T.C.

ERZİNCAN ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON ZİNCİRLERİNİN İRDELENMESİ VE BİR UYGULAMA

Alparslan OĞUZ

107613005

MATEMATİK

ANABİLİM DALI

ERZİNCAN

2014

Yrd.Doç.Dr. Nurettin SAVAŞ danışmanlığında, Alparslan OĞUZ tarafından hazırlanan bu çalışma 27/03/2014 tarihinde aşağıdaki jüri tarafından Matematik Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Başkan : Doç.Dr. Arif DANE *İmza:*

Üye : Yrd.Doç.Dr. Mustafa KUDU *İmza:*

Üye : Yrd.Doç.Dr. Nurettin SAVAŞ (Danışman) *İmza:*

Yukarıdaki sonucu onaylarım.

Doç.Dr. ALİ SÜLÜN

Enstitü Müdürü

ÖZET

Yüksek Lisans Tezi

ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON ZİNCİRLERİNİN İRDELENMESİ VE BİR UYGULAMA

Alparslan OĞUZ

Erzincan Üniversitesi
Fen Bilimleri Enstitüsü
Matematik Anabilim Dalı

Danışman : Yrd. Doç. Dr. Nurettin SAVAŞ

Bu çalışmada, Genelleştirilmiş Doğrusal Modellerin (GLM) ve zincirlerin (splayn) teorik yapısından, birçok alanda kendine uygulama imkanı bulan ve zincir yapısını kullanan Çok Değişkenli Uyarlanabilir Regresyon Zincirleri (MARS) tekniğinin temel kavramlarından, modelin uygulama adımlarından, yapılan bazı çalışmalardan, tekniğin avantaj ve dezavantajlarından bahsedilmiştir. Tezin uygulama bölümünde, öğrencilere yapılan anket den elde edilen veri seti ile SPM 7 paket programının MARS modülü kullanılarak model kurulmuştur. Elde edilen en uygun modele ait değerler ile kullanılan değişkenler arasındaki etkileşimler incelenmiştir. Bu bağlamda, MARS tekniğinin kategorik ve sürekli değişkenler arasındaki etkileşimlerin uygulanabilirliği ve tüm analizlerde istikrarlı ve başarılı tahminler ortaya koyduğu sonucuna ulaşılmıştır.

2014, 74 sayfa

Anahtar kelimeler: MARS, Regresyon, Toplamsal Zincirler

ABSTRACT

Master Thesis

A RESEARCH ON MULTİVARİATE ADAPTİVE REGRESSİON SPLİNE AND A APPLİCATİON

Alparslan OĞUZ

Erzincan University

Graduate School of Natural and Applied Sciences

Department of Mathematics

Supervisor: Asst. Prof. Dr. Nurettin SAVAŞ

In this study was mentioned about advantages and disadvantages of the technique, some studies, application steps of model, main concepts Of Multivariate Adaptive Regression Splines (MARS)'s technique using the splayn structure and finds posssibility of its application in many areas and theoretical structure of Generalized Linear Models (GAM) and spline. In the application section of the thesis, data sets obtained with the students questionnaire were established model used to MARS modul of SPM 7 software package students survey data sets obtained with the SPM software package of 7 MARS model is built using modules

2014, 74 pages

Keywords: MARS, Regression, Additive Spline

TEŞEKKÜR

Bu çalışmada öncelikle araştırma konumun belirlenmesinde ve çalışmamın her aşamasında yardım ve katkılarını esirgemeyen, önerileri ile beni yönlendiren danışman hocam Sayın Yrd.Doç.Dr. Nurettin SAVAŞ'a, juri komitesinde çalışmalarım sonunda önemli katkılarda bulunan ve yönlendiren Sayın Yrd.Doç.Dr Mustafa KUDU ve Doç.Dr. Arif DANE'ye ve çalışmalarım süresince manevi yardımlarını gördüğüm tüm Kemah Meslek Yüksekokulu'nda çalışan mesai arkadaşlarıma saygı ve sevgilerimi sunarım.

Ayrıca çalışmalarım süresince birçok fedakarlıklar göstererek beni maddi ve manevi olarak destekleyen aileme ve eşime en derin duygularla teşekkür ederim.

Alparslan OĞUZ

Mart, 2014

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

β_i : Bilinmeyen Parametreler

μ : Ortalama

θ : Uyum Parametresi

σ^2: Varyans

η : Modelin Doğrusal Kısmı

ϕ : Standartlaştırılmış Normal Değişkenin Birikimli Dağılım Fonksiyonu

Kısaltmalar

BF : Temel Fonksiyon

GAM : Genelleştirilmiş Toplamsal Modeller

GCV : Genelleştirilmiş Çapraz Geçerlilik

$GCVR^2$: Genelleştirilmiş Çapraz Geçerlilik Açıklayıcılık Değeri

GLM : Genelleştirilmiş Doğrusal Modeller

MARS : Çok Değişkenli Uyarlanabilir Regresyon Zincirleri

MSE : Hata Kareler Ortalaması

R^2 : Açıklayıcılık Değeri

ŞEKİLLERİN LİSTESİ

TABLOLARIN LİSTESİ

1.GİRİŞ

Veri analizinde kullanılan ve en çok bilinen istatistik değerlendirme araçlarından biri, doğrusal regresyon modelidir. En basit durum, bir bağlantı değişken (olan *Y*) ve bir açıklayıcı değişken olan (*X)'in* n tane ölçüme sahip olmasıdır. Bu amaç doğrultusunda *Y'*nin ortalaması *X'*in bir doğrusal fonksiyonu olarak ifade edilebilir.

Doğrusal ve doğrusal olmayan regresyon modellerinde istatistiksel analiz yöntemleri bağımlı değişkenin normal dağıldığı varsayımına dayanmaktadır. Bağımlı değişkenin sürekli olmadığı durumları analiz etmek için de genelleştirilmiş modellerde bulunmaktadır. Örneğin; bir hastanın tedaviye verdiği cevap (sonuç) değişkeni 1 ve 0 değerlerini alabilir. Bir başka durumda belli zaman aralığından bir olayın kaç kez tekrarlandığı ilgi konusu olabilir. Bu duruma bir günde meydana gelen ölümle sonuçlanan trafik veya iş kazaları, bir yıl yada bir günde meydana gelen deprem sayısı örnek olarak verilebilir. Bu durumda bağımlı değişken sürekli değildir. Ayrıca değişkenleri sürekli olup da normal dağılım göstermeyen verilerde söz konusu olabilir. Bu tür verilerin analizine imkan verecek geliştirilmiş modeller Genelleştirilmiş Doğrusal Modeller dir (GLM).

GLM, ilk kez (Nelder ve Wedderburn, 1972) tarafından ileri sürülmüş ve çok geniş uygulama alanlarında kullanılmıştır. Bu alanlardaki ilk detaylı kitaplar (McCullagh ve Nelder, 1989) (Aitken vd., 1989) ve (Dobson, 1990) tarafından yazılmıştır ve daha sonraki yıllarda sayısız çalışmalar yapılmıştır. (Cengiz, 1997) genelleştirilmiş doğrusal modellerin genel bir özetini vermektedir.

Genelleştirilmiş doğrusal model analizi üç temel kavrama dayanmaktadır:

1.Bağımlı değişkenin dağılımı (hata yapısı).
2.Açıklayıcı değişkenlerin hatasız dağıldığı.
3.Bağıntı fonksiyonu (doğrusal açıklayıcı değişkenlerle bağımlı değişken ortalamasını ilişkilendiren kısım).

GLM, doğrusal ve doğrusal olmayan modellerinin bir bileşimi olarak ta görülebilir. Ancak burada bağımlı değişkenin normal dağılıma sahip olma zorunluluğu yoktur. Başka bir ifadeyle genelleştirilmiş doğrusal modellerin temel kavramlarından olan bağıntı fonksiyonu doğrusal ve doğrusal olmayan regresyon modellerine sahiptir.
GLM' de bağımsız değişkenle bağımlı değişken arasındaki ilişkiyi tanımlayan bağıntı fonksiyonunu oluşturmak için doğrusal ve doğrusal olmayan regresyon modellerinde kullanılan grafiksel ve istatistikler yöntemlerin benzerleri kullanılmaktadır.

GLM, modelleme için oldukça elverişli ve uygulanabilir araçlar olarak gözlenebilir faktörler ve açıklayıcı değişkenlerin durumlarını tahmin etmek için kullanılırlar. Bu modeller regresyon, varyans ve kovaryans analizi için kullanılan genel doğrusal modelleri kapsayan geniş bir model sınıfını göstermektedir. Dolayısıyla varyans analizi ve klasik regresyon modelleri daha çok genel yapılara izin verirler.

Çoklu regresyon modellerinde bağımlı değişkenin ortalama değeri açıklayıcı değişkenlerin bir kümesinin ve bağımlı değişkenlerin bir gözlenen değeri için elde edilen hata teriminin bir doğrusal fonksiyonu olarak ifade edilir. Klasik metotlar hata terimlerinin bağımsız, sıfır ortalama ve sabit varyanslı normal dağılıma sahip olduğu durumlarda uygulanır. Bağımlı değişkenin nitel olduğu durumlarda bu metot uygulanamaz. Bağımlı değişkenin nitel ve sürekli olduğu durumlarda ise bağımsız değişkenler bağımlı değişken arasındaki ilişki kolaylıkla ifade edilemeyebilir. Örneğin, herhangi bir hastalığın varlığını yada yokluğunu gösteren bir binom değişkeni göz önüne alınsın. Homojen bir grup için değişken, hastalığın sabit olasılığıyla binom dağılımına göre dağılacaktır. Çoğu zaman bireyler homojen bir gruptan gelmez. Hastalığa sahip olma çok şıklı birkaç değişken olabilir. Öyle ise hastalığın olasılığı ile açıklayıcı değişkenler arasında bir ilişki olacaktır. Doğrusal ilişki bazı açıklayıcı değişkenler için (0,1) aralığının dışında bazı olasılıklar vereceğinden bu ilişki doğrusal olmayabilir. Bir doğrusal regresyon uygulamak için yapılacak işlem bir dönüşümdür. Bu dönüşümler de lojistik regresyon ve probit analizlerdir. İkinci bir örnek, bağımlı değişken Poisson dağılımına göre dağıldığında

ve Poisson sürecinin beklenen değerinin açıklayıcı değişkenlerle ilişkisi olduğu durumlardır. Nitel değişkenin analizinde regresyon metotları 30-40 yıldır bilinmesine rağmen son 20 yıldır kullanılmaktadır. GLM bu tür bağımlı değişken yapılarının da modellemesini sağlamaktadır.

Modeldeki açıklayıcı değişkenlerin sayısı çok olduğunda parametrik olmayan metot performans göstermez. Açıklayıcı değişkenin çokluğu tahminlerin varyansını artırır. Açıklayıcı değişken sayısı artırıldığında boyut artacağından varyansın hızlı bir şekilde artması problemi "boyut problemi" olarak bilinir. Bir başka problem ilişkisinin yorumlanmasında ortaya çıkar. Bu zorlukların aşılması noktasında (Stone, 1985) toplamsal modelleri önermektedir. Bu modeller çok değişkenli regresyon fonksiyonları için bir toplamsallık sağlamaktadır. Toplamsal yaklaşımın iki türlü faydası vardır. Birincisi; toplamsal terimin her biri bir tek değişkenli düzleştirici kullanılarak tahmin edildiği için "boyut problemi" ortaya çıkmaz. İkincisi; her bir terimin tahmini bağımlı değişkeninin bağımsız değişkenlerle nasıl değiştiğinin açıklanmasıdır.

Toplamsal modellerin her türlü bağımlı değişken için genelleştirilmiş hali (Hastie ve Tibrishani, 1990), (Hastie, 1991), tarafından önerilmektedir. Bu durum Genelleştirilmiş Toplamsal Modeller(GAM) olarak adlandırılır. Bu modeller, bilinen GLM'nin Toplamsal Modellere modifiye edilmiş halidir. Bağımlı değişkenin ortalamasının bir toplamsal tahmin ediciye bağlı olduğu bu modellerde doğrusal olmayan bir link fonksiyonu kullanılır. GLM'de olduğu gibi GAM, bağımlı değişkenin dağılımının üstel dağılımlar ailesinden olmasını ister. GAM'la GLM arasındaki tek fark GAM'ın doğrusal tahmin edici olarak bilinmeyen düzleştirici fonksiyonları kullanmasıdır. Normal dağılımlı bağımlı değişken için Toplamsal Modellere, parametrik olmayan lojistik modeller ve parametrik olmayan log-doğrusal modeller örnek olarak verilebilir.

Doğrusal modeller her bir tahmin edicide bağımlı değişkeni doğrusal kabul ederken toplamsal modeller sadece düz bir şekilde bağımlı değişkenin her bir tahmin edici

tarafından etkilendiğini varsayar. Genelleştirilmiş Toplamsal Modeller (GAM), tahmin ediciler ve bağımlı değişkenler arasındaki esnek toplamsal olmayan ilişkileri modelleyen, doğrusal modeller ve GLM'nin genelleştirilmiş bir ifadesidir.

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi, o konu ile ilgili tahminler ya da kestirimler yapabilmek amacıyla regresyon modeli olarak adlandırılan matematiksel bir model ile karakterize edilen bir istatistik analiz tekniğidir. Regresyon modeli uydurulduktan sonra modelin yeterli olup olmadığının kontrolü regresyon analizinin en önemli bölümüdür. Uydurulan modelin doğru modele yeterli derecede yaklaştığını garanti etmek ve en küçük kareler regresyon analizinin tüm varsayımlarını sağlayıp sağlamadığını kontrol etmek gerekir. Eğer regresyon modeli yeterli uyum sağlamazsa zayıf veya yanıltıcı sonuçlar verecektir.

Dikkat edilecek olursa basit olmasına rağmen doğrusal modeller sık sık gerçek yaşamda etkilerin genellikle doğrusal olması nedeniyle başarısızlığa uğramaktır. Daha esnek istatistiksel modeller doğrusal olmayan regresyon etkilerini tanımlamak için kullanılabilir. Bu amaçlar için de GAM kullanılabilir.

Toplamsal modeller ve GAM, doğrusal modellerden ve GLM' lerden daha esnek iken parametrik olmayan regresyon modellerinden daha esnek bir yapıya sahiptirler.

Ayrıca toplamsal modelleri ve GAM' ları yarı parametrik fonksiyonlara ve yarı parametrik fonksiyonları da GAM'lara dönüştürmek mümkündür. Bunun için kovaryans ve açıklayıcı değişkenlerin parametrik formlarının modele katılması gerekir (Ruppert ve vd., 2003), (Fengler, 2005) ve (Wood, 2006a, 2006b).

Çok değişkenli uyarlanabilir regresyon zincirleri (MARS), regresyon tipi sorunları çözmek için Friedman tarafından yaygınlaştırılan tekniklerin bir uygulaması ve temel amacı bağımsız ya da belirleyici değişkenlerin kümesinden sonuç değişkeni yada bir sürekli bağımlının değerlerini tahmin etmektedir. Burada sürekli değişkenlerde

uydurma modelleri için mevcut metotlardan, doğrusal regresyonlardan çok sayıda vardır, Örneğin, Çoklu Regresyon, Genelleştirilmiş Doğrusal Modeller(GLM), doğrusal olmayan regresyon(Genelleştirilmiş Doğrusal/Doğrusal Olmayan Modeller), regresyon ağaçlar (Sınıflandırılmış Ve Regresyon Ağaçları), sinir ağları, ve benzeri.

MARS bağımlı ve bağımsız değişkenler arasında fonksiyonel ilişki altında yatanlar hakkında hiçbir varsayım yapmayan ve bir parametrik olmayan regresyon yöntemidir. Bunun yerine, MARS regresyon verisinden tamamen "sürülen" temel fonksiyonları ve katsayılarının bir kümesinden bu ilişkiyi kurar. Bir bakıma, yöntem " böl ve yönet" stratejisine dayanmaktadır ve bölgeler içindeki giriş alanı bölümlerinin her biri onun kendi regresyon denklemi iledir. Bu durumda özellikle daha yüksek girdi boyutları ile problemler için uygun MARS' ı oluşturur, burada çok yüksek boyutlar muhtemelen diğer teknikler için sorun yaratacaktır. MARS tekniği, ilginin bağımlı değişkeni ve belirleyici değişkenler arasındaki ilişkinin sınıfı (örneğin, doğrusal, biçimsel) ya da özellikle herhangi bir türü dayatmadığı yada varsaymadığı için özellikle veri madenciliğinin alanı içinde popüler hale gelmiştir. Bunun yerine yararlı modeller (örneğin, doğru tahminler verecek modeller), parametrik modeller ile yaklaşmak için zor ve monoton olmayan bağımlı değişkenler ve belirleyiciler arasındaki durumlarda bile elde edilebilir.

Regresyon problemleri bir ya da daha fazla bağımsız değişken ve bir bağımlı değişken kümesi arasındaki ilişkiyi belirlemek için kullanılır. Bağımlı değişken, bağımsız değişkenlerin değerlerine dayalı olarak, değerlerini tahmin etmek istediğimiz değişkenlerdir.

2. GENEL BİLGİLER

2.1. Genelleştirilmiş Doğrusal Modeller (GLM)

GLM' de *n* bileşenli *y* gözlem vektörü, bir *Y* tesadüfi değişkeninin değerlerini alır. Bu vektör bazı gözlemlenmiş açıklayıcı değişkene bağımlı olup, μ ortalamalı bağımsız dağılımlı bileşenlere sahiptir. Modelde, sistematik ve tesadüfi olarak iki kısım vardır. Sistematik kısım, bilinmeyen parametreler olan β_i'lerin oluşturduğu μ'leri anlatır ki bu;

$$\mu=\sum_{i=1}^{p} x_i\beta_i \qquad (2.1)$$

şeklinde gösterilir. β_i bilinmeyen parametrelerdir ve genellikle verilerden tahmin edilirler. Matris ifadesiyle

$$\mu=X\beta \qquad (2.2)$$

şeklinde gösterilir. Burada X bağımsız değişken matrisi ve β parametre vektörüdür (McCullagh ve Nelder, 1983). Model matrisi ise, modelde olması düşünülen tüm bağımsız değişkenleri içeren matristir. Tesadüfi kısım ise sabit varyanslı ve beklenen değeri sıfır olan hatalardan oluşur.

Bağımlı değişkenin dağılımı ile bu değişkene ait ortalamayı veren fonksiyonlar birbirinden bağımsız değildir. Çünkü bazı dağılımlar için verilen bazı bağıntı fonksiyonları çok daha uygun yaklaşımlar sağlar. Ayrıca seçilen bağıntı fonksiyonları türevi alınabilen ve monoton olmalıdır.(Myres, Montgomery ve Vining, 2001).

GLM' ler aşağıdaki üç özellikle karakterize edilirler;

- Üstel dağılım ailesinin bir üyesi,
- Bağımlı değişken veya link fonksiyonu,
- Tasarım (dizayn) vektörü.

2.1.1. Üstel Dağılım Ailesi

y, bağımlı değişkeninin olasılık yoğunluk fonksiyonu üstel dağılım ailesine mensup ise,

$$f(y_i \mid \theta, \emptyset) = \exp\{\frac{[\theta y_i - b(\theta)]}{a(\emptyset)} - c(y_i, \emptyset)\} \quad (2.3)$$

dir. Burada θ parametre, $a(\emptyset) = \emptyset/w_i$ bir fonksiyon ve θ uyum parametresidir. w_i, bilinen bir önsel ağırlık olup gözlemden gözleme değişir. Böylece her gözlemin ortalamasının μ olduğu, bağımsız değişkenli normal dağılımlı bir model için $a(\emptyset) = \sigma^2/m$ olur.

Üstel dağılımlı aileye mahsup olmanın genel yapısı normal dağılış üzerinde tanımlanırsa; normal dağılımda Y sürekli tesadüfi değişkeni için olasılık yoğunluk fonksiyonu;

$$f(y; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right.$$

$$= \exp\left[-\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] \quad (2.4)$$

Olarak bulunur. Burada;

$$a(\emptyset) = \sigma^2\,, \theta = \frac{\mu}{\sigma^2}\,, b(\theta) = -\frac{\mu}{\sigma^2}\,, c(y_i, \emptyset) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$$

2.1.2. Link Fonksiyonları

Olasılık yoğunluk fonksiyonu $f(y;\mu)$ olan bir Y tesadüfi değişkeni göz önüne alınsın. Burada Y'nin beklenen değeri μ'dür (E(Y)=μ). Her bir Y gözlemi için $(x_1, \ldots, x_{p)}$ açıklayıcı değişkenlerinin bir kümesinin var olduğu kabul edilirse bu durum μ değişkenlerinin değerlerine bağlıdır. μ'nün bazı dönüşümlerden sonraki değeri olan g(μ) doğrusaldır ve

$$\eta = g(\mu)=\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad (2.5)$$

Olur. Bu şekilde tanımlanmamış bağımlı değişken Y ve bağımsız değişkenler arasındaki ilişki olan genelleştirilmiş doğrusal modeldeki g(μ) dönüşümü, link (bağıntı) fonksiyonu olarak isimlendirilir. Bunun sebebi, modelin doğrusal kısmı olan η ile tesadüfi kısmı olan μ arasındaki bağlantıyı sağlamasıdır. Doğrusal fonksiyon η doğrusal tahmin edici y'nin dağılımı olan $f(y;\mu)$, hata dağılımıdır.

2.1.3. Normal Dağılım

Hata dağılımı normal olduğunda, klasik regresyon modeli aşağıdaki gibidir.

$$E(Y)=\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \quad (2.6)$$

Eşitlik (2.5)'de $\eta = \mu$ (veya g(μ)=μ) ise link fonksiyonu özdeş (birim)dir. Çoklu doğrusal regresyon, genelleştirilmiş doğrusal modeller ailesinin bir üyesidir.

Klasik doğrusal modellerle, ortalama ve doğrusal doğrusal tahmin ediciler benzerdir ve birim link, hem η hem de μ reel eksende aynı değeri aldığında uygundur.

2.1.4. Poisson Dağılımı

Bir Poisson değişkeninin beklenenini pozitif yani μ>0 ve aralık değeri (0,∞)'dur. Böylece birim link tercih edilebilirliğini kaybeder. Bir ölçüde μ bu şartla uymazsa η negatif olabilecektir. Bu aralığı (-∞,∞)' dönüştüren dönüşüm g(μ)=logμ'dür.

$$g(\mu)=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p \tag{2.7}$$

modeline log-doğrusal model denir.

2.1.5. Binom Dağılımı

Eşitlik (2.5)'deki doğrusal tahminci sınırsız bir aralığı içerdiğinden link fonksiyon Binom olasılığı μ'yü (0,1) aralığından (-∞,∞) aralığına dönüştürmelidir. En çok kullanılan dönüşümlerden biri probit dönüşümünü olup

$$g(\mu)=\phi^{-1}(\mu) \tag{2.8}$$

şeklinde yazılır. Burada ϕ standartlaştırılmış normal değişkenin birikimli dağılım fonksiyonudur. Diğer dönüşüm olan lojistik dönüşüm ise

$$g(\mu)=\ln\left[\mu/(1-\mu)\right] \tag{2.9}$$

bu şekilde ifade edilir (Hosmer ve Lemeshow, 2000). Bu dönüşüm lojistik regresyon modelidir. Yani

$$\ln\left[\frac{\mu}{1-\mu}\right]=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p$$

dir.

Probit dönüşümü biyolojik çalışmalarda uygun olmasına rağmen lojistik dönüşüm önemli avantajlara sahiptir. Lojistik dönüşümde hesaplama kolaydır ve sadece logaritma fonksiyonuna ihtiyaç duyulurken probit dönüşümde ise normal dağılım çizelgelerine ihtiyaç duyulur. İkincisi ve daha önemlisi, görevli orantıların kullanımıdır.

Sonuçta üç tane temel link fonksiyonu düşünülebilir.

i) logit ; $\eta = \log [\mu/(1-\mu)]$

ii) probit ; $\eta = \phi^{-1}(\mu)$ (2.10)

iii) tamamlayıcı log-log ; $\eta = \log [-log/(1-\mu)]$

Genelleştirilmiş doğrusal modellerin önemli bir yönü modeldeki uyum iyiliği istatistiğindeki değişmeler farklı modellerdeki açıklayıcı değişkenlerin alt setlerinin katkısını değerlendirmek için kullanılır. Maksimum erişilebilir log-olasılığı ve düşünülen modelin log-olasılığı arasındaki farkın iki katı şeklinde tanımlanan istatistik uyum iyiliğinin bir ölçüsü olarak sıkça kullanılır. Maksimum erişilebilir log-olasılığı her gözlem için bir parametreye sahip olan tam model ile başarılır. Genelleştirilmiş modeller bir modelin ifadesinde spesifik olarak maksimum olarak kullanılan terimlerle modellerin bir sıraya uyumunu anlamayı da sağlar. Modellerin her bir başarılı kısmı arasında log olasılıklarındaki farklı tablo halinde özetler.

Genelleştirilmiş modellerin diğer özellikleri aşağıda tanımlanmaktadır.

- Kullanıcı tarafından belirlenmiş farklılıkların olasılıklı oran istatistikleri yani parametrelerin doğrusal fonksiyonları ve onların asimptotik ki-kare dağılımlarına dayanan *p* değerleri.
- Kullanıcı tarafından belirlenmiş tahmin edici değerler, standart hatalar ve güven sınırları ve ayrıca en küçük kareler ortalaması.

- Prosedürün pek çok Çizelge ile gösterilen bir SAS data verilerini yaratabilme kabiliyeti.
- Model parametreleri için ya profil olasılık fonksiyonu veya asimptotik normal dağılımına dayanan güven aralıkları.

Üstel dağılım ailesine mensup, en çok kullanılan dağılımlara ait bilgiler Çizelge 2.1' de verilmiştir.

(McCullagh ve Nelder, 1989)

Tablo 2.1. Üstel dağılım ailesi bazı yaygın tek değişkenli dağılımların özellikleri.

	Normal	Poisson	Binominal	Gamma	Invers Gaussian
Dağılım ifadesi	$N(\mu,\sigma^2)$	$P(\mu)$	$B(m,\pi)$	$G(\mu,v)$	$IG(\mu,\sigma^2)$
y'nin değişim aralığı	$(-\infty,\infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0,\infty)$	$(0,\infty)$
Uyum parametresi ($\emptyset$)	$\emptyset=\sigma^2$	1	$1/m$	$\emptyset=v^{-1}$	$\emptyset=\sigma^2$
Birikimli fonksiyon:$b(\theta)$	$\emptyset^2/2$	$\exp(\theta)$	$\log(1+e^{\theta})$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y;\emptyset)$	$-\frac{1}{2}(\frac{y^2}{\emptyset}+\log(2\pi\theta))$	$-\log y!$	$\log\binom{\mathrm{m}}{\mathrm{my}}$	$vlog(vy)-\log y - log\Gamma(v)$	$-\frac{1}{2}\left\{\log(2\pi\emptyset y^3)+\frac{1}{\emptyset y}\right\}$
$\mu(\theta)=E(Y;\theta)$	θ	$\exp(\theta)$	$e^{\theta}/(1+e^{\theta})$	$-1/\emptyset$	$(-2\theta)^{1/2}$
Kanonik link:$\theta(\mu)$	Özdeş (birim)	log	logit	$1/\mu$	$1/\mu^2$
Varyans fonksiyonu:$V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

2.2. Zincir (Splayn)

En basit anlamda bir zincir aşağıdaki gibi tanımlanabilir. $a = t_0 < t_1 < \cdots < t_n = \mathrm{b}$ olsun. S(x) fonksiyonu aşağıdaki şartlara sahipse birinci dereceden zincir denir.

S(x) $[a,b]$ aralığı üzerinde sürekli

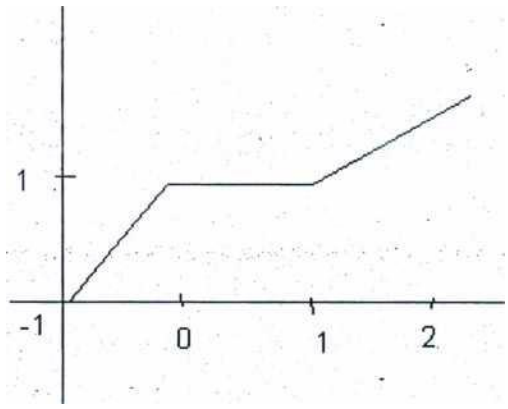
S(x) herbir $[\boldsymbol{t_i}, \boldsymbol{t_{i+1}}]$ $i = 0,1,2 \ldots, n-1$ aralığında doğrusal fonksiyonlardır.

Böylece, S(x) fonksiyonu n sayıda kırık doğrulardan oluşan fonksiyondur. $S(x) = a_i(t - t_i) + b_i$, $t_i \leq t \leq t_{i+1}$.

Bu tanım basit bir örnek üzerinde gösterilirse, S(x) sonksiyonu

$$S(x) = \begin{cases} x, & \boldsymbol{x} \in [-\mathbf{1}, \mathbf{0}] \\ 1, & \boldsymbol{x} \in (\mathbf{0}, \mathbf{1}) \\ 2x - 2, & \boldsymbol{x} \in [\mathbf{1}, \mathbf{2}] \end{cases} \tag{2.11}$$

şeklinde tanımlansın. Bu fonksiyonun zincir grafiği;



Şekil 2.1. Bir Dereceli Zincir

2.2.1.Basit Bir Zincirin Oluşması

$\boldsymbol{t_0, t_1, \dots, t_n \; ve \; y_0, y_1, \dots, y_n}$ verileri için, [a,b]=[t_0, t_n] aralığında aşağıdaki iki denklemin sağlanması gereklidir (Bartels vd., 1998). Bunlar;

1. $S_i(t_i) = y_i \; , i = 0, \dots, n$
2. $S_i(x) = \; a_i x + b_i \; , \; i = 0, \dots, n$

   $S_i(x) = \; y_i + m_i(x - t_i)$

$= y_i + \left(\frac{y_{i+1} - y_i}{t_{i+1} - t_i}\right)(x - t_i)$ (2.12)

(Christian 1967).

Adımlar:

$(n + 1)$ adet veri noktası girilir. $(t_0, y_0), (t_1, y_1), \dots, (t_n, y_n)$

Her bir aralıkta $S_i(x) = a_i x + b_i$ polinomu tanımlanır.

Her $[t_i, t_{i+1}]$ aralığında iki bilinmeyen veya toplam 2n bilinmeyen

$S_i(t_i) = y_i$ adet kısıt konur. Her aralık için *n*-1 =2 kısıt konursa toplam *2n* kısıt eklenir.

2.2.2. İki Dereceli Zincir (Splayn)

S(x) fonksiyonu 2 dereceli bir fonksiyondur. Eğer,

- S(x) in tanım aralığı [a,b] ise
- $S(x)$, [a b] üzerinde sürekli ise
- $S'(x)$, [a, b] üzerinde sürekli ise
- Her bir $[t_i, t_{i+1}]$ alt aralığında S(x)'in kuadratik olduğu bir
  a $= t_0 < t_1 < \cdots < t_n = b$ parçalanması vardır.

Her bir i için,

$S_i(x) = a_i x^2 + b_i x + c_i$ dir.

$$\mathrm{S(x)} = \begin{cases} S_0(x), & \boldsymbol{x} \in [t_0, t_1] \\ S_1(x), & \boldsymbol{x} \in [t_1, t_2] \\ & \vdots \\ S_{\mathrm{n}-1}(x), & \boldsymbol{x} \in [t_n, t_{n-1}] \end{cases} \tag{2.13}$$

Bunlar;

- Her bir aralıkta 3 bilinmeyen ve toplam $3n$ bilinmeyen vardır
- $2n$ kısıt
- $(n-1)$ tane iç nokta birinci türevler süreklidir. $S_i'(x_{i+1}) = S_{i+1}'(x_{i+1})$
- Ancak sadece $(n-1)$ tane kısıt
- $(3n-1)$ toplam kısıt
- Ekstra kısıt, örneğin $S'(x_0) = sabit$.

vardır.

2.2.3. Kübik Zincir (Splayn)

S fonksiyonu aşağıdaki 3 şartı sağlıyorsa k serbestlik dereceli bir zincirdir.

1. S'nin tanım aralığı $[a, b]$ dir.
2. $S, S', S'', \dots, S^{(k-1)}$ süreklidir.
3. $\mathrm{a} = x_0 < x_1 < \cdots < x_n = b$ olacak şekilde x_i noktası (S'nin düğüm noktaları) vardır, öyle ki her bir $[x_{\mathrm{i}}, \mathrm{x}_{i+1}]$ aralığında S, k dereceli polinomdur.

(x_i, y_i) değer çiftinin $(n+1)$ gibi veri noktasını oluşturduğu varsayılırsa,

x	$x_0, \dots, x_n$
y	$y_0, \dots, y_n$

interpolasyon işlemi için aşağıda tanımlanan kısıtın eklenmesi gerekir.

4. $S(x_i) = y_i\ ,\ i \in [0, n]$

En çok kullanılan zincirin derecesi üçtür. Dolayısıyla sonuçta ortaya çıkan zincirler kubik zincir olarak isimlendirilir. Kübik zincirlerin çok sık kullanılma nedenleri:

- $S, S', ve S''$ süreklidir. Sırt noktaları göze düzleştirilmiş gözükür.
- Engelleyici bir şekilde düzleştirilmiş değillerdir.
- Çoğu durumda uygulanabilir ve "left-over" koşulları sağlar.
- Tek dereceli polinomlar daha iyi özelliklere sahiptir.
- Kübik zincirler tek dereceye sahip olduklarından en iyi şekilde elde edilebilenlerdir.

Her bir $[x_\mathrm{i}, \mathrm{x}_{i+1}]$ aralığı için $S_i(x)$ kübik polinomu oluşturarak bunların bileşimlerinden bir $S(x)$ zinciri elde edilebilir. Üç zincir koşullarını ve $n+1$ veri noktasının interpolasyonu sağlanacak şekilde n polinom bir araya getirildiğinde,

$$\mathrm{S}(x) = \begin{cases} S_0(x)\,, & x_0 \le x \le x_1 \\ S_1(x)\,, & x_1 \le x \le x_2 \\ \vdots & \\ S_{\mathrm{n}-1}(x)\,, & x_{n-1} \le x \le x_n \end{cases}$$

olur.

Bunlar n tane kübik polinomdur. Her bir polinom 4 katsayıya sahiptir. Toplam $4n$ tane parametre vardır. Bu parametrelerin $4n-2$ tanesi veri noktalarını interpolasyon yapması için kübik zinciri zorlamak, fonksiyona ve fonksiyonun ilk iki türevine göre

süreklilik şartlarını sağlamak için kullanılır (Press vd.,1992). Bu koşullar aşağıdaki gibidir.

Tablo 2.2. Üstel dağılım ailesi bazı yaygın tek değişkenli dağılımların özellikleri

Koşul	Aralık	Kısıt sayısı
$S_i(x_i) = y_i$	$i = 0,1,\dots,n-1$	n
$S_i(x_{i+1}) = y_i$	$i = 0,1,\dots,n-1$	n
$S_i^{'}(x_i) = S_{i+1}^{'}(x_i)$	$i = 0,1,\dots,n-2$	$n-1$
$S_i^{''}(x_i) = S_{i+1}^{''}(x_i)$	$i = 0,1,\dots,n-2$	$n-1$

Birkaç ortak seçenek vardır.

- Uç noktalardaki sabit eğimler $S'(x_0) = C_0$, $S'(x_n) = C_n$
- Doğal zincir $S''(x_0) = S''(x_n) = 0$
- Bir düğüm koşul olmaması S''', x_1 ve x_{n-1}'de süreklidir.

Aşağıdaki x ve y değerleri verilsin.

x	-1	0	1
y	1	2	-1

Bu noktalar için doğal $S(x)$ zinciri bulunmak istenirse, bunun için yapılması gerekenler:

- Birinci adım aşağıdaki gibi zinciri yazmaktır.

$$S(x) = \begin{cases} S_0(x) = a_0x^3 + b_0x^2 + c_0x + d_0 \quad , \quad x \in [-1,0] \\ \\ S_1(x) = a_1x^3 + b_1x^2 + c_1x + d_1 \quad , \quad x \in [0,1] \end{cases} \tag{2.14}$$

- İkinci adım ise zincir koşullarını kullanmak için $a_0, b_0, c_0, d_0, a_1, b_1, c_1, d_1$ katsayılarını bulmaktır. Veri noktalarının interpolasyonunu yapmak için

$S_0 = 2 \Rightarrow\ d_0 = 2$

$S_0(-1) = -1 \Rightarrow\ -a_0 + b_0 - c_0 = -1$

$S_1(0) = 2 \Rightarrow\ d_1 = 2$

$S_1(1) = -1 \Rightarrow\ a_1 + b_1 + c_1 = -3$

Birinci türevin sürekliliği nedeni ile

$$S'(x) = \begin{cases} S_0'(x) = 3a_0x^2 + 2b_0x + c_0 \quad , \quad x \in [-1,0] \\ \\ S_1'(x) = 3a_1x^2 + 2b_1x + c_1 \quad , \quad x \in [0,1] \end{cases}$$

$S_0'(0) = S_1'(0) \Rightarrow c_0 = c_1$

İkinci türevin sürekliliğinden,

$$S''(x) = \begin{cases} S_0''(x) = 6a_0x + 2b_0 \quad , \quad x \in [-1,0] \\ \\ S_1''(x) = 6a_1x + 2b_1 \quad , \quad x \in [0,1] \end{cases}$$

$S_0''(0) = S_1''(0) \Rightarrow b_0 = b_1$

olur.

Doğal zincir koşullarını sağlaması için,

$S_0''(-1) = S_1''(1) = 0$

$S_0''(-1) = 0 \Rightarrow 3a_0 = b_0$

$S_1''(1) = 0 \Rightarrow 3a_1 = -b_1$

olmalıdır ve bütün doğrusal eşitlikler çözüldüğünde $a_0 = -1, b_0 = -3, c_0 = 1,$ $d_0 = 2,$ ve $a_1 = 1, b_1 = -3, c_1 = -1, d_1 = 1$ bulunur. Bulunan değerler yerine konursa,

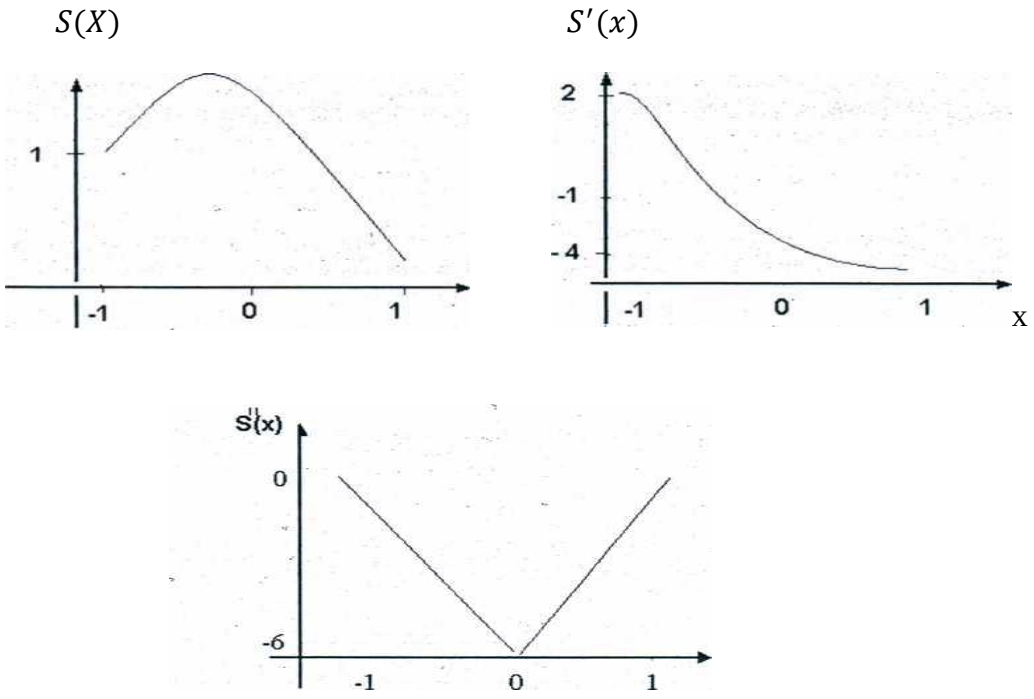
$$S(x) = \begin{cases} S_0(x) = -x^3 - 3x^2 - x + 2 \quad , \quad x \in [-1,0] \\ \\ S_1(x) = x^3 - 3x^2 - x + 2 \quad , \quad x \in [0,1] \end{cases} \tag{2.15}$$

$$S'(x) = \begin{cases} S_0'(x) = -3x^2 + 6x - 1 \quad , \quad x \in [-1,0] \\ \\ S_1'(x) = 3x^2 - 6x - 1 \quad , \quad x \in [0,1] \end{cases} \tag{2.16}$$

$$S''(x) = \begin{cases} S_0''(x) = -6x - 6 \quad , \quad x \in [-1,0] \\ \\ S_1''(x) = 6x - 6 \quad , \quad x \in [0,1] \end{cases} \tag{2.17}$$

olur.

Kübik zincirlerin, birinci ve ikinci mertebeden türevlerinin grafikleri aşağıdaki gibidir.



Şekil 2.2. Kübik zincirler

Otomatik olarak zincir oluşturmanın pek çok yolu vardır. Tekrar doğal zincir gözden geçilirse,

$$\mathrm{S}(x) = \begin{cases} S_0(x) \ , & x_0 \le x \le x_1 \\ S_1(x) \ , & x_1 \le x \le x_2 \\ \quad \vdots & \\ S_{\mathrm{n}-1}(x) \ , & x_{n-1} \le x \le x_n \end{cases} \tag{2.18}$$

- $S(x_i) = y_i$
- $S'(x)$ süreklidir.
- $S''(x)$ süreklidir.
- $S''(x_0) = S''(x_1) = 0$ dır.

$z_i = S''(x_i)$ olsun. Doğal zincirler kullanılmasından dolayı $z_0 = z_n = 0$'dır. Kabul edelim ki bütün $i \in [1, n-1]$ olsun. S'nin ikinci türevleri sürekli olan kübik polinomların bir kümesi olduğundan, $S''(x), (x_i, z_i)$ noktaları üzerinde bir doğrusal

interpolasyon zinciri oluşturur. Önceki örnekte $S''(-1,0),(0,-6),(1,0)$ noktaları üzerinde bir zincirdi. S'' 'nün doğrusal fonksiyonlarından birini basitçe aşağıdaki şekilde yazmak mümkündür. Dikkat edilirse S_i'' soldaki son noktada z_i , sağdakinde z_{i+1} ' dir. Bu da bir toplamsal modeldir ve

$$S_i'' = \frac{z_{i+1}}{x_{i+1}-x_i}(x-x_i) + \frac{z_i}{x_{i+1}-x_i}(x-x_i)$$

$$= \frac{z_{i+1}}{h_i}(x-x_i) + \frac{z_i}{h_i}(x_{i+1}-x) \qquad (2.19)$$

olarak yazılır. İki kez integrali alınarak,

$$S_i(x) = \frac{z_{i+1}}{6h_i}(x-x_i)^3 + \frac{z_i}{6h_i}(x_{i+1}-x)^3 + (x+d)$$

$$= \frac{z_{i+1}}{6h_i}(x-x_i)^3 + \frac{z_i}{6h_i}(x_{i+1}-x)^3 + c_i(x-x_i) + d_i(x_{i+1}-x) \qquad (2.20)$$

bulunur (Stone ve Koo, 1985).

İkinci türevler sürekli olmalıdır. Öyle ki;

- $S(x_i) = y_i \quad (i \in [0,n])$.
- $S'(x)$ süreklidir.

$S_i(x_i) = y_i$ koşulu uygulandığında

$$S_i(x_i) = y_i \Rightarrow 0 + \frac{z_{i+1}}{6h_i}h_i^3 + 0 + d_i h_i$$

$$\Rightarrow d_i = \frac{y_i}{h_i} - \frac{z_i}{6}h_i$$

$$S_i(x_{i+1}) = y_{i+1} \Rightarrow \frac{z_{i+1}}{6h_i} h_i^3 + 0 + c_i h_i + 0$$

$$\Rightarrow c_i = \frac{y_{i+1}}{h_i} - \frac{z_{i+1}}{6} h_i$$

elde edilir. ilk türevlerin sürekliliğinin uygulanması S_i ve S_{i-1} ile ilgili eşitlikleri verir.

$$S_i'(x) = \frac{z_{i+1}}{2h_i}(x - x_i)^2 + \frac{z_i}{2h_i}(x_{i+1} - x)^2 + \frac{y_i}{h_i} - \frac{z_i}{6} h_i - \frac{y_{i+1}}{h_i} + \frac{z_{i+1}}{6} h_i$$

$$S_i'(x_i) = -\frac{h_i}{6} z_{i+1} - \frac{h_i}{3} z_i + b_i \qquad \left(b_i = \frac{1}{h_i}(y_{i+1} - y_i) \right)$$

$$S_{i-1}'(x_i) = \frac{h_{i-1}}{6} z_{i+1} + \frac{h_{i-1}}{3} z_i + b_{i-1}$$

Süreklilik gereği,

$$S_{i-1}(x_i) = S_i(x_i) \Rightarrow h_{i-1} z_{i-1} + 2(h_{i-1} + h_i) z_i + h_i z_{i+1} = 6(b_i - b_{i-1}) \qquad (2.21)$$

olur. Bu ifadeler matris formatında yazılmak istenirse,

$$u_i = 2(h_{i-1} + h_i)$$

ve

$$v_i = 6(b_i + b_{i-1})$$

yazılarak,

$z_0 = 0$

$h_{i-1}z_{i-1} + u_i z_i + h_i z_{i+1} = v_i \qquad i \in [1, n-1]$

$z_n = 0$

olur. Bu ifadeler bir matrise dönüştürülürse,

$$\begin{bmatrix} 1 & 0 & . & . & . & . \\ h_0 & u_1 & h_1 & . & . & . \\ . & h_1 & u_2 & h_2 & . & . \\ & \ddots & \ddots & \ddots & . & . \\ : & & & h_{n-2} & u_{n-1} & h_{n-1} \\ . & . & . & . & 0 & 1 \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \\ z_{n-1} \\ z_n \end{bmatrix} = \begin{bmatrix} 0 \\ v_1 \\ v_2 \\ \vdots \\ v_{n-1} \\ 0 \end{bmatrix} \tag{2.22}$$

birinci ve sonuncu eşitliklerden $(n-1)$'nci dereceden bir üçgen matris elde etmek için bir satır ve sütun silinir (Wahba, 1987).

$$\begin{bmatrix} u_1 & h_1 & & & \\ h_1 & u_2 & h_2 & & \\ \ddots & \ddots & \ddots & & \\ & & h_{n-2} & u_{n-1} & h_{n-1} \\ & & & h_{n-2} & u_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n-1} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_{n-1} \end{bmatrix} \tag{2.23}$$

elde edilir. sistemi çözmenin diğer bir yolu Gauss eliminasyon yöntemini kullanmaktır.

1. İleriye doğru yerine koymak ($i = 2,3, \dots, n-1$ $için$)

$u_i \leftarrow u_i - \frac{h_{i-1}^2}{u_{i-1}}$

$v_i \leftarrow v_i - \frac{h_{i-1}v_{i-1}}{u_{i-1}}$

2. Geriye doğru yerine koymak:

İlk olarak z_{n-1} değeri atanır.

$z_{n-1} \leftarrow \frac{v_{n-1}}{u_{n-1}}$ olur.

Daha sonra kalan değerler atanır.

$i = n-2, n-3, \dots, 1$

$$z_i \leftarrow \frac{v_i - h_i z_{i+1}}{u_i}$$

Algoritma:

Veri girişi(input): $n+1)$ tane interpolasyon noktaları $(x_0, y_0), \dots, (x_n, y_n)$,

Ekran çıktısı(output): Kübik interpolasyon zincir $S(x)$

1. $i = 0,1 \dots, n-1$ için

$$h_i = x_{i+1} - x_i$$

$$b_i = \frac{y_{i+1} - y_i}{h_i}$$

hesaplanır.

2. $u_1 = 2(h_0 + h_1)$

$v_1 = 6(b_1 - b_0)$ olsun.

$i = 2,3, \dots, n-1$ için

$$u_i = 2(h_i + h_{i+1}) - \frac{h_{i-1}}{u_{i-1}}$$

$$v_i = 6(b_i + b_{i-1}) - \frac{h_{i-1} v_{i-1}}{u_{i-1}}$$ bulunur.

3. $z_0 = 0,\ z_n = 0$ olsun ve $i = n-1, n-2, \dots, 1$ için

$$z_i = \frac{v_i - h_i z_{i+1}}{u_i}$$

hesaplanır.

4. Bütün katsayılar S_i'de yerlerine konur.

$$S_i(x) = \frac{z_{i+1}}{6h_i}(x - x_i)^3 + \frac{z_i}{6h_i}(x_{i+1} - x)^3 + \left(\frac{y_{i+1}}{h_i} - \frac{z_{i+1}}{6}h_i\right)(x - x_i) + \left(\frac{y_i}{h_i} - \frac{z_i}{6}h_i\right)(x_{i+1} - x)$$

Bütün ***i***'ler için u_i'lerin asla sıfır olmadığı gösterilebilir.

3. ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON ZİNCİRLERİ (MARS)

İki ya da daha fazla değişken arasındaki ilişkilerin incelenmesinde regresyon modelleri sıkça kullanılmaktadır. Modelde etkilenen değişkene bağımlı, etkileyen değişkene ise; bağımsız değişken adı verilir. Regresyon modeli değişkenler arasında ki ilişkiyi ortaya koymak amacıyla kullanılan modeller için doğrusal, doğrusal olmayan ve karışık yapıda olmak üzere üçe ayrılmaktadır. Kurulan modellerin gerçek yapıyı daha iyi yansıtması için özellikle son yıllarda çok sayıda bağımsız değişken bir arada kullanılmaktadır. (Chatterje vd., 2000; Örekici vd., 2005)

Teknolojik gelişmeler ile birlikte çok sayıda değişkeni bir arada değerlendirebilen karmaşık algoritmaları kullanan regresyon yöntemleri kullanılmaktadır. Bunlardan biride Friedman (1991) tarafından tanıtılan çok değişkenli parametrik olmayan sınıflandırma/regresyon tekniği olan MARS'tır (Multivariate Adaptive Regression Splines). MARS regresyon için uyarlanabilir bir işlemdir ve yüksek boyutlu problemler içinde uygundur. MARS bağımlı ve bağımsız değişkenler arasındaki temel fonksiyonel bir ilişki hakkında herhangi bir önsel varsayım gerektirmez ve herhangi bir matematiksel ilişki aramaz. Bunun yerine sebep-sonuç değişkenleri arasında dinamik bir ilişki geliştirir. MARS tekniği, her bağımsız değişkenin bağımlı değişkenle olan ilişkilerini incelemenin yanı sıra, bağımsız değişkenlerin birbirleri arasındaki etkileşimleri belirler ve etkileşimlerin bağımlı değişken üzerindeki etkisini de ortaya koymaktadır. (Hastie vd., 2001; Tunay, 2001)

MARS'ın amacı bağımsız açıklayıcı değişkenler kümesinden sürekli bağımlı değişkenlerin değerlerini tahmin etmektir.

MARS, adımsal bir regresyon yöntemidir ve regresyon kümesinin performansını geliştirmek için tekrarlamalı ayırma metodunun ve adımsal doğrusal regresyonun genelleştirilmiş hali olarak görülebilir. (Kolyshkina ve Sylvia, 2004)

MARS tekniği bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan ilişkileri doğrusal yapıya dönüştürme amacıyla uygun dönüşümler bulmada ve bağımsız değişkenler arasındaki etkileşimleri belirleme de ideal bir yeniliktir. (Diechman vd., 2002)

MARS parçala-yönet yöntemi üzerine kurulmuştur. Bu yöntem ile veri alanı önce bölgelere ayrılır ve her biri için regresyon yöntemleri için regresyon eşitliği oluşturulur. Bu ise MARS'ı diğer regresyon yöntemleri için yüksek boyutluluk sorunu oluşturabilecek çok değişkenli regresyon problemleri için uygun bir çözüm yöntemi yapmaktadır.

MARS'a temel oluşturan zincir (spline), karmaşık eğri çizimlerinde ve fonksiyon tahminlerinde yeni bir matematiksel süreç olarak göz önüne alınabilir. Zincir (spline) düzleştirme yöntemi iki yada daha üst düzeyli polinomlar kullanıldığında elde edilen ve parametrik olmayan hata varyansının kontrol edilmesini sağlayan bir yöntemdir. (Kaki vd.,2004)

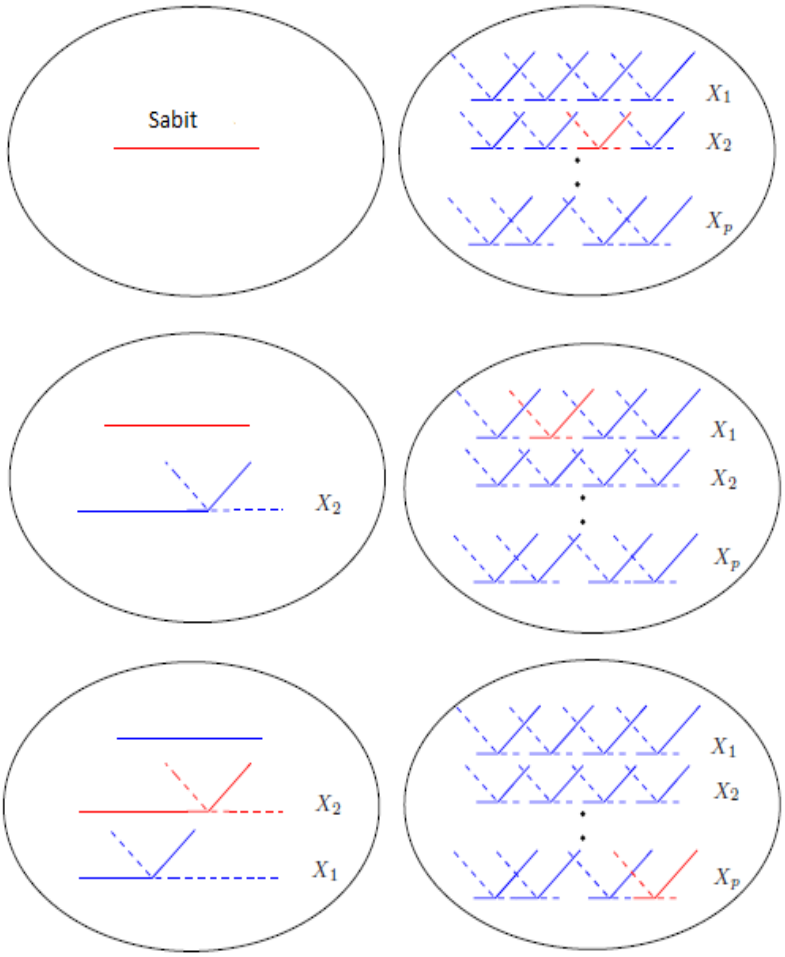
Genellikle zincirler olarak da adlandırılan parçalı polinomların birbirine bağlı düzgün parçaları vardır. MARS terminolojisinde polinomların katılma noktalarına düğümler ya da düğüm adı verilir. Yeterli sayıda düğüm sayısı ile bir model tahmin edilebilir. İki boyutlu zincir çizmek oldukça kolay iken çok boyutlu zincirlerin çizimi zorlaştığından herbir zincir matematiksel temel fonksiyonlarla ifade edilmektedir. (Hastie vd., 2008, MARS User Guide, 2007)

MARS tekniği birçok alanda kullanılmasına karşın; özellikle veri madenciliği alanında popüler hale gelmiştir. 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;

- Hangi bölgeye bakılacağı ve bu bölgenin sınırlarına doğru karar vermek ve
- Her bir değişken için kaç tane aralığın olacağına doğru karar vermeye ihtiyacı vardır. (Kolyshkina ve Sylvia, 2004)

MARS oldukça sadeleştirilmiş bir karar ağacı oluşturur. Parçalı temel fonksiyonlar ve bunların birleşimlerini kullanarak ve regresyon modellerinde ki hem ileri hem de geriye doğru ilerleme algoritmalarından yararlanarak MARS modeli elde edilir. Model kurulumu iki aşamada gerçekleşir.

1. Aşama: MARS sadece sabit terimle modele başlar ve sürekli olarak çiftler halinde temel fonksiyonları ekler. Temel fonksiyonlar sürekli eklenmesi karmaşık ve esnek bir model oluşturur. Temel fonksiyon sayısı en üst seviyeye ulaşıncaya kadar ekleme devam eder. Temel fonksiyonlar oluşturulurken aynı değişkene ait ileride tanımlanacak olan ayna temel fonksiyonu, bağımlı değişken ile bağımsız değişken arasındaki dağılımın düğüm noktasında eğimin değiştiğini ve düğüm noktasına kadar olan eğimin sıfır olduğunu gösterir.



Şekil 3.1. MARS ileri doğru adım prosedürünün şematik gösterimi. (Hastie vd., 2008)

Şekil 3.1'de soldaki model temel fonksiyondur. Başlangıçta, sabit fonksiyon $h(x) = 1$'dir. Sağdakinde modelin oluşması için tüm temel fonksiyonlar adaydır. Bunlar parçalı doğrusal temel fonksiyonların çiftleridir. Burada eşsiz tüm t düğümleri ile her belirleyici X_j'nin x_{ij} değerleridir. Her aşamada model içindeki temel fonksiyon ile bir aday çiftin tüm ürünlerini göz önüne alınmaktadır. Kalan hatanın en aza indirgediği ürün geçerli model içine ilave edilmektedir. Kırmızı renkle seçilmiş fonksiyonlar ile de prosedürün ilk üç adımı gösterilmektedir.

2. Aşama: Bu adıma budama denir ve bu adımı gerçekleştirmek için MARS geriye doğru adım algoritması kullanır. Birinci adımda modele en fazla sayıda temel fonksiyon eklendiğinden model overfit (aşırı tahminleme) sorunu ile karşı karşıya

kalmaktadır. En iyi alt model bulunana kadar her aşamada modele katkısı en az olan temel fonksiyonlar atılır. Birinci adımda çift olarak eklenirken ikinci aşamada genellikle çiftin bir tarafı yok sayılır ve böylece terimler çoğu zaman son modelde çift değildir. Temel fonksiyonlar eklenerek oluşturulan en son model budanır. Yani önemli bağımsız değişkenler ve bu değişkenlerin karşılıklı etkileşimleri belirlenerek, hata kareler toplamı en az olan en uygun model oluşturulur. Budama algoritması yaygın olarak Craven ve Wahba (1979) tarafından tanıtılan ve Friedman (1991) tarafından MARS için genişletilen Genelleştirilmiş Çapraz Doğrulama (GCV) ile yapılır.

GCV hem artıkların hatasını hem de model karmaşasını hesaba katar ve GCV ;

$$GCV(M) = \frac{1}{n}\frac{\sum_{m=1}^{n}(y_i - \hat{f}_m(x_i))^2}{(1-(C(M)/n)^2} \tag{3.1}$$

$$C = 1 + cd \tag{3.2}$$

eşitliklerinden hesaplanır.

Eşitlikteki,

n : Veri setindeki denek sayısını,

d : Etkili serbestlik derecesi olup bağımsız temel fonksiyonların sayısını,

C : Eklenen temel fonksiyonların maliyet-karmaşıklık (costcomplexity) ölçüsünü ve

M : MARS modelinin kurduğu regresyon modeli sayısını göstermektedir.

Hesaplamalar sonucu C değeri için 2<d<3 değerinin en iyi olduğu bulunmuştur. (Briand vd, 2000)

GCV' nin payı hata kareler toplamını, paydası ise modelin karmaşıklığını hesaplamaktadır. MARS algoritmasının ilk adımda oluşturulan en büyük modelin yorumlanması ve kullanımı kolay olmadığından dolayı ikinci adımda en büyük model budanarak, yani önemli bağımsız değişkenler ve bu değişkenlerin etkileşimleri belirlenerek, GCV ölçüsü en küçük olan model elde edilir. (Yerlikaya vd., 2007)

3.1. Mars Modeli

Temel fonksiyonlar ile model parametreleri (en küçük kareler yöntemi ile tahmin edilen); veri girişlerini veren belirleyicilerin sonuçlarından oluşur. Genel MARS modeli 3.3 nolu eşitlikteki gibidir.

$$Y = \beta_0 + \sum_{k=1}^{K} a_k \beta_k(X_t) + \varepsilon_i \qquad (3.3)$$

Burada;

k : Düğüm sayısını,

K : Temel fonksiyon sayısını,

X : Bağımsız değişkeni,

a_k: k. Temel fonksiyonun katsayısı,

β_0 : Modeldeki sabit terim ve,

$\beta_k(X_t)$: t. Bağımsız değişken için k. temel fonksiyondur. (Statsoft, 2013)

Bu fonksiyon kesim parametresi (β_0) ve bir veya daha fazla temel fonksiyonun ağırlıklı toplamından oluşur. Aynı zamanda modeli; her bir belirleyicinin tüm değerlerini karşılayan temel fonksiyon kümelerinden, bu temel fonksiyonların ağırlıklı toplamı olarak da düşünülmektedir. Bu durumda MARS algoritması değişkenler arasındaki etkileşimlerin yanı sıra belirleyici değişkenler ve tüm girişler üzerinde araştırma yapar. Bu araştırma sırasında, temel fonksiyonların gittikçe daha da artan sayısı genel olarak en küçük kareler uyum iyiliği kriterini en üst düzeye

çıkarmak için muhtemel olan temel fonksiyonlar kümesinden modeli ekler. Bu çalışmaların bir sonucu olarak, MARS eklenenlerin arasında ki en önemli etkileşimlerin aynı sıra ile bağımsız değişkenleri otomatik olarak belirler. MARS için;

Kategorik Belirleyiciler: Uygulamada hem sürekli hem de kategorik belirleyiciler kullanılabilir ve çoğu zaman yararlı sonuçlar verecektir. Ancak temel MARS algoritması değişkenlerin doğada sürekli olduğunu varsayar. İki değer alan değişkenlere göre sınıflandırılmış verilerde daha başarılı sonuçlar elde eder.

Çoklu Değişkenler: MARS algoritması çoklu bağımlı değişkenler için uygulanabilir. Bu durumda, algoritma belirleyicilerdeki temel fonksiyonların bir ortak kümesini belirler. Fakat her bağımlı değişken için farklı bir sabit oluşturur ve bu sinir ağları yapısında da böyledir.

Sınıflandırma Problemleri: MARS çoklu bağımlı değişkenleri kullanabildiğinden dolayı, bunun yanı sıra sınıflandırma problemlerinde algoritmayı kolayca uygular. İlk olarak çoklu gösterge değişkenlerine kategorik yanıt değişkenlerinin sınıfları kodlanır. Sonra bir model uydurmak için algoritma uygulanır ve edinilen değerler veya puanlar hesaplanır. Son olarak tahmin edilen en yüksek puan sınıfına her durumda atanır.

Model seçimi ve budama: doğrusal olmayan modeller, esnek bir yapıya sahiptirler. Bu nedenle önlem alınmadığı durumlarda gereğinden fazla yaklaşma çabası içine girebilirler. Denenmiş verilerde neredeyse sıfır hata yapmalarına rağmen yeni gözlemler ve durumlarla karşılaştıklarında zayıf bir performans gösterebilirler. Diğer pek çok metotta olduğu gibi MARS da veriye aşırı uyum (karmaşıklığı arttırma, over-fit) yapma eğilimindedir. Bunu engellemek için budama yöntemi kullanılır ve modelin karmaşıklığı azaltılır.

3.2. Temel Fonksiyonlar

Genellikle regresyon denklemleri, veriler arasındaki ilişkiyi tek bir fonksiyon kullanarak belirlemeye çalışırlar. Fakat MARS parçalı polinomik bir fonksiyon kullanır. Böylece bütün değerlere en yakın noktalardan geçebilecek (bu sayede artıkları da en aza indirebilecek) regresyon kesitleri oluşturulabilir. Regresyon kesit fonksiyonları parçalı polinomik temel fonksiyonların düğümlerde birleştirilmesi ile elde edilmiş sürekli bir fonksiyondur. Temel fonksiyonlardaki sabitler en küçük kareler yöntemi ile bulunur. Temel fonksiyonlar aşağıdaki gibi tanımlanmıştır.

$$B_k(x) = \prod_{j=1}^{J_k} \left[s_{kj}(x_{wkj} - t_{kj}) \right] \qquad k = 1,2,\dots,K \tag{3.4}$$

Buradaki;

J_k:İnteraksiyon derecesini göstermektedir.

$[.]_+ = \max [0, .]_+$

$s_{kj}: \in [\pm 1]$

t_{kj}: düğüm değeri ve

x_{wkj}: bağımsız değişken değerini göstermektedir. (Statsoft, 2013)

Temel fonksiyonlar bağımsız değişkenlerin doğrusal olmayan dönüşümleri olabilirler. Fakat bağımlı değişken temel fonksiyonların doğrusal bir dönüşümüdür. MARS regresyon modeli bağımsız değişkenlerin farklı aralıklarını fit etme (uydurma) temel fonksiyonları tarafından inşa edilmiştir. Genellikle zincirler olarak da adlandırılan parçalı polinomların birbirine bağlı düzgün parçaları vardır. MARS terminolojisinde, polinomların katılma noktalarına arıza noktaları, düğümler veya düğüm olarak adlandırılır. Bunu küçük t harfi ile göstereceğiz.. MARS $(x - t)_+$ ve $(t - x)_+$ formunu parçalı doğrusal temel fonksiyonlarda ki genişlemeleri kullanır. Parametre temel fonksiyonların düğümüdür. Bu düğümler aynı zamanda $(t - x)$

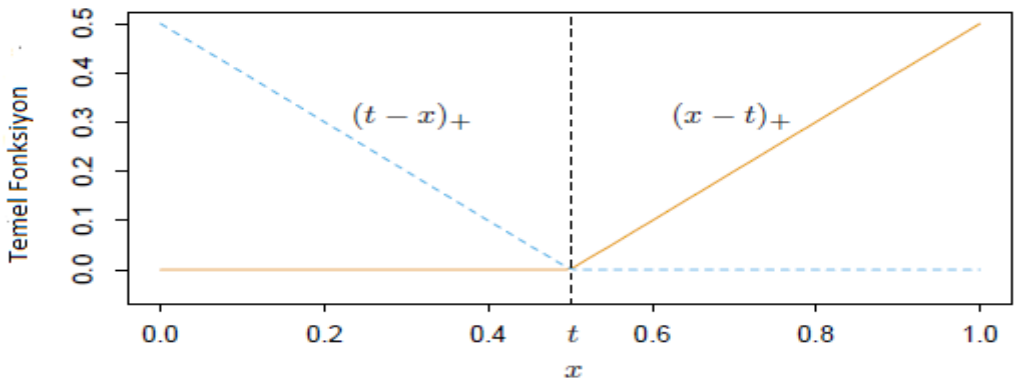
ve$(x-t)$ terimlerinin yanındaki (+) işareti eşitliğin sonucunun yalnızca pozitif olduğunu göstermektedir. Aksi halde her bir fonksiyon sıfır noktasında değerlendirilir.

Böylece;

$$(x-t)_+ = \begin{cases} (x-t), & eğer\ x < t, \\ 0 \quad , & diğer, \end{cases} \tag{3.5}$$

$$(t-x)_+ = \begin{cases} (t-x), & eğer\ x \geq t, \\ 0 \quad , & diğer. \end{cases} \tag{3.6}$$

eşitlikleri kullanılmaktadır.(Hastie vd., 2008)



Şekil 3.2.Temel fonksiyonlar $(\boldsymbol{x-t})_+$ ve $(\boldsymbol{t-x})_+$ MARS tarafından kullanılır.

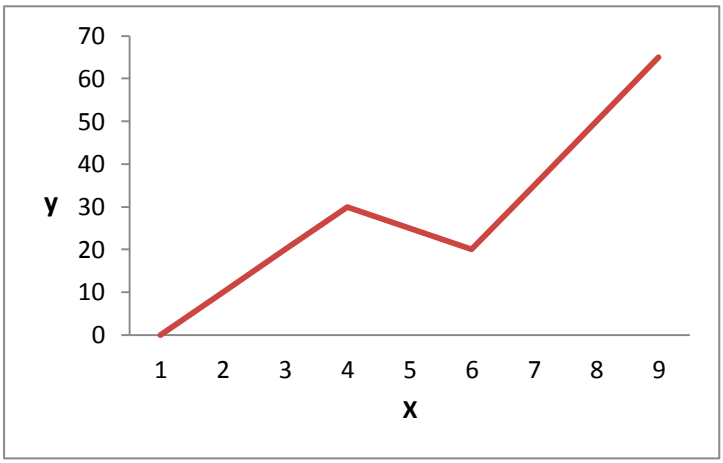
Örnek olarak, $(x-0{,}5)_+$ ve $(0{,}5-x)_+$ Şekil 3.2' de gösterilmiştir.

Her bir fonksiyon, değeri t de bir düğüm ile parçalı doğrusaldır ve bunlar doğrusal zincirlerdir.

3.3. Düğüm Noktası (Knot)

MARS regresyon modeli bağımsız değişkenlerin farklı aralıklarını fit etme (uydurma) temel fonksiyonları tarafından inşa edilmiştir. Genellikle zincirler olarak da adlandırılan parçalı polinomların birbirine bağlı düzgün parçaları vardır. MARS terminolojisinde, polinomların katılma noktalarına düğümler veya düğüm olarak adlandırılır.

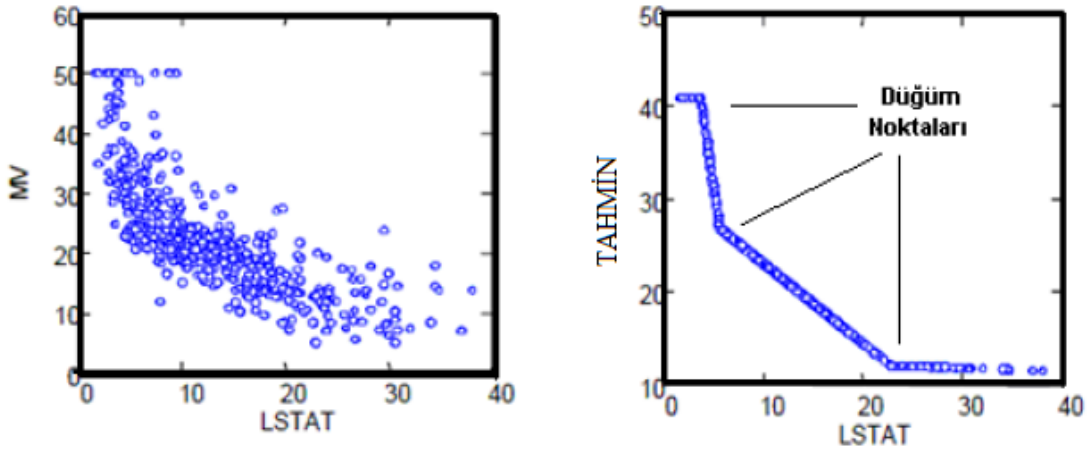
Chen vd., 2005'e göre; MARS parçalı doğrusal regresyon kullanarak esnek modeller meydana getirir ve doğrusal olmayan durumları ortadan kaldırmak için bağımsız değişkenin farklı aralıklarında ayrı regresyon eğilimleri kullanır. Regresyon eğiminin değiştiği ve bir aralıktan diğerine geçildiği noktalara düğüm denir. Şekil 3.3.'te iki düğüm noktası bulunan parçalı-doğrusal regresyon gösterilmiştir.



Şekil 3.3.İki düğüm noktalı parçalı-doğrusal regresyon örneği

Zincir altında yatan ana kavram düğümdür. Düğümler bir veri bölgesinin bitişi iken diğerinin başlangıcıdır. Yani düğüm noktası fonksiyon değişim davranışının olduğu yerdedir. Noktalar arasında model genel olabilir(doğrusal regresyon). Klasik bir zincirde düğümler, önceden belirlenmiş ve eşit aralıklıdır. Halbuki MARS'ta düğümler bir arama prosedürü tarafından belirlenir. MARS modeli ihtiyaç duyulandan fazla düğüm noktası bulur. Eğer bir düz çizgi uyumlu ise burada yeni bir

düğüm aranmaz. Ancak MARS'ta belirleyicinin en küçük gözlenen değerine karşılık en az bir sözde düğüm bulunur. Şekil 3.4'de gerçek verilerden oluşan üç düğüm noktası MARS zincirini göstermektedir.

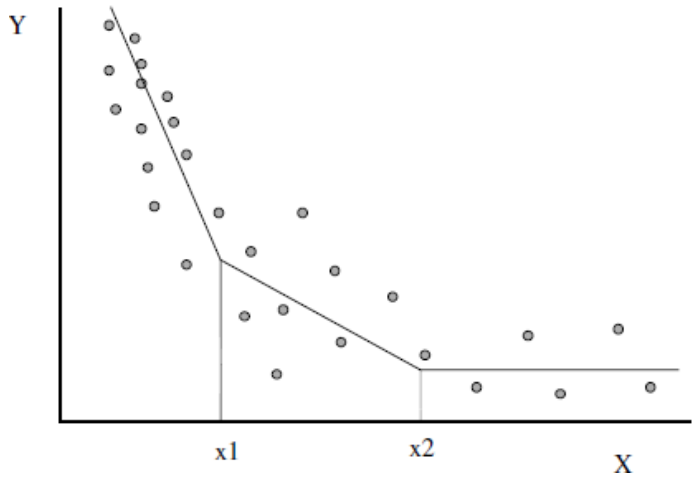


Şekil 4.4.Kesit ve düğüm kullanılarak MARS veri tahmini (solda gerçek veri) (MARS kullanım kılavuzu)

Basit bir regresyonun en iyi noktasını bulmak kolay bir problemdir. Çok sayıda potansiyel düğüm incelenerek en iyi R^2 ile biri seçilir. Ancak en iyi çifti bulmak için çok daha fazla hesaplama gerektirir ve ihtiyaç olan gerçek sayı bilinmediğinden düğümlerin en iyi kümesini bulmak daha da zor olmaktadır. MARS ileriye doğru adım algoritması kullanarak ihtiyaç duyulan noktaların yerini ve sayısını bulur. İlk olarak çok fazla nokta içeren overfit (karmaşıklığı artırma) oluşturulur ve sonrasında geri adım ile uyuma en az katkıda bulunan noktalar kaldırılır. İleri adım ile oluşacak yanlış düğüm yerleri geri adım ile modelden silinir.(Abraham vd, 2001)

3.3.1. Düğüm Değerinin Elde Edilmesi

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 19 monoton dönüşümler (logaritmik, üstel, vb.) yapmadan aralarındaki ilişkiyi doğrusal hale getirirler. Şekil 3.5'de iki farklı düğüm değerine sahip bir değişkenin bağımlı değişkenle olan ilişkisi sunulmuştur. X_1 ve X_2 düğüm değerleridir. Başlangıç noktasından X_1'e kadar doğrunun eğimi aynıdır. Benzer şekilde X_1 ve X_2 arasında da doğrunun eğimi aynıdır. Fakat X_1'den kadar uzaklaşsak bile doğrunun eğimi değişir. İşte doğrunun eğimini değiştirmeyen değer bir düğüm değeridir.



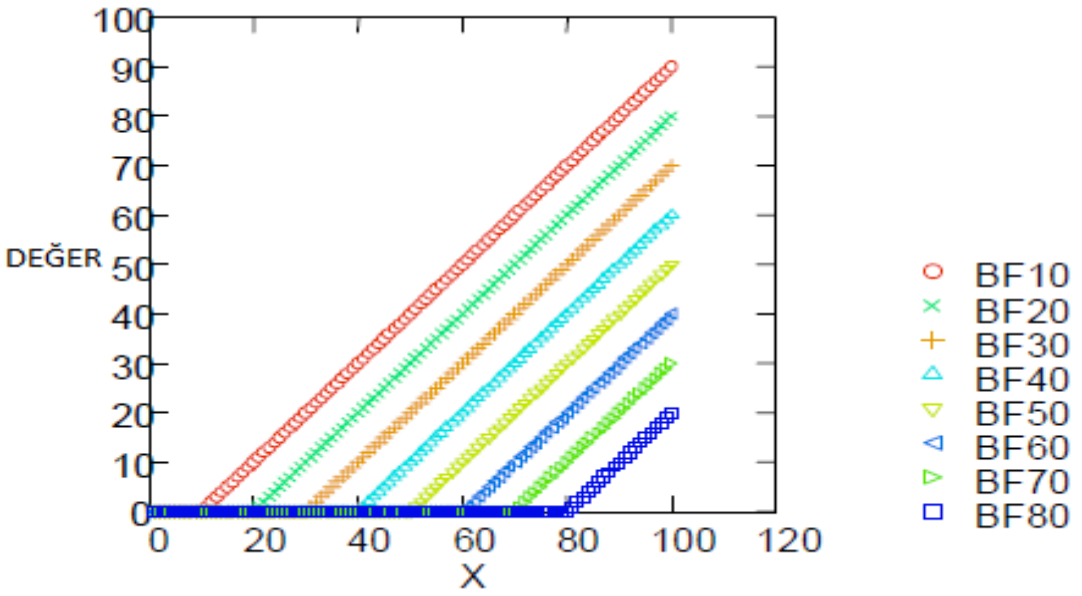
Şekil 3.5.MARS'ta örnek noktalar. (Briand vd., 2004)

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. Temel Fonksiyonlar (BF) düğümler için genelleştirilmiş aramalarda kullanılan mekanizmalardır. BF bir veya daha fazla değişken içinde yer alan bilgileri temsil etmek için kullanılan fonksiyonların bir kümesidir. Ana bileşenler gibi BF temelde hedef değişken ile belirleyici değişkenlerin ilişkisini yeniden ifade eder. BF hokey sopası, MARS modelinin temel yapı bloğu tek değişkenlilerde birden çok kez uygulanır. Hokey sopası fonksiyonu değişken X'ten , X*'a dönüşür.

Maks (0,X-c) yada Maks (0, c-X)

Burada X* , eşik değeri c'nin altında X değerlerinin tümü için 0 değerini alır ve c'den büyük bütün X değerleri için ise X' e eşittir. Yani c değerini aşan X değerini alır. İkinci form birincisinin ayna görüntüsünü oluşturur. Şekil 3.6'da X belirleyici değişkeninin 0-100 aralığı için c değişkeninin 10'ar birimlik değişimlerine karşılık temel fonksiyonlarda meydana gelen değişimleri göstermektedir.



Şekil 3.6.X değişkeni 0-100 aralığı için c=10'dan 80'e kadar temel fonksiyonların değişimi (hokey sopası) (MARS kullanım kılavuzu)

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

MARS ikili parçalarından oluşan temel fonksiyonlar meydana getirir: Standart temel fonksiyonlara ayna görüntülerini sürekli ekler. Eklenen ikililer olabilecek çoğutemel fonksiyonlar gibi ayrı veri değerleri oluşturur. Ayna görüntüsü sağ ve sol ana kırıklara ayrılmış düğümlerdir. Bu fonksiyonlar doğrusal olarak bağımsız değildirler, fakat modelin esnekliğini artırılabilirler. Temel fonksiyonların eklenen belirli ayna görüntülerini değiştirmek final modeli etkilememektedir. Basit regresyonda en iyi düğüm noktasını bulmak kolay bir sorundur. En iyi R^2 ile çok sayıda potansiyel düğüm noktalarını bulur ve bunlardan birini seçer. Halbuki, değişken sayısı arttığında; düğümlerin en iyi çiftini bulmak çok uzun hesaplamalar gerektirir. Gerçek sayıya ihtiyaç duyulduğunda düğümlerin en iyi setini bulmak bilinmeyen bir iştir. MARS ihtiyaç duyulan düğümlerin yerlerini ve sayılarını ileriye / geriye doğru adımsal yöntemle bulur.

3.5. MARS İle İlgili Yapılan Çalışmalar

Çok değişkenli uyarlanabilir regresyon zincirleri (Multivariate Adaptive Regression Splines – MARS), ilk olarak Jerome Friedman (1991) tarafından tanıtılmıştır. MARS yöntemi günümüze kadar ekonomi, bankacılık, yaşam çözümlemesi, sosyal bilimler ve fen bilimleri gibi birçok alanda kullanılmaktadır. Bununla birlikte veri madenciliği alanında çok fazla uygulama yapılmıştır. doğrusal ve lojistik regresyonun yöntemlerinin yetersiz kaldığı ya da varsayımları sağlanamadığı için uygulanamadığı durumlarda MARS alternatifi olmayan bir yöntem olarak öne çıkmıştır.

MARS yöntemi kullanılarak Türkiye'de ve yurtdışında yapılmış birçok çalışma bulunmaktadır. Özellikle son yıllarda Türkiye'de kullanılan yöntem ile ilgili yapılan bazı çalışmalar;

Tunay'ın (2001), Türkiye'de paranın gelir dolaşım hızlarının MARS yöntemiyle tahminini yapması bu alanda ilk çalışmalardandır.

Temel vd (2005) yaptıkları çalışmada, MARS tekniğinin temel özellikleri ve uygulama adımları üzerinde durmuşlardır. Ayrıca behçetli hastaların ailelerinden tespit edilmiş bir takım davranış ve psikolojik test sonuçlarının bu kişilerdeki beck depresyon durumunu tahmin etmede kullanılabilecek bir MARS modeli elde etmişlerdir. Bu modelin özellikleri açıklanmış ve tahmindeki başarısı oldukça yüksek bulunmuştur. Hesaplamalarda MARS 2.0 paket programı kullanılmıştır.

Yerlikaya (2008) MARS üzerinde bir takım düzenlemeler yaparak oluşturduğu yeni modeli veri madenciliği uygulamaları için kullanmıştır.

Kan ve Yazıcı (2010) yaptıkları çalışmada; 4 farklı istatiksel yöntemlerin (kesirli faktöriyel deneyleri, faktöriyel deneyleri, regresyon ağaçlarını ve MARS tekniği) F-4 uçaklarının yakıt tüketimleri üzerindeki önemli etkilerini ve farklı metodlardan bulunan sonuçlarını karşılaştırmışlardır. Uygulanan 4 farklı metoda göre; ana faktörlerden D (akım girişi) ve E (hava basıncı)'nin istatistiki olarak önemli olduğunu tespit etmişlerdir.

Kayri (2010) çalışmasında internet bağımlılığının gerçek etkilerinin belirlenmesinde, tarafsız ve sağlam istatistiksel yöntemlerin kullanımı önem arz ettiğinden MARS tekniğini kullanarak analiz etmiştir. MARS tekniğinin performansını incelemek amacıyla da MARS ve CART'tan elde edilen veriler karşılaştırmıştır. Bu çalışma ile bağımlılık düzeyi tahmininde MARS'ın, CART'dan farklı veriler elde edildiğini bildirilmiştir.

Tunay (2010) çalışmasında Türkiye'de olası bankacılık krizlerini öngörmekte kullanılabilecek bir erken uyarı modeli geliştirmeyi hedeflemiştir. Model MARS ile tahmin edilmiştir. Tahmin sonuçlarını istatistik anlamlılık ve açıklama gücü

açılarından son derece başarılı bulmuştur ve bulgular Türkiye'de banka krizlerinin büyük oranda dış kaynaklı değişkenlerden ileri geldiğini göstermiştir.

Tunay (2011) de yaptığı çalışmasında; Türkiye'de durgunlukların ve kestirimlerinin yapılmasını amaçlamıştır. Durgunluk olaylarına dair gözlemleri kullanarak MARS yöntemiyle örneklem içi kestirimler yapmıştır. MARS önemli üstünlükler sunmuş ve modelin kestirim performansı da bir hayli yüksek çıkmıştır.

Topak (2011) Türkiye'de kurumsal başarısızlığı modellemek için mars tekniğini kullanmıştır. Türkiye'deki finansal başarısızlık örneklerine bir model oluşturmak amacıyla parametrik olmayan mars tekniğini kullanmıştır. Finansal başarısızlık durumunu bir yıl önceden tahmin edebilmek amacıyla 1994 ve 2003 yılları arasında 114 firmaya ilişkin 665 adet yıllık gözlem yapmıştır. 39 u finansal, 2 si finansal olmayan 41 bağımsız değişkenle yapılan çalışma sonucunda, 8 orijinal değişkene dayalı 10 temel fonksiyonlu bir modele ulaşılmıştır. Ortaya konan bu model ile, %81,8 oranında doğru sınıflandırma başarısına sahip olduğunu ve yalın bir modelden çok daha üstün olduğu görülmüştür.

Akyol (2011) Yaşam çözümlemesinde kullanılan tekniklere alternatif olarak kullanılabilecek yeni bir yöntemin geliştirilmesi ve yaşam çözümlemesinde kullanılabilirliğinin ispatlanması amacı ile çalışmalar yapmıştır. MARS tekniğinin kullanılması ile yaşam çözümlemesinde bilinen yöntemlerin eksiklerinin giderilebileceği ya da farklı bir bakış açısı ile yeniden değerlendirilebileceği göstermiştir.

Yurtdışında yapılan bazı çalışmalar;

Kim (2000) gençlerin uyuşturucu kullanımı ile ilgili yaptığı çalışma sonucunda, bağımlı değişkenin kategorik olduğu durumlarda da MARS'ın iyi neticeler verdiğini göstermiştir.

Kuhnert vd., (2000) parametrik olmayan modelleri (CART ve MARS) lojistik regresyonla kıyaslamıştır. Motor kazalarındaki yaralanma verilerine uygulanmış olan bu çalışma için MARS modelinin diğer ikisine göre daha iyi performans gösterdiği belirtilmiştir.

Abraham ve Steinberg (2001); yeni ve esnek regresyon modeli olan MARS'ı, üç farklı derinlikte toprak sıcaklığı benzetimi çalışmalarında kullanmışlardır. Diğer açıklayıcı hesaplama yöntemleriyle karşılaştırıldığında, MARS hızlı, esnek ve modelin önemli girdilerini belirlemede oldukça 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.

Nash ve Bradford (2001)'un yaptığı çalışmada belirli bir bölgedeki bir kurbağa türünün varlığı lojistik regresyon ve MARS yöntemiyle tahmin edilmiş ve iki yöntemin sonuçları değerlendirilmiştir..

Kolyshkina ve Brookes (2002) sigorta riskini veri madenciliği yaklaşımları (regresyon ağaçları ve MARS) ve klasik lojistik regresyonla tahmin etmeye çalışmıştır.

Chen ve arkadaşları (2003), MARS yöntemini kullanarak havayolu karlılığını arttırmaya yönelik olarak çalışmışlardır. Chen ve arkadaşları ise 20 şehir ve 31 uçuş rotasında çalışan yerel hava yolu taşımacılığı yapan şirketin kar yönetimine bir çözüm önerisi oluşturabilmek için MARS'ı stokastik dinamik programlama metodunun bir türevi olarak kullanılmışlardır.

Dieterle (2003) zamana bağlı analitik veriler üzerine hazırladığı doktora tezinde yapay sinir ağları, genetik algoritmalar, CART ve MARS'ı karşılaştırmıştır.

Lee vd. (2004) kredi skorlama ile ilgili çalışmalarında, ayırma analizi, lojistik regresyon, CART ve MARS'm doğru sınıflama oranlarını ve hatalarını karşılaştırmıştır.

Xiong ve Meullent (2004), müşterilerin peynir çubukları (Cheese sticks) tercihlerini etkileyen sebepleri araştırmışlar ve tercihleri belirleyen esas unsurları MARS ile bulmaya çalışmışlardır. MARS, tahmin ediciler arasındaki ilişkiyi ortaya koyabilmiş, parçalı regresyon fonksiyonları ile de tahmin edici değişkenlerin bağımlı değişken (peynir çubuklarını tercih) üzerine etkilerini saptayabilmiştir. Xiong ve Meullent çalışmaları ile MARS'ın karmaşık veriden, elde edilebilecek mümkün olan en iyi modeli oluşturulabildiğini göstermişlerdir.

Stokes ve Lattyak (2005) MARS yöntemini ekonometrik bazı sistem ve yazılımlar ile geliştirmiş ve kullanmıştır.

Verzilli ve arkadaşları (2005) İngiltere'de yaptıkları çalışmada, DNA kodlamasında genetik çok şekillilik olan bölgelerde taşıyıcılarda fenotip etkiler görülebileceğini (örneğin hastalığa bağlı hassasiyet etkisi gibi) belirtmişler ve analiz için bağ tutucu ve lizozom proteinlerindeki başlangıç mutasyonu verileri kullanılarak bir model oluşturmuşlardır. Oluşturdukları modelin performansını sınamak için hiyerarşik Bayesci MARS Yöntemini kullanmışlardır.

Leathwick ve arkadaşları (2006), Yeni Zelanda akarsularında yaşayan balıkların dağılıma ilişkin bir veritabanını kullanarak 15 tatlı su balığı türünün bulunma olasılıkları ile bu canlıların akarsulardaki yaşam alanlarını belirleyen çevresel değişkenler arasındaki ilişkileri MARS ile incelemişlerdir.

Kriner (2007) çalışmasında yaşam analizini MARS yöntemini kullanarak yapmıştır.

Quiros vd (2009) çok değişkenli uyumlu regresyon eğrilerini arazi örtüsünün uydu görüntülerinden yararlanarak sınıflandırılması için kullanmıştır.

Mina ve Barrios (2009), yoksulluk profilinin çıkartılmasında MARS yöntemini kullanmışlar ve bazı koşullar altında lojistik regresyondan daha etkili olduğunu belirtmiştir.

Mina (2010) özürlü kişilerin iş seçimi ile ilgili yaptığı çalışmada lojistik regresyon ve MARS yöntemlerini kullanmış ve sonuçları değerlendirmiştir.

Samui ve Kothari (2011) depolardaki buharlaşma kayıplarının tahminini MARS ile yapmış ve sonuçları yapay sinir ağları ile kıyaslamıştır.

3.6. MARS Tekniğinin Avantajları ve Dezavantajları

3.6.1. MARS Tekniğinin Avantajları

- MARS modelinin anlaşılması ve yorumlaması oldukça kolaydır.
- MARS tekniğinde, bağımlı ve bağımsız değişkenlerin dağılımı kategorik veya sürekli olabilir. (Friedman, 1991)
- MARS doğrusal modellere göre daha esnektir ve az veri gerektirir.
- Kayıp ve aşırı uç değerlerden çok az etkilenir. (Salford, 2013)
- MARS modeli oluşturulurken çok az hiç veri hazırlığı gerektirir yada hiç veri hazırlığı gerektirmez.
- MARS büyük veri setleri için uygundur.
- MARS bağımsız değişkenler arasındaki etkileşimleri tanımlar ve bu etkileşimlerin anlaşılmasını sağlayacak grafikler sunar. (Staicu, 2005)

3.6.2. MARS Tekniğinin Dezavantajları Ve Sınırlılıkları

- MARS modelinin uygulanmasında genellikle büyük veri setlerine ihtiyaç duyulur.

- Model terimlerinin oluşumunda kısıtlama mevcuttur. Yani her bir veri sadece bir kez ürün olarak bulunabilir (Staicu, 2005).

- Yeni uygulanan bir teknik olduğundan modelin yorumlanırken dikkat edilmesi gerekir.

4.UYGULAMA

4.1. Araştırmanın Amacı

Bu çalışmada Erzincan Üniversitesi Kemah Meslek Yüksek Okulu öğrencilerinin not ortalamalarını etkileyen unsurların belirlenmesi amaçlanmıştır. Hazırlanan anket öğrencilere uygulanarak elde edilen sonuçlar ile veri seti oluşturulmuştur.

4.2. Araştırma Örneklemi ve Anakütlesi

Çalışma veri seti Kemah Meslek Yüksek Okulu öğrencilerine anket uygulanarak elde edilmiştir. Anakütleyi oluşturan 643 öğrenci içerisinden Anakütleyi temsil eden 142 öğrenci seçilmiştir. Örneklem büyüklüğü güven düzeyi %99 doğruluk değeri sapma payı 10 olarak belirlenmiştir. Veri seti oluşturulurken öğrencilerin yaşı, cinsiyeti, öğrenim türü, öğrenim gördükleri program, mezun oldukları lise türü, nerede ikamet ettikleri, okuyan kardeş sayıları, devletten öğrenim kredisi yada burs alma durumları, çalışma durumları, gelir düzeyleri, ailesinin ikamet ettiği yer, YGS tercih sırası, bağımsız değişkenler olarak belirlenmiştir. Öğrencilerin not ortalaması bağımlı değişken olarak belirlenmiştir.

Tablo 4.1. Kategorik değişkenler ve cevap sayıları dağılımları

Değişkenler	Şıklar	Sayı
Cinsiyet	Bayan	87
	Bay	55
Tür	Örgün	80
	İkinci	50
	Uzaktan	12
Program	Banka	43
	Çağrı	63
	Bilgisayar	13
	Muhasebe	9
	İşletme	5
	Emlak	8
Lise	Düz	86
	Meslek	40
	Anadolu	10
	Fen	2
	Diğer	4
Burs	Var /Yok	86/56
İş	Var	25
	Yok	117
Yerleşim	Büyükşehir	28
	İl	32
	İlçe	38
	Belde	6
	Köy	38
Barınma	Devlet yurdu	105
	Özel yurt	3
	Öğrenci evi	24
	Diğer	10

4.3. MARS Modelinin Kurulması

Elde edilen veri seti Salford şirketinin SPM7 programına yüklenmiştir. SPM7 programı içerisinde bulunan MARS paket programı kullanılmıştır. MARS a ek olarak program içerisinde CART, TreeNet ve Random Forest programları da bulunmaktadır.

MARS modelinin kurulumunda kullanılan veri seti için değişkenler ve özellikleri Tablo 4.2. de verilmiştir.

Tablo 4.2. MARS modelinin kurulumunda kullanılan değişkenler ve özellikleri

Değişkenler	Kısaltma	Türü	Değerler ve aralık
Not ortalaması	Not Ort	Sürekli	1,07-4
Burs/Öğrenim durumu	Burs	Kategorik	0= Var 1= Yok
Cinsiyeti	Cinsiyet	Kategorik	0= Bayan 1= Erkek
Gelir Düzeyi	Gelir	Sürekli	100-1600 (TL)
Barındığı yer	İkamet	Kategorik	0= Devlet Yurdu 1= Özel Yurt 2= Öğrenci Evi 3= Diğer
Çalışma Durumu	İş	Kategorik	0= Var 1= Yok
Okuyan kardeş sayısı	Kardeş	Sürekli	0-13
Mezun lise türü	Lise	Kategorik	0= Düz Lise 1= Meslek Lisesi 2= Anadolu Lisesi 3= Fen Lisesi 4= Diğer
Devam ettiği program	Program	Kategorik	0= Banka 1= Bilgisayar 2= Çağrı 3= Emlak 4= İşletme 5= Muhasebe
YGS tercih sırası	Tercih	Sürekli	1-30
Öğrenim türü	Tür	Kategorik	0=Örgün 1= İkinci 2= Uzaktan
Yaşı	Yaş	Sürekli	18-29
Aile yerleşim yeri	Yerleşim	Kategorik	0= Büyükşehir 1= İl 2= İlçe 3=Belde 4= Köy

Temel fonksiyon sayısı, etkileşim sayısı, düğümler arası gözlem sayısı ve düğüm optimizasyonu için serbestlik derecesi değerleri modelin kurulumunu belirlemektedir. MARS programının özelliği gereği modelin kurulum aşamasında bu değerlerin tercihi kullanıcıya bırakılmıştır.

Verilerle birçok sayıda model oluşturulabilir. Bu modellerden en iyi olanı en az uyum iyiliğine sahip olandır. GCV değeri en küçük olan uyum eksikliği en az olan modeldir. Yapılan birçok denemeden sonra;

Temel fonksiyon sayısı: 45

Maksimum etkileşim sayısı: 5

Düğümler arası en az gözlem sayısı: 0

Düğüm optimizasyonu için serbestlik derecesi: 3

değerleri için en küçük GCV=0,30729 değeri elde edilmiştir. Modeli oluşturacak değişkenlere ait bazı istatistikler tablo 4.3. de verilmiştir.

Tablo 4.3. Değişkenlere ait bazı değerler

Değişken	N	Ortalama	Minimum	Maksimum	Standart sapma
Burs	142	---	0	1	---
Cinsiyet	142	---	0	1	---
Gelir	142	403,65	100	1600	244,53
İkamet	142	---	0	1	---
İş	142	---	0	1	---
Kardeş	142	1,87	0	13	1,58
Lise	142	---	0	3	---
Not Ort	142	2,58	1,07	4	0,58
Program	142	---	0	5	---
Tercih	142	11,66	1	30	8,96
Tür	142	---	0	2	---
Yaş	142	21,23	18	29	2,05
Yerleşim	142	---	0	4	---

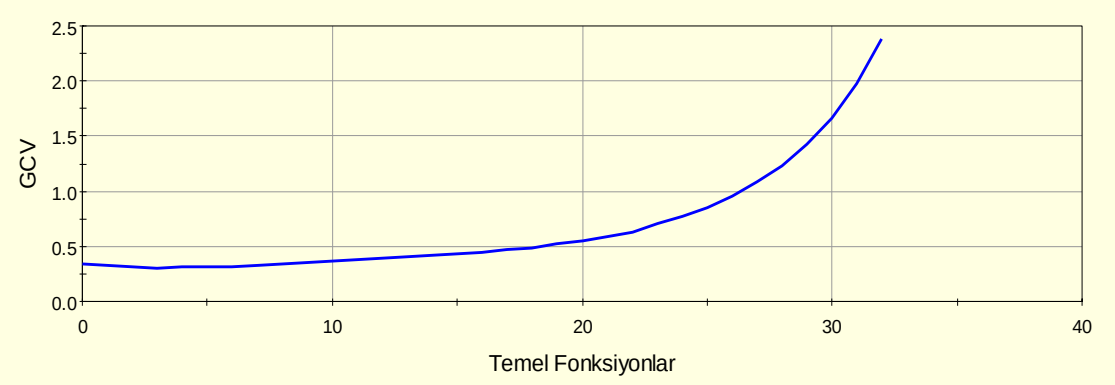
5. BULGULAR

MARS modeli inşa ederken mümkün olan tüm temel fonksiyonlar maksimum karmaşıklığa ulaşıncaya kadar eklenir. Maksimum karmaşıklığa ulaştıktan sonra prosedür budama işlemini yapar. Bu iki aşama ile uygun model GCV değeri ile belirlenir. Kurulan modele ilişkin temel fonksiyonlar ve bunlara ait bilgiler tablo 5.1. te gösterilmiştir.

Tablo 5.1. Kurulan model için sonuç değerleri

Temel fonksiyonlar	Belirleyici sayısı	Kullanılan değişken sayısı	Etkin parametler	GCV	$GCVR^2$
32	11	11	108,00	2,37877	-6,00207
31	11	11	104,66	1,97196	-4,80459
30	11	11	101,31	1,66209	-3,89247
29	11	11	97,97	1,42427	-3,19244
28	11	11	94,63	1,23487	-2,63492
27	11	11	91,28	1,08024	-2,17976
26	11	11	87,94	0,95721	-1,81761
25	10	10	84,59	0,85615	-1,52014
24	10	10	81,25	0,76781	-1,26010
23	10	10	77,91	0,70233	-1,06735
22	10	10	74,56	0,63478	-0,86851
21	10	10	71,22	0,58772	-0,72998
20	10	10	67,88	0,54947	-0,61740
19	10	10	64,53	0,51707	-0,52204
18	10	10	61,19	0,48731	-0,43442
17	10	10	57,84	0,46746	-0,37600
16	9	9	54,50	0,44404	-0,30707
15	9	9	51,16	0,43254	-0,27322
14	9	9	47,81	0,42128	-0,24007
13	9	9	44,47	0,40936	-0,20498
12	9	9	41,13	0,39089	-0,15062
11	9	9	37,78	0,37986	-0,11815
10	9	9	34,44	0,36758	-0,08199
9	9	9	31,09	0,35083	-0,03269
8	9	9	27,75	0,33758	0,00632
7	9	9	24,41	0,32113	0,05472
6	8	8	21,06	0,31803	0,06385
5	8	8	17,72	0,31486	0,07318
4	7	7	14,38	0,31204	0,08149
*3	6	6	11,03	0,30729	0,09548
2	5	5	7,69	0,31333	0,07770
1	3	3	4,34	0,32355	0,04761
0	0	0	1,00	0,33972	

Kurulan model için oluşan Tablo 5.1. de ilk sütunda her bir regresyon modelindeki temel fonksiyon sayısı, ikinci sütunda regresyon modelindeki fonksiyonların yapısında bulunan toplam belirleyici sayısı, üçüncü sütunda kullanılan değişken sayısı, dördüncü sütunda tüm etkin parametreler, dördüncü sütunda modelden elde edilen GCV değerleri ve son sütün da $GCVR^2$ değerleri verilmiştir. Modelin ileri adımsal prosedürü ile en karmaşık yapıya sahip olan model 32 temel fonksiyondan oluşmuştur. Oluşturulan modelde GCV değerleri 0,30727-2,3787 arasında değişmektedir. En uygun model GCV değerinin en küçük olduğu (ideal) ve 3 temel fonksiyondan oluşan ideal modeldir. Başlangıç modelinde ise GCV değeri ise 0,33972 temel fonksiyon sayısı sıfırdır.



Şekil 5.1.Temel fonksiyonlara karşılık GCV değerlerinin grafiksel gösterimi.

Şekil 5.1. de temel fonksiyonlara karşılık gelen GCV değerleri çizgi grafik ile gösterilmiştir.

5.1. İdeal Modele İlişkin Sonuçlar

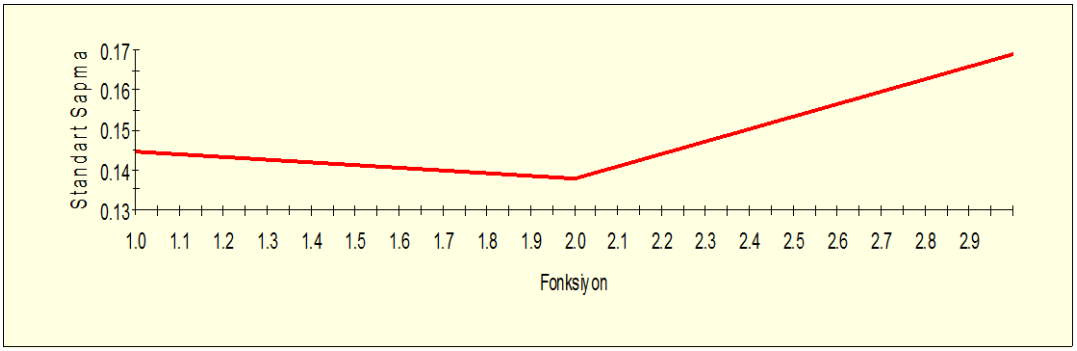
İdeal model belirlenirken en az uyum eksikliğine sahip olan seçilir. En düşük GCV değerine göre seçilen ideal modele ilişkin bilgiler Tablo 5.2. de verilmiştir.

Tablo 5.2. İdeal Modele İlişkin Bilgiler

Temel fonksiyonlar	Katsayılar	Değişken	Kaynak	Düğüm Noktası
0	2,6188			
11	0,0001	TERCIH	CINSIYET	1,0000
12	-0,0013	LISE_	CINSIYET	SubSet1
36	-0,1469	TERCIH	YERLESIM	7,0000

Tablo 5.2. de dört bağımsız değişkenin etkileşimi ile oluşan temel fonksiyonlar, katsayılar ve düğüm noktası değerleri gösterilmiştir.

İdeal modele ilişkin genelleştirilmiş çapraz geçerlilik değeri GCV=0,30729, ve genelleştirilmiş çapraz geçerlilik açıklayıcılık değeri $GCVR^2 = 0{,}09548$ dir. Model için açıklayıcılık değeri $R^2 = 0{,}2196$ ve modelin hata kareler ortalaması MSE= 0,2614 olarak hesaplanmıştır.



Şekil 5.2.İdeal modele ilişkin varyans çözümleme grafiği.

İdeal modele ilişkin Şekil 5.2. de üç temel fonksiyonun varyans çözümlemesi grafik olarak gösterilmiştir. Modele ilişkin varyans çözümlemesi bilgileri Tablo 5.3. de verilmiştir.

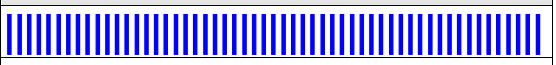
Tablo 5.3. İdeal model için varyans çözümlemesi

Fonksiyon	Standart Sapma	İhmal etme terimi	Temel Fonksiyonları	Değişkenler
1	0,14477	0,31530	1	CİNSİYET,GELİR,TERCİH
2	0,13804	0,31333	1	CINSIYET, LISE_, GELIR
3	0,16882	0,32370	1	PROGRAM, YERLESIM, TERCIH

Tablo 5.3. e göre CİNSİYET, GELİR, TERCİH için modelden ihmal etme terimi 0,31530, standart sapması 0,14477 ve etkili parametre sayısı 3,344 olarak belirlenmiştir. Benzer şekilde diğer değişkenlere ait bilgiler verilmiştir.

İdeal modele ait değişkenler ve bu değişkenlerin göreceli önemlilikleri Tablo 5.4. de verilmiştir. Buna göre TERCIH değişkeni %100 öneme sahiptir. YERLESIM, PROGRAM, GELİR, CINSIYET VE LISE değişkenleri modele alınması gereken diğer değişkenlerdir. YAS, IKAMET, IS, KARDES, BURS ve TUR değişkenlerinin modele alınmasının gereksiz(önemsiz) olduğu belirlenmiştir

Tablo 5.4. İdeal Model İçin Göreceli Önemlilik Yüzdeleri

Değişken	Önem Yüzdesi	
TERCIH	100,00	
YERLESIM	76,37	
PROGRAM	76,37	
GELIR	76,02	
CINSIYET	76,02	
LISE_	46,32	
YAS	0,00	
IKAMET	0,00	
IS	0,00	
KARDES	0,00	
BURS	0,00	
TUR	0,00	

Buna göre NOT ORTALAMASI ile TERCIH değişkeni arasındaki ilişkiyi %100 olarak açıklarken YERLESIM ve PROGRAM değişkenlerinin %76,37, GELİR ve CINSIYET değişkenlerinin %76,02 LISE değişkeninin %46,32 oranında açıkladığı görülmüştür. Yani NOT ORTALAMASI ile TERCİH Değişkenleri arsındaki ilişki tam iken YERLESIM, PROGRAM, GELİR ve CINSIYET değişkenleri arasındaki ilişki kuvvetli ve LİSE değişkeni arasındaki ilişiki zayıf olduğu görülmüştür.. Modele alınması gereken diğer değişkenlerdir. YAS, IKAMET, IS, KARDES, BURS ve TUR değişkenlerinin modele alınmasının gereksiz(önemsiz) olduğu belirlenmiştir

İdeal model için temel fonksiyonlar ve sonuç fonksiyonu aşağıda şekilde ortaya çıkmıştır;

Subsets for PROGRAM
SubSet1 = { "0", "3", "5" }
Subsets for CINSIYET
SubSet1 = { "0" }
SubSet2 = { "1" }
Subsets for LISE_
SubSet1 = { "0" }
Subsets for YERLESIM
SubSet1 = { "1", "2" }

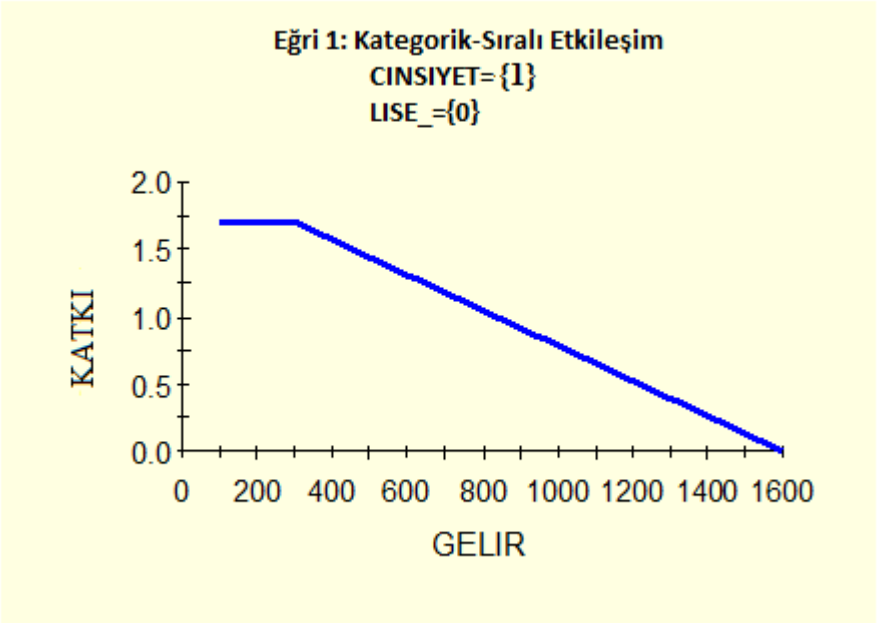
Temel Fonksiyonlar;

BF1 = (PROGRAM is in SubSet1);
BF7 = max(0, GELIR - 300);
BF9 = (CINSIYET in ("1")) * BF7;
BF10 = (CINSIYET in ("1")) * BF7;
BF11 = max(0, TERCIH - 0.999999) * BF10;
BF12 = (LISE_ is in SubSet1) * BF9;
BF28 = (YERLESIM is in SubSet1) * BF1;
BF36 = max(0, 7 - TERCIH) * BF28;

```
Y = 2.61885 + 0.000141606 * BF11 - 0.00130403 * BF12
        - 0.146938 * BF36;
```

```
MODEL ORTALAMA = BF11 BF12 BF36;
```

Sabit terim: β_0 = 2.61885



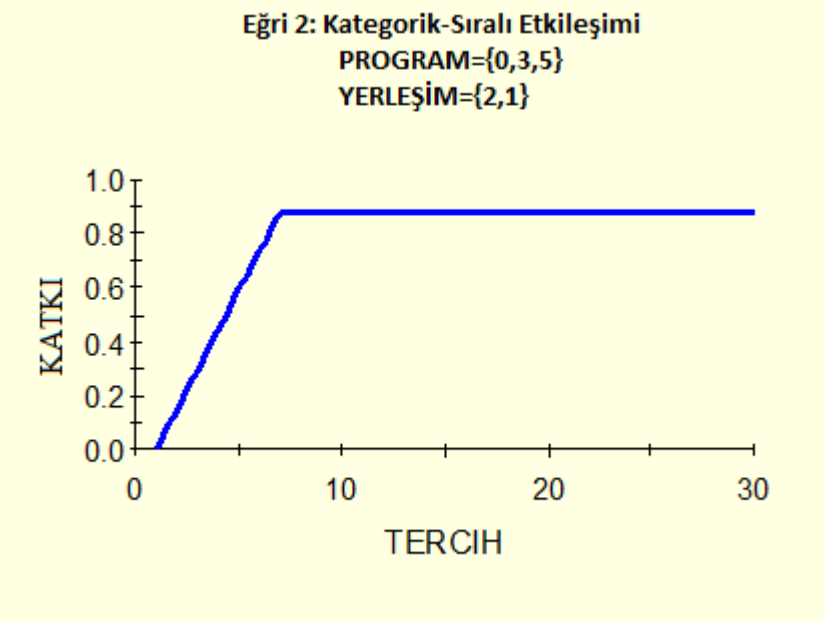
Şekil 5.3.GELIR ile CINSIYET={1} ve LISE_={0} arasındaki ilişki ve düğüm değeri.

BF7 = max(0, GELIR - 300);

BF9 = (CINSIYET in ("1")) * BF7;

BF12 = (LISE_ is in SubSet1) * BF9;

Bu modelde geliri 300 ve daha küçük olan öğrencilerde CINSIYET değişkeni 1 ve LISE_ değişkeni 0 olanların NOT ORT. bağımlı değişkenine etkisi sıfır iken, geliri 300-1600 arası olan öğrencilerde ters yönde bir ilişki oluşturmaktadır.



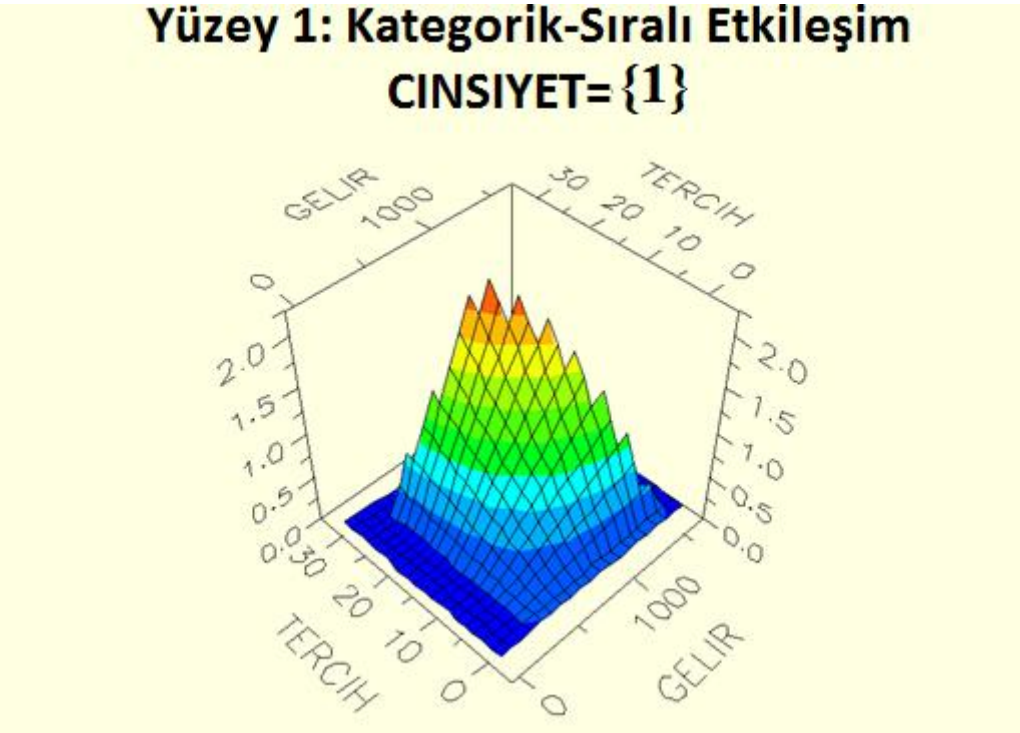
Şekil 5.4.TERCIH ile PROGRAM={0, 3, 5} ve YERLESIM={2, 1} arasındaki ilişki ve düğüm değeri.

BF1 = (PROGRAM is in SubSet1);

BF28 = (YERLESIM is in SubSet1) * BF1;

BF36 = max(0, 7 - TERCIH) * BF28;

Bu modelde PROGRAM değişkeni 0,3 ve 5 inci gruba dahil olan ve YERLESIM değişkeni 2 ve 1 inci gruba dahil olanlar ile tercihi 7 den küçük olanlar arasında doğru yönde bir ilişki var iken, tercihi 7 ve daha büyük olanlar için bağımlı değişkene etkisi sıfır olarak belirlenmiştir.



Şekil 5.5.GELIR ve TERCIH değişkenlerinin etkileşimi ile NOT ORT. bağımlı değişken arasındaki ilişki.

Modelde CINSIYET={1} olduğu durumda, gelirin 300-1000 Türk Lirası arasında ve tercihin 7 ve 7 den büyük olduğu bölgeler de NOT ORT. değişkenin arttığı, gelirin 300 den ve tercih sırasının 7 den küçük olduğu bölgeler de NOT ORT. bağımlı değişkenin azaldığı görülmektedir.

6. SONUÇ VE ÖNERİLER

Veri analizlerinde kullanılan ve yaygın olarak bilinen istatistik değerlendirme araçlarından biri doğrusal modeldir. Genelleştirilmiş Doğrusal Modeller (GLM) bağımlı değişkenin sürekli olmadığı ve sürekli olup normal dağılım göstermediği durumlarda verilerin analizine imkan sağlamaktadır. Bağımlı değişkenin nitel ve sürekli olduğu durumlarda ise bağımsız değişkenler ile bağımlı değişkenler arasındaki ilişki kolay ifade edilmeyebilir.

MARS Modeli, her bağımsız değişkenin bağımlı değişken ile olan ilişkilerini incelemekle birlikte, bağımsız değişkenlerin birbirleri arasındaki etkileşimleri de belirler ve etkileşimlerin bağımlı değişken üzerindeki etkisini de ortaya koyar. Bu modelin temelini zincirler oluşturur. MARS modeli, bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan ilişkileri doğrusal yapıya dönüştürerek uygun çözümler elde eder ve bağımsız değişkenler arasında da etkileşimleri belirler.

Bu çalışmada bağımlı değişkenle bağımsız değişkenlerin ilişkisi ölçülmüş, bağımlı değişken olan not ortalamasına en fazla etkiyi öğrencilerin tercih sıraları göstermiştir. Etki sırasına göre aile yerleşim yeri, öğrenim görülen program, gelir düzeyi ve mezun oldukları lise türü bağımlı değişkene etki eden diğer bağımsız değişkenlerdir. Yaş, ikamet edilen yer, ek iş durumu, okuyan kardeş sayısı, burs durumu ve öğrenim türünün not ortalaması üzerine etkisinin olmadığı görülmüştür.

Yapılan istatistiksel analizlerde SPM7 hazır paket programı kullanılmıştır. Elde edilen sonuçlar daha önceki çalışmalarla karşılaştırıldığında birbirini destekler nitelikte olduğu görülmüştür. Devamı niteliğinde modelde yer alan değişkenler ve yeni değişkenler eklenerek eğitim alanında çalışmalar yapılabileceği düşünülmektedir. Diğer yandan bu çalışmada sadece MARS Modeli üzerinde çalışılmıştır. Teorik bazda bahsedilen diğer zincirlerin bu verilere ve diğer çalışmalara örneğin CART modeline uygulaması yapılacaktır.

7. KAYNAKLAR

1. Abraham, A., Steinberg, D., "MARS: Still an Alien Planet in Soft Computing?" , ***Lecture Notes in Computer Science***, 2074:235-244 (2001).

2. Aitken, M., Anderson, D., Francis, B. and Hinde, J., "Statistical Modelling in GLIM", ***Clarendon Pres.,*** Oxford, (1989).

3. Akyol, M, "Yaşam Çözümlemesine Yeni Bir Yaklaşım: MARS", Doktora Tezi, ***Türkiye Cumhuriyeti Ankara Üniversitesi Sağlık Bilimleri Enstitüsü,*** Ankara, (2011)

4. Briand, L.C, Freimut B., Vollei, F., "IESE;Using Multiple Adaptive Regresyon Splines to Understand Trends in İnspection Data And İdentify Optimal Inspection Rates", ***Software Engineering Research Network Technical Report***, Germany, 5-10 (2000).

5. Cengiz, M.A., **"**Bivariate Logistic Regression Analysis.", ***Thecnical Report, The University of Salford,*** MCS-97-11. **(**1997)

6. Chatterjee S, Hadi A, Price B.," Regression Analysis By Example. 3rd edition", ***John&Wiley***, Canada, 2000.

7. Chen I, Lee T., "A Two-Stage Credit Scoring Model Using Artificial Neural-Networks and Multivariate Adaptive Regression Splines", 28:743-752(2005).

8. Craven, P., Wahba, G.,"Smoothing Noisy Data With Spline Functions", ***Numer. Math***, 31:377-403(1979).

9. Deıchman, J., Eshgı, 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).

10. Dobson, A.J., **"**An Indtroduction to Generalized Linear Models, Second ed." ***Chapman and Hall,.*** London, (1990).

11. Friedman, J.H.," Multivariate Adaptive Regression Splines", ***Annals of Statistics***, 19 (1): 1-67(1991).

12. Hastie, T. J.**,** "Generalized Additive Models," in Statistical Models in S, ed. J. M. Chambers pand T. J. Hastie, ***Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software***, 249–307. (1991).

13. Hastie, T. and Tibshirani, R. **"**Exploring the nature of covariate effects in the proportional hazards model". ***Biometrics 46,*** 1005-1046, (1990).

14. Hastie,T., Tibshirani, R., Friedman, J., "The Elements of Statistic al Learning; Data mining, Inference and Prediction", ***Springer Verlag***, New York (2001).

15. Hastie, T. and Tibshirani, R., Friedman, J., "The Elements Of Statistical Learning, Second Ed.", ***Springer***, New York, 321-328, (2008).

16. İnternet: Copyright StatSoft "Multivariate Adaptive Regression Splines" http://www.statsoft.com/Textbook/Multivariate-Adaptive-Regression-Splines, (2013)

17. İnternet: MARS User Guıde © Salford Systems "Multivariate Adaptive Regression Splines" , http://media.salford-systems.com/pdf/spm7/IntroMARS.pdf, (2013)

18. İnternet: Staicu, A. M. "Classification Trees and MARS", http://www.utstat.utoronto.ca/reid/sta450/MARS16pdf, (2013)

19. Kaki, B.,Yeşilova, A., Şen, C.,"Yarı Parametrik Regresyon Yönteminin Hayvancılıkta Kullanılması", ***4. Ulusal Zootekni Bilim Kongresi Sözlü Bildiriler Programı***, Van, 26-32 (2004).

20. Kayri, M., "The Analysis of Internet Addiction Scale Using Multivariate Adaptive Regression Splines", ***Iranian J Publ Health***, 39: 51-63(2010).

21. Kolyshkina, I., Sylvia, W., "Enhancing Generalised Linear models with Data Minin", ***Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society – Arlington,*** Virginia, 279-290 (2004).

22. Lionel C. Briand, Bernd Freimut, Ferdinand Vollei," Using Multiple Adaptive Regression Splines To Support Decisionmaking İn Code İnspections", ***The Journal of Systems and Software***, 73, 205–217, (2004).

23. McCullagh, P. and Nelder, J. A. "Generalized Linear Models." ***Chapman and Hall.,*** London, (1983).

24. McCullagh, P. and Nelder, J.A. "Generalized Linear Models, Second Edition", ***Chapman and Hall***, London, (1989).

25. Montgomery, D.C., Peck, E.A, and Vining, G.G., "Introduction to Linear Regression Analysis, Third ed.", ***John Wiley***, New York, (2001).

26. Nelder, J.A. and Wedderburn, R.W.M. "Generalized Linear Models", ***J.Roy. Statist. Soc. Ser. A***, 135,370-384. (1972).

27. Press, W.H., Teukolsky, S.A., Vetterling, W.T. Flannery B.P., "Numerical Recipes in C: The Art of Scientific Computing, Second Edition.", ***Cambridge University Press,*** New York, (1992).

28. Reinsch, C. H., "Smoothing by spline functions." ***Numer. Math. 10***, t 77-183 (1967).

29. Ruppert, D., Wand, M. P. and Carroll, R.J., " Semiparametric Regression". ***Cambridge University Press***. New York,(2003).

30. Stone, C. J.**, "**Additive Regression and Other Nonparametric Models", ***Annals of Statistics, 13***, 689–705, (1985).

31. Stone, C.J. and Koo, C.Y. "Additive splayns in statistics.", ***Proceedings of the Stat. Comp. Sec***, ASA 45-8 (1985).

32. Temel, G.O., Çamdeviren, H., Yazıcı, A.C.,"Regresyon modellerine alternatif bir yaklaşım: MARS", ***VIII. Ulusal Biyoistatistik Kongresi***, Bursa, 105-123(2005).

33. Topak, M., S., "Kurumsal Başarısızlığı Modellemek İçin Türkiye Üzerine Yapılan Ampirik Bir Çalışma: Çok Değişkenli Uyumlu Regresyon Uzanımları(MARS) Tekniği Kullanılarak Geliştirilen Model Önerisi", Namık Kemal Üniversitesi Sosyal Bilimler Metinleri, 15(2011).

34. Tunay, K.B., "Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini", ***METU Studies in Development***, Ankara, 28(2):1-23(2001).

35. Tunay, K., B., "Bankacılık Krizleri Ve Erken Uyarı Sistemleri: Türk Bankacılık Sektörü İçin Bir Model Önerisi", ***BDDK Bankacılık ve Finansal Piyasalar Dergisi***, 4(1), 9-46(2010).

36. Tunay, K., B., "Türkiye'de Durgunlukların MARS Yöntemi İle Tahmini ve Kestirimi", ***Marmara Üniversitesi İ.B.B.F Dergisi***, 30(1), 71-91(2011).

37. Verzıllı, C.J., Whıttaker, 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),*** 54 (1):191–206(2005).

38. Yerlikaya, F., "A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Applications to Data Mining for Quality Control in Manufacturing", M.Sc. Thesis, ***Middle East Technical University***, Ankara, 102(2008).

ÖZGEÇMİŞ

Alparslan OĞUZ 05/04/1986 tarihinde Erzincan Yaylabaşı Beldesinde doğdu. İlköğrenimi Erzincan Cumhuriyet İlköğretim Okulunda, orta öğrenimini Erzincan Milliyet Anadolu Öğretmen Lisesinde tamamladı. 2010 yılında Atatürk Üniversitesi Kazım Karabekir Eğitim Fakültesi Matematik Öğretmenliği bölümünden mezun oldu. Aynı yıl Erzincan Üniversitesi Fen Bilimleri Enstitüsü Matematik anabilim dalı Uygulamalı Matematik bölümünde yüksek lisansa başladı. 2011 yılında Erzincan Üniversitesi Kemah Meslek Yüksek Okulunda öğretim görevlisi olarak akademik kariyerine başladı ve halen bu görevi yürütmektedir. İngilizce biliyor