**Çok Değişkenli Uyarlamalı Regresyon Uzanımları (MARS) Yöntemi ile İlgili Sağlık Alanında Yapılan Çalışmalar**

**Studies on the Field of Health, Related to Multivariate Adaptive Regression Splines (MARS) Method**

Şenol ÇELİK\*

\*Bingöl Üniversitesi Ziraat Fakültesi Zootekni Bölümü Biyometri ve Genetik ABD, Bingöl

**ÖZET** Bu çalışmada, Çok Değişkenli Uyarlanabilir Regresyon Uzanımları (MARS) yönteminin özellikleri, konu ile ilgili yapılan çalışmalar ve uygulama alanları üzerinde durulmuş ve böyle karmaşık algoritmaların kullanımının getireceği avantajlar açıklanmıştır. MARS algoritmalarının tahmin doğruluğunu belirlemek için uyum kriterleri (GCV-Genelleştirilmiş Çapraz geçerlilik katsayısı, R2 ve Düzeltilmiş R2) açıklanmıştır. GCV'yi hesaplayan etkileşim dereceleri ve uygun terim sayısının belirlenmesi, MARS algoritması için çok önemlidir. MARS modelinin uygulanabilirliğini göstermek için yapılan çalışmalar değerlendirilmiştir. Genelde Lojistik regresyon, doğrusal olmayan regresyon modeli, çok değişkenli regresyon modeli, yapay sinir ağları, CART ve CHAID gibi modeller ile MARS modelleri karşılaştırmalı olarak incelenmiştir. MARS yönteminin bu çalışmalarda diğer yöntemlerden daha iyi sonuçlar verdiği saptanmıştır.

**Anahtar Kelimeler:** MARS Algoritmaları, temel fonksiyonlar, genelleştirilmiş çapraz geçerlilik

**ABSTRACT** The features of Multivariable Adaptive Regression Splines (MARS) method, the studies related to the subject, and the application areas of the method were emphasized, and the advantages of using such complex algorithms were expressed in this study. (GCV-Generalized Cross-Validation coefficient, R2 and Adjusted R2) were defined to determine the accuracy of the estimation of MARS algorithms. The determination of the degree of interaction, and of the number of appropriate terms that were used in the calculation of GCV are very important for MARS algorithm. The studies conducted to demonstrate the applicability of the MARS model were evaluated. The models such as logistic regression, nonlinear regression model, multivariate regression model, artificial neural networks, CART and CHAID, and MARS models were comparatively studied in general. It was determined that MARS method produced better results in these studies compared to other methods.

**Keywords:** MARS Algorithm, basic functions, generalized cross validation

**GİRİŞ**

Çok değişkenli uyarlamalı regresyon uzanımları (MARS) yöntemi tıp, ziraat, ekonomi, biyoloji gibi birçok alanda kullanılmaktadır. Uygulamada kolaylık sağladığı için veri madenciliği alanında popüler olmuştur. MARS bir düzleştirme algoritmasıdır. Bu nedenle de her türlü veri analizinde rahatlıkla kullanılabildiği gibi diğer analizler için de verilerin istenilen şekilde düzenlenmesine imkan vermektedir. Bu yöntem sebep ve sonuç değişkenleri arasında bir varsayım gerektirmez.

Diğer regresyon modellerinde bazı varsayımlar ve örneklem büyüklüğü aranmasına rağmen, MARS modelinde böyle varsayımların gerekmemesi önemli bir avantajdır. MARS modeli veriden elde edildiğinden, parametrik yöntemlerle tahmini güç olan doğru öngörüleri, dinamik yapısı nedeniyle tahmin başarısı yüksek model kurulmasını sağlamaktadır.1

Çok Değişkenli Uyarlanabilir Regresyon Uzanımları (Multivariate Adaptive Regression Splines: MARS) yöntemi Parametrik Olmayan Regresyonda mevcut olan yeni bir yöntemdir. MARS yöntemi, 1991’de istatistikçi Jerome H. Friedman tarafından geliştirilen parametrik ve doğrusal olmayan bir tekniktir.2,3 Parametrik ve doğrusal olmayan bu yöntem, değişkenlerin alt kümelerini dikkate almaktadır4. Bir başka deyişle tahmin edici değişkenlerin oluşturduğu uzay, birbirleriyle örtüşük birçok bölgeye ayrılıp, uzanım fonksiyonlarını (spline functions) oluşturmakta ve bölgesel olan bu regresyon uzanımları da temel fonksiyon (basis function) olarak adlandırılmaktadır.5

MARS yönteminde; değişkenler arasındaki ilişkideki değişimlerin meydana geldiği noktalar belirlenir ve bu noktalar arasında her bir tanım aralığında uygun dönüşümlerle düz bir doğru elde edilmeye çalışılmaktadır. Bu tanım aralıkları arasındaki kaymalar kontrol altına alınarak düzleştirme yapılmaktadır. Bu kaymalar “düğüm noktası” olarak adlandırılır. Düğüm noktaları arasında tanımlı olan ilişki (karesel, kübik, parabolik vb) ne olursa olsun bu ilişki doğrusal hale getirilir. MARS yönteminin algoritması, tüm değişkenler karşısında olduğu kadar tüm değişkenler arasındaki mümkün olan tüm etkileşimleri içeren her düğüm konumunu belirlemeye yöneliktir.6

MARS yöntemi ile yapılmış çalışmalar vardır. Türe ve ark. (2005)7, Sevimli (2009)8, Ünal (2009)9, Özfalcı (2008)10, Eyduran (2016)11, Eyduran ve ark. (2017)12, Grzesiak ve Zaborski (2012)13 ve Turi ve ark. (2017)14 gibi çalışmalar bu çalışmalara birer örnektir.

Bu çalışmanın amacı, sağlık alanında ve diğer değişik alanlarda kullanılan MARS yönteminin tanıtılması ve literatür taraması sonucu elde edilen bilgilerden yararlanarak bu yöntemin yaygın kullanım alanlarının belirlenmesidir.

**MARS YÖNTEMİ**

MARS algoritması “sınıflandırma ve regresyon ağaçları - classification and regression trees” (CART)’ın bir uzantısıdır ve her iki yöntem de iki simetrik temel fonksiyon (TF)’un düğüm yerinde oluşturulduğu aralıkların bölünmesi işlemi bakımından benzerdir. MARS algoritmasında sürekli parçalı doğrusal fonksiyonlar kullanılır ve doğrusal olmayan ilişkiler daha etkili modellenebilen sürekli bir model oluşturulur. TF’lerin seçimi veriye dayalı ve çalışılan probleme özgüdür. Bu da MARS’ı çok boyutlu problemlerin çözümünde uyarlanabilir bir regresyon yöntemi yapmaktadır. MARS modeli oluşturulurken, kısmi doğrusal TF’ler bağımlı değişkeni belirlemek için bağımsız değişkenlerin katkısal ve etkileşimli etkileri birlikte birbirine eklenir.

MARS algoritması aşağıdaki kesik kısmi doğrusal baz fonksiyonlarının açılımlarını kullanır:

 (1)

 (2)

Eşitlik (1) ve (2)'de, t tek değişkenli düğüm noktası olup, bu iki fonksiyon *yansıyan çift* olarak adlandırılır ve '+' simgesi sadece pozitif parçaların kullanıldığını, diğer durumda sıfır olduğunu gösterir. Genel regresyon modelindeki bağımsız değişkenler ile bağımlı değişken arasındaki ilişki aşağıdaki ifadeyle tanımlanır:

 (3)

Burada *Y* bağımlı değişkeni, bağımsız değişkenler vektörünü, ise ilave sıfır ortalama ve sonlu varyansa sahip stokastik bileşeni göstermektedir.

Eşitlik (3)’te yer alan , ve kesişim noktalı doğrusal bir kombinasyon olarak Eşitlik (4)'de olduğu gibi ifade edilebilir:

 (4)

Eşitlik 4'te, her *Bi*  ifadesi birer temel fonksiyondur. MARS modelinin i. temel fonksiyonu Eşitlik (5)'de verildiği gibi

 (5)

şeklindedir. Burada Kj i. temel fonksiyonun çarpılan doğrusal fonksiyon sayısı; s+1 veya -1; i. temel fonksiyondaki j. değişkendir.15

MARS modelinde, temel fonksiyonlar düğüm belirleme araştırmasının genelleştirilmesi için kullanılan araçlardır. Temel fonksiyonlar, bir veya daha fazla değişkende yer alan bilgileri ve bağımsız değişkenler ile bağımsız değişkenler arasındaki ilişkileri yeniden ifade etmek için kullanılan işlevlerdir. MARS modeli, temel işlevleri çiftler halinde oluşturur.16 Mesela, temel fonksiyonu (BF1) MARS tarafından şöyle tanımlanır:

ve

olarak tanımlanabilir. Söz konusu değişken modele 1150 düğüm noktasında BF1 ve BF2 temel fonksiyonlarla girmiş olsun.

BELGELENEN TOP GELIR değişkeni için oluşturulan temel fonksiyonlar BF1 = (BELGELENEN\_TOP\_GELIR - 1150) ve BF2 = (1150 - BELGELENEN\_TOP\_GELIR) olarak tanımlanmaktadır. Bu durumda BF1 belgelenen toplam gelir tutarının 1150 TL’den büyük olduğu tutarları ve BF2’de bunun tam tersi olarak belgelenen toplam gelir tutarının 1150 TL’den küçük olan tutarları göstermektedir. Bu örnek Ünal (2009)'un yüksek lisans tezinde ayrıntılı olarak açıklanmıştır.9

Bu temel fonksiyonlar ya tek değişkenli ya da çok değişkenli etkileşim terimlerinden oluşmaktadır. En uygun düğüm noktalarını ve temel fonksiyonları belirlemek için Eşitlik (6)'de tanımlanan genelleştirilmiş çapraz doğrulama (GCV) ölçüsü kullanılmaktadır.17,18

(6)

Burada M()=u+dK; N: örneklem genişliği; u: bağımsız temel fonksiyonların sayısı; K: seçilen düğüm sayısı; d: temel fonksiyonların maliyetidir.

GCV’nin payı hata kareler toplamını, paydası ise modelin karmaşıklığını hesaplamaktadır. MARS algoritmasında önemli bağımsız değişkenler ve bu değişkenlerin etkileşimleri belirlenerek, GCV ölçüsü en küçük olan model elde edilir. Elde edilen bu model en uygun modeldir. Bu anlatılan tüm kısım MARS yönteminin "İleri doğru adım algoritması" nı oluşturmaktadır.15

MARS algoritmasının temel adımları aşağıdaki gibi özetlenmiştir.19

* Modelin hesaplanabilen en fazla karmaşıklığı seçilir ve d parametresi tanımlanır.
* En basit modelle, sadece sabit katsayı ile başlanır.
* Her açıklayıcı değişken için temel fonksiyonların kümesi araştırılır.
* Tahmin hatasını minimum yapan temel fonksiyonlar ikilileri belirlenir ve modele dahil edilir.
* Modelin önceden belirlenen tanımlanmış karmaşıklığa ulaşana kadar "Her açıklayıcı değişken için temel fonksiyonların kümesi araştırılır" adımına gidilir.
* Sabit terim hariç tüm temel fonksiyonların kümesi araştırılır ve GCV kriteri kullanılarak modele en az katkıda bulunan temel fonksiyonlar silinir. GCV kendi minimumuna ulaşana kadar bu adım tekrarlanır.

MARS yönteminin diğer aşaması ise "Tikhonov Düzenlemesi" adı verilen "Geriye doğru adım algoritması" dır. Bu aşamada hata kareler toplamına (RSS) ceza parametresi eklenerek hem modelin doğruluğu (RSS) hem de karmaşıklığı kontrol altına alınır. RSS’ye ceza eklenerek elde edilen cezalı hata kareler toplamı (PRSS) Eşitlik (7)'de olduğu gibidir.

 (7)

Burada , m. temel fonksiyonuyla ilişkili değişken kümesidir. , m. temel fonksiyonuna katkıda bulunan değişken vektörünü temsil eder. Üstelik, , için

dir. İntegral içinde karesi alınan türev ise yine m. temel fonksiyona ait bağımsız değişkenlerin kısmi türevleridir.20

Oluşturulan PRSS ölçütü modelin doğruluğu ile karmaşıklığı arasında bir tercihe dayanmaktadır. Bu tercih Eşitlik (7)'de gösterilen PRSS denkleminde kullanılan ceza parametresi ile sağlanmaktadır.20

MARS yöntemi, değişkenlerin tek tek veya kombinasyon halinde model olarak tanımlandığında, düşük veya yüksek dereceli modellerin karşılaştırılmasına imkan sağlamaktadır. Friedman (1991), düzeltilmiş R2’yi bir karşılaştırma ölçütü olarak önermiştir.21 Düzeltilmiş R2 değeri yüksek olan model tercih edilir. Bunun dışında standart sapma oranı (SDratio) ve Hata kareler ortalamasının karekökü (RMSE) diğer uyum iyiliği kriterlerdendir.22 Bu kriterler Tablo 1’de verilmiştir.

Tablo 1. Uyum iyiliği kriterleri

|  |  |
| --- | --- |
| Uyum iyiliği kriterleri | Hesaplama  |
| Belirleme katsayısı  |  |
| Düzeltilmiş Belirleme katsayısı  |  |
| Standart sapma oranı |  |
| Hata kareler ortalamasının karekökü (RMSE) |  |

: gözlenen değer, : tahmin edilen değer : gözlenen değerlerin ortalaması, : hata terimi, : hata terimlerin ortalaması, n: Toplam gözlem sayısı

**TARTIŞMA**

Bu çalışmada, regresyon modellerine ve CHAID, Exh. CHAID ve CART gibi diğer veri madenciliği algoritmalarına alternatif olan ve veri madenciliğinin popüler konularından biri olan MARS yöntemi üzerine bir derleme sunulmuştur.

Uzunlamasına veriler için MARS modeli çalışması Zhang (1996)'dan incelenebilir.23 Müşteri kredisi madenciliği için CART ve MARS algoritmaları çalışması Lee (2006) incelenebilir.24 Yeni Zelanda'nın tatlısu diadromlu balıklarının dağılımlarını tahmin etmek için MARS modeli kullanımı (Leathwick, 2005) makalesinden incelenebilir.25

Sağlık alanında da çeşitli çalışmalar yapılmıştır. Mallick ve arkadaşları (1999), 38 böbrek enfeksiyonlu hastanın enfeksiyon zamanlarını içeren bir veri seti üzerinde MARS modelinden başarılı sonuçlar elde etmişlerdir.26 York ve arkadaşları (2006), ilaç tedavilerinde cinsiyetin etkisi ya da genotip ile çevre etkileşimlerini alt grupları da dikkate alarak en iyi şekilde açıklayabilecek modelin MARS modeli olduğunu ileri sürmüşlerdir.27 Kriner (2007), Alman Kalp Merkezi verilerinden kalp krizi geçirmiş 2376 hastanın bilgilerini kullanarak kalp krizine etki eden faktörleri belirlemek için MARS yaklaşımını kullanmıştır.28

Hastaların depresyon durumunu önemli olarak etkileyen faktörler üzerine MARS yöntemi uygulanmıştır. Bu konudaki çalışma için Temel ve ark. (2010) makalesi incelenebilir.29 Yeniden kabul edilen hastaların risk tahmininde sınıflandırma ağaçları, lojistik regresyon ve MARS yöntemleri karşılaştırmalı olarak araştırılmıştır ve Demir (2014)'ün makalesi incelenebilir.30 MARS modeli ile diş hekimliğinde sıkça kullanılan klinik çalışma yöntemlerinden biri olan bölünmüş ağız tasarımı (split-mouth design) çalışmaları için derlenen verilerle istatistiksel bir tahmin modeli elde edilmiştir ve regresyon analizinden daha iyi sonuçlara ulaşılmıştır (Sevimli, 2009).8 Turi ve ark. (2017), diyabet hastaları üzerine elde ettikleri verilerle MARS yöntemini kullanmışlardır.14 Mina ve Barros (2010), 0-5 yaş arasındaki çocuk ile gebeliğe bağlı nedenlerle kadın ölümlerinin oranı tahminini MARS modeli ile incelemişlerdir.31

**SONUÇ**

Bu çalışmada MARS yöntemi ile temel kavramlar hakkında bilgi verilmiştir. MARS modelinde sürekli ve kesikli değişkenlerden oluşan bağımsız değişkenler kullanılarak bağımsız değişkene etkileri önceki çalışmalarda araştırılmıştır. Diğer regresyon yöntemlerinin çoğunda bu uygulamayı yapmak mümkün değildir. Birçok alanda geniş olarak uygulanması mümkün olan MARS modeli ile yapılan çalışmalarda başarılı sonuçlara ulaşıldığı görülmüştür. Sağlık alanındaki uygulamalarda MARS modeli ile veri madenciliğinde kullanılan sınıflandırma ağaçları, yapay sinir ağları, regresyon modeli, lojistik regresyon gibi modeller karşılaştırmalı olarak deneme sonuçları değerlendirilmiştir. MARS modelinin bu çalışmalardan daha iyi neticeler verdiği saptanmıştır. Yapılan çalışmalarda MARS yöntemi ile hem etkileşimli hem de etkileşimsiz olarak modeller oluşturulduğu için, çok sayıda temel fonksiyondan oluşan aşırı uyumlu (overfit) modeller bulunmuştur.

MARS yönteminde Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation: GCV) kriterine en az katkısı olan temel fonksiyonlar elenmiştir ve sabit temel fonksiyon hariç modele giren bütün temel fonksiyonlar kümesi araştırılmıştır. Bu yöntemde R2 değerlerinin daha yüksek değerler aldığı tespit edilmiştir.

Genel olarak gerek sağlık alanında gerekse diğer alanlarda yapılan çalışmalarda MARS yönteminin diğer yöntemlerden daha iyi sonuçlar verdiği ortaya çıkmıştır.

**KAYNAKLAR**

1. Friedman JH. Multivariate adaptive regression splines (with discussion). Annals of Statistics 1991; 19(1): 1-141.

2. Chou SM, Lee TS, Shao YE, Chen IF. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications 2004; 27(1):133-142.

3. Mukhopadhyay A, Iqbal A. Prediction of mechanical property of steel strips using multivariate adaptive regression splines. Journal of Applied Statistics 2009; 36(1): 1-9.

4. Xu QS, Daeyaert F, Lewi PJ, Massart DL. Studies of relationship between biological activities and HIV reverse transcriptase inhibitors by multivariate adaptive regression splines with curds and whey. Chemometrics and Intelligent Laboratory Systems 2006; 82(1-2): 24-30.

5. Put R, Xu QS, Massart DL, Vander Heyden. Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. Journal of Chromatography A 2004; 1055 : 11–19.

6. Sephton P. Cointegration Tests on MARS. Computational Economics, 1994; 7(1), 23-35.

7. Türe M, Kurt I, Kurum AT, Özdamar K. Comparing Classification Techniques For Predicting Essential Hypertension. Expert Systems with Applications 2005; 29(3):583-588.

8. Sevimli, Y. Çok değişkenli uyarlanabilir regresyon uzanımlarının bir split-mouth çalışmasında uygulaması. Yüksek Lisans Tezi, Marmara Üniversitesi Sağlık Bilimleri Enstitüsü 2009; p.77.

9. Ünal, B. Çok değişkenli uyarlamalı regresyon uzanımları. Yüksek Lisans Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü 2009; p.122.

10. Özfalcı, Y. Çok değişkenli uyarlamalı regresyon kesitleri: MARS. Yüksek Lisans Tezi, Gazi Üniversitesi Fen Bilimleri Enstitüsü 2008; p.58.

11. Eyduran E The possibility of using data mining algorithms in prediction of live body weights of small ruminants. Adv Cal Anal, 2016; 1(1): 18-21.

**12.** Eyduran E, Akkuş O, Kara MK, Tırınk C, Tariq MM. Use of Multivariate Adaptive Regression Splines (MARS) in predicting body weight from body measurements in Mengali Rams. **International Conference on Agriculture, Forest, Food Sciences and Technologies (ICAFOF 2017) 2017; 15-17 May, Cappadocia / Turkey.**

13. Grzesiak, W, Zaborski, D. Data Mining Applications in Engineering and Medicine. Examples of the Use of Data Mining Methods in Animal Breeding. 2012; 303-324. <http://dx.doi.org/10.5772/50893>

14. Turi KN, Buchner DM, Grigsby-Toussaint DS. Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors. Preventing chronic disease public health research, practice, and policy 2017; 14(3): 1-13.

15. Taylan, P, Weber, GW, Yerlikaya, F, Batmaz, I, Köksal, G. MARS algoritmasında Tikhonov düzenlemesi ve çok amaçlı optimizasyon kullanımı, in ¨the Proceedings of Operational Research and Industrial Engineering Annual Conference (YA/EM’08) Galatasaray University, Istanbul, Turkey, 2008; June 30 - July 2.

16. Erick, K. C. A multivariate adaptive regression splines approach to predict the treatment outcomes of tuberculosis patients in Kenya. Head o f Statistics and Operations Research School o f Mathematics University of Nairobi, research project 2012; p.70.

17. Sokołowski A, Pasztyła A. Data mining in forecasting the requirement for energy carriers. StatSoft Poland, Kraków 2004; pp. 91–102.

18. Zakeri IF, Adolph AL, Puyau MR, Vohra FA, Butte NF. Multivariate adaptive regression splines models for the prediction of energy expenditure in children and adolescents. Journal of Applied Physiology 2010; 108(1): 128–136.

19. Xu QS, Daszykowski M, Walczak B, Daeyaert F, Jonge de MR, Heeres J, Koymans LMH, Lewi PJ, Vinkers HM., Janssen PA, Massart DL. Multivariate adaptive regression splines studies of HIV reverse transcriptase inhibitors. Chemometrics and Intelligent Laboratory Systems 2004; 72: 27– 34.

20. Yerlikaya, F. A New Contribution to Nonlinear Robust Regression and Classification with Mars and its Applications to Data Mining for Quality Control in Manufacturing, Master Thesis METU, 2008.

21. Friedman J. Discussion The Π Method for Estimating Multivariate Functions From Noisy Data Technometrics 1991; 33(1):145-148.

22. Ali M, Eyduran E, Tariq MM, Tirink C, Abbas F, Bajwa MA, Baloch MH, Nizamani AH, Waheed A, Awan MA, Shah SH, Ahmad Z, Jan S. Comparison of artificial neural network and decision tree algorithms used for predicting live weight at post weaning period from some biometrical characteristics in Harnai Sheep. Pakistan J. Zool 2015; 47(6):1579-1585.

23. Zhang, H. Multivariate adaptive splines for analysis of longitudinal data. Public Health Service Grant 5 R01 CA59039-20, Technical Report 1996; No. 182.

24. Lee TS, Chiu CC, Chou YC, Lu, CJ. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. Computational Statistics and Data Analysis 2006; 50(4): 1113–1130.

25. Leathwick JR, Rowe D, Richardson J, Elith J, Hastie T. Using multivariate adaptive regression splines to predict the distributions of New Zealand’s freshwater diadromous fish. Freshwater Biology 2005; 50(12): 2034–2052

26. Mallick BK, Denison DGT, Smith AFM. Bayesian Survival Analysis using a MARS Model, Biometrics 1999; 55:1071-1077.

27. York TP, Eaves LJ, Oord EJCG. Multivariate adaptive regression splines: a powerful method for detecting disease–risk relationship differences among subgroups, Statist. Med. 2006; 25(8):1355–1367.

28. Kriner M. Survival Analysis with Multivariate Adaptive Regression Splines, Yayınlanmamış doktora tezi, Almanya, 2007. Erişim: (edoc.ub.uni-muenchen.de/7369/1/Kriner\_Monika.pdf ) Erişim Tarihi: 12.03.2017.

29. Temel GO, Ankaralı H, Yazıcı, AC. Regresyon Modellerine Alternatif Bir Yaklaşım: MARS. Türkiye Klinikleri J Biostat. 2010; 2(2): 58-66.

30. Demir EA decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. Decision Sciences 2014; 45(5): 849-880.

31. Mina, D., Erniel, B. Barrios. Profiling Poverty with Multivariate Adaptive Regression Splines Christian. Philippine Journal of Development 2010; 69(2): 55-97.