşük AIC değerine sahip olan düğüm sayısı optimum düğüm sayısı olarak belirlenir.

Fazla kapsama (overfitting) problemi burada en ciddi problemdir. İstatistik otoriteleri bu problemi ortadan kaldırmak için, parametrik olmayan bir regresyon tekniği olan ve fazla kapsama olasığını minimize etme yöntemine dayanan cezalı splaynları (penalized spline) kullanmayı önermişlerdir.

Fazla Kapsama (overfitting) problemi hem parametrik hem de parametrik olmayan regresyonda sıkça görülmektedir. Modeldeki parametre sayısını AIC gibi kriterlere bağlı kalarak düşürmek bu sorunu çözmek için önerilen yöntemlerden biridir. Önerilen bir diğer

[3] AIC: Akaike's Information Criterion.

yöntemde, cezalı en küçük kareler regresyonudur. Bu modelde kullanılan her bir parametre için bir ceza (penalty) eklenir. Çoklu regresyon analizinde modele ek bir parametre eklendiğinde, parametrenin modele olan katkısını ölçmek için kullanılan ve R_D^2 ile ifade edilen düzeltilmiş belirlilik katsayısı cezalı tahmin yöntemini esas alarak hesaplanmaktadır.

Bir y bağımlı değişkeni ve bu bağımlı değişkenle ne tür bir ilişki içerisinde olduğu bilinmeyen ve bir x bağımsız değişkenin yer aldığı parametrik olmayan regresyon modeli eşitlik (2.19)'da görülmektedir.

$$y_i = f(x_i) + \varepsilon_i \qquad \mathrm{a} < \mathrm{x}_1 < \ldots < x_n < b \qquad (2.19)$$

Burada ,

$f : [\mathrm{a}, \mathrm{b}]$ aralığında bilinmeyen bir pürüzsüz fonksiyon ya da düzgünleştirilmiş fonksiyon

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

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

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

Parametrik olmayan regresyonda amaç bilinmeyen gerçek $f(x)$ fonksiyonunu tahmin etmektir (Wahba, 1990: 107). Dikkate alınacak esas problem, gözlenen verilerden (2.19) modeline uygun $f(x)$'in tahminidir. $f(x)$ fonksiyonunu tahmin etmek için kullanılan yöntemlerden biri doğrusal regresyondur. Doğrusal regresyonda $f(x)$ fonksiyonunu $\hat{f}(x) = a + bx$ şeklinde tahmin edilir. Sırasıyla a ve b sabit ve eğim kestiricileri, $\hat{f}(x) = \alpha + \beta x$ şeklindeki tüm fonksiyonlar için

$$RSS(f) = \sum_{i=1}^{n} \left[y - f(x) \right]^2 \qquad (2.20)$$

Model (2.20) ile verilen hata kareler toplamını minumum yaparak elde edilir. Burada $RSS(f)$ hata kareler toplamını ifade etmektedir. (2.19) modeline uygun gözlem verileri için

f fonksiyonu yaklaşık olarak doğrusalsa tahmin için doğrusal regresyon yaklaşımı uygun olmaktadır. f fonksiyonu doğrusal değilse veri uydurma (tahmin) başarılı olamaz. Başarılı bir veri uydurma süreci için , doğrusal regresyon modelindeki tahmin koşulları değişmeli ve değişen eğilimli f fonksiyonları üzerinde $RSS(f)$ ifadesinin diğer bir deyişle hata kareler toplamının minimizasyonu dikkate alınmalıdır (Aydın, 2005: 39).

Bu durumda, hata kareler toplamını minumum yapacak fonksiyonlar kümesine parametrik kısıtlar yüklemeksizin, eğrinin pürüzlüğü için bir ceza uygulanır. Splayn modelleri için kullanılan ceza (penalty) terimi (2.21) eşitliğinde gösterildiği gibidir ve bu eşitlikte PS pürüzsüzlük cezasını göstermektedir.

$$PS = \lambda \int_a^b [f''(x)]^2 dx \qquad (2.21)$$

Bu bağlamda splayn düzeltme yöntemi tekrar ele alındığında bu düzeltme yönteminin esası, (2.22) modeli ile belirtilen $S(f)$ "cezalı en küçük kareler kriterini" minimum yapmaktır.

$$S(f) = \sum_{i=1}^{n} [y - f(x)]^2 + \lambda \int_a^b [f''(x)]^2 dx \qquad (2.22)$$

Bu modeldeki ilk terim hata kareler toplamını gösterir ve bu ifade uyumdan yoksunluğu cezalandırır. Diğer bir deyişle, uyumun verilere yakınlığını ölçer. İkinci terim pürüzsüzlük cezasını gösterir ve bu pürüzsüzlüğe bir ceza yükler. Diğer bir deyişle, fonksiyondaki eğriliği cezalandırır. İkinci terimde yer alan λ ise düzeltme parametresini belirtir ve bu parametre $\int_a^b [f''(x)]^2 dx$ ile ölçümlenen eğrinin pürüzsüzlüğü ve $\sum_{i=1}^{n} [y - f(x)]^2$ ile ölçümlenen verilere uyumu dengeler (Aydın, 2005: 40).

λ arttıkça düşük varyanslı ancak yanlı, λ azaldıkça yüksek varyanslı ancak yansız uyumlar elde edilir. λ parametresi çok küçük ya da çok büyük değerler almadığında, ara değerler söz konusu olduğunda veriye uygulanan düzgünleştirmenin ölçüsüne yorumlanabilir

nitelikte bir etkisi olmaz. Bu problemin çözümü için λ parametresi serbestlik derecesi yaklaşımı ile dönüştürebilir (Keele, 2008: 65).

Splayn düzeltme için serbestlik derecesi, en küçük kareler ve lokal polinomiyal regresyonla aynı şekilde hesaplanmaktadır. Ancak splayn düzeltmede ceza terimi bazı karmaşıklıklara neden olmaktadır. Doğrusal regresyonda serbestlik derecesi, (2.23) eşitliğinde görülen H matrisinin izi ile hesaplanmakta ve $tr(H)$ olarak ifade edilmektedir.

$$H = X(X'X)^{-1}X' \tag{2.23}$$

Splayn modellerinde de serbestlik derecesi $tr(H)$ ile hesaplanmaktayken, cezalı splaynlarda serbestlik derecesini hesaplayabilmek için H matrisinin genelleştirilmesi gerekmektedir.

Cezalı splayn modellerinde serbestlik derecesini hesaplamak için ilk olarak, (2.22) eşitliğini matris formuna dönüştürmek gerekmektedir. (2.22) eşitliğinin ikinci terimi matris formunda doğrusal regresyona dönüştürülebilir. Bu dönüştürme işleminde (2.22) eşitliğindeki ceza terimi β'nın karesi olarak yazılabilmektedir (Ruppert, Wand ve Carroll, 2003: 75). Bu işlem ceza teriminin (2.24) ile görülen matris formuna dönüştürülmesini sağlamaktadır.

$$\int_a^b \left[f''(x)\right]^2 dx = \beta' D \beta \tag{2.24}$$

(2.24) eşitliğindeki D matrisi ise (2.25) eşitliğinde görüldüğü gibidir. D matrisinde k değeri düğüm sayısını ifade etmektedir.

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times \mathrm{k}} \\ 0_{\mathrm{k}\times 2} & I_{k\times k} \end{bmatrix} \tag{2.25}$$

Bu aşamadan sonra (2.22) eşitliği matris formuna dönüştürülerek (2.26) eşitliği yazılır.

$$S(f) = \left\| y - X\beta \right\|^2 + \lambda \beta' D \beta \tag{2.26}$$

Cezalı splaynlar için şapka matrisi (hat matrix) ya da düzgünleştirme matrisi (2.27) eşitliğine dönüşmektedir (Ruppert, Wand ve Carroll, 2003: 75). Bu matristeki p değeri baz fonksiyonundaki polinomun derecesini ifade etmektedir.

$$S_\lambda = X(X'X + \lambda^{2p} D)^{-1} X' \qquad (2.27)$$

Şapka matrisinin izi splayn modelindeki serbestlik derecesini ifade etmektedir ve $tr(S_\lambda)$ olarak ifade edilmektedir. $tr(S_\lambda)$ splayn modelinde kullanılan parametrelerin sayısına yaklaşık eşit olmaktadır.

Son olarak, (2.28) eşitliği kullanılarak splayn modelleri tahmin edilir. Burada ifade edilen cezalı splaynlar, splayn düzeltme (smoothing spline) olarak ifade edilmektedir.

$$\hat{y} = S_\lambda y \qquad (2.28)$$

Splayn düzeltme tekniği uygulanılırken düğüm sayıları genellikle sabit bırakılır. Optimal düğüm sayısını belirlemek için bazı algoritmalar ve değişik yazılımlar bulumaktadır. Ancak birçok yazılım kullanıcıya düğüm sayısını kontrol etme imkanı verir (Keele, 2008: 67).

Düğüm sayısının cezalı splayn (splayn düzeltme) modellerinde etkisin araştırmak üzere bir çok simülasyon çalışması yapılmıştır. Bu simülasyon çalışmalarında, serbestlik derecesi optimal düzeyde sabit tutulmuş ve değişik düğüm sayıları denenmiştir. Bu denemeler sonucunda düğüm sayılarının etkisinin çok az olduğu saptanmıştır. Kübik splaynlarda düğüm sayısının gereğinden fazla olması fazla kapsama problemine yol açarken, splayn düzeltmede λ yani düzeltme parametresininin düzgünleştirme miktarını kontrol etmesinden dolayı düğüm sayısının etkisi ortadan kalkmaktadır.

Sonuç olarak splayn düzeltme tekniği fazla kapsama (overfitting) problemini tamamen ortadan kaldırmasa bile, bu olasılığı en aza indirgemektedir. Ayrıca splayn düzeltme iki boyuttan daha fazla boyutta düzleştirme yapabilmektedir. Bu özelliği ile diğer yöntemlerden ayrılmaktadır ancak bu durumda da yorumlama zorlukları ortaya çıkmaktadır. Bu nedenden dolayı düzgünleştirme genellikle iki boyutta kullanılmaktadır.

Splayn düzeltme tekniğinde en önemli durum λ'nın seçimidir. λ'nın seçimi için bir çok yöntem bulunmaktadır. Bu seçim teknikleri diğer bölümde ele alınacaktır.

2.4.4. Splayn Modelleri İçin Çıkarım

Splayn modellerinde çıkarım lokal polinomiyal regresyon modellerinde çıkarım ile aynıdır. Güven bantlarını hesaplamak için lokal polinamiyal regresyon modellerinde olduğu gibi standart hataların hesaplanması gerekmektedir.

Güven bantlarını hesaplamak için, S matrisinin (şapka matrisi) önceki bölümlerde açıklanan şekillerde hesaplanması gerekmektedir. λ'nın sabit olduğu ve S matrisinin düzeltme matrisi olduğunu kabul ederek, $\hat{f}$ ($\hat{f} = Sy$)'nın kovaryans matrisi eşitlik (2.29)'da görüldüğü gibidir.

$$\mathrm{cov}(\hat{f}) = SS'\sigma^2 \qquad (2.29)$$

(2.29) eşitliğinde σ^2 'nin tahminine ihtiyaç vardır. f'in tahmini yansız ise σ^2 'nin yansız tahmicisi (2.30) eşitliğinde görüldüğü gibidir. Bu eşitlikte RSS hata kareler toplamını ve df_{res} değeri ise hataların serbestiklik derecesini ifade etmektedir . df_{res} değeri (2.31) eşitliğinde gösterildiği gibidir.

$$\hat{\sigma}^2 = RSS / df_{res} \qquad (2.30)$$

$$df_{res} = n - 2tr\left[S + tr(SS')\right] \qquad (2.31)$$

(2.31)[4] eşitliğinde görülen n yani gözlem sayısı, küçük bir değer ise normal dağılım kritik değeri olan 2, k serbestlik dereceli (S matrisinin izi diğer bir değişle modeldeki parametre sayısı) t- dağılımı kritik değeri ile değiştirilmelidir. Daha önemli bir nokta ise f'in tahminini yansız olarak varsaymaktır. Bu varsayımın geçerliliğini test etmek zordur. Ancak, parametrik olmayan regresyondan bilindiği üzere f'in tahmininde belirli bir seviyede yanlılık söz konusudur. Parametrik olmayan regresyon yanlılık ve varyans dengesini sağladığından, varyansı düşürücek belli bir miktar yanlılık kabul edilebilmektedir (Hastie,

[4] tr ifadesi matrisin izini ifade etmektedir.

Tibshirani, 1999: 42). f'in tahmininde ve güven bantlarının oluşturulmasında yanlılık miktarının dikkate alınması çok iyi sonuçlar sağlar. Splayn düzeltme yönteminde düzeltilmiş güven bantları tahmin edilebilmektedir. Bu tahmin için bayesgil düzgünleştirme ve karma model yapıları kullanılmaktadır (Hastie, Tibshirani, 1999: 63).

Splayn modellerinin geçerliliğine karar vermek için F testi yaklaşımı kullanılmaktadır. F test istatistiği (2.32) eşitliğindeki gibi hesaplanmaktadır.

$$F = \frac{\left(RSS_1 - RSS_2\right)/(df_{res2} - df_{res1})}{RSS_2/\left(n - df_{res2}\right)} \sim F_{dfres2-dfres1,n-dfres2} \qquad (2.32)$$

(2.32) eşitliğinde, RSS_1 değeri kısıtlı modelin (doğrusal model) hata kareler toplamını gösterirken RSS_2 değeri splayn modelinin hata kareler toplamını göstermektedir. Aynı şekilde, df_{res1} değeri kısıtlı modelin serbestlik derecesini, df_{res2} değeri splayn modelinin serbestlik derecesini ifade etmektedir.

F testi yaklaşımı, doğrusal olamayan ilişkilerin ortaya çıkartılması ve üstel dönüştürmenin yeterli olup olmadığına karar vermek için sıkça kullanılmaktadır (Keele, 2008: 78). F testi sonucunda splayn model ile basit model arasında farklılık olmadığı sonucuna ulaşılırsa cimrilik prensibi yaklaşımından dolayı basit olan modelin tercih edilmesi uygundur.

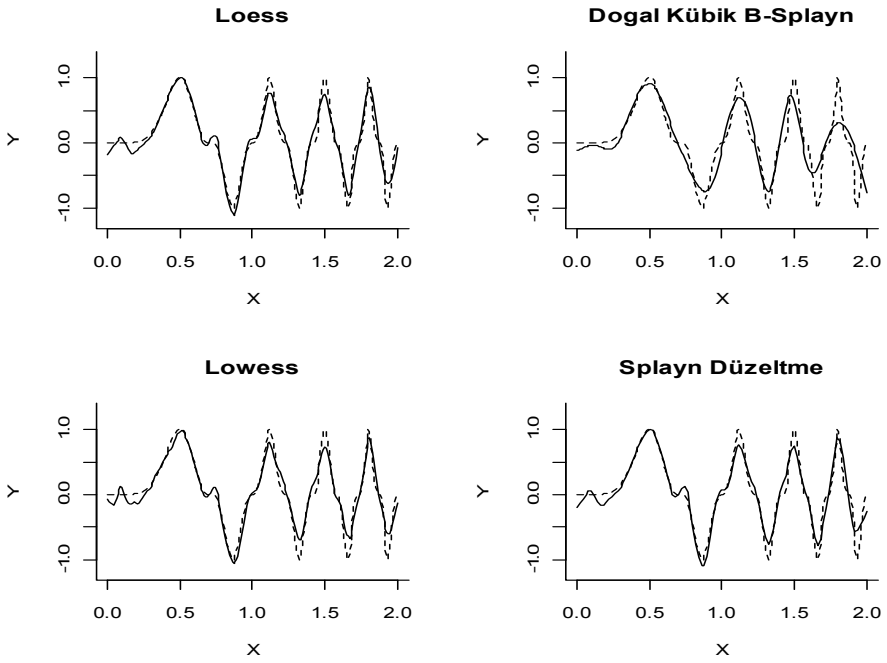
Cimrilik prensibi (rule of parsimony), verinin en iyi şekilde yansıtılabilmesi için en az sayıda parametre yani değişken kullanılmasını önerir (Box ve Jenkins, 1970: 17). Modeldeki parametre sayısı arttıkça yanlılık azalır ancak varyans büyür. Fakat bir modelde çıkarımlar açısından varyansın büyük olması sorunlara yol açmaktadır. Cimrilik prensibi egilim ve varyans arasındaki uygun dengeyi başarmaya çalışmaktadır.

2.5. Düzeltme Tekniklerinin Karşılaştırılması

Bir çok araştırmada hangi düzeltme tekniğinin tercih edileceğine karar vermek gerekmektedir. Kuşkusuzdur ki splayn düzeltme diğer tekniklere nazaran bir çok üstünlüğe sahiptir. Değişkenler arasında yüksek derecede ilişki olduğu durumda, düzeltme tekniklerinin performanslarının karşılaştırılması amacıyla Luke Keele tarafından 2008 yılında bir simülasyon çalışması yapılmıştır. Bu çalışmada oluşturulan fonksiyon eşitlik (2.33)'de gösterildiği gibidir. Bu fonksiyona dört teknik uygulanmış şekil (2.2) oluşturulmuştur[5].

[5] Bu fonksiyon için oluşturulan R kodları EK-A'da verilmiştir.

$$y = \sin^3(2\pi x^2) + \varepsilon \tag{2.33}$$



Şekil 2. 3: Düzeltme Tekniklerinin Karşılaştırılması

Şekil (2.3)'deki bütün grafiklerde noktalı çizgiler doğru fonksiyonel formu, kesiksiz çizgiler ise parametrik olmayan tahminleri göstermektedir.

Sol üst kısımda bulunan parametrik olmayan model, Loess ile tahmin edilmiştir. Aralık (span) değeri denemeler sonucu 0.1 olarak alınmıştır. Loess grafiği incelendiğinde en yüksek ve en düşük değerlerde gerçek durumun altında (undersmooth) tahmin yapıldığı ve bunun dışında yapılan tahminin iyi olduğu görülmektedir. Sağ üst kısımda bulunan parametrik olmayan model, düğüm sayısı AIC Kriterine göre seçilmiş doğal kübik B- splaynla tahmin edilmiştir. Doğal kübik B-splayn grafiği incelendiğinde, genellikle olduğundan daha düşük tahminler yapılmış ve özellikle x değişkenin yüksek değerleri için zayıf bir tahmin süreci söz konusu olduğu görülmektedir. Sol alt kısımda bulunan parametrik olmayan model Lowess ile tahmin edilmiştir. Lowess grafiği incelendiğinde, bu tahminin Loess tahminiyle çok küçük farklılıklar haricinde aynı olduğu görülmektedir. Lowess tahmininde kullanılan ağırlıkların tahminde çok küçük farklılıklar yarattığı tekrar görülmektedir. Son olarak sağ alt kısımda parametrik olmayan model, splayn düzeltme ile tahmin edilmiştir. Serbestik derecesi denemeler sonucu 31 olarak alınmıştır. Grafik incelendiğinde, splayn düzeltmenin doğrusal

olmayan ilişkiyi neredeyse birebir tahmin ettiği bir diğer ifade ile iyi bir uyum elde edildiği görülmektedir.

Sonuç olarak, splayn düzeltmenin diğer tekniklere göre daha iyi tahminler yarattığı görülmektedir. Bu nedenden dolayı splayn düzeltme, doğrusal olmayan ilişkilerin saptanmasında belirsizliğin ortadan kaldırmasını sağlarlar.

3. OTOMATİK DÜZELTME TEKNİKLERİ

İkinci bölümde anlatıldığı üzere hangi düzeltme tekniğinin kullanılacağına karar verildikten sonra düzeltme parametresinin değerinin belirlenmesi sorunu ortaya çıkmaktadır. Lokal polinomiyal regresyon modellerinde aralık (span) değeri, splaynlarda düğüm sayısı ve serbestlik derecesinin belirlenmesi gerekmektedir. Düzeltme parametresinin değeri, varyans ve sapma arasındaki dengeyi sağlamak açısından çok büyük önem taşımaktadır. Düzeltme parametresinin değerinin en basit seçimi deneme- yanılma yöntemini kullanmaktır. Bu seçim tekniği görsel olduğundan kişiden kişiye değişme riski taşımaktadır. Diğer bir yaklaşım ise, verilere dayalı olarak elde edilen düzeltme parametresinin seçimidir. Düzeltme parametresinin bu yolla seçimi için otomatik düzeltme teknikleri geliştirilmiştir. Otomatik düzeltme tekniğinde araştırmacının, tahmin aşamasında hiçbir karar vermesi gerekmemektedir.

Bu bölümde local polinomiyal regresyon modelleri ve splaynlar için otomatik düzeltme tekniklerinden en çok kullanılanları (çapraz geçerlilik, genelliştirilmiş çapraz geçerlilik) anlatılmış ve karşılaştırmalar yapılmıştır.

3.1. Çapraz Geçerlilik ile Aralık (Span) Değerinin Bulunması

Lokal Polinomiyal regresyon modellerinde yanlılık ve varyans değerini minimum yapan aralık parametresinin seçimi çok önemlidir. Bu parametre, bağımsız bir değişken olduğundan, farklı aralık değerleri ile oluşturulan modeller söz konusu olmaktadır. Hangi aralık değerinin optimum olduğuna karar verebilmek için söz konusu bu modellerin karşılaştırılması gerekmektedir. Bu durum ise model seçim kriterlerinin kullanımını gerektirmektedir.

Çapraz geçerlilik yöntemi (CV)[6] en yaygın kullanılan model seçim kriterlerindendir. Çapraz geçerliliğin esas düşüncesi, veri noktalarından herhangi birini atmak ve kayıp veri noktaları altında geri kalan veriler tarafından en iyi kestirilen λ parametresini seçmektir (Wahba ve Wold, 1975: 1-17).

Çapraz geçerlilik yönteminde ilk olarak veri seti rassal olarak iki parçaya ayrılır. İlk veri seti eğitim verisi olarak adlandırılır ve bu veri setinden belirli kriterlere göre (AIC, adımsal seçim vs.) bir model belirlenir. Belirlenen bu model, yeni örneklemi ne kadar iyi tahmin ettiğini kestirmek adına ikinci veri setine uygulanır. Çapraz geçerlilik yöntemi için veri setini bölmenin bir çok yöntemi vardır. Ancak bu yöntemler genellikle çok büyük örnek

[6] CV: Cross Validation.

büyüklüğüne ihtiyaç duyarlar. Veri setini bölümlemenin en çok kullanılan ve her örnek büyüklüğüne uygulana bilen yöntemlerinden biri "birini dışarıda bırak" (leave-one-out) çapraz geçerlilik yöntemidir. Bu yöntemde, bir gözlem tesadüfi olarak seçilir ve veri setinden atılır. Kesilmiş bu veri seti için bir model tahmin edilir ve uyum ölçüsü hesaplanır. Bu işlem herbir verinin veri setinden çıkarılması gerçekleşene kadar devam eder. Hesaplanan çapraz geçerlilik skoru, ortalama uyum ölçüsüdür ve farklı modellerin karşılaştırılması için kullanılır.

Birini dışarıda bırak çapraz geçerlilik yöntemi, lokal polinomiyal regresyon modellerinde aralık parametresini seçmek için sıkça kullanılmaktadır (Keele, 2008: 86). Lokal poliomiyal regresyon modellerinde s (span-aralık) parametresini seçmek için eşitlik (3.1) kullanılmaktadır. Eşitlik (3.1)'de $\hat{f}_s(x_{-i})$ değeri i. gözlem atıldıktan sonra kalan $(n-1)$ gözlem için tahmin edilen fonksiyonu ifade etmektedir. s parametresinin çapraz geçerlilik tahmini (3.1) modelini minimum yapan değerdir.

$$CV(s) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_s(x_{-i}) \right]^2 \qquad (3.1)$$

(3.1) eşitliğinde s parametresinin her bir değeri için n tane parametrik olmayan model tahmin edilmelidir. Bu durumda, büyük veri setleri için birini dışarıda bırak çapraz geçerlilik yöntemi, hesaplamaların karmaşıklığından dolayı kullanışsız duruma gelmektedir. Bu problemden kurtulabilmek için 1975 yılında Craven ve Wahba tarafından, çapraz geçerlilik yöntemine alternatif olarak genelleştirilmiş çapraz geçerlilik yöntemi (GCV)[7] önerilmiştir. Genelleştirilmiş çapraz geçerlilik değeri eşitlik (3.2)'deki gibi bulunmaktadır.

$$GCV(s) = \frac{\sum_{i=1}^{n}\left[y_i - \hat{f}_s(x_i) \right]^2}{(n - df)^2} \qquad (3.2)$$

Eşitlik (3.2)'deki df değeri parametrik olmayan modelin serbestlik derecesini ifade etmektedir. GCV değeri, değişik s parametreleri için hesaplanır ve en düşük GCV değerinini sağlayan s parametresi seçilir. Genelleştirilmiş çapraz geçerlilik yöntemi, çapraz geçerlilik yönteminin serbestlik derecesi ile düzeltilmiş biçimidir (Keele, 2008: 87).

[7] GCV : Generalized Cross Validation.

Lokal polinomiyal regresyon modelinde veri seti çok büyük ise, genelleştirilmiş çapraz geçerlilik yöntemi normalin altında tahmin (underestimate) yapmaktadır. Ancak yapılan çalışmalarda buna rağmen s parametresinin seçimi için en iyi yöntemin GCV olduğu ve otomatik düzeltme tekniklerinin bu modellerde fazla kapsama problemine yol açtığı belirlenmiştir. Otomatik düzeltme teknileri splayn modellerinde çok iyi sonuçlar verdiğinden, lokal polinomiyal regresyon modellerinden çok splayn modellerinde tercih edilmektedirler (Loader, 1999: 120-122).

3.2. Splaynlar ve Otomatik Düzeltme

Otomatik düzeltme teknikleri bütün splayn modellerine uygulanabilmektedir. Ancak splayn düzeltmede çok daha güvenilir sonuçlara ulaşılmaktadır. Lokal polinomiyal regresyon modellerinde ve standart splayn modellerinde aralık ve düğüm sayısı modelde bağımsız değişken olarak bulunmaktadır. Splayn düzeltmede düzeltmenin miktarı modeldeki bir parametredir. Düzeltme parametresi olan λ'nın optimum değerinin ne olacağını belirlemek için standart model seçim kriterleri kullanılabilmektedir. λ'nın optimum değeri gerçek hata kareler ortalamasını (MSE)[8] minimum yapan değerdir ve (3.3) eşitliğindeki gibi tahmin edilebilir. (3.3) eşitliğinde $f_\lambda(x_i)$ belirli bir λ parametresi için splayn düzeltmeyi ifade etmektedir.

$$MSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - f_\lambda(x_i)\right]^2 \qquad (3.3)$$

Eşitlik (3.3)'den görüldüğü üzere, gerçek regresyon doğrusu olan $f(x_i)$ bilinmediğinde λ parametresi tahmin edilemeyecektir. Bu durumda λ parametresi çapraz geçerlilik yöntemi ile tahmin edilebilmektedir.

3.3. Splayn Düzeltme ve Çapraz Geçerlilik

Çapraz geçerlilik yöntemi ile λ parametresinin seçimi, çapraz geçerlilikle aralık değerinin seçimi ile aynı süreci izlemektedir.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_\lambda(x_{-i}) \right]^2 \qquad (3.4)$$

[8] MSE : Mean Square Error.

λ parametresinin çapraz geçerlilik tahmini (3.4) eşitliğini minumum yapan değerdir. λ parametresinin genelleştirilmiş çapraz geçerlilik tahmini ise (3.5) eşitliğini minumum yapan değerdir.

$$GCV(s) = \frac{\sum_{i=1}^{n}\left[y_i - \hat{f}_{\lambda}(x_i) \right]^2}{\left[1 - n^{-1} tr(S)\right]^2} \tag{3.5}$$

(3.5) eşitliğindeki S düzeltme matrisini, $\hat{f}_{\lambda}(x_i)$ ise tahmin edilen splayn düzeltmeyi ifade etmektedir. Otomatik düzeltme tekniğinde, genelleştirilmiş çapraz geçerlilik sıradan çapraz geçerliğe göre üstünlük sağlamaktadır. Lokal polinomiyal regresyonda, genelleştirilmiş çapraz geçerlilik yöntemi kullanıldığında fazla kapsama probleminin ortaya çıktığı görülmekteydi. Splayn düzeltmede splayn düzeltme cezası bu problemin önüne geçmektedir.

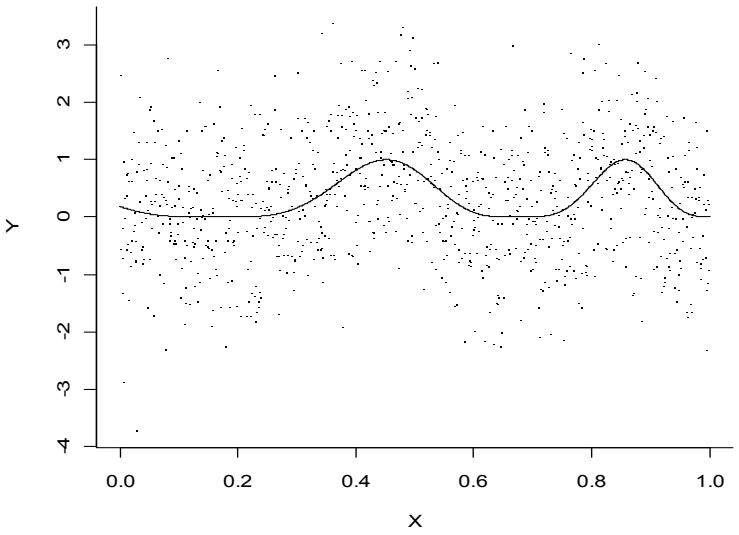
3.4. Otomatik Düzeltme Tekniği Simülasyon Çalışması

Araştırıcının deneme-yanılma yöntemi ile karar verdiği düzeltme parametresinin ve otomatik olarak seçilen düzelme parametresinin performanslarını karşılştırmak amacıyla Luke Keele tarafından 2008 yılında bir simülasyon çalışması yapılmıştır. Bu çalışmada eşitlik (3.6)'da gösterilen bir fonksiyon belirlenmiştir[9].

$$y = \cos^4(4e^{x^3}) \tag{3.6}$$

Bu fonksiyonda hata terimlerinin normal dağıldığı ve sabit varyanslı olduğu bilinmektedir. (3.6) eşitliğindeki x değerlerine 0 ile 1 arasında değerler verilerek 1000 birimlik bir örnek oluşturulmuştur. x ve y değişkenleri arasındaki gerçek ilişki şekil 3.1'de gösterilmiştir. Şekil 3.1 incelendiğinde ilişkinin yüksek derecede doğrusal olmadığı görülmektedir.

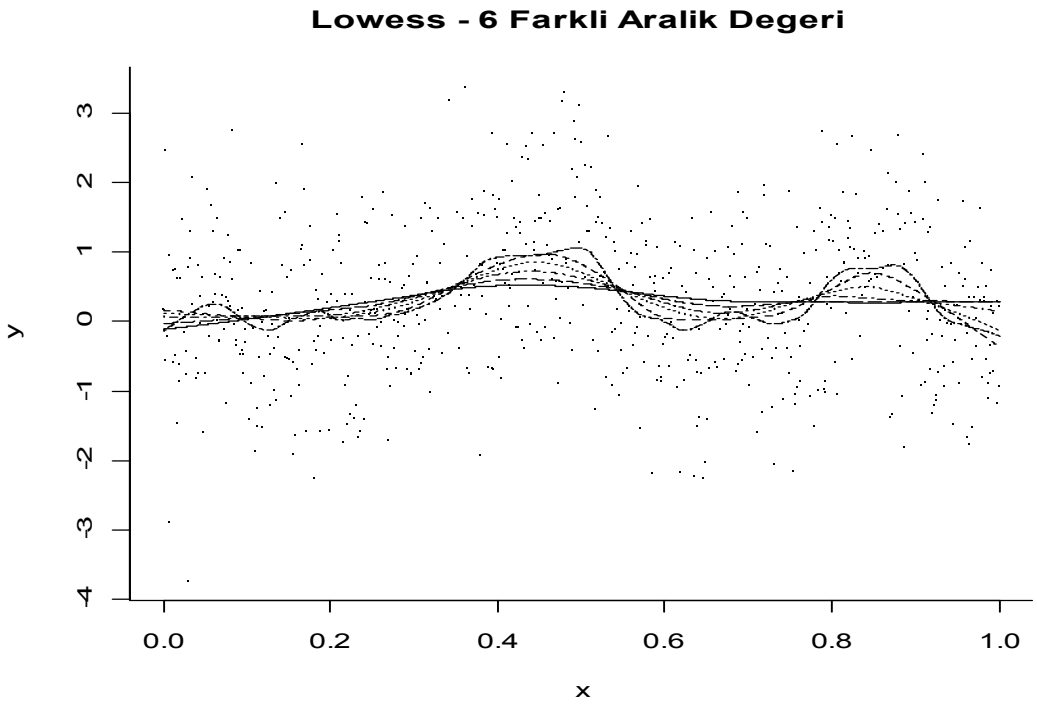
[9] Bu fonksiyon için oluşturulan R kodları EK-B'de verilmiştir.



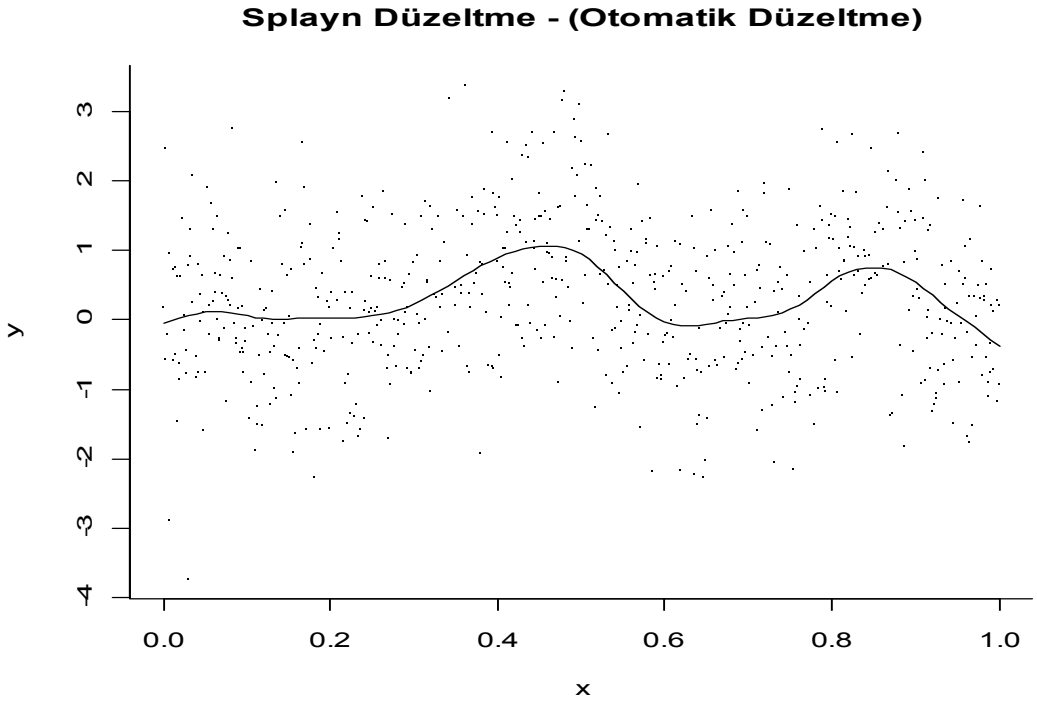
Şekil 3. 1: (3.6) Fonksiyonunda Belirtilen X ve Y Değişkenleri Arasındaki Gerçek İlişki

Çalışmada ilk olarak, 0.1 ile 0.6 arasında altı ayrı aralık (span) değeri seçilmiş ve lowess düzeltme tekniği uygulanmış ve Şekil 3.2'de gösterilmiştir. İkinci aşamada ise otomatik düzeltme tekniği ile splayn düzeltme tekniği uygulanmış ve Şekil 3.3'de gösterilmiştir. Şekil 3.2'deki lowess tahmini incelendiğinde, bu tahminlerin çoğu gerçek fonksiyonel formu yakalamaktadır. Ancak, tahminlerden biri çok pürüzlü iken bir diğeri çok düzdür. Geriye kalan tahminler ise araştırıcıya göre tercih edilebilir düzeydedir. Şekil 3.3 incelendiğinde, splayn düzeltme tahmini gerçek ilişkiyi çok az bir farkla tahmin edebilmiştir.

Sonuç olarak, aralık değeri deneme-yanılma yöntemi ile seçildiğinde gerçek ilişkiye çok yakın olan tahmin yüksek derecede pürüzlü olduğu için tercih edilmeyebilir ve aralık değeri arttırılabilirdi. Ancak otomatik seçim, gerçek fonksiyonel formu denemeler yapmadan ve kesin bir şekilde yakalayabilmiştir.



Şekil 3. 2: Lowess Tahmini (6 Farklı Aralık Değeri İle)



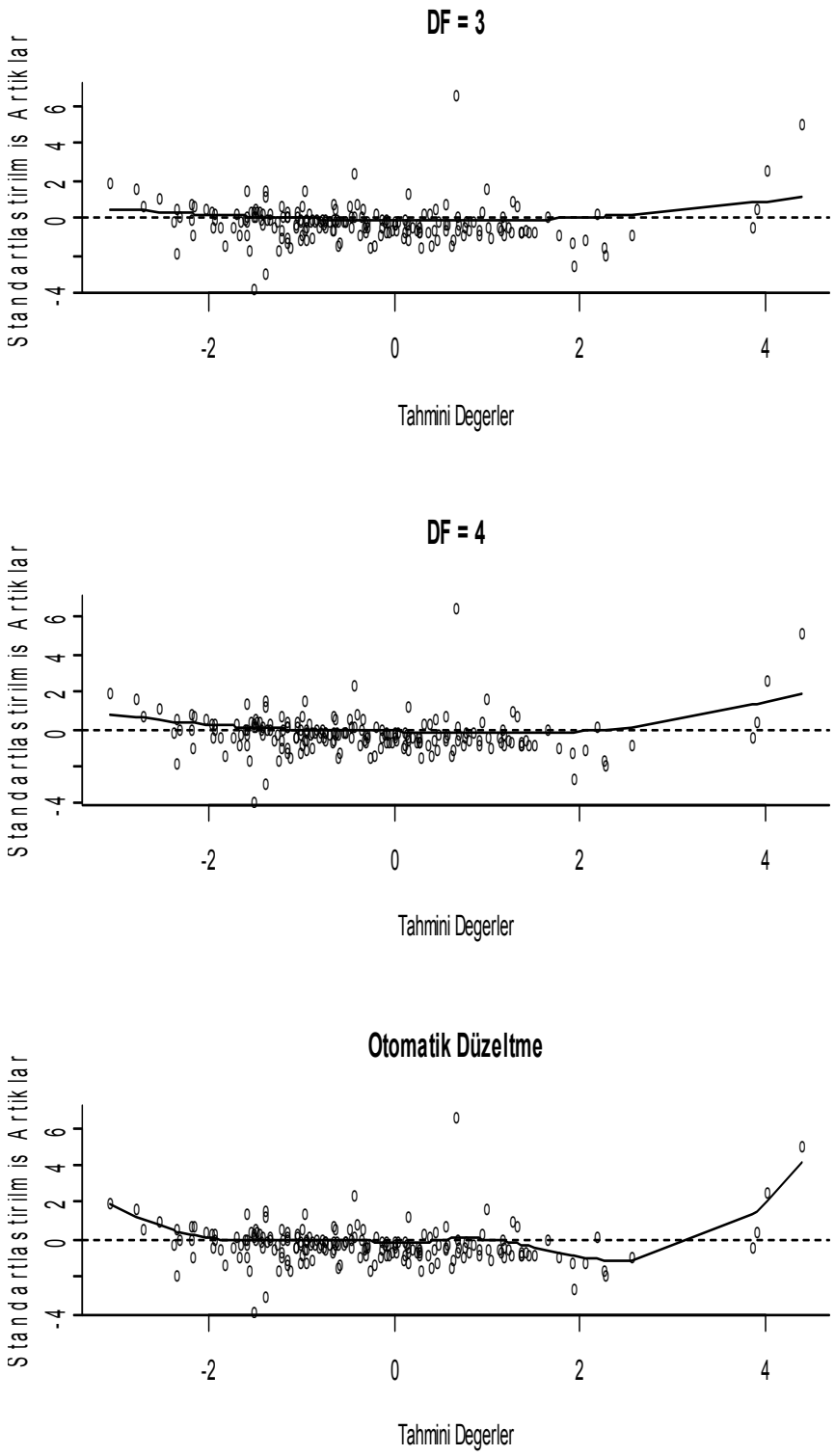
Şekil 3. 3: Otomatik Düzeltme ile Splayn Düzeltme Tahmini

3.5. Otomatik Düzeltme Tekniğinin Artık Grafiklerinde Kullanımı

Doğrusal regresyon modellerinde, tahmini değerlerlere $(\hat{y})$ karşı model artıklarının grafikleri çizilir. Böyle bir grafik, değişen varyans, doğru olmayan fonksiyonel form ve unutulmuş değişkenlerin olup olmadığının kontrolunu sağlamaktadır. Grafik çizilirken genellikle, artıkların kendi tahmininin standart sapmasına bölünmesiyle elde edilen standartlaştırılmış artıklar kullanılır. Standartlaştırılmış artıkların kullanılması sabit varyans varsayımının geçerli olup olmadığını net bir şekilde görmeyi sağlamaktadır. Böyle bir grafiğe düzeltici (smoother) eklendiğinde yukarıda sözü geçen problemlerin varlığı daha net görülebilir. Artık grafiğine düzeltici eklenirken, otomatik düzeltme güvenilir sonuçlara ulaşılmasını sağlamaktadır.

Otomatik düzeltme tekniğinin artık grafiklerinde kullanımına ilişkin bir çalışma yapılmıştır. Çalışmaya ilişkin veriler Li ve Revenue tarafından 2006 yılında yayımlanan bir makaleden alınmıştır (Keele, 2008: 104). Çalışmada öncelikle doğrusal regresyon modeli tahmin edilmiş ve Şekil 3.4'de görülen artık grafikleri elde edilmiştir. Şekil 3.4'deki birinci grafikte, üç serbestlik dereceli splayn düzeltme modeli tahmin edilmiştir. Birinci grafikte görüldüğü üzere, 0 ekseninden uzaklaşma ihmal edilecek kadar az olduğundan modelin doğru olduğuna karar verilir. İkinci grafikte dört serbestlik dereceli splayn düzeltme modeli tahmin edilmiş ve bir trendin olma ihtimali görülmüştür. Bu aşamadan sonra trendi ortadan kaldırmak için bir serbestlik derecesi eklenebilir. Ancak trendi ortadan kaldırıp kaldıramayacağı kesin değildir. Böyle bir durumla karşılaşıldığında otomatik düzeltme trendin varlığına ilişkin bilgiyi kesin olarak göstermektedir. Bu amaçla en son grafikte görülen, otomatik düzeltme yöntemi olan genelleştirilmiş çapraz geçerlilik ile splayn düzeltme modeli tahmin edilmiştir. Grafik incelendiğinde, bir trendin varlığı açıkça görülmektedir. Bu trendin varlığı modelin doğru tanımlanmadığını göstermektedir.

Sonuç olarak, otomatik düzeltme veriye dayalı olarak düzeltme parametresinin seçimini gerçekleştirir. Otomatik Düzeltme görsel metodların ve elle seçim kriterlerinin bir çok olumsuzluğunu ortadan kaldırmıştır. λ parametresinin genelleştirilmiş çapraz geçerlilik ile seçimi aşamasında veri setinin büyüklüğü önem kazanmaktadır. Özellikle çok küçük veri setlerinde (50'den küçük) çok pürüzlü veya yüksek derecede doğrusal olmayan ilişkiler tahmin edilmektedir. Bu durumda görsel metodlar kullanılmalıdır ve düzeltme parametresinin seçimi elle (manually) yapılmalıdır. Veri seti büyüdükçe GCV gerçek değerine yaklaşmakta ve güvenilir sonuçlara ulaşılmaktadır (Hardle, Hall ve Marron: 1998, 145).



Şekil 3. 4: Standartlaştırılmış Artık Grafikleri

4. TOPLAMSAL VE SEMİPARAMETRİK REGRESYON MODELLERİ

Şu ana kadarki açıklamalarda parametrik olmayan regresyon modelleri incelenmiştir. Bu modeller sadece x ve y değişkenleri arasındaki doğrusal olmayan ilişkiyi incelemektedir. Bu durum uygulamada karşılaşılan çok sayıda problem için yetersiz kalmaktadır. Uygulamada genellikle bağımlı değişken, bir çok bağımsız değişkenden eş zamanlı olarak etkilenmektedir. Bu durumda çoklu regresyon analizine ihtiyaç duyulmaktadır. Parametrik olmayan regresyon modellerinde birden fazla değişkenin yer aldığı durum eşitlik (4.1)'de gösterildiği gibidir. Eşitlik (4.1)'de k sayıda bağımsız değişken yer almaktadır.

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon \qquad (4.1)$$

Parametrik olmayan çoklu regresyon modellerinin tahmini, çok boyutluluğun yarattığı sıkıntı (curse of dimensionality) nedeniyle zorlaşmaktadır (Hastie ve Tibshirani, 1999: 83). Bu sorunun çözümü için toplamsal modeller kulllanılabilmektedir. Bu durumda da bağımlı değişken bazı açıklayıcı değişkenlerle doğrusal, fakat diğer bazı açıklayıcı değişkenlerle doğrusal olmayan ilişki içerisinde bulunabilmektedir. Bu sorunun çözümü için ise bu çalışmanın konusu olan semiparametrik regresyon modelleri kullanılmaktadır. Bu bölümde, semiparametrik regresyon modellerinin temelini oluşturan toplamsal modeller açıklanmış ve ardından bu çalışmanın konusunu olan semiparametrik regresyon modelleri ayrıntılarıyla ele alınmıştır.

4.1. Toplamsal Modeller

Parametrik olmayan regresyon analizinde k sayıda bağımsız değişkenin yer aldığı durum eşitlik (4.1)'de gösterilmişti. Eşitlik (4.1)'in çözülmesi, boyutluluğun yarattığı sıkıntı ve yorumlama sıkıntısı nedeniyle zorlaşmaktadır. Bu durumda, eğer bağımsız değişkenler arasındaki ilişki toplamsal ya da toplamsal kabul edilebilirse toplamsal modeller kullanılabilmektedir. Toplamsal modelller eşitlik (4.2)'de gösterildiği gibi ifade edilmektedir.

$$y_i = \alpha + f_1(x_1) + \ldots + f_k(x_k) + \varepsilon \qquad (4.2)$$

Eşitlik (4.2)'de görüldüğü gibi, düzeltme toplamsal modellerde yeralan herbir bağımsız değişken için ayrı ayrı yapılmaktadır. Genel bir ifadeyle, toplamsallık varsayımına

sahip parametrik olmayan modeller toplamsal modeller olarak ifade edilirler (Hastie ve Tibshirani, 1999: 89). Toplamsallık varsayımı araştırmacıya yorumlama aşamasında çok büyük kolaylıklar sağlamaktadır. Eşitlik (4.3)'te üç bağımsız değişkene sahip bir toplamsal model gösterilmektedir.

$$y_i = \alpha + f_1(x_1) + f_2(x_2) + f_3(x_3) + \varepsilon \qquad (4.3)$$

Eşitlik (4.3)'de, $\hat{f}_1$ tahmin edilirken üç bağımsız değişken arasındaki kovaryans dikkate alınır. Böylelikle x_1'in y_i üzerindeki etkisi x_2 ve x_3 değişkenlerinin etkisini sabit tutarak yorumlanabilir. Toplamsallık, parametrik olmayan regresyon modellerindeki esnekliği, parametrik regresyon modellerinin ise yorumlama aşamasını sağlamaktadır (Keele, 2008: 112).

Toplamsal modeller çoklu regresyon modellerine göre daha kısıtsız modellerdir ancak parametrik olmayan regresyon modellerinden daha kısıtlı modellerdir. Toplamsal modeller bağımsız değişkenler arasındaki ilişkiyi dikkate almazlar, bu durum toplamsal modeli parametrik olmayan modellerden daha kısıtlı hale getirmektedir. Toplamsal modellerin genel gösteriminde bağımsız değişkenler arasındaki ilişki dikkate alınmaz ancak bağımsız değişkenler arasında ilişki olabilir. Bağımsız değişkenler arasında ilişki yoksa, her bir bağımsız değişkenin bağımlı değişkene karşı düzeltilmesiyle elde edilen fonksiyonların toplamı toplamsal modelin tahminini oluşturur. Bağımsız değişkenler arasında ilişki olduğu durumda ise modelin tahmininde bu ilişkinin de dikkate alınması gerekir. Bu durumda kısmi regresyonlar yardımıyla hesaplanan kısmi artıklar kullanılır (Çağlayan, 2002: 63).

Toplamsal modeller ile parametrik modeller F testi yaklaşımı ile ya da olabilirlik oran testi yardımıyla karşılaştırılabilirler.

Doğrusal regresyon analizinde kategorik bağımlı değişkenlerin olması durumunda genelleştirilmiş doğrusal modeller (GLM)[10] kullanılmaktadır. Genelleştirilmiş doğrusal modeller stokastik bileşen, sistematik bileşen ve bağlantı fonksiyonu olmak üzere üç bileşenden oluşmaktadır. İlk olarak, bağımlı değişkenin örneklem dağılımını ifade eden stokastik bileşen (binom, poisson, negatif binom, vb.) seçilir. Stokastik bileşen belirlendikten sonra ise sistematik bileşen seçilmektedir.

[10] GLM: Generalized Linear Models

Sistematik bileşen bağımsız değişkenleri içermektedir. Model eşitlik (4.4) ile belirtilen şekilde kurulur.

$$\eta = X\beta \tag{4.4}$$

Son olarak ise $g(.)$ bağlantı fonksiyonu belirlenir. Bağlantı fonksiyonları genellikle logit, probit, log ve benzeri fonkisyonlardır ve sistematik bileşen ile stokastik bileşen arasında bağlantı sağlar. Ayrıca bağlantı fonksiyonu diferansiyellenebilen monoton bir fonksiyondur (Akay, 2007: 9).

Genelleştirilmiş doğrusal model tahminleri, bağlantı fonksiyonu uygulanana kadar sistematik bileşende doğrusaldır. Probit modeller düşünülecek olursa, probit modelden yapılan tahminler birikimli normal dağılım fonksiyonu uygulana kadar doğrusaldırlar.

Genelleştirilmiş doğrusal modellerde, gerçek model değişkenlerde doğrusal değil ise bu doğrusal olmama durumu, modelin sistematik bileşeninde modellenmelidir. Çünkü bağlantı fonksiyonu bu tip bir doğrusal olmama durumu için işe yaramamaktadır. Örneğin, sistematik bileşenin gerçek yapısı eşitlik (4.5)'deki gibi ise,

$$g(\eta_i) = \alpha + \beta_1 X_{i1} + \beta_2 X^2{}_{2i} \tag{4.5}$$

Bu modelin sağ tarafına, hangi bağlantı fonksiyonu seçildiğini dikkate almaksızın karesel terim eklenmelidir. Böyle bir modele karesel terim eklenmediğinde ise modelde tanımlama hatası yapılmış olur. Modelde tanımlama hatası yapmamak için olası doğrusal olmama durumu modelin sistematik bileşenine eklenmelidir. Genelleştirilmiş doğrusal modellerde bu tür bir doğrusal olmama durumunu modellemek için parametrik olmayan yöntemler kullanılabilmektedir. Bu durum genelleştirilmiş toplamsal modellerin (GAM)[11], oluşmasına olanak sağlamıştır.

Eşitlik (4.6)'da gösterilen genelleştirilmiş doğrusal modelde, $g(\eta_i)$ araştırmacı tarafından seçilen bağlantı fonksiyonunu ifade etmektedir.

$$g(\eta_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{4.6}$$

Eşitlik (4.6) genelleştirilmiş toplamsal model olarak tekrar yazıldığında eşitlik (4.7) elde edilmektedir.

[11] GAM: Generalized Additive Models.

$$g(\eta_i) = \alpha + f_1(X_{1i}) + f_2(X_{2i}) \qquad (4.7)$$

Eşitlik (4.7)'de doğrusal bağımsız değişkenler yerine veriden tahmin edilmiş pürüzsüz fonksiyonlar kullanılmıştır. Genelleştirilmiş toplamsal modellerdeki toplamsallık varsayımı, parametrik olmayan tahmine yorum esnekliği sağlamaktadır.

Genelleştirilmiş doğrusal modellerin tahmininde olasılık fonksiyonu maksimize edilir. Bu maksimizasyon işlemi Newton-Raphson algoritması ile ya da tekrarlı yeniden ağırlıklandırılmış en küçük kareler yöntemi (IRLS)[12] kullanılarak gerçekleştirilebilir. IRLS algoritması genelleştirilmiş toplamsal modellerin tahmininde de sıkça kullanılmaktadır. Bu durumda düzgünleştirilmiş (düzeltme yöntemi uygulanmış) bileşenlerin tahmini için ileride anlatılacak olan backfitting algoritması da her iki model için geçerli olmaktadır.

Genelleştirilmiş toplamsal modellerin tahmini, tekrarlı yeniden ağırlıklandırılmış en küçük kareler yaklaşımı ile toplamsal modellerin tahmininden daha zor olmamaktadır. Sadece bilgisayar çözümlemelerinde biraz daha fazla zaman almaktadır. Örneğin, toplamsal model için backfitting algoritması bir kere çalıştırılırken genelleştirilmiş toplamsal modellerde, yeniden ağırlıklandırılmış en küçük kareler algoritmasının her tekrarı için bir toplamsal model tahmin edilir. Özellikle büyük veri setleri için genelleştirilmiş toplamsal modeller tahmin edilmek istendiğinde bilgisayar programlarının daha yavaş sonuca ulaştığı görülmektedir.

Genelleştirilmiş toplamsal modeller ile parametrik modeller, toplamsal modellerden farklı olarak sadece olabilirlik oran testi yardımıyla karşılaştırılabilirler. Olabilirlik oran test istatistiği eşitlik (4.8)'de gösterilmiştir.

$$LR = -2(Logolabilirlik_0 - Logolabilirlik_1) \qquad (4.8)$$

Eşitlik (4.8)'de gösterilen $Logolabilirlik_0$ değeri daha kısıtlı olan parametrik modelden tahmin edilen olabilirlik oranının logaritmasını, $Logolabilirlik_1$ değeri ise genelleştirilmiş toplamsal modelden tahmin edilen olabilirlik oranının logaritmasını ifade etmektedir. Bu koşullar altında olabilirlik oran test istatistiği χ^2 dağılımı ile karşılaştırılmaktadır. Modellerin serbestlik derecesi ise (4.9) [13] eşitliğinden elde edilmektedir.

$$df_{err} = n - tr(2R - R'\sum R\sum^{-1}) \qquad (4.9)$$

[12] IRLS: Iteratively Reweighted Least Squares.

[13] Eşitlik (4.9)'daki R matrisi semiparametrik regresyon modellerinde ayrıntılı olarak ele alınacaktır.

(4.9) eşitliğinden, her iki model için hatalarının serbestlik derecesi hesaplanır ve birbirlerinden çıkartılarak olabilirlik oran χ^2 serbestlik derecesi elde edilir. Olabilirlik oran test istatistiğinin yaklaşık χ^2 dağıldığı ancak tam olarak belirlenemediği simülasyon çalışmaları ile belirlenmiştir (Hastie ve Tibshirani, 1999:158). Bu durumda özellikle otomatik düzeltme aşamasında sorunlar doğabilmektedir. Dolayısıyla, genelleştirilmiş toplamsal model karşılaştırılmalarında bootstrap tahmin yöntemini uygulamak geçerli sonuçlar sağlayacaktır (Keele, 2008: 142).

4.2. Semiparametrik Regresyon Modelleri

Semiparametrik regresyon modelleri bağımlı değişkenin bazı bağımsız değişkenlerle doğrusal, fakat diğer bazı açıklayıcı değişkenlerle doğrusal olmayan ilişki içeresinde olduğu regresyon modelleridir. Semiparametrik regresyon modelleri standart regresyon tekniklerini genelleştiren ve her bir değişkenin etkisinin açık bir şekilde yorumlanmasını sağlayan toplamsal modellerin özel bir durumudur (Aydın, 2005: 48). Bu durumda, semiparametrik regresyon modelleri toplamsal modellere parametrik bileşen eklenerek oluşturulan modellerdir. Semiparametrik regresyon modelleri çok boyutluluğun yarattığı sıkıntı nedeniyle parametrik olmayan modellere tercih edilmektedir.

Semiparametrik regresyon modelinde değişkenler arasındaki ilişkiler, çoklu regresyonda bağımsız değişkenler ile bağımlı değişkenler arasındaki bir kısmı doğrusal bir kısmının ise eğrisel olduğu parametrik regresyon modelindeki ilişkiye benzetilebilir (Çağlayan, 2002: 72). Modelin eğrisel kısımlarını açıklamada mümkün fonksiyonel şekiller doğrusal olmayan ilişkiyi yeterince açıklamayabilir. Bu durumda parametrik doğrusal olmayan matematiksel kalıplar kullanılarak oluşturulan kısmi doğrusal çoklu regresyon modelleri kullanılabilir. Ancak ilişkiyi açıklamakta bu modellerde yetersiz kalabilmektedir. Kısmi doğrusal modeller yerine kısmen parametrik olmayan semiparametrik regresyon modelleri kullanmak ilişkiyi yeterince açıklayamama durumunu tamamen ortadan kaldıracaktır.

Anlaşılacağı gibi modelin parametrik kısmı doğrusal ilişkiyi, modelin parametrik olmayan kısmı ise doğrusal olmayan ilişkiyi ifade etmektedir. Bu nedenle semiparametrik regresyon modellerine yarı doğrusal modeller adı da verilmektedir (Lee, 1990: 203).

Sonuç olarak, semiparametrik regresyon modelleri karmaşık verileri anlaşılabilir ve özetlenebilir bir duruma indirger. Bu modeller, verilerin gerekli olanlarını içerirken önemsiz detayları model dışında bırakır ve sağlıklı kararlar verilmesine olanak sağlar.

Semiparametrik regresyon modeli,

$$y_i = \alpha + f_1(x_1) + ... + f_j(x_j) + \beta_1 x_{j+1} + ... + \beta_k x_k + \varepsilon \qquad (4.10)$$

Biçiminde ifade edilir. (4.10) modelindeki j tane değişkenin y bağımlı değişkeni üzerinde doğrusal olmayan etkisi bulunmaktadır ve modelin parametrik olmayan bölümünü oluşturmaktadır. Diğer değişkenlerin ise y bağımlı değişkeni üzerinde doğrusal etkisi bulunmaktadır ve modelin parametrik bölümünü oluşturmaktadır. Modelin parametrik bölümünde kukla değişkenler gibi kesikli değişkenlere yer verilebilmektedir. Ayrıca parametrik modellerde olduğu gibi değişkenler arasındaki etkileşimlerde incelenebilir.

$$y_i = \alpha + f_1(x_1) + f_2(x_2) + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \qquad (4.11)$$

(4.11) modeli incelendiğinde x_3 değişkeni kukla değişken, x_1, x_2, x_4 değişkenleri ise sürekli değişkenlerdir. Böyle bir semiparametrik regresyon modelinde bir çok etkileşim modele dahil edilebilir. Örneğin, x_1 ve x_2 değişkenleri arasındaki doğrusal olmayan ilişki tahmin edilebilir. Bu durumda, çoklu parametrik olmayan regresyon modellerinde olduğu gibi üç boyutlu bir grafik elde edilecektir. Araştırmanın konusuna ve içeriğine bağlı olarak x_3 ve x_4 değişkenleri arasındaki ilişkiler de incelenebilir. Ayrıca modelin parametrik ve parametrik olmayan bölümündeki değişkenlerin birbirleriyle etkileşimi de analiz edilebilmektedir.

Bir regresyon modeli kurma aşamasında ilk olarak değişkenler belirlenmektedir. Değişkenler belirlendikten sonra ise modelin fonksiyonel şeklinin veya matematiksel yapısının belirlenmesi gerekmektedir. Matematiksel kalıp oluşturken öncelikli olarak yapılması gereken grafiklerin incelenmesidir. Bağımlı değişken ile herbir bağımsız değişkenin ayrı ayrı çizilecek grafikleri incelenerek, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin yapısı hakkında fikir sahibi olunabilir. Matematiksel kalıp ile ilgili tereddütler söz konusu olduğunda, farklı şekillerin denenmesi en uygun sonucu elde etmek için yararlı olacaktır. Herbir değişkenin ilişkisine tek tek bakıldıktan sonra bağımsız değişkenlerin bir veya bir kaçı için parametrik olmayan, diğerleri için parametrik ilişki uygun ise semiparametrik regresyon modeli en uygun model olarak tercih edilecektir (Çağlayan, 2002: 75).

Semiparametrik regresyon modellerinde parametrik kısım doğrusal olabileceği gibi ikinci bölümde anlatılan dönüşüm yöntemleri uygulanarak doğrusallaştırılabilinen gerçekte

doğrusal yapıda da olabilir. Semiparametrik modelin parametrik kısmının belirlenmesinde farklı modeller tahmin edilebilir. Tahmin edilen bu modellerden artık kareler toplamını minimum yapan model semiparametrik modelin parametrik kısmı olarak tahmin edilebilir.

Değişkenler arasındaki ilişkiyi en iyi şekilde açıklayacak model çeşitli denemeler sonucunda da bulunabilir. Özellikle değişkenler arasında şekli tam belirlenemeyen ilişkiler varsa farklı matematiksel kalıplar veya farklı değişkenler için parametrik olmayan ilişkileri kapsayan modellerin denenmesi ve en uygun olanının seçilmesi gerekecektir (Çağlayan, 2002: 76).

4.2.1. Semiparametrik Regresyon Modellerinin Tahmini: Backfitting Algoritması

Toplamsal modellerin ve semiparametrik regresyon modellerinin tahmininde tekrarlı (iterative) algoritmalara ihtiyaç duyulmaktadır. Bu modellerin tahmini için geliştirilen bir çok algoritma bulunmaktadır ve bu algoritmalar bir çok değişik bilgisayar programında yer almaktadır. Özellikle, bu çalışmanın uygulama aşamasında kullanılan R programı bir çok algoritmayı desteklemektedir. Bu algoritmalardan en çok kullanılanlar Newton- Raphson algoritması, backfitting algoritmasıdır. Backfitting algoritması Hastie ve Tibshirani tarafından 1990 yılında tanıtılmıştır. Bu algoritma parametrik olmayan ve parametrik bileşenleri tahmin edebilen en kolay yöntem olarak bilinmektedir.

Model tahmin etme aşamasında, x değişkenleri birbirlerine dik iseler modelin parametrik kısmı iki değişkenli modeller serisi olarak sıradan en küçük kareler yöntemini kullanarak tahmin edilebilir. Parametrik olmayan bileşenlerin tahmininde ise lowess ya da splaynlar kullanılabilir. Bağımsız değişkenler arasında korelasyon bulunmaması durumuna genellikle rastlanmamaktadır. Bu durumda toplamsal modelleri ya da semiparametrik modelleri tahmin ederken bağımsız değişkenler arasındaki ilişkiyi dikkate alacak yöntemlere ihtiyaç duyulmaktadır. Backfitting algoritması parametrik ve parametrik olmayan bileşenleri tahmin ederken bağımsız değişkenler arasındaki korelasyonu dikkate almak üzere tasarlanmıştır.

Backfitting algoritması kısmi regresyon fonksiyonları fikrini önermektedir. Eşitlik (4.12)'de iki bağımsız değişkenli toplamsal bir model görülmektedir.

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \qquad (4.12)$$

Bu modelde f_2'nin gerçek fonksiyonel formunun bilindiği ancak f_1'in bilinmediği varsayılsın. Bu durumda f_1'in tahmini için (4.12) modeli kısmi regresyon fonksiyonu olarak eşitlik (4.13)'deki gibi yeniden düzenlenmelidir.

$$y - \alpha - f_2(x_2) = f_1(x_1) + \varepsilon \tag{4.13}$$

(4.13) eşitliğinde x_1'e karşı $y - \alpha - f_2(x_2)$'nin düzgünleştirilmesi $f_1(x_1)$'in tahminini elde etmeyi sağlamaktadır. Bu nedenle, bir kısmi regresyon fonksiyonunu bilmek diğer kismi regresyon fonksiyonunu tahmin etmeye olanak sağlamaktadır. Gerçek durumda, hiçbir regresyon fonksiyonunu bilmek mümkün olmamaktadır. Ancak f 'lerden herhangi biri için bir başlangıç değeri belirlenirse, toplamsal modellerin tahmini için kismi regresyon fonksiyonları tekrarlı yöntemler ile çözümlenir. Model (4.14) tahmin edilmek istensin:

$$y_i = \alpha + f_1(x_1) + ... + f_k(x_k) + \varepsilon \tag{4.14}$$

(4.14) eşitliğinde S_j, sütunları f_k tahminlerinden oluşan bir matrisi ifade etmektedir. X ise kolonları x değişkenlerinden oluşan model matrisini ifade etmektedir. Toplamsal modellerin tahmini için backfitting algoritması aşağıdaki adımlardan oluşmaktadır (Hastie ve Tibshirani, 1999: 118-119).

1. Adım: $\alpha = \bar{y}$ ve $S_j = X\ (j = 1,...,m)$ başlangıç değerleri olarak seçilir.

2. Adım: Her x değişkeni için kismi artıklar hesaplanır. x_1 değişkeni için tahmin edilen kismi artıklar eşitlik (4.15)'te görüldüğü gibidir.

$$\hat{e}_p^j = y_i - \sum_{i=2}^{k} S_j - \alpha \tag{4.15}$$

3. Adım: x_1 değişkeni civarında e_p^j düzgünleştirilir. Bu aşama için parametrik olmayan regresyon modeli seçilmelidir (ikinci bölümde bahsedilen splaynların özelliklerin dolayı bir çok bilgisayar yazılımı backfitting algoritmasının üçüncü aşaması için splaynları kullanmaktadır).

4. Adım: S_j 'deki x_1 değişkeni, x_i 'nin düzgünleştirilmiş tahminleri ile değiştirilir.

5. Adım: 2'den k'ya kadar olan her x değişkeni için 2-4 adımları tekrarlanır.

6. Adım: Eşitlik (4.16)'da görülen model artık kareler toplamı hesaplanır.

$$RSS = \sum_{i=1}^{n}\left[\left(yi - \sum_{i=1}^{k} S_j \right)^2\right] \tag{4.16}$$

7. Adım: Artık kareler toplamındaki değişim belirli bir tolerans seviyesinde ise model yakınsar ve algoritma durur. Eğer değilse, bu işlem artık kareler toplamındaki değişim belirli bir tolerans seviyesine gelene kadar devam eder.

Backfitting algoritması durduğunda S_j'nin her sütünü x değişkeninin parametrik olmayan tahminini içerir. Bu tahminler x değişkenleri arasındaki ilişkiyi dikkate alır. Dolayısıyla, üç x değişkenine sahip bir toplamsal model tahmin edildiğinde $\hat{f}_1$'in grafiği, x_2 ve x_3 değişkenlerinin etkisi sabit tutulduğunda x_1'in y_i üzerindeki etkisi olarak yorumlanabilir. Backfitting algoritmasının bir çok varyasyonu bulunmaktadır. Bu varyasyonlardan en çok kullanılanlardan biri ise başlangıç değeri olarak sıradan en küçük kareler tahmincilerini kullanmaktır.

Sıradan en küçük kareler tahmincilerini başlangıç değeri olarak kullanan backfitting algoritması aşağıdaki adımlardan oluşmaktadır.

1. Adım: Her bir değişkenin kendi ortalamasından çıkartılmasıyla oluşan doğrusal regresyon modeli (4.17) eşitliğinde görüldüğü gibi oluşturulur ve tahmin edilir.

$$y_i - \bar{y} = \beta_1(x_1 - \bar{x}_1) + ... + \beta_k(x_k - \bar{x}_k) + \varepsilon \tag{4.17}$$

(4.17) eşitliği kısaca (4.18) eşitliğinde olduğu gibi gösterilebilir.

$$y^* = \beta_1 x_1^* + ... + \beta_k x_k^* + \varepsilon \tag{4.18}$$

Modellerdeki $\beta_1,...,\beta_k$ parametreleri tekrarlı backfitting algoritması için başlangıç değeri olarak görev yapar.

2. Adım: x_1 için kısmi artıklar (4.19) eşitliğinden tahmin edilir.

$$\hat{e}_{px1} = y^* - \beta_2 x_{i2}^* + ... + \beta_k x_k^* \tag{4.19}$$

Kısmi artıkların tahmini y ile x_2 değişkenleri arasındaki doğrusal bağlılığı ortadan kaldırır. Ancak en küçük kareler artıklarında ($j = 1,...,m$ için) y ile x_1 arasındaki doğrusal ya da doğrusal olmayan ilişki korunur.

3. Adım: Bir sonraki adımda f_1'in tahminini elde etmek için kısmi artıklar ($\hat{e}^{j}_{px_1}$) x_1'e karşı düzgünleştirilir. Bu aşamada kullanılan düzgünleştiricinin etkisi büyük ölçüde bulunmamaktadır.

4. Adım: $\hat{f}_{x_2}$ 'nin tahmini için eşitlik (4.20) oluşturulur.

$$e_{px_2} = y^* - \hat{f}_{x_1} x_1^* + ... + \beta_k x_k^* \qquad \textbf{(}4.20\text{)}$$

5. Adım: $\hat{f}_{x_2}$ 'nin tahmini için (4.20) eşitliğindeki kısmi artıklar x_2 değişkenine karşı düzgünleştirilir.

6. Adım: $\hat{f}_{x_2}$ 'nin yeni tahmini, x_3 için hesaplanacak olan yeni kismi artıkların hesabında kullanılır. Her bir f_k için başlangıç tahminleri yapılır ve süreç tekrarlanır.

7. Adım: Bu tekrarlı süreç, tahmin edilen kısmi regresyon fonksiyonlarının artık kareler toplamındaki değişimin belli bir tolerans seviyesine ulaşmasına kadar tekrarlanır.

Bu süreç tamamlandığında, x değişkenlerinin y_i değişkeni üzerindeki kısmi etkileri tahmin edilmiş olur.

Backfitting algoritması semiparametrik regresyon modelleri için aynı adımları içermektedir. Öncelikle, modeldeki her bir bağımsız değişken için kısmi artıklar oluşturulur. Eğer seçilen değişkenin doğrusal olmayan uyumu söz konusu ise bu değişken için kısmi artıklar, aynı bağımsız değişkene karşı düzgünleştirilirler. Eğer seçilen değişkenin doğrusal uyumu söz konusu ise düzgünleştirme yöntemi yerine sıradan en küçük kareler yöntemi kullanılır. Bacfkitting algoritması, algoritmada yapılabilen değişikliklerden dolayı bir çok regresyon modelinin tahmininde kullanılmaktadır.

4.2.2. Semiparametrik Regresyon Modellerinde Çıkarım

Semiparametrik regresyon modellerinde çıkarım, doğrusal modellerde çıkarım ile parametrik olmayan modellerde çıkarımın birleşiminden oluşmaktadır. Modeldeki doğrusal

olmayan değişkenler için güven bantları hesaplanır. Modeldeki doğrusal bileşenler için ise güven aralıklarını oluşturmak ve hipotez testlerini uygulamak için standart hatalar hesaplanır.

Parametrik olmayan değişken için oluşturulacak olan güven bantları ve standart hataların hesaplanması için varyans-kovaryans matrisinin tahminine ihtiyaç duyulmaktadır. Semiparametrik regresyon modellerinde varyans-kovaryans matrisinin tahmini parametrik olmayan regresyon modellerindeki tahmin ile çok benzer ancak daha karmaşıktır.

Semiparametrik regresyon modellerinde çıkarım yapabilmek için bazı varsayımların sağlanması gerekir (Aydın, 2005: 62). Bu varsayımlar:

- Bağımsız değişkenler arasında korelasyon yoktur.
- y bağımlı değişkeni bağımsız ve normal olarak dağılır.
- Parametrik olmayan regresyon tahmincisi f, incelenen y bağımlı değişkeni bakımından doğrusaldır.
- Hata varyansı anakütle varyansının uygun tahminidir.
- Bağımlı ve bağımsız değişkenler sürekli ölçekle ölçülür.

Semiparametrik modellerde parametrik kısımda bulunan ε_i'nin klasik doğrusal regresyon varsayımlarını sağlaması gerekmektedir. Bu durumda semiparametrik regresyon modellerinin hata teriminin (ε_i) klasik doğrusal regresyonun tüm varsayımlarına sahip olması gerekir (Fox, 2000: 35). Bu varsayımların gerçekleşmesi tutarlı tahminlerin elde edilmesini sağlayacaktır. Literatürde, varsayımların geçerliliğini incelemek için bazı testler bulunmaktadır. Ancak, semiparametrik regresyon modellerinde varsayımlar çoğunlukla artık grafikleri yardımıyla incelenmektedir.

Semiparametrik regresyon modellerinin tahmininde etkinlik kavramı da çok büyük önem taşımaktadır. Bu modellerde etkinlik kaybı söz konusu olmaktadır. Etkinlik kaybının artması semiparametrik regresyon kullanmama kararına neden olabilir. Bu durumda etkinlik, farklı modellerin karşılaştırılmasında da yardımcı olabilir. Bu durum etkinlik sınırları kavramını ortaya çıkartmaktadır. Etkinlik sınırlarının genişlemesi etkinlik kaybının artması anlamına gelmektedir. Sabit varyans varsayımı geçerli olduğunda etkinlik sınırları daralmaktadır. Semiparametrik etkinlik sınırları belirlenebilirse de, bunu doğrudan hesaplama yöntemi bulunmamaktadır. Bu durumda β'nın kısıtlı tahmin edilmesi etkinlik sınırının belirlenmesinde faydalı olabilmektedir (Çağlayan, 2002: 90).

4.2.2.1. Semiparametrik Regresyon Modellerinde Güven Bantları ve Standart Hataların Hesaplanması

Semiparametrik regresyon modelinde S matrisi düzgünleştirme matrisi olarak ifade edilmektedir. Bu matris doğrusal regresyon modelinde, şapka matrisi olarak ifade edilen H matrisine benzemektedir. Semiparametrik regresyon modellerinde f'in tahmini (4.21) eşitliğindeki gibi ifade edilmektedir.

$$\hat{f} = Sy \tag{4.21}$$

S matrisi elde edildikten sonra, standart hatalar $\hat{\sigma}^2 SS'$ varyans-kovaryans matrisini kullanarak en küçük kareler yöntemindeki gibi tahmin edilir.

Toplamsal modeller eşitlik (4.22)'deki gibi ifade edilmektedir (Hastie ve Tibshirani, 1999: 109).

$$\begin{bmatrix} \mathrm{I} & S_1 & S_1 & \dots & S_1 \\ S_2 & \mathrm{I} & S_2 & \cdots & S_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_k & S_k & S_k & \cdots & \mathrm{I} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix} = \begin{bmatrix} S_1 y \\ S_2 y \\ \vdots \\ S_k y \end{bmatrix} \tag{4.22}$$

Eşitlik (4.22)'de I matrisi $(n \times n)$ boyutlu birim matrisi ve $S_1, \cdots, S_k$ ise her bir X değişkeni için düzeltme matrisi olarak ifade edilmektedir.

Teorik olarak (4.22) eşitliği QR analizi gibi tekrarlı olmayan (noniterative) yöntemlerle çözülebilmektedir. Ancak, bu denklem sistemi tekrarlı olmayan yöntemlerle çözülmek için çok büyük olduğundan backfitting algoritmasının kullanımını zorunlu kılmaktadır (Hastie ve Tibshirani, 1999: 109).

Toplamsal ve semiparametrik regresyon modellerinde güven bantlarının oluşturulması için varyans-kovaryans matrisini elde edilmesi gerekmektedir. Bu durumda, (4.22)'deki denklem sistemi eşitlik (4.23)'deki gibi ifade edilebilir.

$$\hat{S}f = \hat{Q}y \tag{4.23}$$

(4.23) eşitliği yeniden düzenlendiğinde (4.24) eşitliği elde edilmektedir.

$$\hat{f} = \hat{S}^{-1}\hat{Q}y \tag{4.24}$$

(4.24) eşitliği yeniden düzenlendiğinde ise (4.25) eşitliğine ulaşılmaktadır.

$$\hat{f} = Ry \tag{4.25}$$

(4.25) eşitliğinde $R = \hat{S}^{-1}Q$ olarak ifade edilmektedir. Eğer gözlemler bağımsız ve aynı dağılıma sahipler ise (4.26) eşitliği oluşturulabilmektedir.

$$V(\hat{f}) = \sigma^2 RR' \tag{4.26}$$

(4.26) eşitliğindeki σ^2 (4.27) eşitliği ile yer değiştirir ve $V(\hat{f})$ değerine ulaşılır.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{df_{res}} \tag{4.27}$$

Artıkların serbestlik derecesi (df_{res}) ise eşitlik (4.28) ile elde edilmektedir.

$$df_{res} = n - tr(2R - RR') \tag{4.28}$$

Güven bantları, $\sigma^2 RR'$ matrisinin köşegen elemanlarının karekökleri ile ± 2'nin çarpımından elde edilir (Keele, 2008: 118). Semiparametrik modellerde R matrisinin köşegen elemanları β tahminlerinin varyansını ifade etmektedir. R matrisini tahmin etmek için tekrarlı yöntemlere ihtiyaç duyulmaktadır. Backfitting algoritması R matrisini tahmin etmek için kullanılabilmektedir. Varyans-kovaryans matrisinin bu yolla tahmini $\hat{f}$'daki yanlılığı düzeltememektedir (Hastie ve Tibshirani, 1999: 119-120). Bu durumda bayesgil güven bantları gibi alternatif yöntemler kullanılabilmektedir. Ayrıca bir çok bilgisayar programı ve özellikle R programı, toplamsal ve semiparametrik regresyon modelleri için yanı düzeltimiş varyans-kovaryans matrislerini (bias adjusted variance-covariance matrices) hesaplamaktadır.

4.2.2.2. Semiparametrik Regresyon Modellerinde Hipotez Testleri

Semiparametrik regresyon modellerinde hipotez testleri herhangi bir karmaşıklık içermemektedir. Modeldeki parametrik bileşenlerin istatistiksel açıdan anlamlı olup olmadığı

araştırılmak istendiğinde, R matrisinden β tahminlerinin standart hataları hesaplanır ve bilinen t testleri uygulanır.

t testleri için hipotezler ise:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

şeklinde ifade edilir.

Semiparametrik regresyon modellerinde, parametrik olmayan bileşen için hipotez testleri iki amaçla uygulanmaktadır. Birinci hipotez testinin amacı, x bağımsız değişkenin y bağımlı değişkeni üzerindeki etkisininin istatistiksel olarak anlamlı olup olmadığını ortaya çıkartmaktır. İkinci hipotez testinin amacı ise, incelenen değişkenin parametrik olmayan bileşen olarak modelde yer almasının parametrik bileşen olarak modelde yer almasından üstün olup olmadığını belirlemektir. Kısaca amaç, model uyumunun hangi durumda en iyi olduğunu belirlemektir. Her iki hipotez testinde de kısmi F testi ve olabilirlik oran testi kullanılabilmektedir.

Her iki hipotez testini incelemek için eşitlik (4.29)'da görülen iki değişkenli bir toplamsal model oluşturulmuştur.

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \qquad (4.29)$$

x_2 değişkeninin istatistiksel olarak anlamlı olup olmadığını test etmek için, eşitlik (4.29), eşitlik (4.30)'a karşı test edilmelidir.

$$y = \alpha + f_1(x_1) + \varepsilon \qquad (4.30)$$

f_2 değişkenin doğrusal olup olmadığını test etmek için ise eşitlik (4.29), eşitlik (4.31)'e karşı test edilmelidir.

$$y = \alpha + f_1(x_1) + \beta_1 x_2 + \varepsilon \qquad (4.31)$$

F testinin temeli artık kareler toplamından oluşmaktadır. Herhangi bir toplamsal ya da semiparametrik regresyon modelinin artık kareler toplamı eşitlik (4.32)'de görüldüğü gibi hesaplanmaktadır.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{4.32}$$

RSS_0 kısıtlı modelin artık kareler toplamı, RSS_1 ise toplamsal veya semiparametrik regresyon modelinin artık kareler toplamı ise F test istatistiği eşitlik (4.33)'deki gibi hesaplanır.

$$F = \frac{RSS_0 - RSS_1 / [tr(R) - 1]}{RSS_1 / df_{res}} \tag{4.33}$$

Bu test istatistiği F dağılımına yakınsamaktadır.

Tahmin için backffitting algoritması kullanmak yerine, yeniden ağırlıklandırılmış en küçük kareler ya da kısıtlı maksimum olabilirlik yöntemlerini kullanan bilgisayar programları olabilirlik oran ya da sapma fark (diffence of deviance) testlerini kullanmaktadırlar (Keele, 2008: 119). Toplamsal ya da semiparametrik regresyon modelleri için olabilirlik oran testi eşitlik (4.34) yardımıyla uygulanmaktadır.

$$LR = -2(Logolabilirlik_0 - Logolabilirlik_1) \tag{4.34}$$

Eşitlik (4.34)'de $Logolabilirlik_0$ değeri kısıtlı modelin logaritmik olabilirlik değerini, $Logolabilirlik_1$ değeri ise kısıtsız model olan toplamsal ya da semiparametrik regresyon modelinin logaritmik olabilirlik değerini ifade etmektedir. Dikkat edilirse bu test istatistiği iki modelin sapmaları arasındaki fark dikkate alınarak hesaplanır. H_0 hipotezi altındaki test istastiği yaklaşık χ^2 dağılımı göstermektedir. Bu dağılım için serbestlik derecesi ise, iki modelin parametre sayıları arasındaki fark olarak hesaplanmaktadır.

Semiparametrik regresyon modellerinde parametrik olmayan bileşenin tahmininin otamatik düzeltme teknikleri ile yapılması durumunda test istatistiştiğinin dağılımının yaklaşık χ^2 dağılımı gösterdiği düşünülmektedir. Ancak bu yakınlaşmanın ne kadar olduğu örneklem büyüklüğünden ciddi derecede etkilenmektedir (Hardle, Müller, Sperlich ve Werwatz, 2004: 127). Hipotez testlerinde H_0 hipotezini güvenle reddebilmek için tahmin edilen p olasılıklarının mümkün olduğunca küçük olması gerekmektedir. Ancak, tahmin edilen p olasılığı H_0 hipotezini zorlukla reddedebilecek seviyede ise bu durumda otomatik düzeltme teknikleri yerine diğer düzeltme tekniklerinin (manual smoothing) kullanılması daha

kesin sonuçlara ulaşılmasına olanak sağlayacaktır. Yeniden örnekleme tekniklerinden olan bootstrap yöntemide (parametrik dağılım varsayımı gerektirmeyen bootstrap yöntemi) bu tür bir problemin çözümüne olanak sağlamaktadır (Hastie ve Tibshirani, 1999: 293).

5. SİTE İÇERESİNDEKİ DAİRELERİN SATIŞ FİYATLARINI ETKİLEYEN ÖZELLİKLERİN İNCELENMESİ

Son yıllarda site içerisinde konut alımı ve yaşantısı önem kazanmaktadır. Ancak daire satın alınırken sadece site içerisinde olması değil diğer bir çok özelliklede dikkate alınmalıdır. Öncelikle daire satın alırken dikkat edilmesi gereken unsurlardan bahsetmek gerekmektedir. Tempo dergisinin 2009 yılında yaptığı bir araştırmaya göre daire satın alınırken 40 özellik dikkate alınmaktadır (Çevrimiçi: www.satilikdairearıyorum.blogcu.com; 29.03.2010). Bu özellikler, inşaat kalitesi, arsa alanı, İnşaat alanı, toplam kullanım alanı, net alan (oda, salon, koridor, balkon), sosyal tesisler, yeşil alan, mevki, doğa manzarası, güvenlik sistemi, spor kompleksi, kapıcı dairesi, asansör, hidrafor ve su deposu, açık otopark, kapalı otopark, şömine ve barbekü, balkon, oda sayısı, satış kabiliyeti, havuz, ulaşım, semt özelliği, tapu durumu, zemin durumu, ısınma, konutun bulunduğu kat, prim getiri potansiyeli, deniz manzarası, alış veriş merkezine yakınlığı, aidat, malzeme kalitesi, yapım yılı, bina özelliği (apartman ya da bağımsız ev olması), güneş alma durumu, özel dekorasyon, kira geliri, depreme dayanıklılık, net alanla brüt alan arasındaki fark, referanslar olarak belirlenmiştir.

Görüldüğü gibi dairelerin satış fiyatlarını etkileyen bir çok unsur bulunmaktadır. Bu bölümde bu değişkenlerden bazıları ve değişkenlerin satış fiyatı ile ilişkisi incelenenecektir.

5.1. Uygulamada Kullanılan Veri ve Değişkenler

Veriler, İstanbul Çekmeköy İlçesinde, birbiriyle yakın mesafe içerisinde bulunan siteler içerisindeki 81 daireden elde edilmiştir. Bahsedildiği üzere dairelerin satış fiyatı üzerinde, sitenin bulunduğu mevkinin önemi çok büyüktür. Ancak bu çalışmada mevki gibi özelliklerin değil daha çok fiziksel özelliklerinin satış fiyatı üzerinde etkisi araştırılmak istenmiştir. Ayrıca incelenen bütün daireler için açık havuz, kapalı havuz, açık otopark, kapalı otopark, spor salonu, koşu alanı, kafeterya, güvenlik ve çocuk parkı bulunmaktadır. Böylelikle site özelliklerin satış fiyatı üzerindeki ortadan kaldırılmış ve sadece aşağıda adı geçen ev özelliklerinin etkisini ortaya çıkarmak hedeflenmiştir.

Uygulamada kullanılan değişkenler;

- Salon Büyüklüğü (m^2)
- Yatak Odası Büyüklüğü (m^2)

- Banyo Büyüklüğü (m^2)
- Koridor Büyüklüğü (m^2)
- Balkon Büyüklüğü (m^2)
- Dairelerin Satış Fiyatı (TL)
- Oda Sayısı: Oda sayısı için kukla değişken : 2 odalı, 3 odalı, 4 odalı, 5 odalı, 6 odalı.
- Cephe: Cephe durumu için kukla değişken : Site içi cepheli, site dışı cepheli, site içi ve site dışı cepheli.
- Su Deposu: Su deposu için kukla değişken: Su deposu var, su deposu yok .

Görüldüğü gibi incelenen değişkenler site içerisindeki dairelerin fiziksel özellikleridir. Dairenin site içerisinde bulunması ya da bağımsız olması dairelerin satış fiyatını etkileyecektir. Ayrıca sitenin diğerlerinden çok farklı ve ayrıcalık yaratacak özellikleri de olabilmektedir. Ancak bu uygulama için seçilen siteler bölge olarak ve site özelliği olarak çok büyük benzerlikler göstermektedir.

Burada dikkat edilmesi gereken nokta, bazı değişkenlerin fiyat değişkeni ile parametrik ilişkide bulunması, bazı değişkenlerin ise parametrik olmayan ilişki içerisinde bulunmasıdır. Örneğin dairenin koridorunun büyüklügü ile fiyat arasında kesin bir doğrusal ilişki olduğunu söylemek doğru değildir. Ayrıca kukla değişkenler modele parametrik olarak dahil edilmek durumundadırlar. Çünkü bu tip değişkenler fonksiyonun eğriliğini etkilemezler. Bu nedenle, bu tip ilişkileri analiz ederken hem parametrik bölüm hemde parametrik olmayan bölümü bir bütün olarak ele alan semiparametrik regresyon yöntemi uygun olabilmektedir.

5.2. Uygulamanın Aşamaları

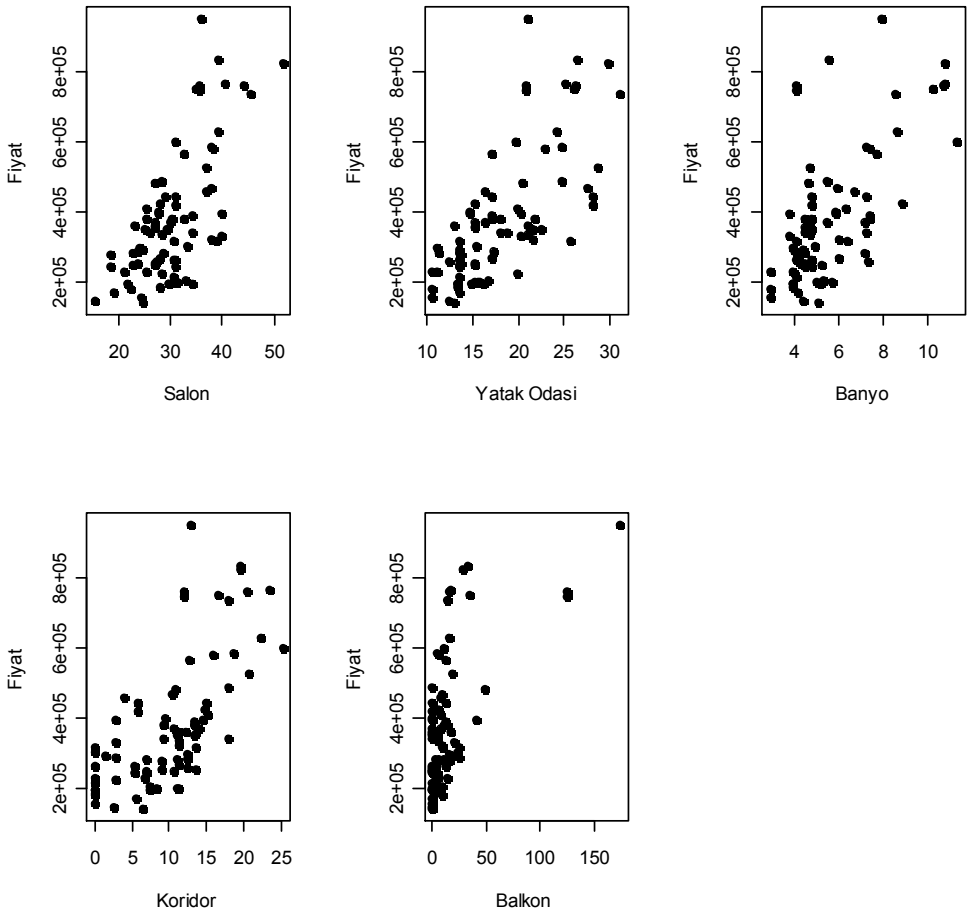
Uygulamanın ilk aşamasında değişkenlerin orijinal grafikleri incelenmiş ve fiyat bağımlı değişkeni ile diğer değişkenlerin nasıl bir ilişkisi olduğu görsel olarak saptanmıştır. İkinci aşamada ise, görsel olarak saptanan ilişkiler semiparametrik regresyon modeli ile istatistiksel olarak belirlenmiştir. Bu aşamadan sonra, düzgünleştirilerek (pürüzsüz bir fonksiyon olarak) modele dahil edilecek değişkenler için, düzgünleştirmenin logaritmik ya da quadratik dönüşüme üstünlüğü, İkinci bölümde bahsedilen cimrilik prensibi yaklaşımından

dolayı test edilmiştir. Son aşamada ise uygulanmasına karar verilen model ve bu modele ilişkin varsayımlar incelenmiştir.

Uygulamada, semiparametrik regresyon modellerindeki β katsayıları ve s(x) bilinmeyen fonksiyonları İkinci bölümde incelenen splayn düzeltme tekniği kullanarak R bilgisayar ortamında yazılan bir programla ile elde edilmiştir. λ düzeltme parametresinin seçiminde genelleştirilmiş çapraz geçerlilik (GCV) yöntemi kullanılmış ve otomatik düzeltme gerçekleştirilmiştir. Yazılan R kodları EK-C'de gösterilmektedir.

5.2.1. Değişkenlerin Orijinal Grafiklerinin İncelenmesi

İncelenen değişkenlerin fiyat bağımlı değişkeni ile olan ilişkisini gösteren grafik Şekil 5.1'de gösterilmiştir. Bu grafikler incelendiğinde, bütün değişkenler için fiyat ile ilişkinin kesinlikle doğrusal ya da kesinlikle eğrisel olduğunu söylemek doğru olmayacaktır. Dolayısıyla, hangi değişkenlerin modele düzgünleştirilerek dahil edilmesi gerektiğini saptamak için sadece grafikler yeterli olmamaktadır.



Şekil 5. 1: Değişkenlerin Orijinal Grafikleri

5.2.2. Fiyat İle Doğrusal Olmayan İlişki İçerisinde Bulunan Değişkenlerin Belirlenmesi

Fiyat ile doğrusal olmayan ilişki içerisinde bulunan değişkenlerin belirlenmesi amacıyla sürekli olan bütün değişkenler splayn düzeltme yaklaşımı kullanılarak modele dahil edilmiştir. Bu modeldeki öncelikli amaç, değişkenlerin hangilerinin dairelerin satış fiyatı üzerinde doğrusal olmayan etki içerisinde olduklarını ortaya çıkartmaktır. Ancak, bir diğer amaç ise, fiyat bağımlı değişkeni üzerinde istatistiksel olarak anlamsız olan değişkenleri ortaya çıkartmaktır. Bilindiği üzere, uygulamada kukla değişkenler söz konusudur. Semiparametrik regresyon yönteminde kukla değişkenler modele parametrik olarak dahil edilmektedir. Oluşturulan semiparametrik regresyon modeli sabit terim içerdiğinden kukla değişkenler içerisindeki bir kategori, referans kategori olarak seçilir. Sabit terim içeren bir modelde her kategori için oluşturulan kukla değişkenlerin modele dahil edilmesi durumunda tam çoklu doğrusal bağıntı problemi ortaya çıkacağından regresyonun tahmini gerçekleşemeyecektir. Semiparametrik regresyonda da bu varsayımların sağlanması gerekmektedir.

Tablo 5. 1: Fiyat İle Doğrusal Olmayan İlişki İçerisinde Bulunan Değişkenlerin Belirlenmesi İçin Oluşturulan Model Sonuçları

Değişkenler	*Parametre Tahmini*	*Standart Hata*	*t -İstatistiği*	*Olasılık*
Sabit	303332	21419	14.162	< 2e-16 ***
Oda Sayısı 3	58448	13749	4.251	7.47e-05 ***
Oda Sayısı 4	129796	19044	6.816	4.95e-09 ***
Oda Sayısı 5	242663	29482	8.231	1.85e-11 ***
Su Deposu	-34082	17463	-1.952	0.0556
İç Cephe Manzaralı	18574	18353	1.012	0.3156
İç ve Dış Cephe Manzarası	18304	20759	0.882	0.3814
Bileşen	*F - İstatistiği*	*Olasılık*		
Salon	6.679	0.01218 *		
Yatak Odası	5.609	0.00036 ***		
Banyo	5.511	0.00113 **		
Koridor	5.028	0.00784 **		
Balkon	69.176	< 2e-16 ***		
R^2(adj)= 0.954	Açıklanan Sapma= %96.5	GCV Değeri = 2.1226e+09		

*: Parametrelerin 0.05 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

**: Parametrelerin 0.01 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

***: Parametrelerin 0.001 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

Uygulamadaki değişkenler incelendiğinde, cephe değişkeni üç kategoriden oluştuğu gözlenmektedir. Ancak 5.1'de görüldüğü gibi, iç cephe ve iç-dış cephe manzara kategorileri

için tahmin yapılmış ve referans kategori olarak dış cephe seçilmiştir. Referans kategori araştırıcının istediğine bağlı olarak belirlenebilmekte ve değiştirilebilmektedir. Kukla değişkenlerin yorumu referans kategoriye göre yapılmaktadır. Örneğin; diğer değişkenlerin etkisi sabit tutulduğunda, dairenin iç cephe manzaralı olması dış cephe manzaralı olmasına göre daire fiyatını ortalama 18574 TL arttırmaktadır. Ancak, Tablo 5.1'de görüldüğü üzere, modele kukla değişken olarak katılan cephe değişkenin her iki kategoriside dış cephe manzaralı olma kategorisine göre istatistiksel olarak anlamsız bulunmaktadır. Ancak referans kategori, bütün kategoriler için denenmiş ve istatistiksel olarak anlamlı sonuçlara ulaşılamamıştır. Dolayısıyla genel olarak cephe faktörünün dairelerin satış fiyatları üzerinde etkili olmadıkları sonucuna varılmaktadır ve yukarıda yapılan yorum gerçeği yansıtmamaktadır. Bu nedenle cephe değişkeni modelden çıkartılmış ve aynı yöntemle yeni bir model kurulmuştur.

Tablo 5.1'de düzgünleştirilen bileşenlerin (Pürüzsüz Fonksiyonların) *F* istatistikleri ve yaklaşık olasılık değerleri görülmektedir. Bu değerler incelendiğinde düzgünleştirilen değişkenlerden yatak odası, banyo, koridor ve balkon değişkenlerinin 0.001 anlam düzeyinde istatistiksel olarak anlamlı olduğu, ancak salon değişkenin 0.01 anlam düzeyinde istatistiksel olarak anlamlı olduğu görülmektedir. Sonuç olarak, bu değişkenlerin modele pürüzsüz bir fonksiyon olarak dahil olması anlamlıdır sonucuna ulaşılmaktadır. Ancak, her bir değişken için yapılacak ayrı analizler sonucunda uygulanması gereken son modele karar verilecektir.

Tablo 5. 2: Fiyat İle Doğrusal Olmayan İlişki İçerisinde Bulunan Değişkenlerin Belirlenmesi İçin Oluşturulan 2. Model (Temel Model) Sonuçları

Değişkenler	*Parametre Tahmini*	*Standart Hata*	*t -İstatistiği*	*Olasılık*
Sabit	313975	18533	16.941	< 2e-16 ***
Oda Sayısı 3	58382	13314	4.385	4.54e-05 ***
Oda Sayısı 4	133183	18167	7.331	5.55e-10 ***
Oda Sayısı 5	244526	28885	8.466	5.89e-12 ***
Su Deposu	-28128	15212	-1.849	0.0692
Bileşen		*F - İstatistiği*	*Olasılık*	
Salon		6.625	0.012446 *	
Yatak Odası		5.778	0.000253 ***	
Banyo		6.317	0.000349 ***	
Koridor		5.045	0.007946 **	
Balkon		75.753	< 2e-16 ***	
R^2(adj)= 0.955		Açıklanan Sapma= %96.5	GCV Değeri = 2.0208e+09	

*: Parametrelerin 0.05 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

**: Parametrelerin 0.01 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

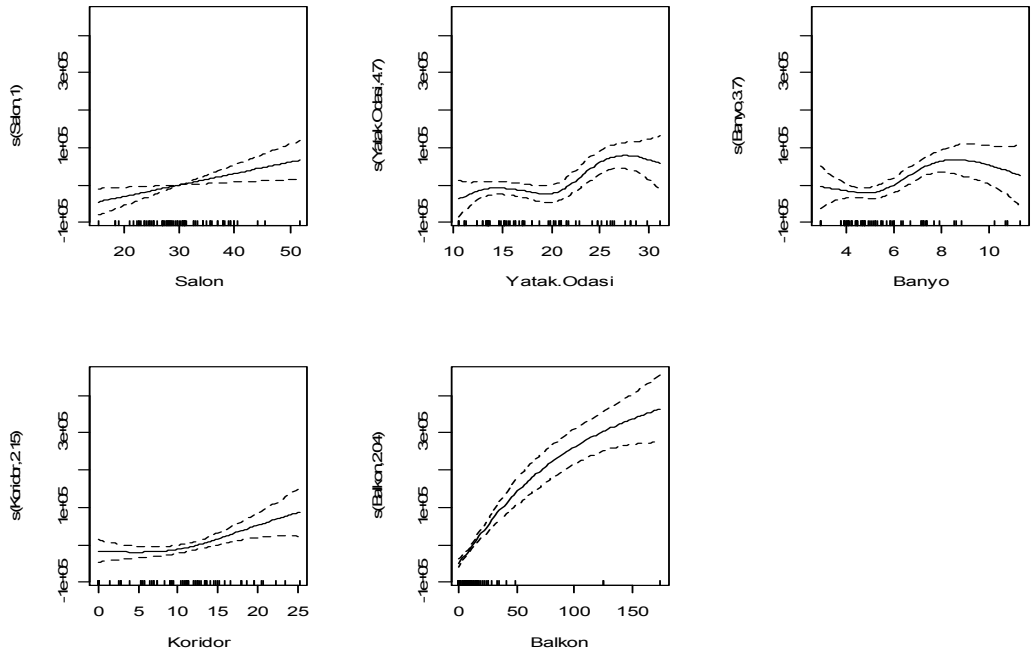
***: Parametrelerin 0.001 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

Fiyat ile doğrusal olmayan ilişki içerisinde bulunan değişkenlerin belirlenmesi amacıyla kurulan ilk modelden cephe değişkenin istatistiksel olarak anlamlı olmadığı sonucuna ulaşıldığından cephe değişkeni modelden çıkartılarak yeni bir model oluşturulmuştur. Bu modelin sonuçları Tablo 5.2'de gösterilmektedir.

Tablo 5.2. incelendiğinde kukla değişkenlerden su deposu değişkenin 0.10 anlam düzeyinde istatistiksel olarak anlamlı olduğu diğer değişkenlerin ise çok düşük anlam düzeyinde bile istatistiksel olarak anlamlı olduğu görülmektedir. Su deposu değişkeninin modelden çıkartılması ileride otokorelasyon problemine neden olabileceğinden, bu değişken modele dahil edilmektedir. Uygulanacak en son modele karar verme aşamasında bu değişken yeniden değerlendirilecektir.

Düzgünleştirilen bileşenlerin *F* istatikleri ve olasıkları incelendiğinde bu değişkenlerin modele pürüzsüz bir fonksiyon olarak dahil olması anlamlıdır sonucuna ulaşılmaktadır.

Bilindiği üzere bu modellerin oluşturulmasındaki temel amaç fiyat ile doğrusal olmayan ilişkisi olan değişkenlerin belirlenmesidir. Bu amaçla Tablo 5.2'deki modelin grafikleri incelenmiştir.



Şekil 5. 2: Fiyat İle Doğrusal Olmayan İlişki İçerisinde Bulunan Değişkenlerin Belirlenmesi İçin Oluşturulan 2. Model (Temel Model)'in Grafikleri

Şekil 5.2 incelendiğinde salon değişkeni hariç bütün sürekli değişkenlerin fiyat bağımlı değişkeni ile doğrusal olmayan bir ilişki içerisinde olduğu görülmektedir. Sonuç

olarak, Tablo 5.2'de oluşturulan model temel model olarak ele alınacaktır. Değişkenlerin modele pürüzsüz bir fonksiyon olarak, doğrusal olarak veya üstel dönüştürme yöntemlerini kullanarak dahil edilmesi aşamasına, tamel model ile yapılan karşılaştırmalar sonucunda kesin olarak karar verilecektir.

5.2.3. Değişkenlerin Fiyat Değişkeni ile Doğrusal ya da Doğrusal Olmayan İlişkisinin Test Edilmesi

Bu aşamada kurulan bütün modeller, tüm sürekli değişkenlerin pürüzsüz bir fonkiyon olarak modele dahil edildikleri ve sonuçları Tablo 5.2'de gösterilen model (temel model) ile karşılaştırılmıştır. Modellerin karşılaştırılması semiparametrik regresyon modelleri için oluşturulan sapma analizi (Analysis of Deviance) ile gerçekleştirilmiştir. Her bir sürekli değişken için doğrusal model oluşturulmuş ve temel model ile karşılaştırılaştırılmıştır.

Yatak Odası (Büyüklüğü) Değişkeni İçin Yapılan Değerlendirme

Yatak odası değişkeni için yapılan değerlendirmede, sadece yatak odası değişkeni modele doğrusal bir fonksiyon olarak dahil edilmiş ve temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur. Görüldüğü gibi H_1 hipotezi temel modeli ifade etmektedir.

H_0 : E(Fiyat) = β_0 + s(Salon)[14] + β_1 Oda Sayısı + β_2 Yatak Odası + s(Banyo) + s(Koridor) + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

H_0 hipotezinin kabul edilmesi durumunda, yatak odası değişkeninin fiyat bağımlı değişkeni ile olan ilişkisinin doğrusal olduğunu kabul edilecektir. H_0 hipotezinin red edilmesi durumunda ise ilişkinin doğrusal olmadığı ve yatak odası değişkenin modele düzgünleştirilerek yani pürüzsüz bir fonksiyon olarak dahil edilemesi gerektiği sonucuna uşalıcaktır. Bu karşılaştırma için sapma analizi uygulanmıştır. Bu analiz sonucunda sapma değeri 1.9865e+10 ve olasılık değeri 0.003089 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük olduğundan H_0 hipotezi reddedilmiş ve

[14] $s_i(x_i)$ olarak ifade edilen fonksiyon splayn düzeltme ile düzgünleştirilmiş bir fonksiyonu ifade etmektedir. Bu fonksiyon önceki bölümlerde $f_i(x_i)$ olarak gösterilmiştir. Buradaki amaç, splayn düzeltmenin kullanıldığını vurgulamaktır.

yatak odası değişkeninin modele pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Banyo (Büyüklüğü) Değişkeni İçin Yapılan Değerlendirme

Banyo değişkeni için yapılan değerlendirmede, sadece banyo değişkeni modele doğrusal bir fonksiyon olarak dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + β_2 Banyo + s(Koridor) + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri -1.4327e+10 ve olasılık değeri 0.2042 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden büyük olduğundan H_0 hipotezi reddedilememiş ve banyo değişkeninin modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmiştir. Şekil 5.2 incelendiğinde banyo değişkenin fiyat değişkeni ile doğrusal olmayan bir ilişkisi olduğu görülmekteydi. Ancak sapma analizi sonuçları istatistiksel olarak anlamlı bir doğrusal olmayan ilişki olmadığını ortaya çıkartmıştır. Bu nedenden dolayı grafikler kesin bir sonuç sağlamamakta, sadece değişkenler hakkında önsel bir bilgi sağlamaktadırlar.

Koridor (Büyüklüğü) Değişkeni İçin Yapılan Değerlendirme

Koridor değişkeni için yapılan değerlendirmede, sadece koridor değişkeni modele doğrusal bir fonksiyon olarak dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + β_2 Koridor + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 7.3790e+09 ve olasılık değeri 0.01325 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük

olduğundan H_0 hipotezi reddedilmiş ve koridor değişkeninin modele pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Balkon (Büyüklüğü) Değişkeni İçin Yapılan Değerlendirme

Balkon değişkeni için yapılan değerlendirmede, sadece koridor değişkeni modele doğrusal bir fonksiyon olarak dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + β_2 Balkon+ β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 1.5055e+10 ve olasılık değeri 0.001987 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük olduğundan H_0 hipotezi reddedilmiş ve balkon değişkeninin modele pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Sonuç olarak, yatak odası, koridor ve balkon değişkenlerinin modele düzgünleştirilerek yani pürüzsüz fonksiyonlar olarak dahil edilmesine, banyo değişkeninin ise modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmiştir. Salon değişkenine ise sapma analizi uygulanmamıştır. Şekil 5.2'de salon değişkenin fiyat değişkeni ile doğrusal bir ilişkisi olduğu açıkça görülmektedir. Böyle bir değişkene sapma analizi uygulandığında olasılık değeri oluşmayacaktır. Bu durumun nedeni, belirgin bir doğrusallık söz konusu olduğunda test edilmesi gereken bir durumun olmamasıdır. Özellikle, düzeltme parametresi otomatik yöntemle seçilindiğinde bu durumla karşılaşılmaktadır (Çevrimiçi: http://r.789695.n4.nabble.com; 10.03.2010).

5.2.4. Logaritmik Modeller İle Temel Modelin Karşılaştırılması

Bu aşamada bütün sürekli değişkenler için logaritmik modeller oluşturulacak ve sapma analizi araclığıyla temel model ile karşılaştıralacaktır. Dönüştürme yöntemleri kullanılarak düzgünleştirilebilecek bir değişken pürüzsüz fonksiyon olarak modele dahil edildiğinde ikinci bölümde açıklanan cimrilik prensibinden dolayı (parsimony) dolayı olumsuz sonuçlar yaratacaktır. Böyle bir sorunla karşılaşmamak için bütün karşılaştırmalar yapılacaktır.

Yatak Odası (Büyüklüğü) Değişkeni İçin Logaritmik Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, yatak odası değişkeninin logaritması alınarak semiparametrik regresyon modeli oluşturulmuş H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + β_2 log(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

H_0 hipotezinin kabul edilmesi durumunda, yatak odası değişkeninin fiyat bağımlı değişkeni ile olan ilişkisinin, yatak odası değişkeninin logaritması alınarak doğrusallaştırılabildiği kabul edilecektir. H_0 hipotezinin red edilmesi durumunda ise ilişkinin logaritma alınarak doğrusallaştırılamadığı ve yatak odası değişkeninin modele düzgünleştirilerek yani pürüzsüz bir fonksiyon olarak dahil edilmesi gerektiği sonucuna ulaşılacaktır. Bu karşılaştırma için sapma analizi uygulanmıştır. Bu analiz sonucunda sapma değeri 2.1815e+10 ve olasılık değeri 0.001227 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük olduğundan H_0 hipotezi reddedilmiş ve yatak odası değişkeninin modele logaritmik olarak değil, pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Banyo (Büyüklüğü) Değişkeni İçin Logaritmik Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, banyo değişkeninin logaritması alınarak semiparametrik regresyon modeli oluşturulmuş H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + β_2 log(Banyo) + s(Koridor) + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri -1.3953e+10 ve olasılık değeri 0.2152 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden büyük olduğundan H_0 hipotezi reddedilememiştir. Yapılan ilk analizde bu değişkenin modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmişti. Bu nedenle logaritması alınmış değişkenle oluşturulan modelin pürüzsüz fonksiyonla oluşturulan modele göre daha üstün olması beklenen bir durumdur. Bu nedenle, oluşturulacak en son modelde bu değişkenin logaritmik olarak modele dahil edilmesi gerekmemektedir.

Koridor (Büyüklüğü) Değişkeni İçin Logaritmik Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, banyo değişkeninin logaritması alınarak semiparametrik regresyon modeli oluşturulmuş H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + β_2 log(Koridor) + s(Balkon) + β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 1.4390e+10 ve olasılık değeri 0.000873 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük olduğundan H_0 hipotezi reddedilememiş ve koridor değişkeninin modele logaritmik olarak değil, pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Balkon (Büyüklüğü) Değişkeni İçin Logaritmik Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, banyo değişkeninin logaritması alınarak semiparametrik regresyon modeli oluşturulmuş H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + β_2 log(Balkon)+ β_3 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 3.4309e+10 olarak bulunmuş ve olasılık değeri hesaplanamamıştır. Olasılık değerinin hesaplanamamasının nedeni, salon değişkenine sapma analizi yapıldığında karşılaşılan durumun aynısıdır. Burada, salon değişkeni gibi doğrusal bir ilişkinin kesinliği değil, doğrusal olmayan bir ilişkinin kesinliği söz konusudur ve test edilecek bir durum söz konusu değildir. Bu nedenle, her koşulda koridor değişkeni modele düzgünleştirilerek yani pürüzsüz bir fonksiyon olarak dahil edilecektir.

Sonuç olarak, ilk yapılan değerlendirmelerle aynı sonuca ulaşılmıştır. Yatak odası, koridor ve balkon değişkenlerinin modele düzgünleştirilerek yani pürüzsüz fonksiyonlar olarak dahil edilmesine, banyo değişkeninin ise modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmiştir. Diğer bir ifade ile, değişkenlerin logaritmalarının alınması fiyat ile olan ilişkilerini doğrusallaştıramamıştır.

5.2.5. Karesel Modeller İle Temel Modelin Karşılaştırılması

Bu aşamada bütün sürekli değişkenler için karesel modeller oluşturulacak ve sapma analizi araclığıyla temel model ile karşılaştıralacaktır.

Yatak Odası (Büyüklüğü) Değişkeni İçin Karesel Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, yatak odası değişkeninin, kendisi ve karesi semiparametrik regresyon modeline dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + β_2 Yatak Odası + β_3 (Yatak Odası) ^{2} + s(Banyo) + s(Koridor) + s(Balkon) + β_4 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 2.0459e+10 ve olasılık değeri 0.002708 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden küçük olduğundan H_0 hipotezi reddedilememiş ve yatak odası değişkeninin modele karesel olarak değil, pürüzsüz bir fonksiyon olarak dahil edilmesine karar verilmiştir.

Banyo (Büyüklüğü) Değişkeni İçin Karesel Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, banyo değişkeninin, kendisi ve karesi semiparametrik regresyon modeline dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + β_2 Banyo + β_3 (Banyo)^2 + s(Koridor) + s(Balkon) + β_4 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri -1.4069e+10 ve olasılık değeri 0.2620 olarak bulunmuştur. Bu sonuçlara göre olasılık değeri 0.05 anlamlılık düzeyinden büyük olduğundan H_0 hipotezi reddedilememiştir. Yapılan ilk analizde ve logaritmik karşılaştırmada bu değişkeninin modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmişti. Bu nedenle karesel modelin pürüzsüz fonksiyonla oluşturulan modele göre daha üstün olması logaritmik karşılaştırmada olduğu gibi beklenen bir durumdur. Bu nedenle, oluşturulacak en son modelde bu değişkenin ne logaritmasının ne de karesinin modele dahil edilmesi gerekmemektedir.

Koridor (Büyüklüğü) Değişkeni İçin Karesel Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, koridor değişkeninin, kendisi ve karesi semiparametrik regresyon modeline dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + β_2 Koridor + β_3 (Koridor)^2 + s(Balkon) + β_4 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri 7.1140e+08 bulunmuş ancak olasılık değeri hesaplanmamıştır. Bu durum, balkon değişkeni için logaritmik model ile temel modelin karşılaştırılmasında ortaya çıkan durumun aynısıdır. Doğrusal olmayan ilişki çok belirgindir.

Bundan dolayı karesel modele karşı test edecek bir durum söz konusu değildir ve koridor değişkeni modele pürüzsüz bir fonksiyon olarak dahil edilmelidir.

Balkon (Büyüklüğü) Değişkeni İçin Karesel Model ile Temel Modelin Karşılaştırılması

Yapılan karşılaştırmada, balkon değişkeninin, kendisi ve karesi semiparametrik regresyon modeline dahil edilmiş ve H_1 hipotezinde görülen temel model ile karşılaştırılmıştır. Bu karşılaştırma için aşağıda görülen hipotezler oluşturulmuştur.

H_0 : E(Fiyat) = β_0 + s(Salon) + β_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + β_2 Balkon + β_3 (Balkon)^2 + β_4 Su Deposu + ε

H_1 : E(Fiyat) = α_0 + s(Salon) + α_1 Oda Sayısı + s(Yatak Odası) + s(Banyo) + s(Koridor) + s(Balkon) + α_2 Su Deposu + ε

Sapma analizi sonucunda sapma değeri -2487193863 bulunmuş ancak olasılık değeri koridor değişkeninde açıklanan nedenden dolayı hesaplanmamıştır. Doğrusal olmayan ilişki çok belirgindir. Bundan dolayı karesel modele karşı test edecek bir durum söz konusu değildir ve balkon değişkeni modele pürüzsüz bir fonksiyon olarak dahil edilmelidir.

Sonuç olarak yapılan bütün karşılaştırmalarda yatak odası, koridor ve balkon değişkenlerinin modele düzgünleştirilerek yani pürüzsüz fonksiyonlar olarak dahil edilmesine, banyo değişkeninin ise modele doğrusal bir fonksiyon olarak dahil edilmesine karar verilmiştir. Diğer bir değişle, değişkenlerin logaritmalarının ya da karelerinin alınması fiyat ile olan ilişkilerini doğrusallaştıramamıştır.

Dönüşüm yöntemleri cimrilik prensibinden dolayı çok büyük önem taşımaktadır. Ancak, karesel dönüşüm yöntemi, splayn düzeltme ile neredeyse aynı serbestlik derecesini kullanmaktadır (Keele, 2008: 123). Ayrıca, ikinci bölümde bahsedildiği gibi dönüştürme yöntemleri bazı durumlarda kullanılamamakta ve yorum zorluklarına neden olmaktadır. Özellikle, bu çalışmada olduğu gibi splayn düzeltme yöntemi kullanılarak oluşturulan semiparametrik regresyon modellerinde karesel dönüşüm ciddi bir fayda sağlamamaktadır.

5.2.6. Uygun Semiparametrik Regresyon Modelinin Belirlenmesi

Uygulanacak modele karar verilmesi amacıyla, ilk olarak istatistiksel olarak anlamlı olan değişkenler ve fiyat değişkeni ile doğrusal olmayan ilişkisi bulunan değişkenler belirlenmiştir. Daha sonra ise, ilgili değişkenler için logaritmik modeller ve karesel modellerin düzgünleştirmeye karşı üstün olup olmadığı test edilmiştir. Bütün işlemler

sonucunda, cephe değişkeninin istatistiksel olarak anlamsız olduğu, yatak odası, koridor ve balkon değişkenlerinin düzgünleştirilerek diğer bir ifade ile pürüzsüz bir fonksiyon olarak modele dahil edilmesi gerektiği, banyo ve salon değişkenin fiyat değişkeni ile doğrusal bir ilişki içerisinde olduğu saptanmıştır. Tablo 5.2'de görüldüğü üzere su deposu değişkeninin 0.10 anlam düzeyinde istatistiksel olarak anlamlı olduğu saptanmıştır. Uygulanacak en son model, bu değişkenin modelde bulunup bulunmaması konusunda yardımcı olacaktır.

Tablo 5. 3: Uygulanmasına Karar verilen Semiparametrik Regresyon Modeli

Değişkenler	*Parametre Tahmini*	*Standart Hata*	*t -İstatistiği*	*Olasılık*
Sabit	133097	44313	3.004	0.00387 **
Salon	3026	1230	2.460	0.01678 *
Oda Sayısı 3	69327	14226	4.873	8.24e-06 ***
Oda Sayısı 4	140149	18993	7.379	5.31e-10 ***
Oda Sayısı 5	237662	28605	8.308	1.35e-11 ***
Banyo	14611	5199	2.810	0.00665 **
Su Deposu	-23276	18377	-1.267	0.21014
Bileşen		*F -İstatistiği*	*Olasılık*	
Yatak Odası		4.045	0.00315 **	
Koridor		4.655	0.00242 **	
Balkon		40.700	< 2e-16 ***	
R^2(adj) = 0.953		Açıklanan Sapma=%96.4	GCV Değeri = 2.1761e+09	

*: Parametrelerin 0.05 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

**: Parametrelerin 0.01 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

***: Parametrelerin 0.001 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

Bu aşamadan sonra modellerin gösterimi için değişkenler, aşağıda gösterilen kısaltmalarla ifade edilecektir.

Salon: SL, Oda Sayısı 3: OD3, Oda Sayısı 4: OD4, Oda Sayısı 5: OD5, Banyo: BY, Su deposu: SD, Yatak Odası: YO, Koridor: KR, Balkon: BN, Fiyat: FY

Tablo 5.3'de sonuçları görülen semiparametrik regresyon modeli eşitlik 5.1'de gösterilmiştir.

$$\begin{aligned} FY = {} & 133097 + 69327OD3 + 140149OD4 + 237662OD5 + 3026SL + 14611BY \\ & - 23276SD + s(YO) + s(KR) + s(BN) + \varepsilon \end{aligned} \quad (5.1)$$

Eşitlik 5.1, parametrik ve parametrik olmayan olarak iki bölümden oluşmaktadır. Bu bölümler için katsayı yorumları ve çıkarımlar ayrı yöntemlerle yapılmaktadır. Semiparametrik regresyon modelinin parametrik bölümü için yorumlar ve çıkarsamalar doğrusal regresyon modelleri aynı olmaktadır. Parametrik olmayan bölüm için yorumlar grafik yardımı ile çıkarsamalar ise F testi yardımı ile yapılmaktadır. Dördüncü bölüm'de bu konu ile ilgili açıkmalardan ayrıntılı bir şekilde bahsedilmiştir. Tablo 5.1 incelenirken su deposu değişkeninin modele dahil edilip edilemeyeceği en son modele bırakılmıştı. Tablo 5.3.

incelendiğinde su deposu değişkenine ilişkin olasılık değeri 0.21014 olarak bulunmuştur. Bu olasılık değerine göre su deposu değişkeni istatistiksel olarak anlamlı bulunmamaktadır. Bu nedenden dolayı su deposu değişkeni modelden çıkartılarak yeni bir model oluşturmalıdır.

Tablo 5. 4: Su Deposu Değişkeni Çıkartıldıktan Sonra Oluşturulan Semiparametrik Regresyon Modeli (Son Model)

Değişkenler	*Parametre Tahmini*	*Standart Hata*	*t -İstatistiği*	*Olasılık*
Sabit	116199	41644	2.790	0.00709 **
Salon	3004	1206	2.490	0.01561 *
Oda Sayısı 3	74093	14225	5.209	2.55e-06 ***
Oda Sayısı 4	148210	17672	8.387	1.24e-11 ***
Oda Sayısı 5	238852	28746	8.309	1.67e-11 ***
Banyo	13713	5190	2.642	0.01054 *
Bileşen		*F - İstatistiği*	*Olasılık*	
Yatak Odası		4.754	0.000366 ***	
Koridor		3.913	0.002349 **	
Balkon		7.466	< 2e-16 ***	
R^2(adj) = 0.955		Açıklanan Sapma= %96.7	GCV Değeri = 2.1324e+09	

*: Parametrelerin 0.05 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

**: Parametrelerin 0.01 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

***: Parametrelerin 0.001 anlam düzeyinde istatistiksel olarak önemli olduğunu ifade etmektedir.

Tablo 5.4'de, (5.1) modelinden su deposu değişkeninin çıkartılmasıyla oluşturulan semiparametrik regresyon modelinin sonuçları görülmektedir.

Tablo 5.4'de sonuçları görülen semiparametrik regresyon modeli eşitlik 5.2'de gösterildiği gibidir.

$$FY = 116199 + 74093OD3 + 148210OD4 + 238852OD5 + 3004SL + 13713BY + s(YO) + s(KR) + s(BN) + \varepsilon \quad (5.2)$$

Eşitlik (5.2)'de oda sayısı, salon ve balkon parametrik olmayan değişken olarak ele alınmış, diğer değişkenler ise parametrik değişken olarak ele alınmıştır. Bilinği üzere oda sayısı kukla değişken olarak ele alımıştır. Oda sayısı kukla değişkeni için referans kategori ise, oda sayısı 2 olarak alınmıştır.

Tablo 5.4 veya eşitlik (5.2) incelendiğinde, değişkenlerin dairelerin satış fiyatı üzerindeki etkisi aşağıda görüldüğü gibidir.

- Diğer değişkenlerin etkisi sabit tutulduğunda, salon metrekaresindeki bir birimlik artış, site içerisindeki dairelerin satış fiyatını ortalama 3004 TL arttırmaktadır.

- Diğer değişkenlerin etkisi sabit tutulduğunda, Banyo metrekaresindeki bir birimlik artış, site içerisindeki dairelerin satış fiyatını ortalama 13713 TL arttırmaktadır.

- Diğer değişkenlerin etkisi sabit tutulduğunda, evin üç odalı olması iki odalı olmasına göre, site içerisindeki dairelerin satış fiyatını ortalama 74093 TL arttırmaktadır.

- Diğer değişkenlerin etkisi sabit tutulduğunda, evin dört odalı olması iki odalı olmasına göre, site içerisindeki dairelerin satış fiyatını ortalama 148210 TL arttırmaktadır.

- Diğer değişkenlerin etkisi sabit tutulduğunda, evin beş odalı olması iki odalı olmasına göre, site içerisindeki dairelerin satış fiyatını ortalama 238852 TL arttırmaktadır.

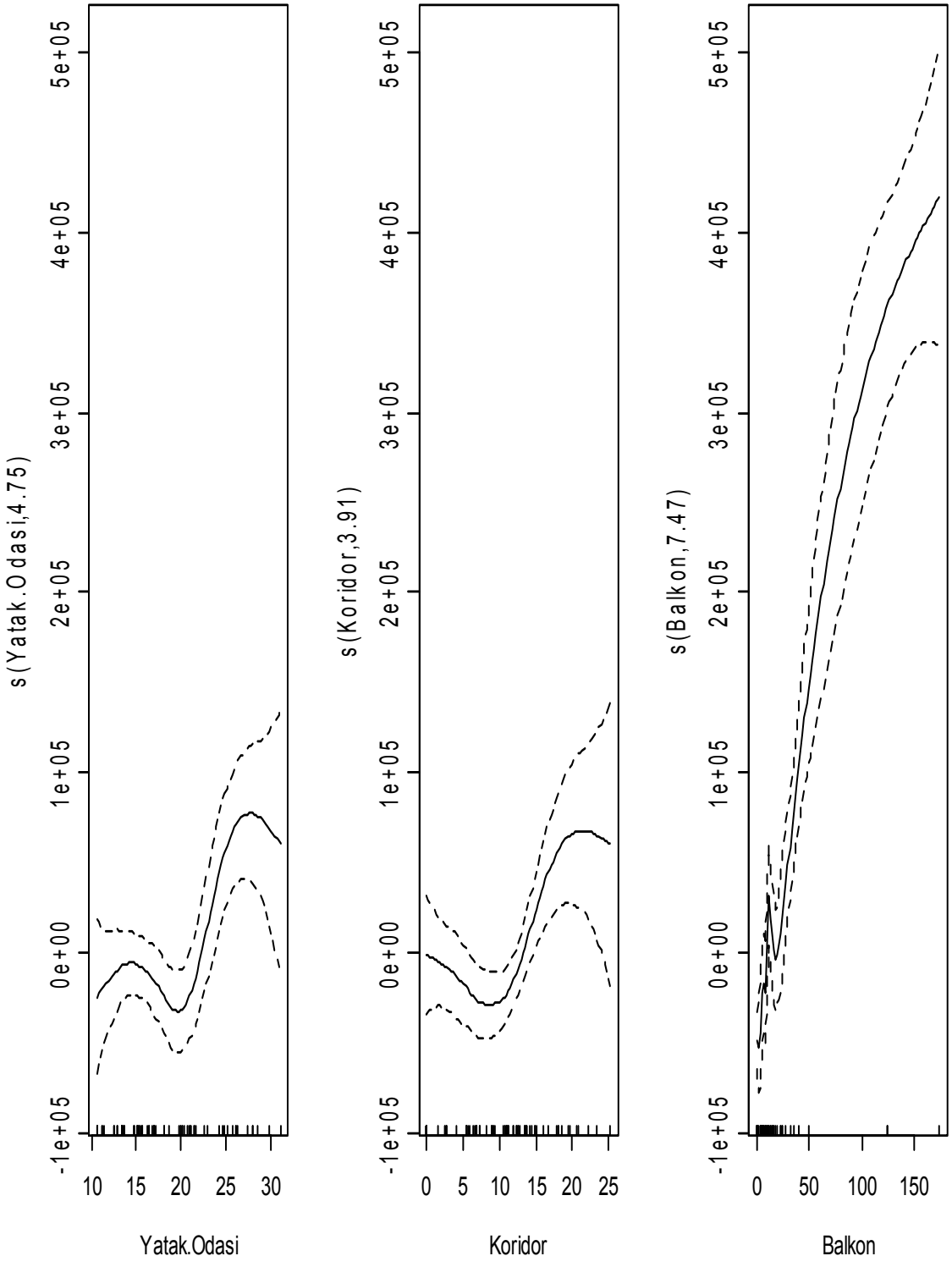
Görüldüğü gibi, tüm değişkenlerle site içerisindeki dairelerin satış fiyatları arasında pozitif yönlü bir ilişki vardır. Ek olarak, Parametrik regresyon modellerinde olduğu gibi semiparametrik regresyon modellerinde de sabit terimin yorumlanması büyük önem taşımamaktadır.

(5.2) eşitliğindeki modelin parametrik olmayan bölümünde bulunan değişkenlerin yorumları ise, model sonucunda oluşan grafikler yardımıyla yapılmaktadır. Düzgünleştirilen fonksiyonların grafikleri şekil 5.3'de görülmektedir.

Şekil 5.3'de sırasıyla yatak odası, koridor ve balkon değişkenlerinin (5.2) modelinden elde edilen grafikleri görülmektedir. Semiparametrik regresyon modellerinde parametrik bölümün yorumlanması sedece grafik yoluyla yapılmaktadır. Bu durumda, sonuçlar yaklaşık olarak elde edilir. Örneğin, yatak odası 20 metrekare civarında olduğunda, site içerisindeki dairelerin satış fiyatı en düşük, 27-28 metrekare civarında olduğunda ise en yüksek olmaktadır. Bu durumda, yatak odası metrekaresi 17 olduğunda, satış fiyatının ne olacağı, yaklaşık olarak bilinmektedir. Koridor ve balkon metrekareleri içinde grafikler aynı şekilde yorumlanır.

Aynı değişkenlerle, parametrik bir regresyon modeli olan çoklu doğrusal regresyon modeli kurulduğunda, Şekil 5.3'de grafikleri gösterilen değişkenlerin dairelerin satış fiyatları ile doğrusal bir ilişkisi olduğu varsayılacaktı. Bu durumda, yatak odası değişkeninin metrekaresi artıkça satış fiyatınında artması beklenecekti. Şekil 5.3'de görüldüğü gibi üç değişken içinde dalgalanmalar söz konusu olmaktadır. Yani, bu ilişkilerin doğrusal olduğunu

kabul etmek, yanlış sonuçlara ulaşılmasına ve dolayısıyla yanlış kararlar verilmesine yol açacaktır.



Şekil 5. 3: Eşitlik 5.2'de Görülen Semiparametrik Regresyon Modelinin Parametrik Olmayan Bileşenlerinin Grafikleri

5.2.6.1. Son Semiparametrik Regresyon Modeline İlişkin Çıkarımlar

Dördüncü bölümde açıklandığı üzere, semiparametrik regresyon modeli parametrik ve parametrik olmayan, olmak üzere iki bölümden oluşmaktadır. Semiparametrik regresyon modelleri için yapılan çıkarımlar iki bölüm için farklı yöntemle yapılmaktadır.

Eşitlik 5.2 ve sonuçları Tablo 5.4'de gösterilen model, uygulanmasına karar verilen son modeli ifade etmektedir. Öncelikle, bu modelin R ortamında yazılan bir programla tahmin edilen parametrik katsayılarının anlamlı olup olmadığı incelenmelidir. Tablo 5.4 incelendiğinde, parametrik katsayılara ilişkin tahminler, t-istatistikleri ve olasılık değerleri görülmektedir. Parametrik regresyonda olduğu gibi t-istatistikleri, parametre tahminlerinin standart hatalara bölünmesi ile elde edilmektedir. t-istatistiklerinin anlamlı olması, incelenen bağımsız değişkenin site içerisindeki dairelerin sayış fiyatları üzerindeki etkisinin anlamlı olduğunu ifade eder. Bu testi gerçekleştirmek için sabit katsayı dahil bütün katsayılar için,

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

hipotezleri test edilmelidir. H_0 hipotezinin kabul edilmesi, katsayının istatistiksel açıdan anlamlı olmadığını ifade eder. Bu durumda değişkenlerin etkisinin yorumlanması anlamsız olacaktır. Örneğin, daha önce yorumlandığı gibi diğer değişkenlerin etkisi sabit tutulduğunda, salon metrekaresindeki bir birimlik artış, site içerisindeki dairelerin satış fiyatını ortalama 3004 TL arttırmaktadır. Bu şekildeki bir yorum, salon değişkenine ait katsayının istatistiksel olarak anlamsız çıkması durumunda yapılamaz.

Tablo 5.4 incelendiğinde, Tüm parametrik bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini gösteren parametrik katsayıların istatistiksel olarak anlamlı oldukları görülmektedir. Bu durumda, semiparametrik regresyon modelinin parametrik bölümü için yapılmış olan bütün yorumların geçerli olduğu sonucuna varılır.

Semiparametrik regresyon modelinin parametrik olmayan bileşenine ait çıkarımlar, dördüncü bölümde analtıldığı gibi, F-testi ve güven bantları yardımıyla yapılır. F-testi, semiparametrik modele dahil edilecek fonksiyonun, doğrusal olarak ya da pürüzsüz bir fonksiyon olarak mı dahil edilmesinin uygunluğunu her değişken için,

$$H_0 : \text{Dogrusal Fonksiyon}$$
$$H_a : \text{Pürüzsüz Fonksiyon}[s(x_i)]$$

hipotezleri ile test eder. H_0 hipotezinin kabul edilmesi, ilgili değişkenin doğrusal bir fonksiyon olarak modele dahil edilmesi gerektiğini ifade eder. Bu doğrultuda Tablo 5.4 incelendiğinde; yatak odası, balkon ve koridor (metrekareleri) değişkenlerinin olasılık değerleri 0.001 anlam düzeyinde bile istatistiksel olarak önemli bulunmuştur. Dolayısıyla, söz konusu değişkenlerin semiparametrik regresyon modeline düzgünleştirilerek (pürüzsüz bir fonksiyon olarak) dahil edilmesi gerektiğini ifade etmektedir.

Pürüzsüz fonksiyonların anlamlılığı güven bantları ile de incelenebilmektedir. İlgili değişkenler için güven bantları Şekil 5.3'de gösterilmektedir. İncelendiği gibi bu grafikler, site içerisindeki dairelerin satış fiyatlarının ilgili değişkene göre değişimini göstermektedir. Şekil 5.3'de üç değişken için kesikli çizgiler güven bantlarını ifade etmektedir. Güven bantlarının, geniş olmayan bir aralıkta ve düz çizgi ile neredeyse aynı doğrultuda olması istatistiksel olarak anlamlılığı ve yapılan tahminlerin kalitesini ifade eder.

Şekil 5.3'de yatak odası metrekaresinin değişim eğrisi ile güven bantları yüksek metrekarelere kadar aynı doğrultuda ve dar bir aralıktadır. Yüksek metrekarelere gelindiğinde üst sınır yön değiştirmektedir. Ancak bu durum fonksiyonun anlamlılığını bozmamaktadır. Sadece düşük metrekarelerde yapılan tahminlerin çok daha güvenilir olduğunu ifade etmektedir. Koridor metrekaresinin ve balkon metrekaresinin grafiği incelendiğinde aynı durum söz konusu olmaktadır. Ancak, balkon metrekaresi grafiğinde diğer değişkenlerin güven bantlarına göre değişim eğrisinden daha az bir sapma görülmektedir.

Yapılan incelemeler sonucunda semiparametrik modelinin, parametrik bölümündeki değişkenlerin ve parametrik olmayan bölümündeki pürüzsüz fonksiyonların istatistiksel olarak anlamlı olduğu sonucuna ulaşılır.

Bu aşamadan sonra, semiparametrik regresyon modelindeki, bağımlı değişkenin bağımsız değişkenlerle yeterince açıklanıp açıklanamadığı incelenmiştir. Bu değişimin ölçüsü, R^2(adj) olarak ifade edilen düzeltilmiş belirlilik katsayısı ve açıklanan sapmadır. Tablo 5.4 incelendiğinde, düzeltilmiş belirlilik katsayısının %95.5 olduğu görülmektedir. Bu durumda, site içerisindeki dairelerin sayış fiyatındaki değişimlerin %95.5'i modeldeki bağımsız değişkenlerce açıklanabildiği sonucuna ulaşılır. Açıklanan sapma değeri %96.7 olarak hesaplanmıştır. Açıklanan sapma değerinin yüksek olması modelin tahmin kalitesinin oldukça yüksek olduğunu ifade etmektedir.

Son olarak, şu ana kadar oluşturulan oluşturulan dört semiparametrik regresyon modelinin düzeltilmiş belirlilik katsayıları ve açıklanan sapma değerleri incelenmiştir ve Tablo 5.5 oluşturulmuştur.

Tablo 5. 5: Semiparametrik Regresyon Modellerinin Düzeltilmiş Belirlilik Katsayıları ve Açıklanan Sapma Değerleri

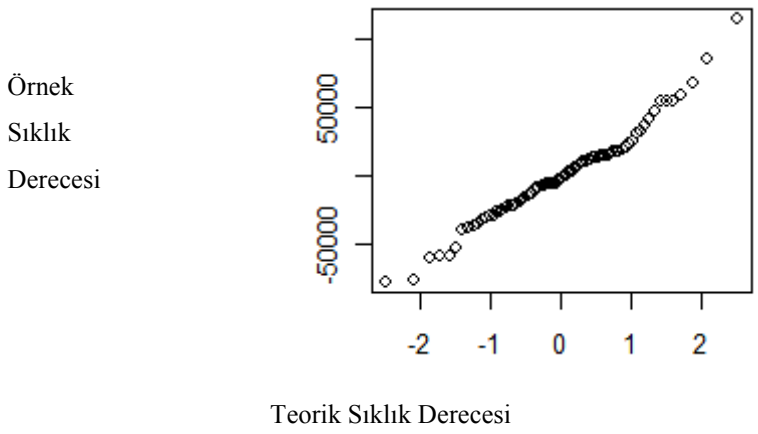
1.Model	R^2(adj)= 0.954	Açıklanan Sapma= %96.5
2.Model	R^2(adj)= 0.955	Açıklanan Sapma= %96.5
3.Model	R^2(adj) = 0.953	Açıklanan Sapma= %96.4
4.Model	R^2(adj) = 0.955	Açıklanan Sapma= %96.7

Tablo 5.5'de görülen 1. Model cephe değişkenini içeren, 2. Model (temel model) cephe değişkeni çıkartılmış ancak bütün sürekli değişkenlerin düzgünleştirildiği, 3. Model sadece yatak odası, balkon ve koridor değişkenlerin düzgünleştirildiği ancak istatistiksel olarak anlamsız olan su deposu değişkenini içeren, 4. Model ise su deposunun çıkartılmasıyla oluşan modelleri ifade etmektedir. Açıklandığı üzere birinci ve ikinci modeller, değişkenlerin yapısını belirlemek için oluşturulmuş modellerdir. İstatistiksel olarak anlamsız bir değişkenin modelden çıkartılması modelin düzeltilmiş belirlilik katsayısını arttırmaktadır. 3. modelde istatistiksel anlamsız olan bir değişkenin modelde bulunması düzeltilmiş belirlilik katsayısını az bir düşüşle etkilemektedir. Son model ise düzeltilmiş belirlilik katsayısının ve açıklanan sapma değerinin birlikte en yüksek olduğu modeldir.

5.2.6.2. Son Semiparametrik Regresyon Modeline İlişkin Varsayımların İncelenmesi

Bölümde anlatıldığı üzere, semiparametrik regresyon modellerinde tahmin, çıkarım yapabilmek ve tutarlı tahminler gerçekleştirmek için parametrik regresyon modellerinde olduğu gibi bazı varsayımların sağlanması gerekir. Semiparametrik regresyon modellerinde, varsayımlardan sapmalar gerçekleşebilmektedir. Ancak, bu modellerde amaç varsayımlardan sapmaların göz ardı edilebilecek düzeyde olmasıdır.

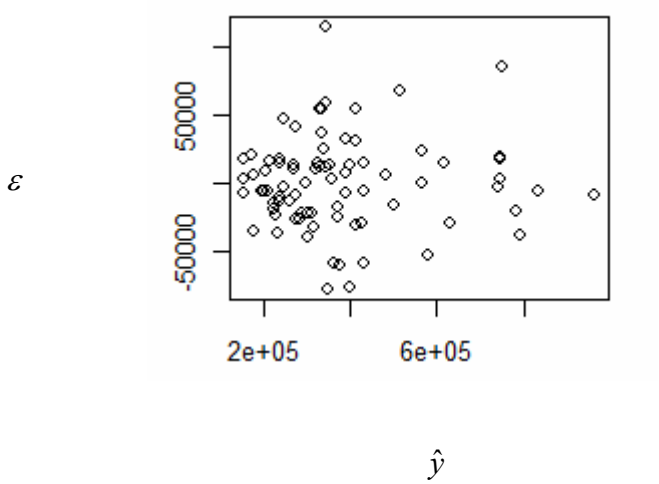
Karar verilen son semiparametrik regresyon modelinin varsayımları R programında "gam.check" komutu ile incelenmiştir. Öncelikle, bağımlı değişkenin normallik varsayımı Şekil 5.4 yardımıyla incelenmiştir.



Şekil 5. 4: Normallik Varsayımı İçin Oluşturulan Q-Q Grafiği

Şekil 5.4 incelendiğinde bağımlı değişkeninin dağılımının yaklaşık normal dağılımdan geldiği kabul edilebilmektedir.

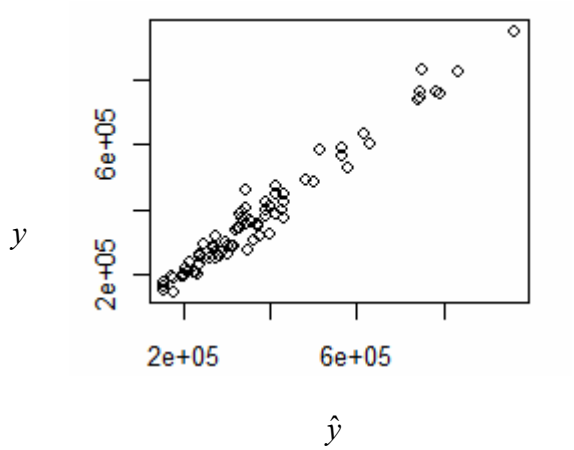
İkinci olarak hata terimlerinin rassallığı varsayımı Şekil 5.5 yardımıyla incelenmiştir.



Şekil 5. 5: Tahmin Değerleri ve Artıklara İlişkin Grafik

Bu amaçla tahmin değerleri ($\hat{y}$) ve artıklar arasındaki ilişki incelenmiştir. Bu grafikteki noktaların tesadüfi olarak dağıldığı ve varsayımın gerçekleştiği sonucuna ulaşılmaktadır.

Son olarak modelin genel değerlendirmesini yapmak amacıyla, gerçek değerlere (y) karşı tahmin değerlerinin ($\hat{y}$) grafiği incelenmiş ve Şekil 5.6'da gösterilmiştir.



Şekil 5. 6: Gerçek Değerlere Karşı Tahmin Değerlerinin Grafiği

Şekil 5.6 incelendiğinde, uygulanan semiparametrik regresyon modeli küçük değerlerde daha iyi tahmin yapmaktadır. Genel olarak modelin güvenilir tahminler yaptığı sonucuna ulaşılmaktadır.

6. SONUÇLAR

Parametrik ya da parametrik olmayan regresyon analizinde amaç, en iyi modeli bulmak ve bu modellerle güvenilir tahminler elde etmektir. Parametrik regresyon analizi, bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkilerin doğrusal ya da ikinci bölümde incelenen üstel dönüştürme yöntemleri ile doğrusallaştırılabileceğini ve parametrik fonksiyonların kullanılabileceğini varsayar. Parametrik olmayan regresyon analizi ise varsayım gerektirmeden, veriye en uygun eğriyi tahmin eder. Ancak, birden fazla bağımsız değişken olması durumunda çok boyutluluğun yarattığı sıkıntı nedeniyle parametrik olmayan regresyon analizi özellikle yorum aşamasında kullanışsız olmaktadır. Bu durumda dördüncü bölümde incelenen toplamsal modeller ve semiparametrik modeller devreye girmektedir.

Toplamsal modeller, bağımlı değişkenle bağımsız değişkenler arasındaki ilişkinin doğrusal olmadığını ve düzgünleştirilerek modele dahil edilmesi gerektiğini ifade eder. Ayrıca, toplamsallık varsayımı modelin yorumlanması aşamasında büyük kolaylık sağlamaktadır. Toplamsal modellerde bütün bağımsız değişkenler modele düzgünleştirilerek dahil edilir. Ancak, bazı değişkenler bağımlı değişken ile doğrusal ilişki içerisinde iken, bazı değişkenler doğrusal olmayan ya da parametrik olmayan ilişki içerisinde bulunabilirler. Böyle bir durumda parametrik ve parametrik olmayan olmak üzere iki bölümden oluşan semiparametrik regresyon analizi devreye girer.

Semiparametrik regresyon analizinde önemli olan, modele düzgünleştirilerek dahil edilecek değişkenlerin, hangi düzgünleştirme yöntemi ile düzgünleştirileceği ve buna bağlı olarak düzeltme parametresinin seçimidir. İkinci bölümde düzeltme teknikleri ayrıntılı bir şekilde incelenmiş ve incelemeler sonucu splayn düzeltme tekniğinin diğer tekniklere göre daha iyi tahminler elde ettiği görülmüştür.

Splayn düzeltme tekniğinde en önemli adım, en iyi düzeltme parametresi olan λ'yı seçmektir. λ parametresinin seçimi için geliştirilen yöntemler üçüncü bölüm olan otomatik düzeltme başlığı altında ayrıntılarıyla incenmiştir. Bu çalışmanın konusu olan semiparametrik regresyon modellerinde kullanılan splayn düzeltme tekniğinde, otomatik seçim, (genelleştirilmiş çapraz geçerlilik kriterinde) en güvenilir sonuçlara ulaşılmasını sağlamaktadır.

Bu çalışmanın uygulama konusu olan, site içerisindeki dairelerin satış fiyatlarını etkileyen özelliklerin incelenmesi beşinci bölümde ayrıntılarıyla yer almıştır. Uygulamada

öncelikle, modele düzgünleştirilerek dahil edilmesi gereken değişkenler belirlenmiştir. Modelde kukla değişkenler olduğundan semiparametrik regresyon analizinin uygulanması uygun bulunmuştur. Bütün değişkenlerin, düzgünleştirilerek modele dahil edilmesi durumunda toplamsal modeller uygulanmalıdır.

Semiparametrik regresyon analizinin önemli bir özelliği, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkilerin şeklini istatistiksel testlerle belirleyebilmesidir. Yani, bir değişkenin düzgünleştirilerek, doğrusal ya da dönüştürme yöntemleri ile doğrusallaştırarak modele dahil edileceğine karar verir. Modellerin birbirlerine olan üstünlüğünü test eder. Semiparametrik regresyon analizi ile modelleme yapılmasa bile, bu analiz tekniğini kullanarak değişkenlerin yapılarının belirlenmesi en iyi tahminlere ulaşılmasını sağlayacaktır. Uygulamada, istatistiksel olarak anlamsız olan değişkenler belirlenmiş, semiparametrik regresyon modelleri ile doğrusal, logaritmik ve karesel dönüşümle oluşturulan modeller karşılaştırılmıştır. Uygulanacak son modele karar verilmesi aşamasında su deposu değişkenin istatistiksel olarak anlamsız olduğu tespit edilmiş ve modelden çıkartılarak yeni bir model oluşturulmuştur. Sonuç olarak, salon, banyo metrekaresi değişkenlerinin ve oda sayısı değişkenlerinin semiparametrik modelin parametrik bölümünde, diğer değişkenlerin ise modelin parametrik olmayan bölümünde yer aldığı model en iyi model olarak seçilmiştir.

Son semiparametrik modelin, parametrik bölümündeki değişkenlerin anlamlılığı t testleri ile, parametrik olmayan bölümdeki değişkenlerin yani düzgünleştirilerek modele dahil edilmiş değişkenlerin anlamlılıkları ise, F testleri ve güven bantlarıyla incelenmiştir.

Bu aşamadan sonra, oluşturulan dört semiparametrik regresyon modelinin düzeltilmiş belirlilik katsayıları ve açıklanan sapma değerleri karşılaştırılmış ve son olarak karar verilen modelin en iyi model olduğuna karar verilmiştir. Karar verilen en iyi modelde, semiparametrik regresyon varsayımları incelenmiş ve tahmini etkilemeyecek düzeyde bazı sapmaların olduğu grafiklerle belirlenmiştir.

Sonuç olarak, semiparametrik regresyon analizi parametrik regresyon modellerinin tahmin ve yorum kolaylığını ve parametrik olmayan regresyon analizinin esneklikliğini birleştirir. Ayrıca, değişkenlerin modele nasıl dahil edilmesi gerektiğini belirlediğinden diğer regresyon modellerine göre ciddi derecede üstünlük sağlar.

KAYNAKÇA

Akay, Kadri, 2007. Genelleştirilmiş Lineer Modeller Yardımıyla Karma Denemelerin Analizi, Doktora Tezi, Marmara Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Aydın, D., 2007. A Comparision of The Nonparametric Regression Models Using Smoothing and Kernel Regression, Proceeding of Worl Academy Of Science, Engineering and Technology, Volume: 26.

Aydın, Dursun, 2005. Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı İle Tahmin ve Çıkarsamalar, Doktora Tezi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

Berk, Richard A., (2006). "Regression Analysis: A Constructive Critique'', Sage Publications, London.

Box, G.E.P. ve Jenkins, G.M., (1970). "Time Series Analysis: Forecasting and Control'', Holden-Day, London.

Çağlayan, Ebru, 2002. Yarı Parametrik Regresyon Modelleri ile Yaşam Boyu Sürekli Gelir Hipotezinin Türkiye Uygulaması, Doktora Tezi, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

Çevrimçi: http://r.789695.n4.nabble.com/Why-there-is-no-p-value-from-likelihood-ratio-test-using-anova-in-GAM-model-fitting-td888781.html (29.03.2010).

Çevrimçi: http://satilikdairearіyorum.blogcu.com/konut-ev-daire-alirken-dikkat-edilmesi-gerekenler/5189921 (10.03.2010).

Delis, M. ve Papanikolaou, N., 2009. Determinants of Bank Efficiency: Evidence From a Semiparametric Methodology, Managerial Finance 35, 260-275.

Eilers, Paul HC., Brian D. Marx., 1996. Flexible Smoothing with B-splines and penalties, Statistical Science 11, 98-102.

Engle, R.F., Granger, C.W.J., Rice, J.A. ve Weiss, A., 1986. Semiparametric Estimates of the Relation Between Weather and Electiricity Sales, Journal of American Statistical Assosiation 81, 461-170.

Erar, Aydın, 2006 "Bağlanım (Regresyon) Çözümlemesi'', Yayınlanmamış Ders Notları, M.S.G.S. İstatistik Bölümü, İstanbul.

Eubank, R.l., (1999). "Nonparametric Regression and Spline Smoothing'', Marcel Dekker İnc., New York.

Fox, John, (2008). " Applied Regression Analysis and Generalized Linear Models'', Sage Publications, California.

Fox, John, 2000. Multiple and Generalized Nonparametric Regression, a Sage University Paper, Thousand Oaks.

Green, P.J. ve Yandell, B., 1985. Semiparametric Generalized Linear Models, Lecture Notes on Statistics 32, 44-55.

Hardle W., Müller M., Sperlich S. ve Werwatz A., (2004). "Nonparametric and Semiparametrik Models, Springer, Berlin.

Hardle, W., (1991). “Applied Nonparametric Regression’’, Cambridge University Press, Economic Society Monographs, No:19, Cambridge.

Hardle, W., Hall ve Marron., 1998. How Far are Automatically Choosen Regression Smoothing Parameters from their Optimum, American Journal of Political Science 38, 601-624.

Hastie, T. ve Tıbshıranı, RJ., (1999). “Generalized Additive Models, Chapman & Hall, London.

Hastie, Trevor, Robert Tibshirani ve Jerome Friedman, (2003). “The Elements of Statistical Learning: Data Mining, İnference and Prediction’’, 3.Edition, Springer Verlag, New York.

Jerome, Gronniger, 2006. A semiparametric Analyisis of the Relationship of Body Mass Index to Mortality, American Journal of Public Health 96, 173-178.

John D., Tobias J., 2001. Nonparametric Density and Regression Estimation, Journal of Economic Perspectives 4, 11-28.

Keele, Luke, (2008). “Semiparametric Regression For The Social Sciences’’, John Wiley & Sons, U.S.A.

Kim, I, Cohen, N.D., ve Carroll, R.J., 2003. Semiparametric Regression Splines in Matched Case-Control Studies 59, 1158-1169.

Lee, Kou Chen, 1990. Avaoiding Misspecifications and İmproving Efficiency in Hedonic and, Consumption Models: Applications of Semiparametric Method, Doktora Tezi, London School of Economics and Political Sciense, London.

Lınn, X. ve Carroll, R.J., 2001. Semiparametric Regression for Clustered Data, Biometrica 88, 1179-1185.

Li, Quan ve Rafeal Reuveny, 2006. Democracy and Enviromental Degredation, İnternational Studies Quarterly 50, 935- 956.

Loader, Clive, (1999). “ Local Regression and likelihood’’, Springer, New York.

Nadarya, E.A., 1964. On Estimating Regression, Theor. Probab. Appl. 9, 141-142.

Robinson, M. P., 1988. Root N-Consistent Semiparametric Regression, Econometrica 56, 931-954.

Ruppert D., Wand, M.P. ve Carroll R.J., (2003). “Semiparametric Regression’’, Cambridge University Press.

Speckman, P. E., 1988 . Regression Analysis for Partially Linear Models, Journal of American Statistical Assosiation 50, 413-436.

Stone, C.J., 1986 . Comment: Generalized Additive Models, Statistical Science 2, 312-314.

Wahba, G. ve Wold, S., 1975. A Completely Automatic French Curve: Fitting Spline Fuction By Cross-Validation, Communication in Statistics, 4, 1-17.

Wahba, G., (1990). “Spline Models of Observational Data’’, University Of Winconsin At Madison, Pensilvenya.

Weisberg, Sanford, (2005). “Applied Linear Regression’’, 3.Edition, Wiley-İnterScience, New Jersey.

EKLER

EK-A: Düzeltme Tekniklerinin Karşılaştırılmasında Kullanılan R kodları

```
# Çalıştırılması Gereken Kütüphaneler
library(SemiPar)
library(splines)
# Çok Dalgalı Bir Fonksiyonel Fonksiyon Oluşturulması
trans<-function(x) {sin(2*pi*x^2)^3}
x<-seq(0,2,by=.01)
y<-trans(x)+.2*rnorm(201)
plot(x,y, pch=1, type="o", lty=1, xlab="X", ylab="Y")
#  Gerçek İlişkinin Grafiğinin Çizilmesi
matplot(x, cbind(y, trans(x)),
  pch=1, type="pl", lty=1,
  xlab="X", ylab="Y")
  loess <- loess(y ~ x, span=0.1)
xhatloess <- x
yhatloess <- fitted(loess)
# Karşılaştırma Grafikleri
par(mfrow = c(2,2))
# Loess
matplot(x, cbind(y, trans(x)),
   type="pl", lty=2, col=1, pch="",
  xlab="X", ylab="Y", main = "Loess", bty = "l")
lines(xhatloess, yhatloess, lwd=1)
# Doğal Kübik B-splayn
matplot(x, cbind(y, trans(x)),
  pch="", type="pl", lty=2, col=1,
  xlab="X", ylab="Y", main = "Natual Cubic B-Spline", bty = "l")
sp1<-lm(y~ns(x, df=15))
lines(x,sp1$fit, lwd=1)
# Lowess
matplot(x, cbind(y, trans(x)),
  pch="", type="pl", lty=2, col=1,
  xlab="X", ylab="Y", main="Lowess", bty = "l")
lines(lowess(x,y, f = 0.05), lwd=1)
```

EK-A: Devam

#Splayn Düzeltme

```
matplot(x, cbind(y, trans(x)),
   pch="", type="pl", lty=2, col=1,
   xlab="X", ylab="Y", main="Smoothing Spline", bty = "l")
fit <- spm(y ~ f(x))
lines(fit, se=FALSE, lwd=1
```

EK- B: Otomatik Düzeltme Tekniği Simülasyon Çalışması R Kodları

```
# Çalıştırılması Gereken Kütüphaneler
library(splines)
library(mgcv)
library(SemiPar)
# Fonksiyonun Oluşturulması
trans<-function(x) {cos(4*exp(x))^4}
x<-seq(0,1,by=.001)
y<- trans(x) + rnorm(1001)
# Değişkenler Arasındaki Gerçek İlişkinin Grafiği
  matplot(x, cbind(y, trans(x)),
 pch=".", type="pl", lty=1, lwd=1, col =1,
  xlab="X", ylab="Y", main="", bty = "l")
# Karşılaştırma Grafikleri
par(mfrow = c(2,1))
plot(x,y, pch=".", main = "Lowess - 6 Farklı Aralık Değeri", cex=0.85, bty="l")
lines(lowess(x,y, f = 0.1), lwd=1, lty=6)
lines(lowess(x,y, f = 0.2), lwd=1, lty=2)
lines(lowess(x,y, f = 0.3), lwd=1, lty=3)
lines(lowess(x,y, f = 0.4), lwd=1, lty=4)
lines(lowess(x,y, f = 0.5), lwd=1, lty=5)
lines(lowess(x,y, f = 0.6), lwd=1, lty=1)
fit <- spm(y ~ f(x))
plot(x,y, pch=".", main = "Splayn Düzeltme –Otomatik Düzeltme", cex=0.85, bty="l")
lines(fit, se=FALSE, lwd=1)
```

EK- C: Site İçerisindeki Dairelerin Satış Fiyatını Etkileyen Özelliklerin Semiparametrik Regresyon Modeli İle İncelenmesine İlşkin R Kodları

 # Çalıştırılması Gereken Kütüphaneler

```
library(foreign)
library(mgcv)
```

 # Veri setinin R Programına Aktarılması

```
site <- read.table("c:/site2.csv", header=TRUE, sep=",",)
site
attach(site)
```

 # Kukla Değişkenlerin Belirlenmesi

```
cephef <- (factor(cephe))
cephef
odaf <- (factor(odasay))
odaf
```

 # Değişkenlerin Orijinal Grafiklerinin Çizilmesi

```
par(mfrow=c(2,3))
plot(mod.1, select=1, rug=FALSE, se=TRUE, ylab="Satis Fiyati", xlab="Salon", residual=FALSE, bty="l", shift=301749)
points(salon, fiyat, pch=".", cex=1.75)
plot(mod.1, select=2, rug=FALSE, se=TRUE, ylab="Satis Fiyati", xlab="Yatak Odasi", residual=FALSE, bty="l", shift=301749)
points(ebyatakod , fiyat, pch=".", cex=1.75)
plot(mod.1, select=3, rug=FALSE, se=TRUE, ylab="Satis Fiyati", xlab="Banyo", residual=FALSE, bty="l", shift=301749)
points(banyo, fiyat, pch=".", cex=1.75)
plot(mod.1, select=4, rug=FALSE, se=TRUE, ylab="Satis Fiyati", xlab="Koridor", residual=FALSE, bty="l", shift=301749)
points(koridor, fiyat, pch=".", cex=1.75)
plot(mod.1, select=5, rug=FALSE, se=TRUE, ylab="Satis Fiyati", xlab="Balkon", residual=FALSE, bty="l", shift=301749)
points(balkon, fiyat, pch=".", cex=1.75)
```

EK- C: Devam

 # Temel Semiparametrik Modelin Oluşturulması

```
base <- gam(fiyat ~ s(salon, bs="cr") + oda + cephe + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr")+ su deposu,  data=site)

summary(base)
```

 # Cephe Değişkeni Çıkartıldıktan Sonra Yeni Modelin Oluşturulması

```
mod.1 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası, bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr")+ su deposu ,data=site)

summary(mod.1)

 plot(mod.1)
```

 # Doğusal Olmayan İlişkilerin Belirlenmesi İçin Oluşturulan Modeller

```
mod.2 <- gam(fiyat ~ salon + oda + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr") + sud eposu ,data=site)

mod.3 <- gam(fiyat ~ s(salon, bs="cr") + oda + yatak odası + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.4 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası , bs="cr") + banyo+ s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.5 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ koridor + s(balkon, bs="cr") + su deposu, data=site)

mod.6 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + balkon + su deposu ,data=site)
```

 # Oluşturulan Modellerin Temel Model İle Karşılaştırılması

```
anova(mod.2, mod.1, test="Chisq")

anova(mod.3, mod.1, test="Chisq")

anova(mod.4, mod.1, test="Chisq")

anova(mod.5, mod.1, test="Chisq")

anova(mod.6, mod.1, test="Chisq")
```

 # Logaritmik Modeller

```
mod.7 <- gam(fiyat ~ log(salon)   + oda + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.8 <- gam(fiyat ~ s(salon, bs="cr") + oda + log(yatak odası) + s(banyo, bs="cr")+ s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.9 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası , bs="cr") + log(banyo)+ s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.10 <- gam(fiyat ~ s(salon, bs="cr") + odaf + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ log(koridor1) + s(balkon, bs="cr") + su deposu ,data=site)

mod.11 <- gam(fiyat ~ s(salon, bs="cr") + oda + s(yatak odası , bs="cr") + s(banyo, bs="cr")+ s(koridor, bs="cr") + log(balkon1) + su deposu ,data=site)
```

EK- C: Devam

 # Logaritmik Modellerin Temel Model İle Karşılaştırılması

anova(mod.7, mod.1, test="Chisq")

anova(mod.8, mod.1, test="Chisq")

anova(mod.9, mod.1, test="Chisq")

anova(mod.10, mod.1, test="Chisq")

anova(mod.11, mod.1, test="Chisq")

 # Karesel Modeller

mod.12 <- gam(fiyat ~ salon + I(salon^2) + s(yatak odası , bs="cr") + oda + s(banyo, bs="cr") + s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.13 <- gam(fiyat ~ s(salon, bs="cr") + yatak odası + I(yatak odası^2) + oda + s(banyo, bs="cr") + s(koridor, bs="cr") + s(balkon, bs="cr")+ su deposu ,data=site)

mod.14 <- gam(fiyat ~ s(salon, bs="cr") + s(yatak odası , bs="cr") + oda + banyo + I(banyo^2) + s(koridor, bs="cr") + s(balkon, bs="cr") + su deposu ,data=site)

mod.15 <- gam(fiyat ~ s(salon, bs="cr") + s(yatak odası , bs="cr") + oda + s(banyo, bs="cr")+ koridor + I(koridor^2) + s(balkon, bs="cr") + su deposu ,data=site)

mod.16 <- gam(fiyat ~ s(salon, bs="cr") + s(yatak odası , bs="cr") + oda + s(banyo, bs="cr")+ s(koridor, bs="cr") + balkon + I(balkon^2) + su deposu ,data=site)

 # Karesel Modellerin Modellerin Temel Model İle Karşılaştırılması

anova(mod.12, mod.1, test="Chisq")

anova(mod.13, mod.1, test="Chisq")

anova(mod.14, mod.1, test="Chisq")

anova(mod.15, mod.1, test="Chisq")

anova(mod.16, mod.1, test="Chisq")

 # Karar Verilen Semiparametrik Regresyon Modeli

mod.17 <- gam(fiyat ~ salon + oda + s(yatak odası , bs="cr") + banyo + s(koridor, bs="cr") + s(balkon, bs="cr")+ su deposu ,data=site)

summary(mod.17)

 # Karar Verilen Son Semiparametrik Regresyon Modeli (Su Deposu Değişkenin Çıkartılması)

mod.18 <- gam(fiyat ~ salon + oda + s(yatak odası , bs="cr") + banyo + s(koridor, bs="cr") + s(balkon, bs="cr") ,data=site)

summary(mod.17)

 # Varsayımların Grafiksel Olarak İncelenmesi

gam.check(mod.18)

EK-D

ÖZGEÇMİŞ

Doğum tarihi	13.11.1985
Doğum yeri	İstanbul
Lise 1999-2003	İstek Özel Acıbadem Lisesi
Lisans 2004-2008	İstanbul Ticaret Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü

Çalıştığı kurumlar

2008- İstanbul Ticaret Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü Araştırma Görevlisi