

ÇOKDEĞİŞKENLİ UYARLANABİLİR REGRESYON KESİTLERİ: MARS

Yasemin ÖZFALCI

**YÜKSEK LİSANS TEZİ
İSTATİSTİK**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

**Temmuz 2008
ANKARA**

Yasemin ÖZFALCI ÜSTÜNER tarafından hazırlanan ÇOKDEĞİŞKENLİ UYARLANABİLİR REGRESYON KESİTLERİ: MARS adlı bu tezin Yüksek Lisans tezi olarak uygun olduğunu onaylarım.

Prof. Dr. Alptekin ESİN
Tez Danışmanı, İstatistik

Bu çalışma, jürimiz tarafından oy birliği ile İstatistik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof. Dr. Hamza GAMGAM
İstatistik, Gazi Üniversitesi

Prof. Dr. Alptekin ESİN
İstatistik, Gazi Üniversitesi

Doç. Dr. Nezir KÖSE
Ekonometri, Gazi Üniversitesi

04/07/2008

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Nermin ERTAN
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

(İmza)

Yasemin ÖZFALCI

ÇOKDEĞİŞKENLİ UYARLANABİLİR REGRESYON KESİTLERİ:MARS
(Yüksek Lisans Tezi)

Yasemin ÖZFALCI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Temmuz 2008

ÖZET

Bu çalışmada, regresyon analizi hakkında bilgi verilmiş ve bilinen regresyon yöntemlerindeki kısıtlılıkları gidermek üzere geliştirilmiş olan Çokdeğişkenli Uyarlanabilir Regresyon Kesitleri– MARS (Multivariate Adaptive Regression Splines) tekniğinin temel özellikleri, konu ile ilgili yapılan çalışmalar ve uygulama adımları üzerinde durulmuş, ayrıca bu tip karmaşık algoritmaların kullanımının getireceği avantajlar açıklanmıştır. Uygulama kısmında, MARS regresyon modeli ile hipotetik olarak geliştirilen kız öğrencilere ilişkin veri seti kullanılarak MARS modeli kurulmuştur. Bu modelin özellikleri açıklanmış, regresyon denklemi oluşturulmuş ve kullanılan değişkenlerin ve bu değişkenlere ait etkileşim terimlerinin modele katkısı incelenmiştir.

Bilim Kodu : 205
Anahtar Kelimeler : MARS, temel fonksiyon, düğüm noktası, etkileşim
Sayfa Adedi : 58
Tez Yöneticisi : Prof.Dr.Alptekin ESİN

MULTIVARIATE ADAPTIVE REGRESSION SPLINES: MARS
(M.Sc.Thesis)

Yasemin ÖZFALCI

GAZİ UNIVERSITY
ART & SCIENCE INSTITUTE
July 2008

ABSTRACT

This article describes regression analysis and Multivariate Adaptive Regression Splines (MARS) development issues to eliminate regression analysis's known limited method and investigated properties of MARS's main techniques, studies, process steps also explained this kind of complex algorithm's advantages. In implementation step, MARS model built up with MARS regression and hypothetic developed girl students' dataset. Model's property explained, equation regression built up, used variables and this variables relation's terms contribution investigated.

Science Code : 205
Keywords : MARS, Basic Function, knot, interaction.
Page Count : 58
Thesis Advisor : Prof.Dr.Alptekin EŞİN

TEŞEKKÜRLER

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Prof.Dr.Alptekin ESİN'e, kıymetli tecrübelerinden ve manevi desteğinden faydalandığım Mesut AKYOL'a, ayrıca manevi desteğiyle beni yalnız bırakmayan eşim, ailem, amirlerim ve çalışma arkadaşlarıma teşekkürü borç bilirim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	viii
ŞEKİLLERİN LİSTESİ	ix
KISALTMALAR	x
1.GİRİŞ	1
2. Çokdeğişkenli Uyarlanabilir Regresyon Kesitleri (MARS)	7
2.1. Temel Fonksiyon.....	14
2.2. Adımsal İlerleme (Modeli Geliştirmek).....	15
2.3. Adımsal Gerileme (Modeli Budamak)	15
2.4. Knot (Düğüm Noktası).....	16
2.4.1. Düğüm Değerinin Elde Edilmesi.....	18
2.5. Ayna-Görüntü Temel Fonksiyonlar (Mirror-Image Basis Functions)	20
2.6. Genelleştirilmiş Çapraz Geçerlilik (GCV)	21
2.7. MARS Modelinin Avantaj ve Dezavantajları	22
2.7.1. MARS Avantajları ve Diğer Yöntemlerle Karşılaştırılması	22
2.7.2. MARS'ın Dezavantajları	23
2.8. MARS ile Yapılmış Bazı Çalışmalar	24
3. UYGULAMA.....	28
3.1. Araştırmanın Amacı	28

Sayfa

3.2. Araştırma Evreni	28
3.3. MARS Modelinin Kurulması	29
4. BULGULAR.....	31
4.1. Minimum Model.....	33
4.2. En Karmaşık Model.....	36
4.3. En Uygun Model	41
5. SONUÇ ve ÖNERİLER.....	53
6. KAYNAKLAR	55
7. ÖZGEÇMİŞ.....	58

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 2.1. Temel Fonksiyonlara Karşılık Ayna Değişken.....	15
Şekil 2.2. Kesit Ve Düğüm Kullanılarak MARS Veri Tahmini.....	17
Şekil 2.3. X ve Y Arasındaki Dağılım Grafiği	19
Şekil 2.4. X Değişkeni 0-100 Aralığı İçin C=10'dan 80'e Kadar Temel Değişkenlerin Değişimi.....	20
Şekil.4.1. GCV Değerlerin Grafikselsel Gösterimi.....	33
Şekil 4.2. GRUP ile FONK_PUAN Arasındaki İlişki ve Düğüm Değeri.....	45
Şekil 4.3. NOT_ORT ile FONK_PUAN Arasındaki İlişki ve Düğüm Değeri.....	46
Şekil 4.4. YAS ile FONK_PUAN Arasındaki İlişki ve Düğüm Değeri.....	47
Şekil 4.5. BMI ile FONK_PUAN Arasındaki İlişki ve Düğüm Değeri	48
Şekil 4.6. YAS ile FONK_PUAN Arasındaki İlişki ve Düğüm Değeri.....	49
Şekil 4.7. YAS ve GELİR Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki	50
Şekil 4.8. YAS ve NOT_ORT Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki	51
Şekil 4.9. BMI ve GELİR Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki	52

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 1.1. Regresyon Modelleri	3
Çizelge 3.1. MARS Modeline Alınan Değişkenler Ve Değişken Türleri.....	29
Çizelge 3.2. Bağımlı ve Bağımsız Değişkenlere Ait Tanımlayıcı İstatistikler	30
Çizelge 4.1. Kurulan Modelin Sonuçları	32
Çizelge 4.2. Minimum Modele İlişkin Bilgiler	34
Çizelge 4.3. Minimum Model Varyans Çözümleme Çizelgesi.....	34
Çizelge 4.4. Modele Alınan Değişkenler Ve Göreceli Önemlilik Yüzdeleri	35
Çizelge 4.5. En Karmaşık Modele İlişkin Bilgiler.....	37
Çizelge 4.6. En Karmaşık Model Varyans Çözümleme Çizelgesi.....	38
Çizelge 4.7. Modele Alınan Değişkenler Ve Göreceli Önemlilik Yüzdeleri	39
Çizelge 4.8. En Uygun Modele İlişkin Bilgiler	41
Çizelge 4.9. En Uygun Model Varyans Çözümleme Çizelgesi.....	42
Çizelge 4.10. En Uygun Modele Alınan Değişkenler Ve Göreceli Önemlilik Yüzdeleri	43

KISALTMALAR

Bu çalışmada kullanılan bazı kısaltmalar açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
MARS	Çokdeğişkenli Uyarlanabilir Regresyon Uzanımları (Regresyon Multivariate Adaptive Regression Splines)
GCV	Genelleştirilmiş Çapraz Geçerlilik
BF	Temel Fonksiyon (Basic Fonksiyon)

1. GİRİŞ

İstatistik biliminin önemli konularından birisini regresyon analizi oluşturmaktadır. Regresyon analizi, araştırma, matematik, finans, ekonomi, tıp gibi bilim alanlarında yoğun olarak kullanılmaktadır. Regresyon analizinin temelinde; gözlenen bir olay değerlendirilirken, hangi etkenlerin etkisi içinde olduğunun araştırılması yatmaktadır. Bu etkenler bir veya birden fazla olabileceği gibi dolaylı veya direkt etkiliyor da olabilirler.

Regresyon analizi yapılırken, gözlem değerlerinin ve etkilenilen olayların bir matematiksel gösterimle yani bir fonksiyon yardımıyla ifadesi gerekmektedir. Kurulan bu modele regresyon modeli denilmektedir [1].

Çoğu kez, bir araştırmacı ya da deney yapan kimse iki ya da daha çok değişken arasında bir ilişki olup olmadığını bulmak ve bu ilişkinin bir denklemlerle nasıl ifade edildiğini göstermek ister. Örneğin, bir ziraatçı buğday verimi ve gübre miktarı arasındaki ilişkiyi, bir mühendis, basınç ve sıcaklık, bir ekonomist gelir düzeyi ve tüketim harcamaları, bir eğitimci öğrencilerin devamsızlık gösterdiği gün sayıları ve başarı dereceleri arasındaki ilişkiyi bilmek isteyebilir. Regresyon, iki (ya da daha çok) değişken arasındaki ilişkinin fonksiyonel şeklini göstermekle kalmaz, değişkenlerden birinin değeri bilindiğinde diğeri hakkında tahmin yapılmasını sağlar [2].

İki ya da daha fazla değişken arasındaki ilişkilerin incelenmesinde regresyon modelleri sıkça kullanılmaktadır. Regresyon modellerinde, gözlenen bir sonuç değişkeni değerlendirilirken hangi etkenlerin sonuç üzerinde etkili olduğunun ve etkili olan değişkenler içerisinde hangi değişkenlerin etkisinin daha fazla olduğunun araştırılması amaçlanmaktadır. Modelde, etkilenen değişken bağımlı, etkileyen değişkenler ise bağımsız değişken olarak adlandırılır [3].

Regresyonda amaçlardan biri bağımlı ve bağımsız değişken(ler) arasındaki ilişkilerin ortaya çıkartılmasıdır. Örneğin Y ile X arasında $Y = \beta_0 + \beta_1 X + \varepsilon$ gibi doğrusal bir ilişki öngörülüyorsa ilk adım, modelin bilinmeyen β_0 ve β_1 parametrelerinin tahmin edilmesidir. Modelin bilinmeyen parametreleri tahmin edildiğinde, bağımsız değişken(ler)in farklı değerleri için bağımlı değişkenin alacağı değeri tahmin etmek regresyonda bir diğer amaçtır. Bağımsız değişken(ler)in her farklı değer(ler)i için bağımlı değişkenin değeri sabit ise ortada araştırılacak bir problem yoktur. Bağımsız değişken(ler)in değerlerine bağlı olarak farklı değerler alan bağımlı değişkenin bir varyansı vardır. Bu varyansın kaynağı nedir? Modeldeki bağımsız değişken(ler) bu varyansın ne kadarını açıklayabiliyor? Regresyonda bu sorulara cevap vardır. Daha sonra belirleme katsayısı olarak bulunacak değer bu soruyu yanıtlamak için kullanılacaktır [4].

Regresyon yöntemleri, modele alınacak değişken sayısına, matematiksel yapıya göre farklı şekillerde sınıflandırılmaktadır. Değişken sayılarına ve matematiksel modele göre regresyon türleri aşağıdadır.

- Bir bağımlı ve bir bağımsız değişkenden oluşan Basit Regresyonlar:
 - Basit Doğrusal Regresyon
 - Polinomial Regresyon
 - Geometrik Regresyon
 - Üssel Regresyon
 - Basit Eğrisel (Non-Linear) Regresyon
- Bir bağımlı ve k sayıda ($k \geq 2$)bağımsız değişkenden oluşan Çoklu Regresyonlar:
 - Çoklu Doğrusal Regresyon
 - Çoklu Eğrisel Regresyon
- m Sayıda ($m \geq 2$) bağımlı ve k sayıda ($k \geq 1$) sayıda bağımsız değişkenden oluşan Çokdeğişkenli regresyonlar:
 - Çokdeğişkenli Doğrusal Regresyon

- Çokdeğişkenli Eğrisel Regresyon [5].

X bağımsız değişken vektörü, Y bağımlı değişken vektörü ve ε hata terimi olmak üzere, X ve Y arasındaki ilişkinin denklem ile ifadesi Eş.1.1 nolu eşitlikteki gibidir.

$$Y = f(X) + \varepsilon \quad (1.1)$$

Bu modelde $f(X)$, doğrusal, logistik veya doğrusal olmayan bir ilişki içinde olabilir. Örnek olarak Çizelge 1.1’de $f(X)$ fonksiyonu farklı regresyon modelleri için tanımlanmıştır.

Çizelge 1.1. Regresyon modelleri

FONKSİYONU	İlişki şekli
$f(X)=\beta_0 + \beta_1 X_1 + \dots$	Doğrusal Model
$f(X)=\beta_0 + \beta_1 X_1 + \beta_2 X_2^2$	Parabolik Model
$f(X)=\beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \dots + \beta_{k-1} X_{k-1}^{k-1} + \beta_k X_k^k$	k. dereceden model

Regresyon modellerinin iki amacı vardır.

1. Bağımlı ve bağımsız değişken arasındaki yapısal ilişkiyi ortaya çıkarmak.
2. Bağımsız değişkenlere ait ölçüm vektöründen, doğru ve güvenilir bir şekilde bağımlı değişkenin değerini tahmin etmek.

Belirleme Katsayısı: Regresyonda amaçlardan birinin bağımlı değişkendeki varyansın kaynağını araştırmak olduğunu belirtmiştik. Bağımlı değişkene ilişkin fark kareler toplamı;

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (1.2)$$

şeklinde tanımlanır. Bu ifadeye $\bar{Y}_i$ değerini hem ekleyip hem de çıkaralım.

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \bar{Y} - \bar{Y}_i + \bar{Y}_i) \quad (1.3)$$

ifadesi gerekli kısaltmalar yapıldığında;

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2 + (\bar{Y}_i - \bar{Y})^2 \quad (1.4)$$

ulaşılın bu eşitliğin sözlü ifadesi;

$$\left(\begin{array}{c} \text{Toplam Kareler} \\ \text{Toplamı} \end{array} \right) = \left(\begin{array}{c} \text{Regresyon Kareler} \\ \text{Toplamı} \end{array} \right) + \left(\begin{array}{c} \text{Hata Kareler} \\ \text{Toplamı} \end{array} \right)$$

Bu kısa açıklamadan sonra, bağımsız değişkenin bağımlı değişkenin varyansının % kaçını açıkladığının göstergesi olan $R_{Y.X}^2 : \bar{Y} = b_0 + b_1 x$ için belirleme katsayısı;

$$R_{Y.X}^2 = \frac{\text{Regresyonla Açıklanan Kısım}}{\text{Toplam Kareler Toplamı}} \quad (1.5)$$

yada

$$R_{Y,X}^2 = 1 - \frac{\text{Regresyonla Açıklanamayan Kısım}}{\text{Toplam Kareler Toplamı}} \quad (1.6)$$

şeklinde ifade edilir. $R_{Y,X}^2$ değerinin bire yakın olması her araştırmacının arzuladığı bir durumdur. Bu durumda bağımlı değişkendeki varyansın büyük bir kısmını modeldeki bağımsız değişkenin açıkladığı sonucuna varılır. Çokdeğişkenli durumlarda matris işlemleri yapılır [4].

MARS (*Multivariate Adaptive Regression Splines*, Çokdeğişkenli Uyarlanabilir Regresyon Kesitleri) modeli, regresyon modeli ile açıklanan hata varyansının oranı hakkında bilgi verir. Regresyon modeli, değişkenler arasındaki ilişkiyi ortaya koymak amacıyla kullanılan modellere göre, doğrusal, doğrusal olmayan veya karışık yapıda olabilir. Kurulan modellerin gerçek yapıyı daha iyi yansıtmaları için özellikle son yıllarda çok sayıda bağımsız değişken bir arada değerlendirilmektedir. Günümüzde bilgisayar teknolojisindeki değişmelere paralel olarak çok sayıda değişkeni bir arada değerlendirebilen karmaşık algoritmaların kullanılabilirliği artmıştır. Bu algoritmaları kullanan regresyon yöntemlerinden birisi de MARS tekniğidir.

MARS tekniği özellikle veri madenciliği alanında popüler hale gelmiştir. Çünkü bu yöntem sebep ve sonuç değişkenleri arasında bir varsayım gerektirmez ve herhangi bir matematiksel ilişki aramaz. Bunun yerine sebep ve sonuç değişkenleri arasında dinamik bir ilişki geliştirir. Parametrik modellerle yaklaşılması oldukça güç olan doğru öngörüler, dinamik yapısı sayesinde yorumlaması kolay, kestirim başarısı yüksek modeller kurarak gerçekleştirir [1].

Diğer regresyon modellerinde belli varsayımlar ve örneklem çapı aranmasına karşın, MARS modelinde bu şekilde varsayımların olmaması önemli bir güçtür. MARS modeli veriden elde edildiğinden, parametrik yöntemlerle tahmini güç olan doğru öngörüler, dinamik yapısı sayesinde yorumlaması kolay, kestirim başarısı yüksek model kurarak bulmamızı sağlamaktadır [6].

2. ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON KESİTLERİ (MARS)

MARS tekniği ilk olarak 1991 yılında Stanford'da fizikçi ve istatistikçi Jerome Friedman tarafından geliştirilmiştir. Sürekli değişkenlere model uyarlayabilmek için kullanılan doğrusal regresyon, doğrusal olmayan regresyon, regresyon ağaçları, yapay sinir ağları gibi birçok model vardır. MARS ise adımsal regresyon yöntemidir, regresyon setinin performansını geliştirmek için tekrarlamalı ayırma metodunun ve adımsal doğrusal regresyonun genelleştirilmiş bir hali olarak düşünülebilir [7].

Kesit (spline) , karmaşık eğri çizimlerinde ve fonksiyon tahminlerinde yeni bir matematiksel süreç olarak göz önüne alınabilir. Spline düzleştirme yöntemi iki yada daha üst düzeyli polonomiyaller kullanıldığında elde edilen ve parametrik olmayan hata varyansının kontrol edilmesini sağlayan bir yöntemdir [8].

Kesit oluşturmak için X ekseninde uygun sayıda bölge oluşturulur. Bu bölgeler arasındaki sınırlar düğüm olarak da bilinir. Yeterli sayıda düğüm sayısı ile model tahmin edilebilir. İki boyutlu kesit çizmek oldukça kolayken ikiden çok boyutlu kesitlerin çizimi zorlaştığından her bir kesit matematiksel temel fonksiyonlarla (Basic Fonctions) ifade edilmektedir. MARS; temel fonksiyonları orijinal veriden elde eden ve tahminde kullanan bir regresyon modelidir. Her bir temel fonksiyon; orijinal verideki değişimi sıfır yapacak bölgeleri seçmekte, MARS ise verinin özellikli alt bölgelerine odaklanmaktadır. MARS; en uygun değişken dönüşümlerini ve kesişimlerini kullanarak çok boyutlu veri içerisinde gizli olan karmaşık veri yapısını ortaya çıkarmada ve yorumlamada diğer yöntemlere göre öne çıkmaktadır [9].

Pek çok veri setinde eldeki bilgileri kullanarak, $y=f(x)$ fonksiyonunu x 'in her bir ayrı bölgesinde y 'yi özetleyerek tahmin etmek uygun değildir. Bazı değişkenler için fonksiyon parçalarının oluşturduğu iki bölgenin özelliklerini

izlemek de yeterli olmayabilir. Örneğin, bağımlı y değişkeninin; bağımsız 35 x değişkeni ile ilişkisi 3. veya 4. bölgede farklı ise test edilmesi gereken bölge sayısı sadece 35 bağımsız x değişkeni için –etkileşimlerle birlikte- yaklaşık 34 milyardır [10].

K.Batu TUNAY, Türkiye’de 1987-2000 döneminde farklı parasal büyüklüklere göre paranın gelir dolaşım hızını etkileyen sebepleri MARS yöntemini kullanarak tahmin etmeye çalışmıştır. Çalışmada, MARS’ın geleneksel yöntemlere göre imkânsız değilse bile son derece zor olan, çok boyutlu verilerin içinde gizlenmiş karmaşık veri yapısını, optimal veri dönüşümlerini ve verilerin karşılıklı etkileşimlerini belirleyebilme avantajı ile regresyon modellemesinde kolaylık sağladığı görülmüştür. Küçük veri tabanları söz konusu olduğunda bile iyi bir regresyon modelinin geliştirilmesi çok zaman ve emek istemektedir. Bununla birlikte, MARS yöntemiyle geniş veri tabanları ve çok karmaşık veri yapıları için regresyon modelleri kolayca geliştirilebilmektedir [11].

MARS bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişkiye dair varsayımlarda bulunmayan, parametrik olmayan bir regresyon yöntemidir. Bunun yerine regresyon verisinden kendisinin çıkardığı temel fonksiyonlara dayanarak kendisi bir ilişki oluşturmaktadır. Bağımsız değişkenlerin farklı aralıklarına karşılık gelen temel fonksiyonları kullanarak esnek bir regresyon modeli kurar.

Başka bir deyişle stratejisi parçala ve çözümü yöntemi üzerine kuruludur. Bu yöntem veri uzayını (yığın) önce bölgelere ayırır ve her birine bir regresyon eşitliği oluşturur. Bu da MARS’ı diğer yöntemler için çok boyutluluk sorunu oluşturabilecek çokdeğişkenli regresyon problemleri için uygun bir çözüm yöntemi olarak öne çıkarır [12].

Mars, bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan ilişkileri doğrusal yapıya dönüştürme amacıyla uygun dönüşüm metotları bulmada ve bağımsız değişkenler arasındaki etkileşimleri belirlemede optimum bir yeniliktir [13].

Regresyon bağımsız değişkenlerle bağımlı değişkenler arasında ilişkiyi incelemek için kullanılır. Doğrusal olmayan regresyon bağımlı değişkenle bağımsız değişkenler arasında bir öngöründe bulunmaz. Doğrusal olmayan regresyon veriden model üretir.

MARS; regresyon yapısını, veri uzayını küçük küçük doğrusal regresyon denklemlerine (temel fonksiyonlar) bölerek oluşturur. Her bölme noktası veriye bağlı olarak bulunur.

MARS modeli, orijinal veriden tahmin edilen temel fonksiyonların özel sınıflarının regresyon kesitidir. MARS'ın üstünlüğü en uygun değişken dönüşümleri ve etkileşimlerine ek olarak çoğunlukla yüksek-boyutta saklanan karmaşık veri yapılarını bulmasından kaynaklanmaktadır.

Kolyshkina'ya göre; pek çok veri madenciliği uygulamalarında verilen kestirimlerin sayısı, x 'in her bir farklı bölgesinde y kesitlerinin genelleştirilmiş fonksiyonu olan $y = f(x)$ yaklaşımı kullanılarak yakınsanamaz. Ne regresyon sayısı ne de düğüm yerleri önceden belirlenemez. Prosedürün bunları yapabilmesi için aşağıdakilere ihtiyacı vardır [14].

- Hangi bölgeye bakılacağı ve bu bölgenin sınırlarına doğru karar vermek,
- Her bir değişken için kaç tane aralığın olacağına doğru karar vermek,

Yine Kolyshkina'ya göre; doğru bir karar vermek için verinin karakteristiğine uyumlu bir bölge seçmek gerekmektedir. Belki birkaç önemsiz değişkeni dışarıda bırakarak (değişken seçimi yapmak) ve aynı anda sadece birkaç

değişkeni hesaba katarak bir çözüm istenebilir. Hatta eğer MARS model için 30 değişken seçerse, aynı anda 30'una da bakamaz. Oldukça sadeleştirilmiş bir karar ağacı oluşturur. Parçalı temel fonksiyonlar ve bunların birleşimlerini kullanarak ve regresyon modellerindeki hem ileri doğru hem de geriye doğru ilerleme algoritmalarından yararlanılarak MARS modeli elde edilir. Model kurulumu iki adımda gerçekleşir [14].

1. Adım: Mümkün olan tüm temel fonksiyonlar oluşturulur. Çok büyük bir model bulununcaya kadar yani modelin karmaşıklığı maksimum seviyeye ulaşınca kadar eklenen temel fonksiyonlarla model geliştirilir. Temel fonksiyonlar ya dönüştürülmüş tek değişkenden ya da çokdeğişkenli karşılıklı etkileşim terimlerinden meydana gelir. Temel fonksiyonlar modeldeki bağımsız değişkenlerin karşılıklı etkileşimlerini ve doğrusal olmayan dönüşümlerini hesaba katan fonksiyonlardır. Temel fonksiyonlar oluşturulurken tüm değişken ve değişkenlerin birleşimleri tek tek ele alınır. Bu durumda, her bir bağımsız değişken ve bu değişkenlerin birbirleri ile olan etkileşimlerinin tanımlı bulunduğu aralıklardaki tüm olası değerler birer düğüm olarak düşünülüp, mümkün olan tüm olası temel fonksiyonlar belirlenir. Temel fonksiyonlar eklenerek model daha karmaşık ve daha esnek bir yapıya kavuşturulur. Temel fonksiyon sayısı maksimum seviyeye ulaşınca kadar bu işlem sürer. Ayrıca temel fonksiyonlar oluşturulurken aynı değişkene ait ileride tanımlanacak olan Ayna Görüntü Fonksiyonu da oluşturulup modele alınabilir. Ayna temel fonksiyonu, bağımlı değişkenle bağımsız değişken arasındaki dağılımın düğüm noktasında eğiminin değiştiğini ve düğüm noktasına kadar olan eğiminin sıfır olduğunu gösterir.

2. Adım: Eğilim ve varyansların optimal bir dengesi bulununcaya kadar temel fonksiyonlar modelden çıkarılır. Bu adıma budama adımı denir ve bu adımı gerçekleştirmek için MARS geriye doğru adım algoritması kullanır. Geriye adım algoritması gerçekleştirebilmek için aşağıdaki aşamalar yerine getirilmelidir. Modele yalnızca temel fonksiyon sabitini içeren en küçük modelle başlanır.

- Tahminin hatasını minimize eden en uygun ölçümle her değişken ve mümkün tüm düğümler için temel fonksiyon uzayı araştırılır.

- Maksimum karmaşıklığa ulaşan modeli oluşturuncaya kadar bütün düğümlere ve değişkenlere bakma işlemi tekrarlanır.
- Son olarak, budama prosedürüne başvurulur.

MARS özellikle ayna değişkenli fonksiyonları doğrusal ve doğrusal olmayan dağılımları temel fonksiyonları olarak kullanır ve etki-tepki değişkenleri arasındaki ilişkiyi tahmin eder.

Kategorik Tahmin Ediciler; Pratikte hem sürekli hem de kategorik tahmin ediciler kullanılabilir. İkisi de başarılı sonuçlar verir. Ancak temel MARS algoritması değişkenlerin sürekli olduğunu öngörür. İki değer alan değişkenlere göre sınıflandırılmış verilerde daha başarılı sonuçlar elde eder.

Çoklu Bağlı Değişkenler; MARS algoritması çoklu bağlı değişkene de uygulanabilir. Burada algoritma verilere bağlı olarak temel fonksiyonlar bulur. Fakat sinir ağı yapılarında olduğu gibi, her bağımlı değişken için farklı bir sabit oluşturur.

Sınıflandırma Problemleri; Çoklu bağlı değişkeni problemleri çözülebildiği için sınıflandırma problemlerine de uygulanabilir. Sınıflandırılmış değişkenleri kodlar, algoritmayı işletir, model uygun ise o kodlarla skorları belirler. Hangi sınıf için en iyi skor elde edilirse kurulan modelde o sınıf yer alır.

Model Seçimi ve Budama; Doğrusal olmayan modeller esnek bir yapıya sahiptir ve eğer bir önlem alınmazsa gereğinden fazla yaklaşıma çabası içine girilebilir. Denenmiş verilerde neredeyse sıfır hata yapmalarına rağmen yeni gözlemler ve durumlarla karşılaştıklarında zayıf bir performans gösterebilir. Diğer pek çok metotta olduğu gibi MARS da veriyi over-fitt (karmaşıklığı arttırma) etme eğilimindedir. Bunu engellemek için budama tekniği kullanılır ve modelin karmaşıklığı azaltılır (sınırlanır).

Tahmin Edicilerin Seçilmesinde MARS: Temel fonksiyonların seçimi ve budanması özelliği tahmin edici seçiminde MARS'ı güçlü bir araç haline getirir.

Mars algoritması sadece, sonuca ölçülebilir bir etki yapan temel fonksiyonları seçer, diğerlerini dikkate almaz. Bunlara bağlı olarak MARS veri madenciliği problemlerinde oldukça fazla kullanılır.

MARS'da istenen model elde edilene kadar tekrar tekrar uygulanan iki adımlı bir prosedür vardır. İlk adımda temel fonksiyonları teker teker ekleyerek önceden belirlenmiş maksimum karmaşıklık seviyesine ulaşıncaya kadar model oluşturulur. Daha sonra geriye dönük bir prosedür uygulanarak sonuca daha az etki eden temel fonksiyonlar modelden çıkarılır. Sabit temel fonksiyona dayalı en basit model oluşturulur. Temel fonksiyonları her değişken ve her olası düğüm için araştırarak tahmindeki hatayı azaltanları modele ekler. En küçük kareler mantığı ile sonuca en az etki eden temel fonksiyonları çıkarır.

MARS algoritmasını çok fazla küçük doğrusal regresyon denklemlerine bölerek oluşturur. Her bölme noktası veriye bağlı olarak bulunur.

MARS özellikle ayna değişkenli fonksiyonları doğrusal ve doğrusal olmayan dağılımları temel fonksiyonları olarak kullanır ve etki tepki değişkenleri arasındaki ilişkiyi tahmin eder.

Bu teknik, pratikte yaygın kullanılan doğrusal ve doğrusal olmayan regresyon modellerinin aksine bağımsız değişkenin farklı aralık değerleri için farklı katsayılar türetir ve modele etkileşim terimlerini de katarak gerçek yapıyı daha iyi yansıtan modeller geliştirir. Ayrıca bağımlı ve bağımsız değişkenlerin dağılımları üzerine her hangi bir varsayım gerektirmez ve değişken tipi de önemli değildir [1].

Elimizdeki bağılı olarak oluşturulan temel fonksiyonlar model parametreleri ile birlikte (en küçük kareler yöntemi) modeli tahmin eder. Genel regresyon modeli;

$$Y=f(x)+\varepsilon \quad (2.1)$$

olarak tanımlanır. Bu gösterime genel olarak regresyon denklemi ya da regresyon modeli denir. Genel MARS modeli 2.2 nolu eşitlikteki gibi tanımlanabilir.

$$Y = \beta_o + \sum_{k=1}^K a_k \beta_k (X_t) + \varepsilon_i \quad (2.2)$$

Burada;

- k : Düğüm sayısını,
- K : Temel fonksiyon sayısını,
- X : Bağımsız değişkeni,
- a_k : k. Temel fonksiyonun katsayısı,
- β_0 : Modeldeki sabit terim,
- $\beta_k(x_t)$: t. Bağımsız değişken için k. temel fonksiyon

Temel fonksiyon ise aşağıdaki gibi tanımlanır.

$$B_m(x) = \prod_{t=1}^{L_m} [s_{l,m}(x_{v(l,m)} - k_{l,m})] + \quad k= 1,2, ..., K \quad (2.3)$$

Buradaki;

- L_m : Etkileşim derecesini,
- $s_{l,m}$: $\in [\pm 1]$
- $k_{l,m}$: Düğüm değerini
- $x_{v(l,m)}$: Bağımsız değişken değerini göstermektedir.

2.1. Temel Fonksiyon

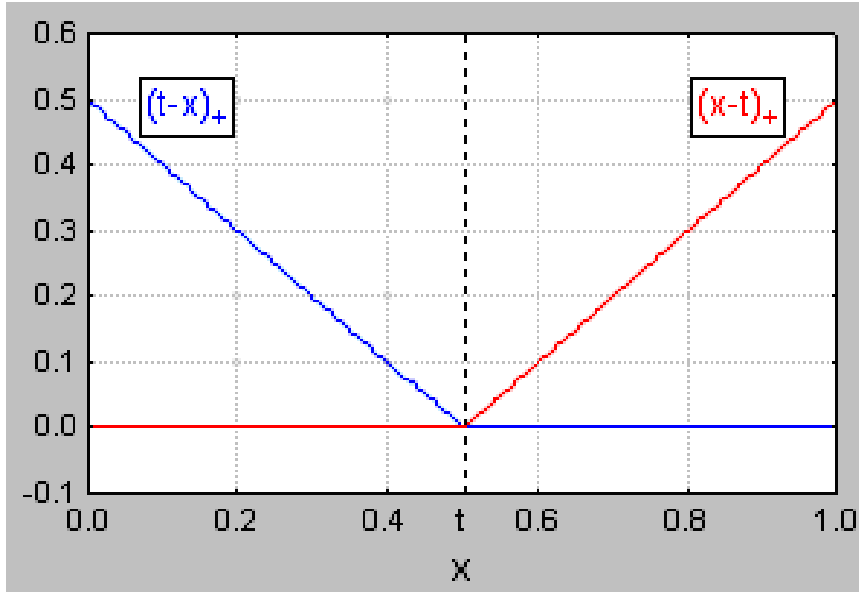
Normalde kullanılan regresyon eşitlikleri veriler arasındaki ilişkiyi tek bir fonksiyon kullanarak belirlemeye çalışırlar. Fakat MARS parçalı polinomik bir fonksiyon kullanır. Böylece bütün değerler net olarak görülebilir. Regresyon kesit fonksiyonları parçalı polinomik temel fonksiyonların düğümlerde birleştirilmesi ile elde edilmiş sürekli bir fonksiyondur ve bu temel fonksiyonlardaki sabitler en küçük kareler yöntemi ile bulunur. Parçalı fonksiyonlar elimizdeki verinin yerel karakterlerine daha uygun olduğu için tek polinom kullanmaya göre daha fazla esneklik sağlar. MARS her duruma yüksek derecede uyum sağlayabilir ve düğüm yerlerini ve derecelerini otomatik olarak seçen bir algoritma oluşturur.

$$Y_t = \beta_o + \sum_{k=1}^M a_k \beta_k(x_t) + \varepsilon_i \quad (2.4)$$

Burada $\beta_k(x_t)$, x_t 'nin k'inci temel fonksiyonudur.

Temel fonksiyonlar, x_t 'nin yüksek oranda doğrusal olmayan dönüşümleri olabilir. Fakat Y_t temel fonksiyonların doğrusal bir fonksiyonudur.

Özellikle doğrusal ve doğrusal olmayan genişleme için temel fonksiyonlar iki yönlü kullanılır. İki temel fonksiyon (t-x) ve (x-t), t parametresi temel fonksiyonların düğüm (knot) noktasıdır. Düğüm noktası veriden elde edilir. Sadece pozitif sonuçları gösterir. Diğer durumlar için 0 (sıfır) değeri alır.



Şekil 2.1. Temel Fonksiyonlara karşılık Ayna Değişken

2.2. Adımsal İlerleme (Modeli Geliştirmek):

Sonuç değişkeninin tüm olası etken öğelerle bağıntısı ayrı ayrı belirlenir. En güçlü bağıntıyı gösteren öğe ile işe başlanır. Bu bağıntının belirleyicilik düzeyi R^2 elde edilir. Denkleme ikinci derecede güçlü bağımsız değişken de katılır, yeni denklemde elde edilen belirleyicilik, daha öncekine göre anlamlı olarak yüksek bulunursa ikinci bağımsız değişken de modele eklenir. Eğer ikinci bağımsız değişkenin olduğu modelde R^2 'de anlamlı bir yükseliş olmaz ise ikinci bağımsız değişken modele alınmaz. Bu şekilde devam edilerek öncelikle tüm bağımsız değişkenler ve sonrasında değişkenler arası etkileşimler ile en karmaşık model kurulur.

2.3. Adımsal Gerileme (Modeli Budamak):

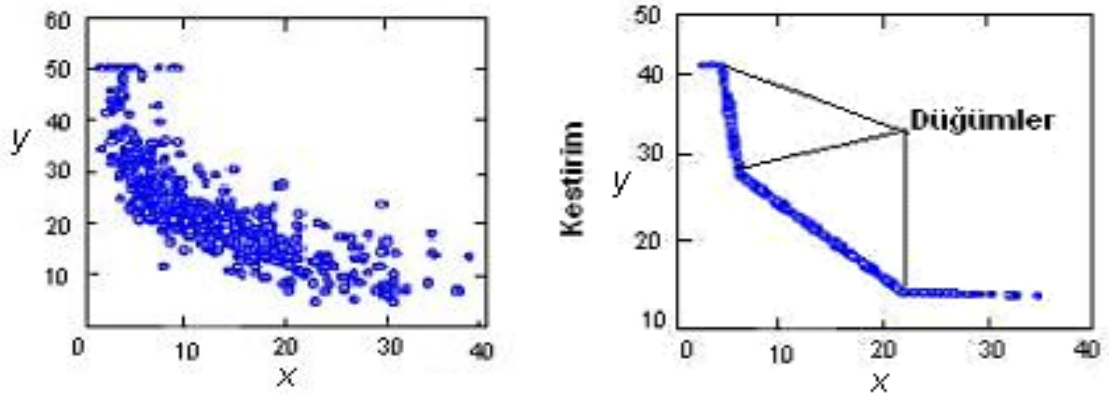
2.2.'de adımsal ilerleme ile elde edilen en karmaşık anlamlı regresyon modeli ile yola çıkılır. Bu modelin belirleyiciliği, en güçsüz olan değişken çıkartıldığı zaman artıyor ise çıkartmalara devam edilir. Bu tür ekleme ve

çıkartmalar ile en az sayıda değişken kullanılarak en yüksek belirleyicilik araştırılır. Bir değişkenin eklenip çıkartılması için karar vermede belirleyiciliğin farklılaşması dışında ölçütler de kullanılabilir. Örneğin, araştırmacı bir değişkenin mutlaka modelde yer almasını isteyebilir. Ayrıca iki değişkenin etkileşimi modelde yer alıyorsa tıpkı lojistik regresyondaki gibi etkileşim içerisinde yer alan bağımsız değişkenlerin de modelde yer alması gerekir.

2.4. KNOT (Düğüm Noktası)

Bağımlı ve bağımsız değişken arasındaki ilişkiler doğrusal, eğrisel ve kübik şeklinde olabilir. Aynı bağımsız değişken üzerinde ilişkinin şeklinin değiştiği bağımsız değişken değerine düğüm değeri denir. Düğüm değerinde hata kareler toplamı en küçük değerini alır. Ardışık iki düğüm değerini birleştirerek çizilen doğrunun eğimi β değerini gösteren regresyon katsayısıdır.

Kesitlerin altında yatan temel düşünce düğümdür (knot). Düğüm; verinin bir bölgesinin sonu ile diğerinin başlangıcını belirler. Bu nedenle, düğüm, fonksiyonun değiştiği yerdir. İki düğüm noktası arasında veri herhangi bir regresyon modeline uygun olabilir (örneğin doğrusal regresyon). Klasik kesitte, düğümler önceden belirlenmiş ve eşit aralıklı olarak ayrılmıştır, oysa MARS modelinde düğümler prosedür (procedure) tarafından araştırılarak bulunur. MARS modeli ihtiyaç duyulandan daha çok sayıda düğüm noktasını bulur. Eğer iki düğüm noktası arasında düz bir çizgi yakalanırsa, bu aralıkta yeni düğüm noktasına ihtiyaç yoktur. Tahmin değeri önemsiz olan gözlemlerin oluşturduğu sahte düğüm noktaları MARS tarafından en aza indirgenir. Şekil 2.2.'de karmaşık gerçek bir veri ve bu veriden oluşturulan iki düğüm noktası olan bir MARS kesiti gösterilmiştir [15].



Şekil 2.2. Kesit ve düğüm kullanılarak MARS veri tahmini (solda gerçek veri)

MARS modeli oluşturulurken modele giren değişkenler ve bağımlı değişken belirlenmelidir. MARS model verisindeki kırıklar, eşit aralıklı temel birçok kesitten oluşur. Her bir kesit, farklı değişkenler arasında yada aynı değişkenin belirli aralıklarında oluşur. MARS modele daha birçok altgruplar ilave eder ve birkaç düğüm noktası oluşturur.

MARS modeli bir regresyon modeli tahminidir. Bu regresyon modeli düzgünleştirilmiş fonksiyonlardan oluşur. Bu fonksiyonlar temel fonksiyonlar olarak adlandırılır. İki düğüm noktası arasında veri karakteristiğine göre doğrusal yada eğrisel regresyon olabilir. Temel fonksiyon herhangi iki düğüm noktası arasında tektir ve her bir düğümde başka bir temel fonksiyon oluşur. İki temel fonksiyon kesişimi de düğüm noktası oluşturmaya devam eder. MARS, verinin oluşturduğu eğrisel regresyon çizgisini, oluşturulan alt gruplarla kesitlere böler. Uygun regresyon modelindeki, komşu iki düğüm noktası arasındaki en kısa uzaklığı önceden belirleyerek alt gruptaki küçük verilerdeki döngüyü önler.

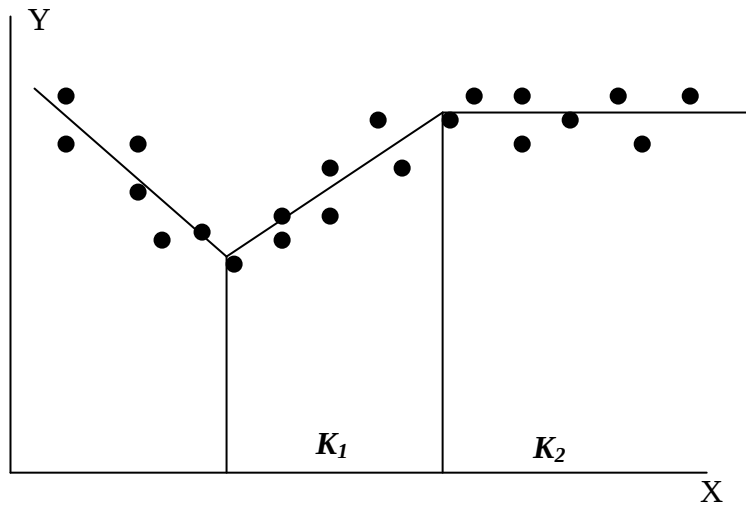
İlk önce birçok düğüm noktasından oluşan bir model (bu aralıktaki veri düz bir çizgi kabul edilir) oluşturulur, sonra modele en az katkısı olan düğümler modelden çıkarılır. Modelin kurulması aşamasında ileriye doğru adımsal yöntem birçok gereksiz (modele az katkısı olan ya da katkı sağlamayan) düğüm oluşturacak fakat gereksiz düğümler geriye doğru adımsal budama yöntemi ile modelden çıkarılacaktır [11].

Basit regresyonda bir tane en iyi düğüm noktasını bulmak basit bir problemdir. Çok sayıda potansiyel düğümleri inceleyerek en iyi R^2 ile biri seçilebilir. Ancak düğümlerin en iyi çiftini bulmak bir çok uzun hesaplamalar gerektirir. İhtiyacımız olan doğru sayıda düğümlerin en iyi setini bulmak zor bir iştir. MARS ise gereken sayıda ve konumda düğüm noktasını ileriye ve geriye adımsal yöntemle bulur. Önce birçok düğüm noktası ile modeli en karmaşık yapıya kavuşturur. Sonra bu düğüm noktalarından katkısı en az olanları modelden çıkarır. Böylece ileriye doğru düğüm seçimi yanlış yerlerdeki birçok düğüm içerecek fakat yanlış düğüm noktaları modelden geriye doğru budama yöntemi ile silinir [16].

2.4.1. Düğüm Değerinin Elde Edilmesi

Bölüm 2.4'te belirtildiği gibi bağımlı ve bağımsız değişken arasındaki ilişkiler doğrusal, eğrisel ve kübik şeklinde olabilir. Aynı bağımsız değişken üzerinde, ilişkinin şeklinin değiştiği bağımsız değişken değerine düğüm değeri denir. Bir başka ifadeyle bağımsız değişken değeri, tanımlı bulunduğu aralıklarda doğrunun eğimini değiştirmeyen en son değer düğüm değeri olarak alınır. Bu düğüm değerlerinde hata kareler toplamı en küçük değerini alır. Ardışık iki düğüm değerini birleştirerek çizilen doğrunun eğimi, Model'de β ile gösterilen regresyon katsayısıdır. Bu haliyle MARS modeli parçalı regresyon modeline benzer. Oluşturulan her temel fonksiyonda bağımsız değişkenden seçilen uygun düğüm değerleri bağımlı değişken ile bağımsız değişken arasında

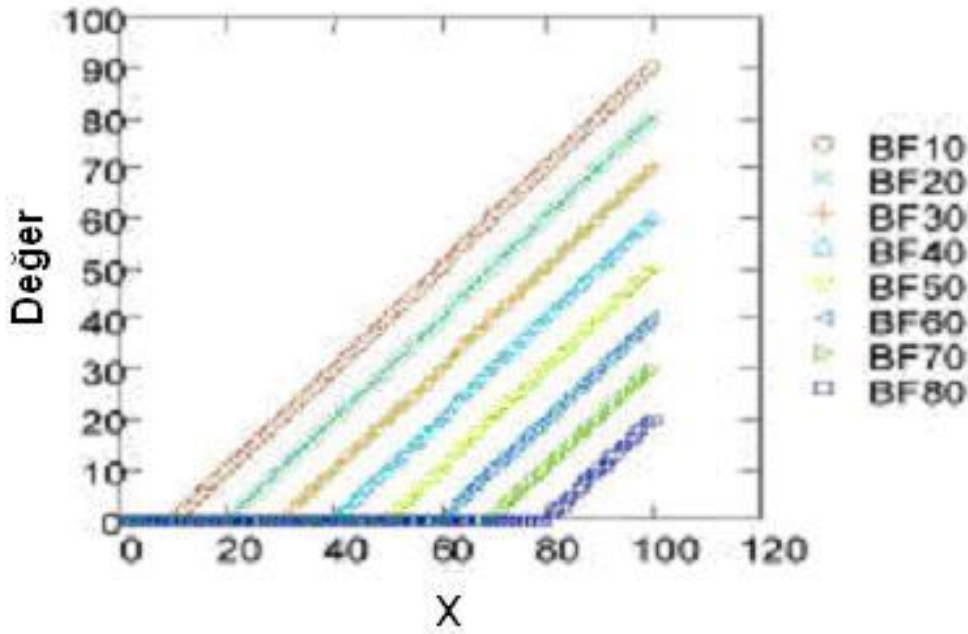
monoton dönüşümler (logaritmik, üstel, vb.) yapmadan aralarındaki ilişkiyi doğrusal hale getirirler. Şekil 2.3'de iki farklı düğüm değerine sahip bir değişkenin bağımlı değişkenle olan ilişkisi sunulmuştur. K_1 ve K_2 düğüm değerleridir. Başlangıç noktasından K_1 'e kadar doğrunun eğimi aynıdır. Benzer şekilde K_1 ve K_2 arasında da doğrunun eğimi aynıdır. Fakat K_1 'den ε kadar uzaklaştıkça bile doğrunun eğimi değişir. İşte doğrunun eğimini değiştirmeyen değer bir düğüm değeridir.



Şekil 2.3. X ve Y arasındaki dağılım grafiği

Düğüm seçiminin şartlarını incelersek tek boyutta kesitleri çok iyi bir şekilde anlayabiliriz. Fakat bu şartlar aynı anda çok sayıda değişkenle çalışıldığında kullanışsızdır. Programlama yapabilmek için kısa ve kolay anlatımlar gerekir. Düğüm yerlerini kullanırken etkileşimlerin nasıl yapılacağı ya da ifade edileceği açık değildir. MARS'ta Temel Fonksiyonlar (BF) düğümlerin araştırılmasını genelleştirmek için kullanılan mekanizmalardır. BF bilgi içeren bir ya da daha çok değişkeni ifade etmekte kullanılan fonksiyonların setidir. Ana bileşenlere benzer şekilde, BF'de hedef değişkenle tahmin edici değişkenlerin ilişkisi aslında daha açıktır. Şekil 2.4 incelendiğinde BF hokey sopası, MARS modelinin yapıtaşı tek değişkenlilerde çoğu zaman uygulanır. Hokey sopası fonksiyonu değişken gösterimi X 'ten X^* 'a dönüşür.

burada X , eşik değeri c 'nin altında bir değerse $X^* 0$ değerini alır. c 'den büyük bütün değerler için ise X değerini alır ($\max(0, X-c)$). İkinci form birincisinin ayna değişkenini oluşturur ($\max(0, c-X)$). Şekil 2.4'te, X kestiriminin 0-100 aralığı için c değişkeninin 10'ar birimlik değişimlerine karşılık temel fonksiyonlarda meydana gelen değişimler gösterilmiştir.



Şekil 2.4. X değişkeni 0-100 aralığı için $c=10$ 'dan 80'e kadar temel değişkenlerin değişimi

2.5. Ayna-Görüntü Temel Fonksiyonlar (Mirror-Image Basis Functions)

MARS ikili parçalarından oluşan temel fonksiyonlar yaratır: Standart temel fonksiyonlara ayna görüntülerini daima ekler. Bu ikililer olası birçok temel fonksiyonlar gibi ayrı veri değerleri oluşturur. Ayna görüntüsü sağ ve sol ana kırıklara ayrılmış düğümlerdir.

Bu fonksiyonlar doğrusal olarak bağımsız değildir, fakat modelin esnekliğini artırılabilir. Düğümlerin verilmiş seti için, temel fonksiyonların sadece bir ayna görüntüsü standart temel fonksiyonlardan doğrusal olarak bağımsız

olacaktır. Modelde kullanılan temel fonksiyonların tümü için, aynı değeri vereceğinden temel fonksiyonun hangi ayna değişkenin ekleneceği önemli değildir. Bununla birlikte herhangi bir düğümde standart temel fonksiyonların ayna görüntüleri yerine model değiştirilebilir. Temel fonksiyonların eklenen belirli ayna görüntülerini değiştirmek final modeli etkilemez.

Basit regresyonda en iyi düğüm noktasını bulmak kolay bir problemdir. En iyi R^2 ile çok sayıda potansiyel düğüm noktalarını bulur ve birini seçer. Oysa, değişken sayısı arttığında; düğümlerin en iyi çiftini bulmak çok uzun hesaplamalar gerektirir. Gerçek sayıya ihtiyaç duyulduğunda düğümlerin en iyi setini bulmak bilinmeyen hatta çok ilgi çekici bir iştir. MARS ihtiyaç duyulan düğümlerin yerlerini ve sayılarını ileriye / geriye doğru adımsal yöntemle bulur.

2.6. Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation – GCV)

Oluşturulan maksimum model budanarak yani önemli bağımsız değişkenler ve bu değişkenlerin karşılıklı etkileşimleri belirlenerek, hata kareler toplamı minimum olan en uygun model oluşturulur. Budama algoritması en yaygın olarak genelleştirilmiş çapraz geçerlilik tekniği ile yapılır. Genelleştirilmiş çapraz geçerlilik hem hata terimini hem de model karmaşasını hesaba katar. Eş.2.5. ve Eş.2.6. yardımıyla hesaplanır.

$$GCV(M) = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \hat{f}_M(x_i))^2}{(1 - C(M)/n)^2} \quad (2.5)$$

$$C = 1 + cd \quad (2.6)$$

Burada,

- N : Veri setindeki denek sayısını,
- d : Etkili serbestlik derecesi olup bağımsız temel fonksiyonların sayısını gösterir ve
- C : Eklenen temel fonksiyonların maliyet-karmaşıklık (cost-complexity) ölçüsüdür.
- M : MARS Modelinin kurduğu regresyon modeli sayısıdır.

Hesaplamalar sonucu C değeri için $2 < d < 3$ değerinin en iyi olduğu bulunmuştur [17].

2.7. MARS Modelinin Avantaj ve Dezavantajları

2.7.1. MARS'ın Avantajları ve Diğer Yöntemlerle Karşılaştırılması

1. MARS yönteminde; bağımlı ve bağımsız değişkenlerin dağılımları ile ilgili herhangi bir varsayım yoktur (değişkenler kategorik veya sürekli olabilir). Doğrusal regresyon analizinde bağımsız değişkenin normal dağılım göstermesi şartı aranırken, Lojistik Regresyon Analizinde böyle bir şart yoktur. Doğrusal regresyon analizinde tahmin edilecek olan bağımlı değişken sürekli iken, Lojistik Regresyon Analizinde bağımlı değişken kesikli bir değer almaktadır.

2. MARS yönteminde; bağımlı ve bağımsız değişkenler arasında doğrusallık varsayımı bulunmaz.

MARS yönteminde; bağımlı değişkeni tanımlamada oluşturulan modelde, hem eklemeli modele hem de bağımsız değişkenlerin etkileşimine yer verilmektedir [18].

3. Mars hızlıdır ve diğer doğrusal modellere göre daha az veri gerektirir [19].

4. Kayıp değerlerden ve aşırı uç değerlerden çok az etkilenir [20].

5. Diğer regresyon yöntemleri için sorun oluşturan çoklu bağlantıdan çok az etkilenir.

6. Diğer çokdeğişkenli regresyon modellerinde yapılan simülasyon çalışmalarına göre başlangıçtaki bağımsız değişkenlerin % 90'nın modele girmesine karşın Mars modelinde daha az bağımsız değişken modele girmiştir [21]. Daha az bağımsız değişkenin modele girmesi, modelin yorumlanmasını oldukça kolaylaştırır ve model kurulması için gerekli süreyi en aza indirir.

2.7.2. MARS'ın Dezavantajları

1. Henüz uygulaması yeni bir yöntem olmasından dolayı modelin yorumlanmasında dikkat etmek gerekir. Özellikle kayıp gözlem durumunda ve etkileşim terimlerinin yorumlanmasında iyi bir bilgi birikimine ihtiyaç vardır.

2. Uygulaması için genellikle büyük veri setlerine ihtiyaç duyulur. Fakat büyük veri seti kavramı görecelidir. Literatürdeki yayınlar incelendiğinde MARS yönteminin 80 – 100 denekten oluşan veri setlerine de uygulandığı görülmektedir. Bu nedenle bazı yayınlarda geçen bu dezavantajın geçerliliği konusunda şüpheler ifade edilmektedir [13].

2.8. MARS ile Yapılmış Bazı Çalışmalar

Mallick ve Arkadaşları; Bayesci bir yaklaşım olan MARS yöntemi ile tek ve çok değişkenli sansürlü yaşam analizi verisi üzerine yaptıkları çalışmada, MARS yönteminin olasılıklı oran gibi alt sınıflarını da içeren tüm modelleri otomatik olarak kurduğunu ve farklılıkları tespit ettiğini bulmuşlardır. Çalışmalarında iki yönlü sıçramalı Markov Zincirini kullanan Monte-Carlo algoritmasını kullanmışlar ve olasılık fonksiyonunun tahminini sağ kalım eğrisi kadar iyi bir şekilde kestirmişlerdir [22].

Verzilli ve arkadaşları İngiltere’de yaptıkları çalışmalarında, DNA kodlamasındaki genetik çok şekillilik olan bölgelerde taşıyıcılarda fenotip etkiler görülebileceğini söylemişlerdir (örneğin hastalığa bağlı hassasiyet etkisi gibi). Bu soruna çözüm olarak bağ tutucu ve lizozom proteinlerindeki başlangıç mutasyonu verileri kullanılarak bir model oluşturmak ve oluşturdukları modelin performansını sınamak için hiyerarşik Bayesci MARS Yöntemini kullanmışlardır. Deneylerinde 12 Aminoasit durumu hakkında bilgi edinebilmek için her bir aminoasidin pozisyonu ve protein işlevselliği etkilerini ölçmüşler ve her bir pozisyondan elde edilen değerleri modelde yerine koyarak kestirimlerde bulunmuşlardır. Öğrenme verisi olarak bağ tutucu verilerini kullanmış ve lizozom değişimleri ile bu değişkenlerin tersleri üzerinde modeli test etmişlerdir [23].

Chen ve arkadaşları, havayolu karlılığını arttırmaya yönelik olarak yaptıkları çalışmalarında MARS yöntemini kullanmışlardır. Kar yönetimi şirket kazancını arttırmayı hedefler. Rekabetçi bir havayolu endüstrisi için kar payı iyi bir kar yönetimi politikasına bağlıdır. Havayolu kar yönetimi üzerinde bolca araştırma yapılmaktadır. Fakat bunlar halen deneysel ve belirli alanlara yöneliktir. 20 şehir ve 31 uçuş kolundan oluşan bir yerel hava yolu taşıyıcısının tek merkezden yönetimli ağı için kar yönetimi merkezine bir çözüm önerisi oluşturabilmek için çalışmalarını gerçekleştirmişlerdir. Hava yolu ağları birbirlerinden genellikle merkez şehirlerle ayrıldığından bu

gerçekçi bir örnek teşkil eder. Çalışmalarında MARS stokastik dinamik programlama metodunun bir türevini kullanılmışlardır. Çalışmada kullanılan metod, kar yönetiminin en son teknolojisini ortaya koymaktadır [24].

LeBlanc ve Crowley; Yaşam analizi verilerinin araştırılması için MARS model yapısı geliştirmişlerdir. Bu metod, kovaryans ve düğümlerin uyarlanabilir seçimi ve MARS tekniğinin ağırlıklandırılmış en küçük kareler versiyonu ile Cox regresyon modelinin bir birleşimidir. Bu yeni teknik kovaryanslar arasındaki etkileşimler ve doğrusal olmayan etkileri gösteren terimlerle en uygun modeli otomatik olarak seçer. Çalışmada temel simülasyon verileri ve miyelom (myeloma: omurilik tümörü) için klinik veriler incelenmiş ve sonuçta miyelom uygulamalarında bilinen hastalık belirti değişkenlerinin bir laboratuvar değişken ile tekdüze olmayan bir ilişki içerdiği gözlenmiştir [25].

Z.Huang ve HaipengShen; Doğrusal zaman serilerinde fonksiyonel regresyon modelinin katsayı tahmini için polinom kesitlerinde düzgünleştirilmiş temel metodu önermişlerdir. Tutarlılık kovaryans oranı sonuçlarının bu tahmin metodlarını desteklediğini söylemektedirler. Mesafe tahminleri ve yoğunluk tahmininde çoklu adım ileri tahmin için kestirilen modelin kullanılabileceğini bildirmişlerdir. Bu yöntemi, simülasyonlar ve iki gerçek örnek veri seti ile denemişlerdir [26].

Leathwick ve arkadaşları, 15 tatlı su balığı türünün bulunma olasılıkları ile bu canlıların akarsulardaki yaşam alanlarını belirleyen çevresel değişkenler arasındaki ilişkileri, Yeni Zelanda akarsularındaki balık dağılımlarına ilişkin dev bir veri setini kullanılarak MARS ile incelemişlerdir. Her tür için yaşam koşullarını en iyi belirleyen çevresel faktörler bu analizler sonucunda saptanmıştır. Analizler sonucunda; her tür için ideal yaşam alanını belirleyen faktörler ortaya çıkarılmış, bu sonuçlar coğrafi bilgi sistemine işlenerek bu türlerin olası yaşam alanlarının haritaları çıkarılmıştır [27].

R.Xiong ve J.F.Meullent, müşterilerin peynir çubukları (Cheese Stick) tercihlerini etkileyen sebepleri araştırmışlar ve etkileri MARS uygulaması ile bulmaya çalışmışlardır. MARS tahmin ediciler arasındaki ilişkiyi ortaya çıkarabilmiş, ayrıca parçalı regresyon fonksiyonu yardımı ile de tahmin edici değişkenlerin bağımlı değişken (peynir çubuklarını tercih sebebi) üzerine etkilerini saptayabilmişlerdir. Bu çalışma MARS'ın karmaşık veriden elde edilebilecek mümkün olan en iyi modeli oluşturulabildiğini göstermiştir [28].

Bigelow ve Dunson; yaptıkları çalışmada hiyerarşik fonksiyonel veri analizine uygun olarak; menstrual dönemde üreme hormon profillerinin çalışmalarını incelemişlerdir. Yaygın metotlarda devir uzunluğu standardize edilir ve devir içindeki yumurtlama zamanları görmezlikten gelinir ki her ikisi de biyolojik bilgi vericilerdir. İlave olarak, birçok devir için veri toplandığı zaman bağımlılığı hesaplamaya ihtiyaç vardır. Çözüm için genelleştirilmiş hiyerarşik Bayesian MARS temel yöntemi önerilmiştir. Kovaryans ilişkisinde; kadına özgü yörünge ve popülasyonun ortalama karakteristiği temel fonksiyonlarının setini bilmediğimiz için bizim formülasyonumuz gerçeği yansıtmayabilir. Sonraki hesaplamalar için iki yönlü sıçramalı Markov Zincirini kullanan Monte-Carlo algoritmasını geliştirmişlerdir. Kuzey Carolina'da yapılan gebeliğin ilk zamanlarındaki çalışmalarında gebe kalma ve gebe kalmama devirleri arasındaki üriner progesteron verilerindeki farklılıkları araştırmışlardır [29].

Abraham ve arkadaşları; yeni ve esnek regresyon modeli olan çokdeğişkenli uyarlanabilir regresyon kesitlerini (MARS), üç farklı derinlikte toprak sıcaklığı benzetimde uygulanmışlardır. MARS otomatik olarak çalışılan, verileri çok sayıda gruplara ayırmada böl ve keşfet (divide and conquer) yaklaşımını kullanmaktadır. Her bir grupta, regresyon doğrusu veya hiper düzlem oluşturulmaktadır. Diğer keşfedici hesaplama teknolojileriyle karşılaştırıldığında; MARS hızlı, esnek ve modelin önemli girdilerini belirlemede etkilidir. Modelin girdileri; yılın günlerini, maksimum ve minimum sıcaklıkları, yağış miktarını ve potansiyel buharlaşmayı içermektedir. Çıktılar

ise, 100, 500 ve 1500 mm derinlikteki toprak sıcaklığını içermektedir. MARS ın performansı yapay sinir ağı (ANN) ile karşılaştırılmıştır. Doğrusal regresyonun korelasyon katsayısı MARS ve ANN’de daima 0.950’nin üzerinde olmuştur. MARS ayrıca yılın gününün, ardından da minimum hava sıcaklığının çıktı üzerindeki en önemli girdi olduğunu göstermiştir. Sonuçlar regresyon teknolojisi olarak MARS’ın potansiyelini tarımsal uygulamalarda göstermiştir [10].

3. UYGULAMA

3.1. Araştırmanın Amacı

Bu çalışmada bir yüksek öğrenim kız yurdundaki farklı sınıf ve okullarda okuyan kız öğrencilerin fonksiyonel olmayan tutum puanlarının kestirebilmesi için bir model geliştirilmesi amaçlanmıştır. Öğrencilerin ilgilenilen değerlerindeki (yaş, gelir vb. bağımsız değişkenler) değişikliklerin, fonksiyonel olmayan tutum puanı üzerine etkilerinin bilinmesi ve olası olumsuzluklara karşılık önceden tedbir alınmasını sağlayabileceği düşünülmektedir. Kullanılan veri seti, sadece bu çalışma için yapay olarak oluşturulmuştur.

3.2. Araştırmanın Evreni

Çalışma verileri yapay olarak oluşturulmakla birlikte, çalışmanın Ankara'daki üniversitelerde okuyan ve yüksek öğrenim yurtlarında kalan kız öğrenciler araştırma evrenini oluşturmaktadır. Örneklem olarak evreni temsil ettiği düşünülen toplam 372 kız öğrenci alınması planlanmıştır. Veri seti oluşturulurken öğrencilerin aile yapısı (AILE), gelir düzeyi (GELIR), aile ilişkileri (ILISKILER), mezun oldukları lise türü (LISE), okudukları bölüm türü (GRUP), ek iş yapıp yapmadıkları (EK_IS), geldikleri şehir yapısı (SEHIR), öğrencinin yaşı (YAS), vücut kitle indeksi (BMI), not ortalaması değişkenleri (NOT_ORT) bağımsız değişken olarak belirlenmiştir. Fonksiyonel olmayan tutum puanı ise bağımlı değişken olarak alınmış ve ilgili literatüre [30] uygun olarak tesadüfi olarak belirlenmiştir.

3.3. MARS Modelinin Kurulması

Çalışmanın model öğrenme grubu verileri Salford şirketinin MARS 2.0 programına yüklenmiştir. Model kurulumu aşamasında Çizelge 3.1.'deki değişkenler yanlarında belirtilen özellikleri ile modele alınmıştır.

Çizelge 3.1. MARS Modeline Alınan Değişkenler ve Değişken Türleri

Sıra No	Değişken Adı	Türü	Kategorileri
1.	Yaş	Sürekli	17 – 22 yaş aralığında
2.	Grup	Kategorik	Okullar (A-D aralığında)
3.	Aile	Kategorik	Çekirdek – Geniş aile
4.	İlişkiler	Kategorik	Aile desteği; yok-yeterli-yetersiz-diğer
5.	Gelir	Sürekli	50-1000 YTL aralığında
6.	Lise	Kategorik	Lise türü; düz-anadolu-Süper-Diğer
7.	Şehir	Kategorik	Gelişmiş-Gelişmekte olan-Az Gelişmiş
8.	Ek iş	Kategorik	Var –Yok
9.	Vücut Kitle İndeksi	Sürekli	17.6 – 36.7 aralığında
10.	Not Ortalaması	Sürekli	29.83 – 99.00 aralığında

Modelin kurulması aşamasında kullanılan toplam 10 adet bağımsız değişken için (yapılan pek çok denemeden sonra) olası tüm ikili etkileşimleri de modele alabilmek için toplam temel fonksiyon sayısı 30 ve etkileşim sayısı 2 olarak belirlenmiştir. Modelin kurulması aşamasında kullanılan programın özelliği gereği oluşturulacak temel fonksiyon sayısı istenildiği kadar verilebilir. Benzer şekilde aynı anda etkileşimde olabilecek değişken sayısı

da belirlenebilir. Bu çalışmada ikiden fazla değişkenin aynı anda etkileşimde olması yorum gücünü yaratacağından ve pratikte çok fazla kullanılamayacağından etkileşim sayısı 2 olarak belirlenmiştir. Bağımlı ve bağımsız değişkenler ve değişkenlere ait tanımlayıcı istatistikler Çizelge 3.2’de sunulmuştur.

Çizelge 3.2. Bağımlı ve Bağımsız Değişkenlere Ait Tanımlayıcı İstatistikler

SIRA NO	Değişkenler	N	Minimum	Maksimum	Ortalama	SS
1.	FONK_PUAN	372	98.00	253.00	173.89	23.95
2.	SINIF	372	4	1	-	-
3.	YAS	372	17.00	23.00	20.05	1.31
4.	BMI	372	17.60	36.679	24.31	4.07
5.	NOT_ORT	372	29.83	99.00	69.18	12.73
6.	AILE	372	2	1	-	-
7.	ŞEHİR	372	3	1	-	-
8.	GELİR	372	50.00	1.000.00	557.60	202.29
9.	LİSE	372	4	1	-	-
10.	İLİŞKİLER	372	3	1	-	-
11.	EK_İŞ	372	2	0	-	-

SS: Standart Sapma

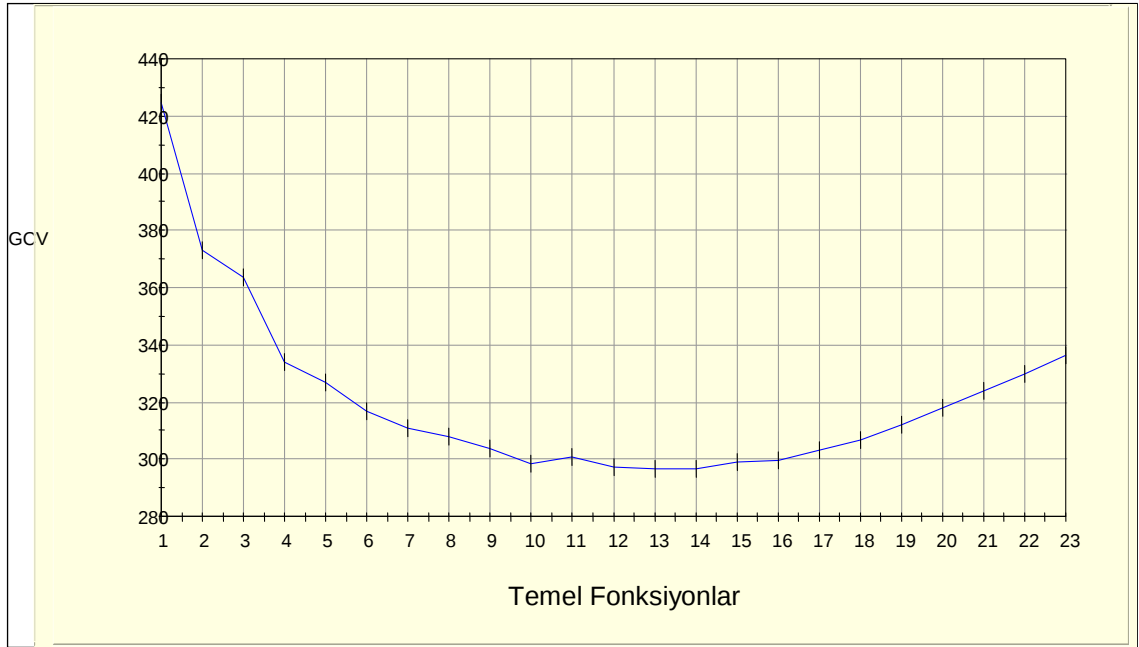
4. BULGULAR

MARS modeli adımsal ilerleme ve gerileme modeli olduğundan önce maksimum karmaşıklığa ulaşınca kadar sonuç fonksiyonuna mümkün olan tüm temel fonksiyonlar basamak basamak eklenir. Maksimum karmaşıklığa ulaşıldığında budama aşaması başlar. Biz de başlangıç modeli olan minimum model ve maksimum modeli inceledikten sonra GCV değerine göre en uygun modeli inceleyeceğiz.

Çizelge 4.1. Kurulan Modellerin Sonuçları

Temel Fonksiyonlar	Toplam Değişken Sayısı	Doğrudan Kullanılan Değişken Sayısı	GCV	GCV R-Square
23	7	7	316.43	0.42
22	7	7	329.96	0.43
21	7	7	323.70	0.44
20	7	7	317.83	0.45
19	7	7	312.21	0.46
18	7	7	306.87	0.47
17	7	7	303.05	0.47
16	7	7	299.82	0.48
15	7	7	298.74	0.48
14	7	7	296.78	0.48
**13	7	7	296.44	0.48
12	7	7	297.43	0.48
11	7	7	300.94	0.48
10	6	6	298.63	0.48
9	6	6	303.76	0.47
8	6	6	307.59	0.47
7	6	6	310.80	0.46
6	6	6	317.00	0.45
5	6	6	326.63	0.43
4	4	4	313.75	0.42
3	3	3	363.44	0.37
2	2	2	373.10	0.35
1	1	1	424.53	0.26

Çizelge 4.1’de ilk sütunda MARS modelimizde oluşturulan her bir regresyon modelindeki temel fonksiyon sayısı, ikinci sütunda her bir regresyon modelindeki temel fonksiyonların yapısında kullanılan toplam bağımsız değişken sayısı, üçüncü sütunda doğrudan kullanılan bağımsız değişken sayısı, dördüncü sütunda GCV ve son süründe GCV- R^2 değerleri verilmiştir. Buradan da görüleceği gibi GCV değeri en küçük olan model 13 temel fonksiyondan oluşan modeldir. Burada en karmaşık yapıda 23 adet, başlangıç modelinde ise 1 adet temel fonksiyon bulunmaktadır.



Şekil.4.1. GCV değerlerin Grafıksel Gösterimi

4.1. Minimum Model

Çalışmadaki 10 bağımsız değişken ve bu bağımsız değişkenlerin ikili etkileşimlerini içeren toplam 23 temel fonksiyon oluşturulmuştur. Başlangıç modeli olarak bulunan modele ilişkin değerler Çizelge 4.2’de gösterilmiştir.

Çizelge 4.2. Minimum Modele İlişkin Bilgiler

Fonksiyon	Standart	Model Dışı	Temel	Değişkenler
	Sapma	Bırakma Maliyeti	Fonksiyon Sayısı	
1	12.51	575.00	1	GRUP

Minimum modelde Çizelge 4.2’de de görüldüğü gibi sadece GRUP değişkeni modele alınmıştır ve bir temel fonksiyon kullanılmıştır. Modelin açıklayıcılık değeri $R^2 = 0.273$, düzeltilmiş açıklayıcılık değeri $R_D^2 = 0.271$ ve genelleştirilmiş çapraz geçerlilik açıklayıcılık değeri GCV $R^2 = 0.262$ olarak belirlenmiştir. Modelin genelleştirilmiş çapraz geçerlilik değeri (daha önce açıklandığı gibi en iyi modeli seçmek için en küçük GCV değerine sahip olan model tercih ediliyordu) GCV = 424.53 olarak belirlenmiştir. Başlangıç modelin hata kareler ortalaması MSE = 417.79 olarak hesaplanmıştır. Minimum model ait varyans çözümlemesi (ANOVA tablosu) çizelge 4.3’te verilmiştir.

Çizelge 4.3. Minimum Model Varyans Çözümleme Çizelgesi

Fonksiyon	Standart Sapma	Model dışı Bırakma Maliyeti	BF Sayısı	Değişkenler
1	12.51	575.00	1	GRUP

Değişkenlerin göreceli önemlilik derecelerine baktığımızda GRUP değişkeninin göreceli önemlilik derecesinin %100 olduğu ve diğer değişkenlerin önemsiz olduğu belirlenmiştir (Çizelge 4.4).

Çizelge 4.4. Modele Alınan Değişkenler ve Göreceli Önemlilik Yüzdeleri

Değişken	İhmal Edilen kalkış	Önemlilik Yüzdesi	Önemlilik Histogramı
GRUP	575.00	100.00	
YAS	424.53	0.00	
BMI	424.53	0.00	
NOT_ORT	424.53	0.00	
AILE	424.53	0.00	
SEHIR	424.53	0.00	
GELIR	424.53	0.00	
LISE	424.53	0.00	
ILISKILER	424.53	0.00	
EK_IS	424.53	0.00	

GRUP değişkenine bağlı olarak kurulan bir temel fonksiyon ve sonuç fonksiyonu aşağıda gösterilmiştir.

$$BF1 = (GRUP = 1 \text{ OR } GRUP = 2);$$

$$Y = 186.00 - 25.02 * BF1;$$

$$\text{model FONK_PUA} = BF1$$

4.2. En Karmaşık Model

En karmaşık modeldeki fonksiyonlar, katsayıları ve düğüm değerleri Çizelge 4.5.'de verilmiştir.

Çizelge 4.5. En Karmaşık Modele ilişkin bilgiler

Temel Fonksiyon	Katsayı	Değişken	Kaynak Değişken	Düğüm Noktası
0	186.46			
1	-30.17	GRUP		0.00
3	0.02	GELİR		280.00
4	-0.01	GELİR		280.00
5	-1.36	BMI		22.59
6	-0.10	BMI		22.59
7	-3.18	YAS		20.00
8	-11.37	YAS		20.00
9	-6.90	LİSE	YAS	0.00
11	0.03	GELİR	YAS	450.00
12	0.02	GELİR	YAS	450.00
13	-2.57	NOT_ORT	YAS	87.73
14	-0.09	NOT_ORT	YAS	87.73
15	23.32	YAS	GRUP	22.00
16	7.06	YAS	GRUP	22.00
17	1.14	NOT_ORT	GRUP	76.48
18	0.08	NOT_ORT	GRUP	76.48
19	-7.19	GRUP	YAS	0.00
21	-0.13	GELİR		870.00
23	0.06	GELİR		650.00
25	-3.77	EK_IS	BMI	0.00
27	-0.09	GELİR	BMI	870.00
28	-0.01	GELİR	BMI	870.00
29	-5.05	LİSE	YAS	0.00

Oluşturulan bu en karmaşık modelde Çizelge 4.5'de de görülebileceği gibi toplam 7 bağımsız değişken ve 7 bağımsız değişkenin ikili etkileşimlerini içeren 23 temel fonksiyon oluşturulmuştur. Oluşturulan modelin açıklayıcılık değeri $R^2 = 0.610$, düzeltilmiş açıklayıcılık değeri $R_D^2 = 0.584$ ve genelleştirilmiş çapraz geçerlilik açıklayıcılık değeri GCV $R^2 = 0.415$ olarak belirlenmiştir. Modelin genelleştirilmiş çapraz geçerlilik değeri GCV = 316.429 olarak belirlenmiştir. En karmaşık modelin hata kareler ortalaması MSE = 238.593 olarak hesaplanmıştır.

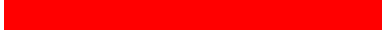
Kurulan modele ilişkin varyans çözümleme çizelgesi oluşturulmuştur. Buna göre sadece GRUP değişkeninin olduğu fonksiyonun standart sapması 15.077, model dışı bırakma maliyeti 384.117 ve etkili parametre sayısı 2.957 olarak belirlenmiştir. Elde edilen varyans çözümleme sonuçları Çizelge 4.6 da verilmiştir.

Çizelge 4.6. En Karmaşık Model Varyans Çözümleme Çizelgesi

Fonksiyon	Standart Sapma	Model dışı Bırakma Maliyeti	BF Sayısı	Değişkenler	Kaynak Değişkenler
1	15.077	384.117	1	GRUP	
2	6.839	325.222	4	GELİR	
3	4.358	310.684	2	BMI	
4	8.154	321.153	2	YAS	
5	3.351	317.758	2	YAS	LISE
6	6.197	316.915	2	YAS	GELİR
7	4.211	316.944	2	YAS	NOT_ORT
8	8.301	319.506	3	GRUP	YAS
9	3.975	315.156	2	GRUP	NOT_ORT
10	4.799	316.600	1	BMI	EK_IS
11	2.831	328.698	2	BMI	GELİR

Modele alınan değişkenler ve bu değişkenlerin göreceli önemlilikleri Çizelge 4.7’de verilmiştir. Burada görülebileceği gibi GRUP, GELİR, YAŞ, BMI, LİSE ve EK İŞ değişkenleri önemli (modele alınması gerekli) olarak bulunurken, NOT ORT, AİLE, ŞEHİR ve İLİŞKİLER değişkenleri önemsiz (modele alınması gereksiz) olarak belirlenmiştir.

Çizelge 4.7. Modele Alınan Değişkenler ve Göreceli Önemlilik Yüzdeleri

Değişken	İhmal Edilen Maliyet	Önemlilik Yüzdesi	Önemlilik Histogramı
GRUP	387.504	100.000	
GELİR	383.240	95.735	
YAS	355.905	61.750	
BMI	326.600	44.623	
LISE	317.758	16.126	
EK_IS	316.600	5.776	
NOT_ORT	328.872	0.000	
AİLE	316.429	0.000	
SEHIR	316.429	0.000	
ILISKILE	316.429	0.000	

Bağımsız değişkenlerin önemliliklerini dikkate alarak oluşturulan temel fonksiyonlar ve temel fonksiyonları kullanarak oluşturulan sonuç modeli;

$$BF1 = (GRUP = 1 \text{ OR } GRUP = 2);$$

$$BF3 = \max(0, GELİR - 280.00);$$

$$BF4 = \max(0, 280.00 - GELİR);$$

$$BF5 = \max(0, BMI - 22.59);$$

$$BF6 = \max(0, 22.586 - BMI);$$

$$BF7 = \max(0, YAS - 20.00);$$

$$BF8 = \max(0, 20.00 - YAS);$$

$$BF9 = (LİSE = 1 \text{ OR } LİSE = 4) * BF7;$$

$$BF11 = \max(0, GELİR - 450.00) * BF8;$$

$BF12 = \max(0, 450.00 - GELIR) * BF8;$
 $BF13 = \max(0, NOT_ORT - 87.73) * BF8;$
 $BF14 = \max(0, 87.73 - NOT_ORT) * BF8;$
 $BF15 = \max(0, YAS - 22.00) * BF1;$
 $BF16 = \max(0, 22.00 - YAS) * BF1;$
 $BF17 = \max(0, NOT_ORT - 76.49) * BF1;$
 $BF18 = \max(0, 76.48 - NOT_ORT) * BF1;$
 $BF19 = (GRUP = 1) * BF8;$
 $BF21 = \max(0, GELIR - 870.00);$
 $BF23 = \max(0, GELIR - 650.00);$
 $BF25 = (EK_IS = 0) * BF6;$
 $BF27 = \max(0, GELIR - 870.00) * BF5;$
 $BF28 = \max(0, 870.00 - GELIR) * BF5;$
 $BF29 = (LISE = 2) * BF8;$

$Y = 186.455 - 30.17 * BF1 + 0.02 * BF3 - 0.01 * BF4 - 1.36 * BF5 - 0.10 * BF6 - 3.18 * BF7 - 11.37 * BF8 - 6.90 * BF9 + 0.03 * BF11 + 0.02 * BF12 - 2.57 * BF13 - 0.09 * BF14 + 23.32 * BF15 + 7.06 * BF16 + 1.14 * BF17 + 0.08 * BF18 - 7.19 * BF19 - 0.13 * BF21 + 0.06 * BF23 - 3.77 * BF25 - 0.09 * BF27 - 0.001 * BF28 - 5.05 * BF29;$

model FONK_PUA = BF1 BF3 BF4 BF5 BF6 BF7 BF8 BF9 BF11 BF12 BF13
 BF14 BF15 BF16 BF17BF18 BF19 BF21 BF23 BF25 BF27 BF28 BF29

şeklinde oluşmuştur.

4.3. En Uygun Model

GCV değerine göre en uygun modele ilişkin temel fonksiyonlar, katsayılar ve düğüm değerleri Çizelge 4.8’da verilmiştir.

Çizelge 4.8. En Uygun Modele İlişkin Bilgiler

Temel Fonksiyon	Katsayı	Değişkenler	Kaynak Değişken	Düğüm Değerleri
0	189.41			
1	-30.23	GRUP		0.00
5	-1.67	BMI		22.59
8	-11.57	YAS		20.00
9	-8.69	LISE	YAS	0.00
11	0.03	GELIR	YAS	450.00
13	-2.30	NOT_ORT	YAS	87.79
16	7.84	YAS	GRUP	22.00
17	1.08	NOT_ORT	GRUP	76.48
19	-7.43	GRUP	YAS	0.00
23	0.09	GELIR		650.00
25	-4.183	EK_IS	BMI	0.000
27	-0.131	GELIR	BMI	870.000
29	-5.912	LISE	YAS	0.000

En uygun modelde Çizelge 4.8’de de görülebileceği gibi toplam 7 bağımsız değişken ve 7 bağımsız değişkenin ikili etkileşimlerini içeren 13 temel fonksiyon oluşturulmuştur. Oluşturulan modelin açıklayıcılık değeri $R^2 = 0.586$, düzeltilmiş açıklayıcılık değeri $R_D^2 = 0.574$ ve genelleştirilmiş çapraz geçerlilik açıklayıcılık değeri GCV $R^2 = 0.484$ olarak belirlenmiştir. Modelin genelleştirilmiş çapraz geçerlilik değeri GCV = 296.435 olarak belirlenmiştir. En uygun modelin hata kareler ortalaması MSE = 246.182 olarak hesaplanmıştır.

En uygun modele ilişkin varyans çözümleme çizelgesi oluşturulmuştur. Buna göre sadece GRUP değişkeninin olduğu fonksiyonun standart sapması 15.105, model dışı bırakma maliyeti 359.389 ve etkili parametre sayısı 2.957 olarak belirlenmiştir. Elde edilen varyans çözümleme sonuçları Çizelge 4.9'da verilmiştir.

Çizelge 4.9. En Uygun Model Varyans Çözümleme Çizelgesi

Fonksiyon	Standart Sapma	Model dışı Bırakma Maliyeti	BF Sayısı	Değişkenler	Kaynak Değişkenler
1	15.105	359.389	1	GRUP	
2	5.457	319.290	1	BMI	
3	9.035	311.552	1	YAS	
4	6.470	326.623	1	GELİR	
5	4.116	305.676	2	YAS	LİSE
6	5.731	307.248	1	YAS	GELİR
7	3.580	301.046	1	YAS	NOT_ORT
8	9.174	309.601	2	GRUP	YAS
9	3.842	301.867	1	GRUP	NOT_ORT
10	5.320	317.686	1	BMI	EK_IS
11	3.751	305.887	1	BMI	GELİR

Modele alınan değişkenler ve bu değişkenlerin göreceli önemlilikleri Çizelge 4.10'da verilmiştir. Çizelge 4.9'dan da görülebileceği gibi GRUP, GELİR, YAŞ, BMI, EK_IS, LİSE ve NOT_ORT değişkenleri önemli (modele alınması gerekli) olarak bulunurken, AİLE, ŞEHİR ve İLİŞKİLER değişkenleri önemsiz (modele alınması gereksiz) olarak belirlenmiştir.

Çizelge 4.10. En Uygun Modele Alınan Değişkenler ve Göreceli Önemlilik Yüzdeleri

Değişken	İhmal Edilen Maliyet	Önemlilik Yüzdesi	Önemlilik Histogramı
GRUP	362.313	100.000	
GELİR	359.360	97.718	
YAS	316.643	78.112	
BMI	314.828	76.329	
EK_IS	317.686	56.788	
LISE	305.676	37.447	
NOT_ORT	298.901	19.327	
AİLE	296.435	0.000	
SEHIR	296.435	0.000	
İLİSKİLE	296.435	0.000	

En uygun modelde bulunan temel fonksiyonlar ve sonuç fonksiyonu aşağıda verilmiştir.

$$BF1 = (GRUP = 1 \text{ OR } GRUP = 2);$$

$$BF5 = \max(0, BMI - 22.586);$$

$$BF6 = \max(0, 22.586 - BMI);$$

$$BF7 = \max(0, YAS - 20.000);$$

$$BF8 = \max(0, 20.000 - YAS);$$

$$BF9 = (LİSE = 1 \text{ OR } LİSE = 4) * BF7;$$

$$BF11 = \max(0, GELİR - 450.000) * BF8;$$

$$BF13 = \max(0, NOT_ORT - 87.728) * BF8;$$

$$BF16 = \max(0, 22.000 - YAS) * BF1;$$

$$BF17 = \max(0, NOT_ORT - 76.483) * BF1;$$

$$BF19 = (GRUP = 1) * BF8;$$

$$BF23 = \max(0, GELİR - 650.000);$$

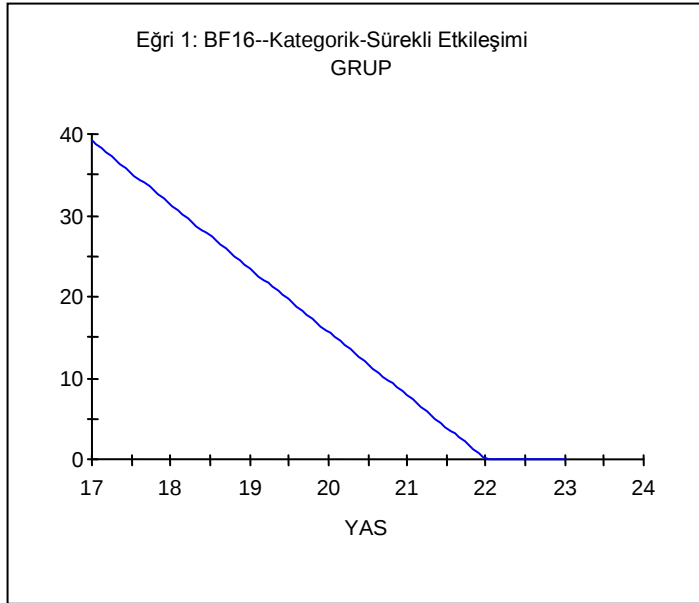
$$BF25 = (EK_İS = 0) * BF6;$$

$$BF27 = \max(0, GELİR - 870.000) * BF5;$$

$$BF29 = (LİSE = 2) * BF8;$$

$$Y = 189.408 - 30.225 * BF1 - 1.673 * BF5 - 11.570 * BF8 - 8.688 * BF9 + \\ 0.031 * BF11 - 2.304 * BF13 + 7.839 * BF16 + 1.082 * BF17 - 7.429 * BF19 + \\ 0.093 * BF23 - 4.183 * BF25 - 0.131 * BF27 - 5.912 * BF29;$$

```
model FONK_PUA = BF1 BF5 BF8 BF9 BF11 BF13 BF16 BF17 BF19 BF23  
BF25 BF27 BF29
```

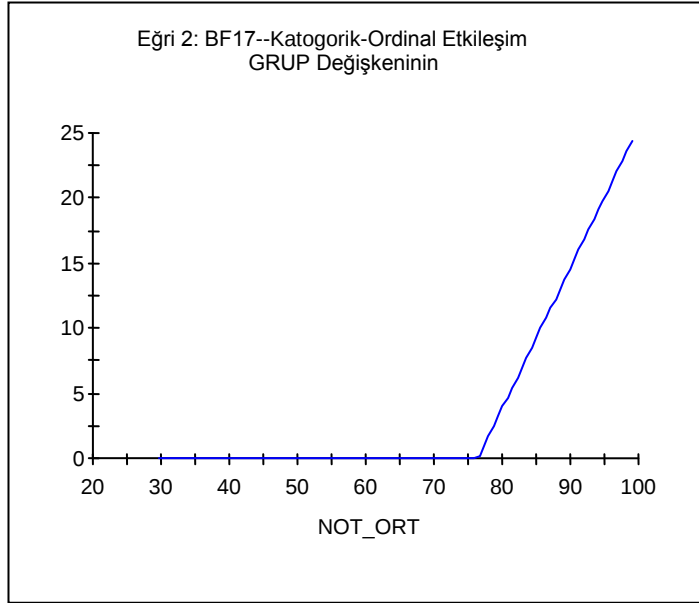


Şekil 4.2. GRUP ile FONK_PUAN arasındaki ilişki ve düğüm değeri

$BF1 = (GRUP = 1 \text{ OR } GRUP = 2);$

$BF16 = \max(0, 22.000 - YAS) * BF1;$

Bu modelde 1 veya 2 nci gruba dahil öğrencilerin yaşı 22'den küçük olanlar bağımlı değişkenle ters yönde bir ilişki gösterirken, yaşı daha büyük olanlarda ilişkisi sıfır değerini alıyor.

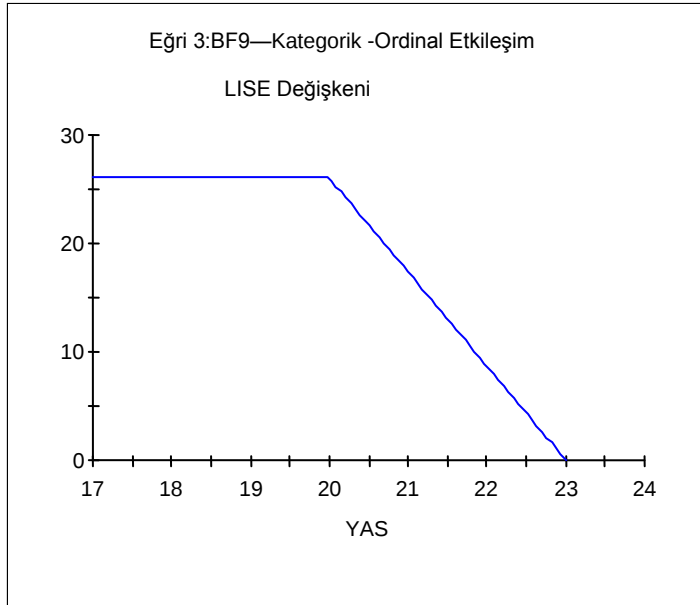


Şekil 4.3. NOT_ORT ile FONK_PUAN arasındaki ilişki ve düğüm değeri

$BF1 = (GRUP = 1 \text{ OR } GRUP = 2);$

$BF17 = \max(0, NOT_ORT - 76.483) * BF1;$

Burada not ortalaması 76.483 ve daha büyük olup 1 ve 2 nci gruba dahil olan öğrencilerin Fonksiyonel Olmayan Tutum Puanlarında artış gözlenmektedir.

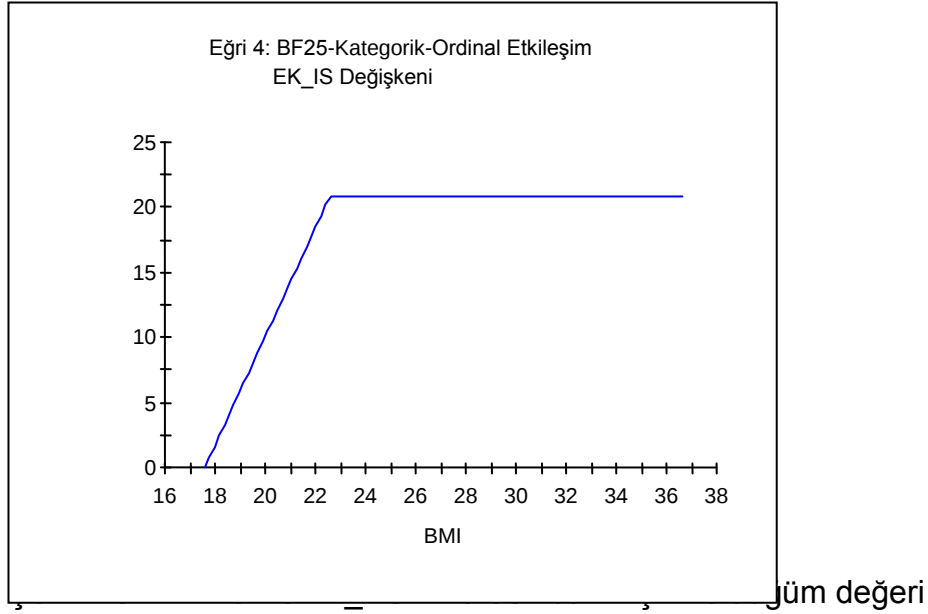


Şekil 4.4. YAS ile FONK_PUAN arasındaki ilişki ve düğüm değeri

$$BF7 = \max(0, YAS - 20.000);$$

$$BF9 = (LISE = 1 \text{ OR } LISE = 4) * BF7;$$

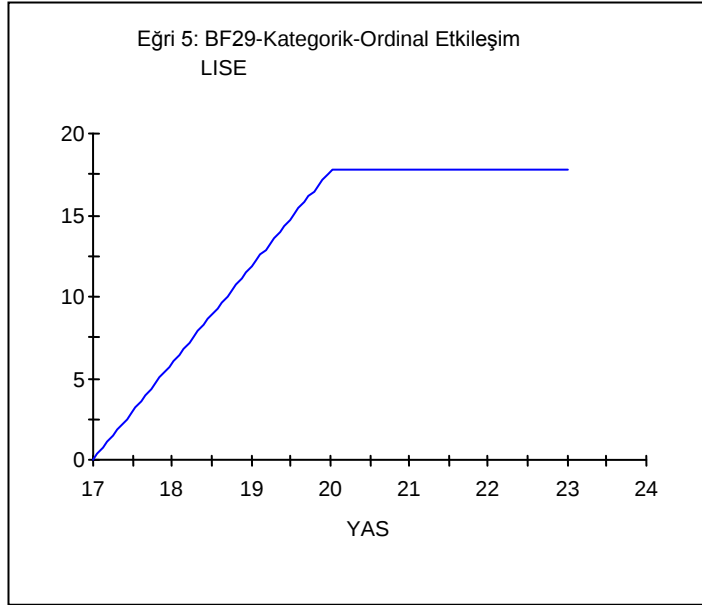
Yaşı 20 ve daha küçük olan öğrencilerden LİSE değişkeni 1 veya 2 nci kategoride olanların FONK_PUAN bağımlı değişkenine etkisi sıfır iken, yaşı 20 ile 23 yaş arası olan öğrencilerde ters yönde bir ilişki oluşturmaktadır.



$$BF6 = \max(0, 22.586 - BMI);$$

$$BF25 = (EK_IS = 0) * BF6;$$

Beden Kitle İndeksi 22.586'dan küçük olan öğrencilerin bağımlı değişkenle ilişkisi artan bir ilişki içerisinde olup, 22.586 değerinde bir düğüm değeri oluşturmaktadır. Bu değerden sonra ilişki sıfırdır.

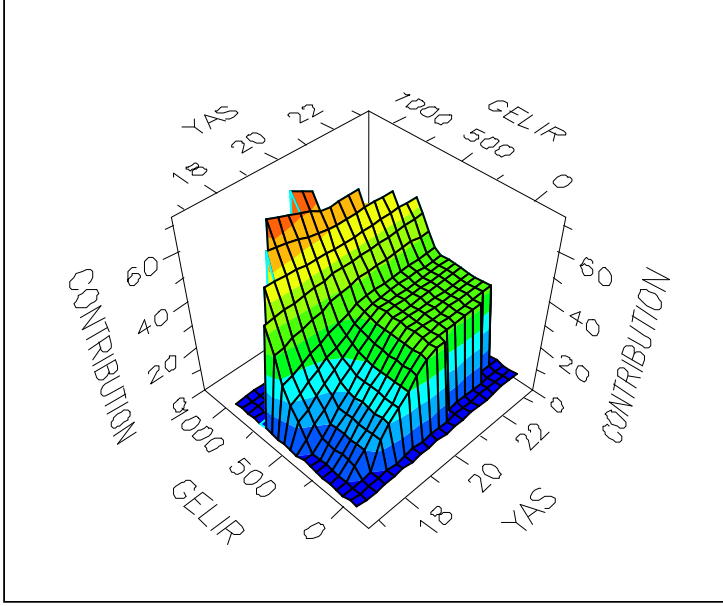


Şekil 4.6. YAS ile FONK_PUAN arasındaki ilişki ve düğüm değeri

$$BF8 = \max(0, 20.000 - YAS);$$

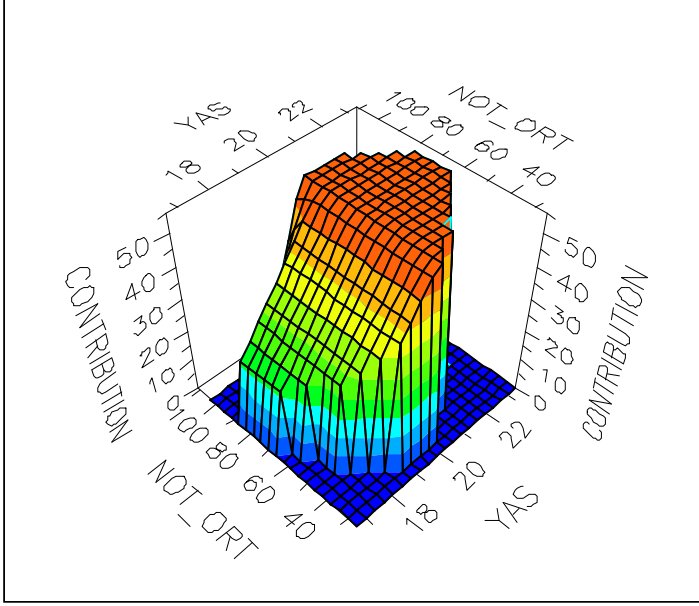
$$BF29 = (LİSE = 2) * BF8;$$

LİSE değişkeni değeri 2 olan öğrencilerin yaşı 20'den küçük olanların FONK_PUAN değişkeni ile artan doğrusal bir ilişki gösterdiği bu noktadan sonra ise ilişkinin sıfır değeri aldığı görülmektedir.



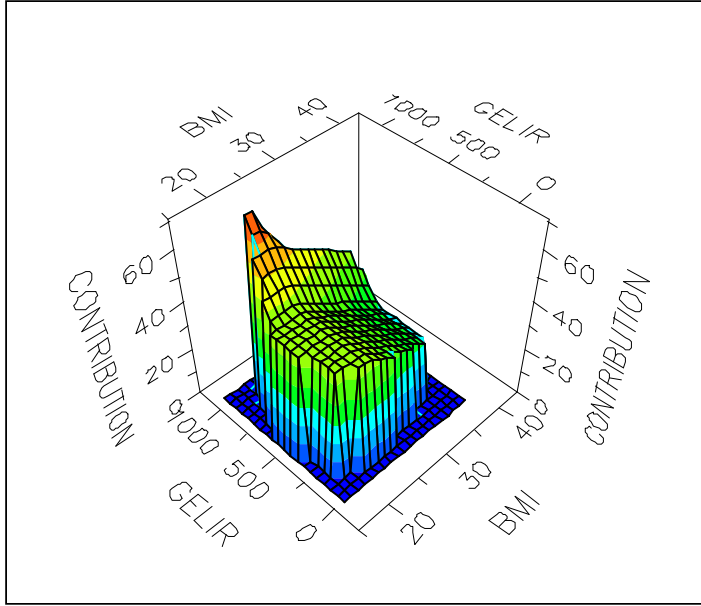
Şekil 4.7. YAS ve GELİR Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki

Burada gelirin 500-1000 YTL ve yaşı ise 17-20 arasında olduğu bölgelerde FONK_PUAN değişkeninin arttığı; Gelirin 0-500 YTL ve 17-20 arasında olduğu bölgelerde ise FONK_PUAN azaldığı gözlenmektedir.



Şekil 4.8. YAS ve NOT_ORT Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki

Bu düzlemde yaşı 20'den büyük olduğu bölgelerde FONK_PUAN oldukça yüksek olduğu, 20'den küçük olan bölgede gittikçe azalan bir değer gösterdiği gözlenmektedir.



Şekil 4.9. BMI ve GELİR Değişkenlerinin Etkileşiminin FONK_PUAN Bağımlı Değişkeni İle Arasındaki İlişki

BMI değişken değerinin 20-25 aralığında ve GELİR değişkeninin yüksek olduğu bölgelerde FONK_PUAN bağımlı değişkenin değerinin arttığı gözlenmektedir.

5. SONUÇ VE ÖNERİLER

Bu çalışmada MARS yöntemi modeli ile ilgili temel kavramlar tanıtılmıştır. Özellikle bağımlı değişkeninin açıklanmasında ve yorumlanmasında önemli sayılabilecek bilgiler sunulmuştur. Bağımsız değişkenlerin etkileşimlerini de modele almak modelin açıklayıcılık gücünü artırır. Uygulama bölümünde üniversitede okuyan ve yurttan kalan kız öğrencilerin Fonksiyonel Olmayan Tutum Puanlarını etkileyen sebepler araştırılmış ve yapay veri oluşturularak tahmin edilmeye çalışılmıştır.

Değişken yapısı farklı ve çokdeğişkenli bir denklem oluşturan regresyon modelinin çözümü oldukça zor bir uygulamadır. MARS modeli ile karmaşık denklem çözümleri yerine; bilgisayar ortamında basit doğrusal denklemlerden oluşan temel fonksiyonların çözümü ile regresyon modeli oluşturulmuştur.

Doğrusal modellerde katsayı değişkenin tüm değerleri için sabit kalırken, MARS modelinde düğüm değerlerine bağlı olarak katsayı farklı değerler almaktadır.

Bu uygulamada sürekli ve kesikli değişkenlerden oluşan bağımsız değişkenler kullanılarak bağımsız değişkene etkileri araştırılmıştır. Diğer regresyon yöntemlerinin çoğunda bu uygulamayı yapmamız imkansızdır.

Bağımlı değişkenimiz olan Fonksiyonel Olmayan Tutum Puanlarını etkileyen en önemli değişkenin oluşturulan grup ve sırası ile aylık gelirleri, yaşları, beden kitle indeksi, ek işinin olup olmadığı, mezun olduğu lise türü, not ortalaması olduğu; aile yapısının, geldiği şehrin ve ailesi ile ilişkisinin ise Fonksiyonel Olmayan Tutum Puanı üzerinde etkisiz olduğu bulunmuştur.

MARS modelinde bağımlı değişken ve bağımsız değişkenlerin değişken yapısı farklı olabilir (sürekli, kesikli). MARS modeli değişken yapısından bağımsızdır. Diğer modellerde değişken yapısının tek tip olması şartı aranır. Bu yönüyle modelin işlem yapabilme yeteneği ve açıklayıcılığı artar.

Bu çalışma ile karmaşık veri yapısına sahip, öngörüle bulunulması güç yada imkansız olan regresyon problemlerinde, bilgisayar teknolojisini de kullanarak hızlı sonuçlar elde eden MARS modelinin kullanımının avantajları gösterilmiştir.

KAYNAKLAR

1. İnternet: Copyright StatSoft “Multivariate Adaptive Regression Spline” www.statsoft.co.uk/SixSigma (2007).
2. İnternet: Vikipedi, Özgür Ansiklopedi, “Regresyon” [ttp://tr.wikipedia.org/wiki /Regresyon](http://tr.wikipedia.org/wiki/Regresyon) (2007).
3. Chatterjee, S., Hadi, A., Price, B., “Regression Analysis By Example”, 3 rd Edition., **John & Wiley**, Canada, 1-20 (2000).
4. Ünver, Ö., Gamgam, H., “Uygulamalı İstatistik”, **Siyasal Kitapevi**, Ankara, 273-274;282-283 (1999).
5. Özdamar, K., “SPSS ile Biyoistatistik”, 5.Bs., **Kaan Yayınevi**, Eskişehir, 419-421 (2003).
6. Friedman, J., “Discussion The Π Method for Estimating Multivariate Functions From Noisy Data” **Technometrics**,33:145-148 (1991).
7. Kolyshkina, I., Sylvia, W., “Enhancing Generalised Linear models with Data Mining”, **Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society**, Arlington, Virginia, 279-290 (2004).
8. Kaki, B.,Yeşilova, A., ŞEN, C., “Yarı Parametrik Regresyon Yönteminin Hayvancılıkta Kullanılması”, **4. Ulusal Zootekni Bilim Kongresi Sözlü Bildiriler Programı**, Van, 26-32 (2004).
9. Friedman, J.H., “Multivariate Adaptive Regression Splines” **Annals of Statistics**, 19(1):1-67 (1991).
10. Abraham, A., Steinberg, D., “MARS: Still an Alien Planet in Soft Computing?” , **Lecture Notes in Computer Science**, Springer Berlin / Heidelberg Pres, 2074/2001:235-244 (2001).
11. Tunay, K.B., “Türkiye’de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini”, **METU Studies in Development**, Ankara, 28(2):, 1-23 (2001).
12. Orekeci Temel, G., Çamdeviren, H., Yazıcı, A.C., “Regresyon Modellerine Alternatif Bir Yaklaşım: MARS”, **VIII. Biyoistatistik Kongresi Sözlü Bildiriler**, Bursa, 105-123 (2005).
13. Deichman, J., Eshgi, A., Haughton, D., Sayek, S., Teebagy, N., ”Application of Multiple Adaptive Regression Splines (MARS) in Direct

- Response Modelling”, *Journal of Interactive Marketing*, 16:15-27 (2002)
14. Kolyshkina, I., Sylvia, W., “Enhancing Generalised Linear models with Data Mining”, *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society*, Arlington, Virginia, 279-290, (2004).
 15. Abraham, A., Steinberg D., “Computational Science - ICCS 2001” *Springer*, Berlin / Heidelberg , 235-244 (2001).
 16. Mukkamala, R., Andrew, H., Sung, “Department of Computer Science”, *Elsevier Science Publishers B. V.*, Amsterdam, The Netherlands, 631-641 (2007).
 17. Briand, L.C, Freimut B., Vollei, F., “IESE; Using Multiple Adaptive Regresyon Splines to Understand Trends in Inspection Data And Identify Optimal Inspection Rates”, *Software Engineering Research Network Technical Report*, Germany, 5-10 (2000).
 18. Sephton, P., “Forecasting Recessions: Can We Do Beter on MARS?”, *Federal Reserve Bank of St.Louis Review*, by Sephton, Peter, 83(2), 39-49 (2001).
 19. Chen, I.F., Lee, T.A, “Two-Stage Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines”, *Elsevier*, 28:743-752 (2005).
 20. Internet: Salford Systems, MARS User Guide, <http://www.salfordsystems.com/424.php>.
 21. Kolyshkina, I., Sylvia, W., “Enhancing Generalised Linear models with Data Mining”, *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society – Arlington, Virginia*, 279-290 (2004)
 22. Mallick, B.K., Denison, D.G.T., Smith, A.F.M., “Bayesian Survival Analysis Using A Mars Model”, *Journal of Food Science*, 69(4): 131-134 (2004).
 23. Verzilli, C.J., Whittaker, J.C., Stallard, N., Chasman, D., “A Hierarchical Bayesian Model For Predicting The Functional Consequences Of Amino-Acid Polymorphisms”, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 191–206 (2005).

24. Chen, V.C.P., Günther, D., Johnson, E.L., "Solving for an optimal airline yield management policy via statistical learning", ***Biometrics***, 55(1):204 (1999).
25. LeBlanc M., Crowley M., "Adaptive Regression Splines in the Cox Model", ***Biometrics*** , 55(1):204-213 (1999).
26. Huang, J.Z., Shen, H., "Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach", ***Scandinavian Journal of Statistics***, 31: 515-534 (2004).
27. Leathwick, J.R., Elith , J. and Hastie, T., "Comparative Performance Of Generalized Additive Models And Multivariate Adaptive Regression Splines For Statistical Modelling Of Species Distributions", ***Ecological Modelling***, 199:188-196 (2006).
28. Xiong, R., Meullent, J.F., "Application Of Multivariate Adaptive Regression Splines (MARS) To The Preference Mapping Of Cheese Sticks" ***Journal of Food Science***, 69(4):131-139 (2004).
29. Jamie L. Bigelow, David B. Dunson, "Bayesian Adaptive Regression Splines for Hierarchical Data",***Biometrics***, 63 (3): 724–732 (2007).
30. Özgüven H.D., "Psikiyatrik Kriz Vakalarında Hatalı Otomatik Düşünceler Ve Fonksiyonel Olmayan Tutumların Sıklığı", ***Kriz Dergisi***, 7(2):9-16, (1999).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : ÖZFALCI ÜSTÜNER, Yasemin
Uyruğu : T.C.
Doğum tarihi ve yeri : 12.08.1976 Samsun
Medeni hali : Evli
Telefon : 0 (312) 304 21 15
Faks : 0 (312) 304 21 00
e-mail : yaseminozfalci@mynet.com.

Eğitim Derece

Eğitim Birimi

Mezuniyet tarihi

Lisans	Gazi Üniversitesi/ İstatistik	1999
Lise	Namık Kemal Lisesi	1993

İş Deneyimi Yıl

Yer

Görev

2000-2002	Anadolubank AŞ.Samsun Şb.	Gişe Asistanı
2002-2008	GATA Personel Ş.Md.lüğü	İstatistikçi

Yabancı Dil

İngilizce

Hobiler

Müzik, Bilgisayar teknolojileri, Yüzme