T.C.
HARRAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ İLE ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON EĞRİLERİ (MARS MODELLEMESİ) YÖNTEMİNİN UYGULANMASI

Fatma TOSUN

ZOOTEKNİ ANABİLİM DALI

ŞANLIURFA 2021

İÇİNDEKİLER

Sayfa No

ÖZET i
ABSTRACT ii
TEŞEKKÜR iii
ŞEKİLLER DİZİNİ iv
1. GİRİŞ 1
2. ÖNCEKİ ÇALIŞMALAR 4
3. MATERYAL ve YÖNTEM 8
3.1. Materyal 8
3.2. YÖNTEM 8
4. ARAŞTIRMA BULGULARI ve TARTIŞMA 12
5. SONUÇLAR ve ÖNERİLER 20
KAYNAKLAR 21
EKLER 22

ÖZET

Yüksek Lisans Tezi

VERİ MADENCİLİĞİ İLE ÇOK DEĞİŞKENLİ UYARLANABİLİR REGRESYON EĞRİLERİ (MARS MODELLEMESİ) YÖNTEMİ UYGULANMASI

Fatma TOSUN

Harran Üniversitesi
Fen Bilimleri Enstitüsü
Zootekni Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Zeki DOĞAN
YIL: 2021, Sayfa: 23

Bu çalışmada, Friedman'ın 1991'de geliştirmiş olduğu çok değişkenli uyarlanabilir regresyon eğrileri (Multivariate Adaptive Regression Splines-MARS) modellemesi kullanılmıştır. MARS modelinin avantajlarına değinilerek, Holstein- Montofon hayvanların ırk, yaş, laktasyon süresi, canlı ağırlık ve belli bir aralıktaki süt verimi bağımsız değişkenleri kullanılarak ortalama süt verimini (OSV) MARS algoritması kullanılarak modellenmesi amaçlanmıştır. MARS, adım fonksiyonlarına benzer kesme noktalarını (düğümler) değerlendirerek polinom regresyonunun doğrusal olmayan yönünü yakalamak için uygun bir yaklaşım sağlar. Çalışmada Şanlıurfa'da yerel bir çiftlikteki 25 hayvanın ırk, yaş gibi değişkenlerinin yanı sıra 2020 yılında Haziran-Eylül ayları arasındaki aylık süt verimi, kayıtlardaki laktasyon süreleri ve elde edilen canlı ağırlık verileri kullanılmıştır. Ortalama süt verimi için, maksimum etkileşimde olan modellerin denenip en uygun modelin belirlenmesi sağlanmıştır. Belirleme katsayısı (R^2) maksimum ve genelleştirilmiş çapraz geçerlilik katsayısı (GCV) minumum olarak tahmin edilmiştir. Buna göre, OSV maksimum etkileşimin derecesi arttığında, GCV değerinin azaldığı ve 2'li etkileşimde sabitlendiği görülmektedir. OSV için en uygun olan modelin; 3 temel fonksiyonla 2'li etkileşimde, R^2 değeri 0.999, düzeltilmiş R^2 değeri 0.996 ve GCV değeri 1.578'dir. Modelin içine dahil edilen tüm temel fonksiyonların %99 seviyesinde önemliliği görülmüştür. OSV tahmin etmek için kurulan MARS modellerinde en etkin değişkenin canlı ağırlık olduğu tahmin edilmiştir. Mevcut araştırmanın sonuçlarına göre MARS yönteminin OSV'nin istatistiksel olarak modellenebilmesi ve değerlendirilmesi için uygun bir örnek olduğu görülmektedir.

ANAHTAR KELİMELER: Mars, Regresyon, Modelleme, Laktasyon

ABSTRACT

MSc Thesis

APPLICATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS MODELING) METHOD WITH DATA MINING

Fatma TOSUN

Harran University
Graduate School of Natural and Applied Sciences
Department of Animal Husbandry

Supervisor: Assist. Prof. Dr. Zeki DOĞAN
Year: 2021, Page: 23

In this study, Multivariate Adaptive Regression Splines (MARS) modeling developed by Friedman in 1991 was used. By referring to the advantages of the MARS model, it is aimed to model the mean milk yield (MMY) of Holstein-Montofon animals using the independent variables of race, age, lactation period, body weight and a certain range of milk yield using the MARS algorithm. MARS provides a convenient approach to capture the nonlinear aspect of polynomial regression by evaluating breakpoints (nodes) similar to step functions. In the study, in addition to the variables such as race and age of 25 animals in a local farm in Sanliurfa, monthly milk yield between June and September in 2020, lactation periods in the records and live weight data were used. The most suitable model was determined by testing maximum interactive models for average milk yield. The coefficient of determination (R^2) was estimated as the maximum and the generalized cross-validity coefficient (GCV) as the minimum. Accordingly, it is seen that when the MMY maximum interaction degree increases, the GCV value decreases and is fixed in 2-way interaction. The most suitable model for MMY; with 2 interactive, 3 basic functions, the GCV value is 1.578, the R^2 value is 0.999 and the corrected R^2 value is 0.996. It has been observed that all basic functions included in the model are important at 99% level. Live weight was estimated to be the most effective variable in MARS models established to predict OSV. According to the results of the present study, it is seen that the MARS method is a suitable example for statistical modeling and evaluation of MMY.

KEY WORDS: Mars, Regression, Modeling, Lactation

TEŞEKKÜR

Tezin konusunun seçiminde, uygulamasında ve çalışmamda yardımlarını esirgemeyen danışmanım sayın Dr. Öğr. Üyesi Zeki DOĞAN'a ve tez çalışmamda beni fikirleriyle doğru yönde yönlendiren Prof. Dr. Mustafa ŞAHİN'e, desteğini hiçbir koşulda ve şartta üzerimden esirgemeyen babam Zr. Müh. Ahmet TOSUN'a, merhametini ve sevgisini bizden hiç esirgemeyen annem Dilber TOSUN'a ve hayatımın her anını güzelleştiren kardeşlerim Dr. Halil TOSUN ve Ecz. Muhammet Fatih TOSUN'a teşekkür ederim.

ŞEKİLLER DİZİNİ

Şekil 4.1. ……..……..……...16
Şekil 4.2. ……...……...…. 16
Şekil 4.3……..……...……. 17
Şekil 4.4 ……...…. 17
Şekil 4.5. ……...……...……. 18
Şekil 4.6……..……...…. 18
Şekil 4.7 ……...……. 19

ÇİZELGELER DİZİNİ

Çizelge 4.1..…..…….……. 13

Çizelge 4.2…….……...……..….13

Çizelge 4.3……...13

Çizelge 4.4……...14

SİMGELER ve KISALTMALAR

CVRSq	Çapraz geçerlilik değeri
GCV	Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation)
GRSq	Modelin tahminleme gücü
OSV	Ortalama süt verimi
R^2	Belirleme katsayısı

1. GİRİŞ

Veri madenciliği uygulamaları bilimsel bilgisayar ağlarının doğal bir sonucu olarak görülebilmektedir. Edinilen bilgiyi depolamayı, uzun zaman alacak uygulamaları yorumlamayı ve ileriye dönük tahminler yapmayı sağlayan bir akışın sonucudur. Bu akış sonucunda verilerin birleştirilmesi, gerekli verinin seçilmesi, verileri elverişli biçime dönüştürme, veri madenciliği yorumları ve uygulamaları şeklinde sıralanabilmektedir.

Regresyon analizi konusu, İstatistik alanının önemli konularından birisidir. Regresyon algoritmaları matematik, ekonomi, araştırma, finans vb. gibi birçok bilim alanında kullanılmaktadır. Regresyon analiziyle bir olay araştırılırken, hangi değişkenlerin etkisinde olduğunun incelenmesine öncelik verilir. Bu değişkenler bir veya sayıca daha fazla olabilecekleri gibi doğrudan veya dolaylı olarak etkilemiş olabilirler.

Regresyon analizinde, gözlemlenen değerler ve etkilenilen durumların matematiksel denklemlerle veya fonksiyonlarla ifade edilmesi gerekir. Bu modellemeye regresyon modeli denir.

Genellikle bilim insanları iki veya ikiden fazla değişkenin arasında olan ilişkileri inceleyip bulmak ve ilişkiyi denklemle nasıl ifade edebileceklerinin yolunu göstermek isterler. Örneğin, bir ziraat mühendisi toprak yapısı ve verim, bir eczacı hastalık ve ilaç etken maddesi, bir muhasebeci gider ve ihtiyaçların birbiriyle ilişkisini incelemeyi amaçlayabilir. Regresyon, iki veya ikiden fazla değişkenin ilişkilerini incelemekle kalmaz, değişkenlerden birinin değeri bilindiği takdirde diğerini tahmin etmemize olanak sağlar.

İki veya daha çok değişken arasındaki ilişkileri gözlemlemekte regresyon analizlerine sık sık başvurulmaktadır. Regresyon modellerinde, bağımlı değişken

incelenirken hangi faktörlerin etkisi altında kaldığını ve etkileyen faktörlerden hangisinin etkisinin daha çok olduğunun saptanması amaçlanmaktadır.

Regresyon modellerinde başlıca amaçlanan bağımlı ve bağımsız değişkenler arasındaki ilişkilerin gösterilmesidir. Regresyon modelleri, analiz edilecek değişken miktarına, matematiksel yapısına göre farklı sınıflandırılmaktadırlar. Bir bağımlı ve bağımsız değişkenlerden meydana gelen basit doğrusal, basit regresyonlar, polinomsal, üssel, doğrusal olmayan (non-linear), geometrik regresyonlar olarak sınıflandırılabilir. Çok değişkenli regresyonlar, iki veya daha fazla bağımsız değişken ve bir bağımlı değişkenden meydana gelir. Çok değişkenli regresyonda modele giren bazı bağımsız değişkenler önemsiz olabilir. Modeldeki bağımlı değişkeni yeterli şekilde izah edecek bağımsız değişkenlerin saptanarak, önemsiz olan değişkenlerin çıkarılması gerekir.

MARS (Multivariate Adaptive Regression Splines) modelini ilk kez, istatistikçi ve fizikçi J. Friedman 1991'de ele alıp üzerinde çalışmıştır. Gündelik olarak pek çok alanda değerlendirilen doğrusal olmayan ve doğrusal regresyon modellerinden farklı olarak, farklı aralıklardaki bağımsız değişken değerlerine değişik katsayıların bulunmasını sağlar. Modele interaksiyon terimleri de ekleyerek gerçek yapıya çok daha yakın modeller geliştirir. Buna ek olarak da bağımlı ve bağımsız değişkenlerinin dağılımıyla ilgili rastgele bir tahmine gerek olmaz ve değişkenlerin çeşitleri de önem arz etmez (Friedman, 1991).

MARS modeli, regresyon yöntemiyle ortaya konulan hata varyansının oranıyla ilgili bilgiler vermektedir. Regresyon yöntemi, değişkenlerin arasındaki ilişkiyi ifade etmek için kullanılan algoritmalara nazaran, doğrusal, doğrusal olmayan veya karmaşık yapıda olabilmektedir. Üzerinde çalışılan modellerde gerçeğe yakın doğru tahminlere ulaşılmak için özellikle son zamanlarda çok fazla bağımsız değişken birlikte incelenmektedir. Regresyon modellerinin doğru sonucu daha iyi tahmin edebilmesi için son yıllarda çok fazla bağımsız değişken birlikte incelenmektedir. Bilgisayar teknolojisinin gelişmesiyle çok fazla bağımsız değişkenin birlikte

inceleyebilen komplike algoritmaların kullanımı artmıştır. Bu algoritmaları değerlendiren regresyon modellerinden biri de MARS modelidir.

MARS modellemesi artık veri madenciliği alanında da popüler olarak kullanılmaktadır. Çünkü MARS modeli neden ve sonuç değişkenlerinin aralarında bir hipotez gerektirmeksizin rastgele bir matematiksel ilişkiye gerek duymaz. Buna karşılık neden ve sonuç değişkenlerinin aralarında güçlü bir bağlantı geliştirir. Parametrik modellerle elde edilmesi epey zorlu doğru tahminleri, etkin yapısıyla kolay yorumlanabilen, kestirim derecesi yüksek modeller kurmayı sağlar.

Regresyon modeli, sürekli bir değişkenin tahmin edilmesi amaçlanan ve yapay sinir ağları kapsamındaki veri madenciliği algoritmaları sınıflayıcı (nominal), sıralı (ordinal) ve sürekli (continuous) bağımsız değişkenler için kolaylıkla kullanılmaktadır. MARS algoritması, bağımlı ve bağımsız değişkenlerin dağılımlarıyla ilgili herhangi bir varsayımının olmaması ve bağımlı değişkenlerle bağımsız değişkenler arasındaki ilişkiyi inceleyerek tahmin denklemlerinin açığa konulmasını sağlamaktadır.

MARS modeli, bilimsel alanlarda yeterince ilgi görmemekte, bilinmemektedir. Bu açığın bir nebze giderilmesi amaçlanan bu tez çalışmasında, hayvancılık alanından alınarak kullanılan bir veri setinin analizinde MARS algoritmasının R yazılım programı ile ayrı ayrı nasıl analiz edileceği, MARS algoritması analizinde dikkat edilmesi gereken hususlara değinerek, elde edilen sonuçların yorumlanmasından bahsedilecektir. Bu çalışmada uygulanan MARS analiziyle elde edilen tahmin denklemleri yardımıyla bağımlı değişken (OSV: ortalama süt verimi) ile bağımsız değişkenlerin birbiriyle oluşturduğu etkileşim hakkında bilgi sahibi olunacaktır.

2. ÖNCEKİ ÇALIŞMALAR

Friedman (1991), MARS modellemesini incelemiştir. MARS modellemesini en ince ayrıntısına kadar ele aldığı çalışmasında, modellemenin iki aşamadan oluştuğu belirtilmiştir. Aşamalar ayrı ayrı açıklanıp örneklerle etkinlikleri ortaya konulmuştur.

Çamdeviren ve ark. (2010), çalışmalarında bir regresyon modeli olan MARS modelinin özelliklerini tanımlayarak diğer regresyon modellerine göre üstün olduğu yönlere değinmiş ve uygulama adımları üzerinde çalışmışlar. MARS modelinin avantajı, bağımlı ve bağımsız değişkenlerin dağılımlarına ilişkin herhangi bir tahmin yoktur. Birden fazla bağımlı ve bağımsız değişkenlerin aynı anda analiz edilmesini sağladığını göstermişlerdir.

Yerlikaya (2008), yaptığı çalışmada MARS modelinin ikinci aşamasını gerçekleştiren geriye dönük eleme yöntemindense penaltı yöntemini önermiş, bundan dolayı da MARS cezalandırılmış hata kareler toplamı kullanarak Tikhonov düzenlemesi problemini oluşturmuşlar. Problemi incelerken, bir tamamlayıcı olarak sürekli optimizasyon ve geriye doğru eleme yöntemine alternatif tekniklerini kullanarak endüstride kalite kontrolü amaçlı MARS uygulamalarının etkisine değinmişlerdir.

Tunay (2001), MARS modelini kullanarak 1987-2000 yılları arasındaki Türkiye'de değişik parasal büyüklüklerine nazaran paranın gelir dolaşım hızını etkileyen nedenleri tahmin etmeye çalışmıştır. MARS modelinin bilindik yöntemlere göre zor olan, verilerin birbiriyle etkileşimlerini, optimal veri dönüşümlerini ve çok boyutlu verilerin içindeki karışık veri yapısını belirleme üstünlüğüyle regresyon algoritmalarında kolaylıklar oluşturmaktadır. Küçük veri setleriyle araştırma yapıldığı zamanlarda dahi iyi bir regresyon modeli kurabilmek için çok fazla zaman ve emek harcamak gerekmektedir. Ayrıca, fazlasıyla karışık olan veri setlerinin ve büyük veri tabanlarının regresyon yöntemi MARS modeliyle kolayca gerçekleştirilmektedir.

Grzesiak ve ark. (2003), yaptıkları çalışmada kısmi laktasyonun kayıtlarıyla 305 güne ait süt verimini tahmin etmeye çalışmışlardır. Araştırmacılar yapay sinir ağları ve çoklu regresyon yöntemlerini kullanarak süt verimini tahmin etmişlerdir. Modellerin karşılaştırırken, hata kareler ortalamasının karekökü, standart sapma oranı, tahmine ilişkin göreceli ortalama, hata ve düzeltilmiş belirleme katsayısı gibi model uyumlu ölçütleri kullanmışlar. Çoklu regresyon modeline göre sinir ağları modelinin süt verimi gibi nicel özelliklerin tahmin edilmesinde iyi bir alternatif yöntem olabileceği sonucuna ulaşmışlardır.

Tunay (2010), Türkiye'de muhtemel bankacılık krizlerini tahmin etmek için uygulanabilecek bir yöntem geliştirmek için doğrusal ve nonparametrik bir tahminleme yöntemi olan MARS modelini kullanmıştır.

Acciani ve ark. (2011), gayrimenkul ve değerlendirme belirleyicileri arasındaki karmaşık desenleri ve etkileşimini ortaya koymak için çalışmalarında MARS modellerinden yararlanmışlardır.

Aytekin ve ark. (2018), farklı ırklardan 103 genç boğanın vücut ölçüleri verileriyle besi süresinden yararlanarak besi sonu canlı ağırlığı tahmin etmek amacıyla MARS modelini kullanmışlardır. Araştırmacılar yaptıkları çalışma sonucunda MARS modelinin besi sonu canlı ağırlığın etkenlerinin saptanmasında, hayvanlara ait genotip standartların tanımlanmasında ve besi süresinin belirlenmesinde faydalı olabileceği sonucuna varmışlardır.

Ertürk ve ark. (2018), Doğu Anadolu bölgesindeki melez sığırların besi performanslarını etkileyen unsurları saptamak için etkileşim terimleri kapsamlı MARS modeli kullanmışlar. Araştırmacılar, çalışmaları sonucunda MARS modellemesinin biyolojik, ekonomik ve sosyal faktörlerin besi performanslarını etkileyen faktörler olduklarını beyan etmişlerdir.

Orhan ve ark. (2016), Çalışmada sağım süresi, kontrol günü, iletkenlik ve hareketlilik bağımsız değişkenleri kullanılan siyah alaca ineklerin, MARS modeliyle

süt verimi tahmin edilmiştir. Bu çalışmanın sonucunda laktasyonlar için kurulan MARS modellerin tahminlerinin yüksek olduğunu ortaya koymuşlardır.

Temel ve ark. (2010), MARS modelinin özellikleri, uygulamanın aşamalarını ve diğer regresyon modellerine göre avantajlarını incelemişlerdir. Behçetli ve kanser hastaların aile bireylerinde saptanmış davranış ve psikolojik testlerin sonuçlarından alınan verilerden, aile bireylerinin depresyon durumunu tahmin etmek amaçlamışlardır. Tahmin için kurdukları MARS modelinin başarılı tahminler yaptığı sonucunu çıkarmışlardır. Değişkenlerin arasındaki doğrusal olmayan ilişkilerinin gözden geçirilip yorumlayabildiğini belirtmişlerdir.

Nacar ve ark. (2017), Doğu Karadeniz Havzası'ndaki Haldizen Deresi günlük akım verilerini tahmin etmek için ÇDURE (Çok Değişkenli Uyarlanabilir Regresyon Eğrileri) modelinin uygulanabilirliği araştırılıp, modelin doğruluğu KRA ile kıyaslanarak değerlendirilmiştir. Model sonuçları incelendiğinde bir gün sonrasının akım değerlerini tahmin etmekte en makul sonuçlara ÇDURE'den ulaşılacağı saptanmıştır.

Kan ve ark. (2010), yaptıkları çalışmada dört farklı yöntemin (kesirli faktöriyel deneyleri, faktöriyel deneyleri, regresyon ağaçlarını ve MARS tekniği) F-4 uçaklarının yakıt tüketimleri üzerindeki önemli etkilerini ve farklı metotlardan bulunan sonuçları kıyaslamışlardır. Kullanılan dört farklı metoda göre; ana faktörlerden D (akım girişi) ve E (hava basıncı)'nın istatistiki olarak önemli olduğu sonucuna varmışlardır.

Akyol (2011), çalışmasında yaşam çözümlemesinde kullanılan yöntemlere alternatif olup uygulanılabilecek yeni bir modelin geliştirilmesi ve yaşam çözümlemesinde uygulanabilirliğinin ispatlanması amacıyla araştırmalar yapılmıştır. MARS modeli kullanılarak yaşam çözümlemesinde kullanılan yöntemlerin eksiklerinin giderilmesinde avantajlı bir alternatif olduğu saptanmıştır.

Geridönmez ve ark. (2010), yaptıkları araştırma da atış zarfı problemi ve parametrik olmayan regresyon yöntemleri kullanılarak bu probleme nasıl

yaklaşılabileceğini ölçümlemişlerdir. MARS modelinin atış zarfı hesaplamasında kullanımının en önemli getirisinin fonksiyonları görece kısa kod parçaları halinde gömülü yazılıma aktarabilme ve daha sonrasında veri tabanı değişse dahi mühimmat kodunda değişikliğe gerek kalmadan güncelleme yapılabilmesi olduğu sonucuna varmışlardır.

3. MATERYAL ve YÖNTEM

3.1. Materyal

MARS modeli algoritmasını uygulamak için alınan veri seti, yerel bir çiftliğin büyükbaş hayvan kayıtlarından temin edilmiştir. Ortalama süt verimlerini tahmin etmek amacıyla Holstein ve Montofon ırklardan, toplamda 25 hayvanın Haziran-Eylül ayları aralığındaki süt verimleri, yaşları, laktasyon süreleri ve canlı ağırlık verileri alınmıştır.

3.2. Yöntem

MARS modeli ilk kez 1991'de Stanford'da istatistikçi ve fizikçi olan Jerome Friedman tarafından geliştirilmiştir. Regresyon yöntemleri, özellikle doğrusal yöntemler, uygulamalı alanlarda büyük bir öneme sahiptir. Ayrıca genellikle gerçek hayatta rastlanılan etkenlerin doğrusal olmayan bir eğilim göstermesinden dolayı, doğrusal modellerin bu etkenleri başarılı temsil etmesi mümkün olmamaktadır. Bu etkenleri karakterize etmek için nonparametrik regresyona başvurulur. Eğer kurulan modelde kullanılan bağımsız değişkelerin sayısı fazlaysa, nonparametrik regresyon türevleri yeterince kullanışlı olamamakla birlikte basitçe yorumlamak da mümkün olmamaktadır.

Fakat, Friedman'ın geliştirdiği çok değişkenli uyarlanabilir regresyon eğrileri (MARS), nonparametrik regresyon modelinden türemiştir. Yukarıda belirtilen dezavantajlara, modelde bağımsız değişkenlerin fazla sayıda olduğu analizlerde sahip değildir. Doğrusal olmayan çok değişkenli yöntemlerin uygulanmasında epey kullanışlı olmaktadır. Model uygulanırken bağımsız değişkenler ve bağımlı değişken arasındaki algoritmik ilişkinin şekline dair varsayımlarda bulunulmaz. Model kullanılan verilere göre taban fonksiyonları ve bu fonksiyonlar bağlantılı katsayılardan meydan getirilir. Bu model bağımsız değişkenlerin değerlerini alanlara bölüp, her alanı bir regresyon eşitliğiyle bağıntılar.

MARS modeli ileriye ve geriye doğru olarak iki adımdan oluşmaktadır (Friedman, 1991):

1. Adım: Modelde olabildiğince bütün temel fonksiyonlar (TF) kurulur. Büyük bir fonksiyon buluncaya yani modelin karışıklığı en üst seviyeyi görene kadar fonksiyonlarla model geliştirilmeye devam edilir. Temel fonksiyonlar, bir değişkenden veya çok değişkenli, birbiriyle etkileşimli terimlerden meydana gelir. Modelimizdeki temel fonksiyonlar, bağımsız değişkenlerin birbiriyle olan etkileşimleri ve doğrusal olmayan dönüşümleri tahminlerini göz önünde bulundururlar. Temel fonksiyonlar meydana gelirken, modeldeki tüm değişkenlerin birleşimlerinin tek tek incelenmesi sağlanır. Böylelikle, bağımsız değişkenlerin tek tek hepsi ve bu değişkenlerin karşılıklı etkileşimleri tanımlanan aralıklardaki tahmini değerlerin tümünün birer düğüm olduğu varsayılarak, mümkün olan tahmini bütün temel fonksiyonlar belirlenir. Temel fonksiyonlarda modele dahil edilerek daha esnek ve daha karışık bir yapı elde edilir. Temel fonksiyonların sayısı en üst seviyelere gelene kadar bu işleme devam edilir. Ayrıca, aynı değişkenlere ait Ayna Görüntü Fonksiyonu da belirlenip, temel fonksiyonlar belirlenirken modele dahil edilebilir. Bağımsız değişkenlerle bağımlı değişken arasında olan dağılım düğüm noktasındaki eğilimin değişimlerini ve eğilimin düğüm noktalarına kadarının sıfır olduğunu ayna görüntü fonksiyonları göstermektedir. (Eşitlik 3.1).

Mars modeli aşağıdaki formülle tanımlanabilir (Eşitlik 3.1)

$$Y = \beta_0 + \sum_{k=1}^{K} \alpha_k \beta_k \, (X) \qquad (3.1)$$

Eşitlikte;
k: Düğüm sayısı
K: Temel fonksiyon sayısı
X: Bağımsız değişkeni
β_0: Modeldeki sabit terim
α_k: k. temel fonksiyon katsayısı

$\beta_k(X)$: bağımsız değişkeni için k. temel fonksiyonu belirtmektedir.

Aşağıdaki şekilde ise temel fonksiyonun tanımlanması yapılmaktadır. (Eşitlik 3.2):

$$\beta_k(X) = \prod_{j=1}^{J_k}\left[S_{k,j}\left(X_{w(k,j)} - t_{(k,j)}\right)\right] \quad (3.2)$$

Burada;

J_k: İnteraksiyon derecesini

$S_{k,j}$: ϵ $[\pm 1]$

$t_{k,j}$: Düğüm değerini

$\chi_{w(k,j)}$: Bağımsız değişken değeridir.

Bağımsız değişkenlerin, doğrusal olmayan dönüşümlerinin temel fonksiyon olması mümkündür. Fakat temel fonksiyonların doğrusal dönüşümleri bağımlı değişkendir.

2. Adım: Temel fonksiyonların tahmin hatasının optimal dengesi bulunana kadar modelde tutulmayıp, çıkarılmaktadır. Bunun gerçekleştiği basamağa budama (pruning) adımı denilir. Bu adımın gerçekleşmesi için MARS modeli geriye doğru basamağının algoritmasını çalıştırır. Geriye doğru basamak algoritmasının uygulanabilmesi için aşağıda verilen adımların gerçekleştirilmesi gerekir. Modele girerken temel fonksiyon sabitinin olduğu en küçük model ile başlanmaktadır.

- En uygun ölçümlerle tahmin hatasını en düşük seviyeye getiren, değişkenler ve tüm düğümlerin temel fonksiyon uzayı araştırılmaktadır.
- En yüksek karmaşıklığa sahip modele ulaşana dek değişkenlere ve bütün düğümlere bakma işlemi tekrarlanmaktadır.
- Son adımda da budama prosedürüne başvurulmaktadır.

Elde edilen en büyük seviyedeki model budanır. Yani, önemli olan bağımsız değişkenlerle, karşılıklı olan interaksiyonları bu bağımsız değişkenler için belirlenir. Daha sonra en uygun model için hata kareler toplamının en düşük olacağı seviye saptanır. En bilindik yöntemle budama modeli, genelleştirilmiş çapraz geçerlilik

tekniği kullanılarak yapılmaktadır. GCV tekniği hem model karmaşıklığını hem de artıkların hatasını göz önünde bulundurur. Eşitlik 3.3. ve Eşitlik 3.4. yardımlarıyla hesaplanmaktadır.

$$GCV(M) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{f}_M(x_i)^2)}{(1-\frac{C(M)}{n})^2} \quad (3.3.)$$

$$C = 1 + cd \quad (3.4.)$$

Burada,

N: Veri setindeki toplam hayvan sayısıdır

d: Bağımsız temel fonksiyonların sayısını gösteren etkili serbestlik derecesidir

C: Eklenen temel fonksiyonların maliyet-karmaşıklık (cost-complexity) ölçüsüdür

Hesaplamalardan çıkarılan sonuca göre en iyi C değeri 2<d<3 bulunmuştur (Friedman, 1991).

4. ARAŞTIRMA BULGULARI ve TARTIŞMA

Şanlıurfa ilindeki yerel bir çiftlikten 25 hayvanın ırk, yaş, laktasyon süresi, canlı ağırlık ve Haziran-Eylül ayları arasındaki süt verimi verileri alınarak, çok değişkenli uyarlanabilir regresyon eğrileri (Multivariate Adaptive Regression Splines – MARS) modeli algoritmasının kullanılmasıyla ortalama süt verimi (OSV) analiz edilmiştir. Çalışma, hayvancılık alanından alınarak kullanılan veri setinin analizinde MARS algoritmasının R programı "earth" paketi (Hastie ve Stephen Milborrow, 2019) kullanılarak ayrı ayrı analiz edilmiştir.

Çok değişkenli uyarlanabilir regresyon eğrileri (MARS modeli) R programı "earth" paketi (Hastie ve Stephen Milborrow ile OSV analiz edilmiştir. OSV için etkileşimi maksimum olan modeller denenmiş, böylelikle de en uygun modelin tespiti yapılmıştır. Modelin uygun olup olmadığı, belirleme katsayısının (R2) en yüksek ve genelleştirilmiş çapraz geçerlilik katsayısının (GCV) en düşük seviyede olduğu kriterlere göre değerlendirilip saptanmıştır. OSV için en uygun maksimum etkileşim seviyesi Çizelge 4.1.'de verilmiştir. Budamadan evvel en büyük terim sayısı (nk) R programında 18 olarak belirlenmiş, fakat bunun 11 tanesi MARS tahmin algoritmasına dahil edilmiştir. Genelleştirilmiş çapraz geçerlilik (GCV) 1.578986 değeri olarak elde edilmiştir. Son çapraz geçerlilik değeri 0.998 gibi yüksek bir değer tahmin edilmiştir. MARS denkleminde genelleştirilmiş çapraz geçerlilik değeri için en iyi model en düşük değeri olandır. Analizde R^2(CVRsq) ve GRsq değerleri yaklaşık 0.998 gibi bir yüksek değer elde edilmiştir.
Standart sapma oranı 0.038 çıkarak 0.10'dan düşük olmasından dolayı modelin çok iyi bir uyumda olduğu görülmüştür.

Orhan ve ark. (2018), 80 ineğin 2006-2011 laktasyon kayıtları kullanılarak yaptıkları çalışmada sağım süresi, kontrol günü, hareketlilik ve iletkenlik bağımsız değişkenleriyle MARS modellenmiştir. Bu çalışmada elde edilen sonuçlar Orhan ve ark. (2017)'nın sonuçlarından daha iyi bulunmuştur.

Kaynak araştırmalarında MARS modelinin daha çok sağlık ve sosyal bilimlerde kullanıldığı görülmektedir. Kooperberg ve ark. (1995), Orhan ve ark. (2017), Muzır (2011), Mallick ve ark. (1999), Temel ve ark. (2010) ve Öztürk ve Sevinç (2013) çalışmaları bunun için örnek olabilir. Zootekni alanında MARS modeli ile ilgili çeşitli çalışmalar yapılsa da ortalama süt verimi modelinde MARS kullanılan çalışmalara rastlanmamıştır. Uyum iyiliği ölçütleri bakımından MARS algoritması ile kurulan modellerin tahmin performansı, çoklu doğrusal modelden çok daha iyi bulunmuştur.

MARS modelinde, bağımlı değişkene (ortalama süt verimi) dair gözlemlenen ve elde edilen değerler arasındaki farkın azalması temel amaçlardandır. Çizelge 4.1.'de maksimum etkileşim derecesi arttığında, GCV değerinin azaldığı ve 2'li etkileşimde sabitlendiği görülmektedir. OSV için en uygun olan modelin; 3 temel fonksiyonla 2'li etkileşimde, R^2 değeri 0.999, düzeltilmiş R^2 değeri 0.996 ve GCV değeri 1.578'dir (Çizelge 4.1. ve Çizelge 4.3.). Bağımsız değişkenlerle nasıl bir sonuç belirleneceğini sınayan sonuçlar Çizelge 4.2.'de verilmiştir.

Çizelge 4.1.Pearson korelasyon sonuçları

GCV	1.578986
RSS	39.47465
GRSq	0.998553
RSq	0.998553
CVRSq	0.2494669

Çizelge 4.2. Çapraz geçerlilik

nterms	sd	nvars	sd
11.6	3.78	4	0
CVRSq	sd	MaxErr	sd
0.249	0.915	-153	67.1

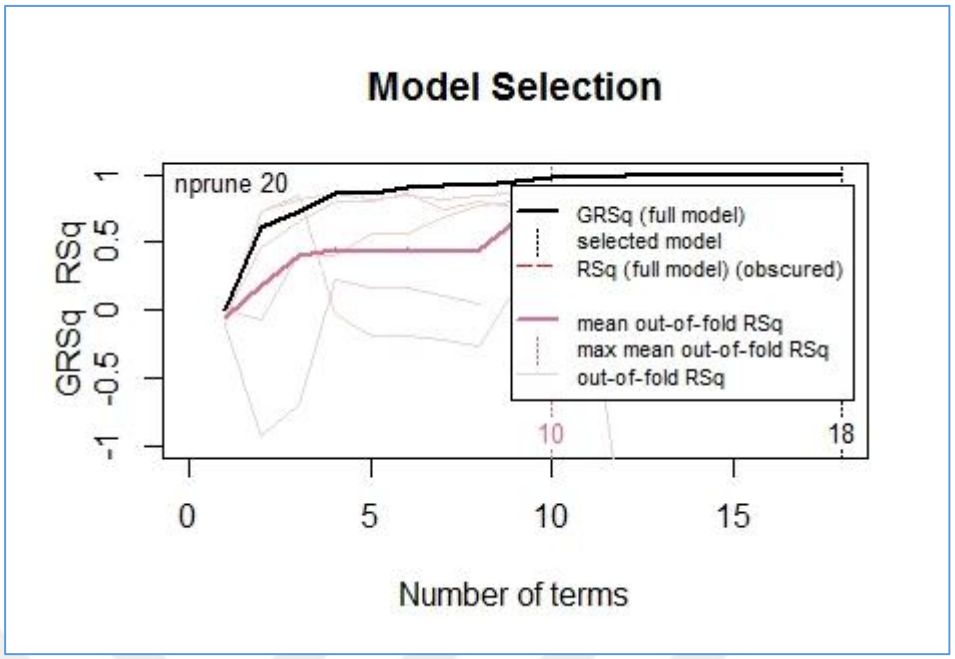
Çizelge 4.3. Pearson korelasyonu analiz sonuçları

t-değeri	125.98
df	23
p-değeri	<0.05
r	0.999
Standard sapma oranı:	0.038
Varyasyon katsayısı:	0.22
RMSE:	1.257
MSE:	1.579
RAE:	0.002
MAPE:	0.158
MAD:	0.921
Rkare:	0.999
Düzeltilmiş R kare:	0.996

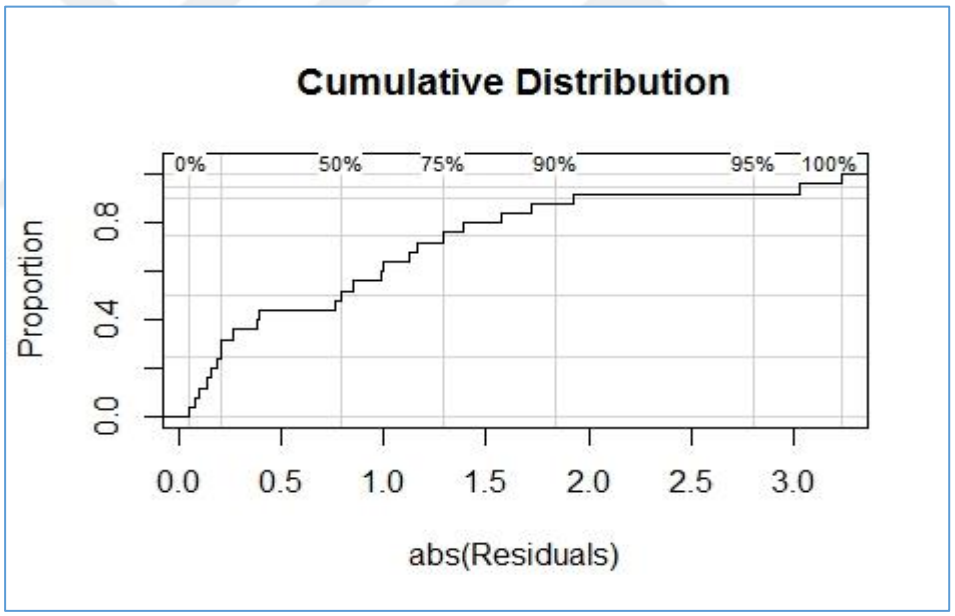
Çizelge 4.4. Mars tahmin eşitliğine ait terimler ve katsayılar

Terimler	Katsayılar
(Intercept)	593.5648
h(43-Yas)	-903.405
h(Yas-43)	424.3956
h(Yas-45)	-347.802
h(300-Laktasyon)	26.9929
h(Laktasyon-300)	8.81711
h(Canli_Agirlik-537)	-5.41285
h(546-Canli_Agirlik)	-0.28235
h(Canli_Agirlik-546)	-1.61678
Irk*h(43-Yas)	-0.79489
Irk*h(300-Laktasyon)	-29.3884
Irk*h(Laktasyon-300)	-2.32603
Irk*h(Canli_Agirlik-546)	28.33082
h(Yas-43)*Laktasyon	-1.40218
h(Yas-45)*Laktasyon	1.15897
h(43-Yas)*Canli_Agirlik	1.64091
Irk*Yas*h(Canli_Agirlik-546)	0.83025
Irk*Laktasyon*h(Canli_Agirlik-546)	-0.19824

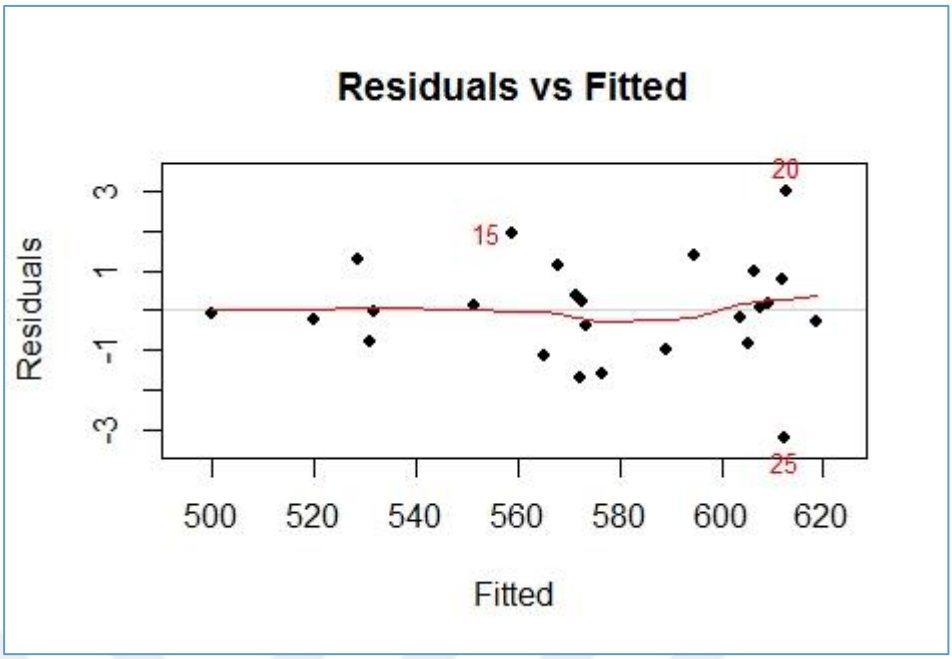
MARS model nesneleri için çizim yöntemi, yararlı performans ve artık grafikler sağlar. Şekil 4.1., GCV'nin grafiğini çizen model seçim grafiğini göstermektedir R^2 (sol el y- eksen ve düz siyah çizgi) modelde tutulan terimlerin sayısına göre (x-axis) belirli sayıda orijinal tahminden (sağ el y eksen).10 ile işaretlenmiş dikey kesikli çizgi, bize GCV' deki marjinal artışlarda tutulan en uygun terim sayısını gösterir. R^2 0.001'den azdır. Ampirik CD, X'e eşit veya daha küçük değerlerin oranıdır. Gözlenen bir değere eşit her X değerinde 1 / N'lik bir dikey sıçramaya sahip artan bir adım fonksiyonudur. CD grafikleri, farklı veri kümelerinin dağılımını karşılaştırmak için kullanışlıdır (Şekil 4.2.). Artıklar ve Uyum grafiğine bakıldığında, kırmızı çizginin (bu sadece bir dağılım grafiğidir, takılan değerin her bir değerinde artıkların ortalama değerini gösteren) tamamen düz olduğunu görürüz. Bu bize kalıntılara yönelik fark edilebilir doğrusal olmayan bir eğilim olmadığını gösterir. Ayrıca artıklar, tüm uyan değerler aralığı boyunca eşit derecede değişken görünmektedir (Şekil 4.3.). QQ grafiği veya nicelden nicel grafiğe, ampirik dağılım ile verilen teorik dağılım arasındaki uyumu test eden bir grafiktir. Bir regresyon modelinin hatalarının normalliğini doğrulamak için kullanılan yöntemlerden biri, artıkların bir QQ grafiğini oluşturmaktır. Noktalar çizgi üzerinde hizalıysa x = y (Şekil 4.4.). Varsayılan olarak en büyük (mutlak değer olarak) artıklara sahip üç değeri tanımlar. Burada bunlar 15,20 ve 25 noktalarıdır, bunların her biri pozitif bir kalıntıya sahiptir yani, modele göre eksik tahminleri temsil ederler (Şekil 4.5.). Bağımlı değişken bakımından gerçek ve tahmin edilen değerler arasındaki uyumluluk değerlendirildiğinde, en iyi uyumun MARS algoritması ile sağlandığı belirlenmiştir (Şekil 4.6.). Ortalama süt verimini kestirmek için en yüksek interaksiyon derecesi 2 değerine göre MARS ile oluşturulan tahmin denklemine ait katsayılar ve temel fonksiyonlar verilmiştir. Şekil 4.7.'yi incelediğimizde, canlı ağırlık ve laktasyon süresi bağımsız değişkenlerinin OSV üzerindeki etkisinin ırka göre değiştiği söylenebilir. Aynı zamanda yaş ve laktasyon süresi değişkenlerinin OSV üzerindeki etkisi birbirine bağlıdır. Yaş değişkeninin CA üzerindeki etkisi ırka bağlı olarak değişebilmektedir (Şekil 4.7.).



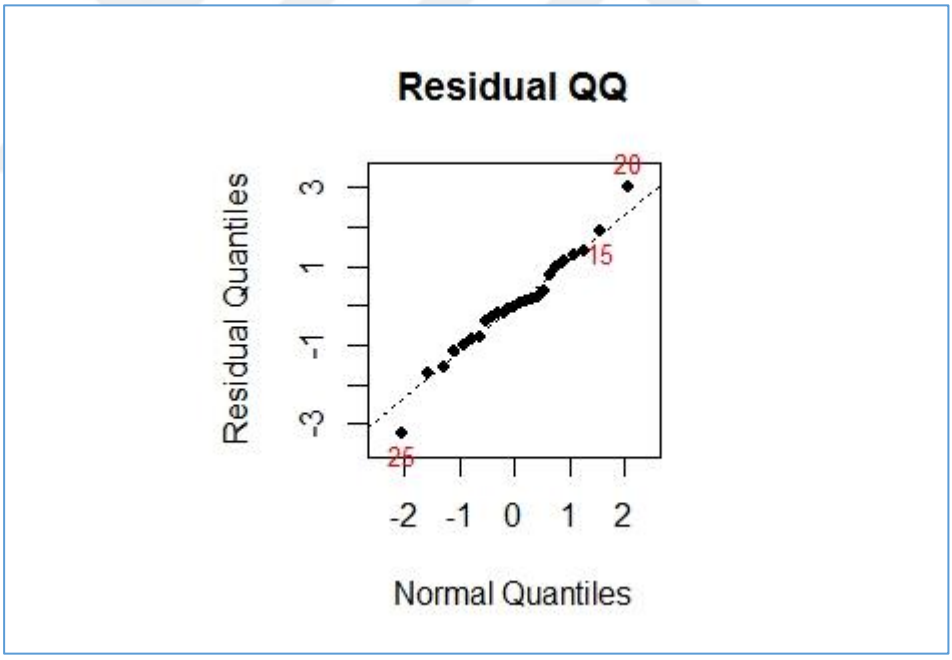
Şekil 4.1. Model seçim grafiği



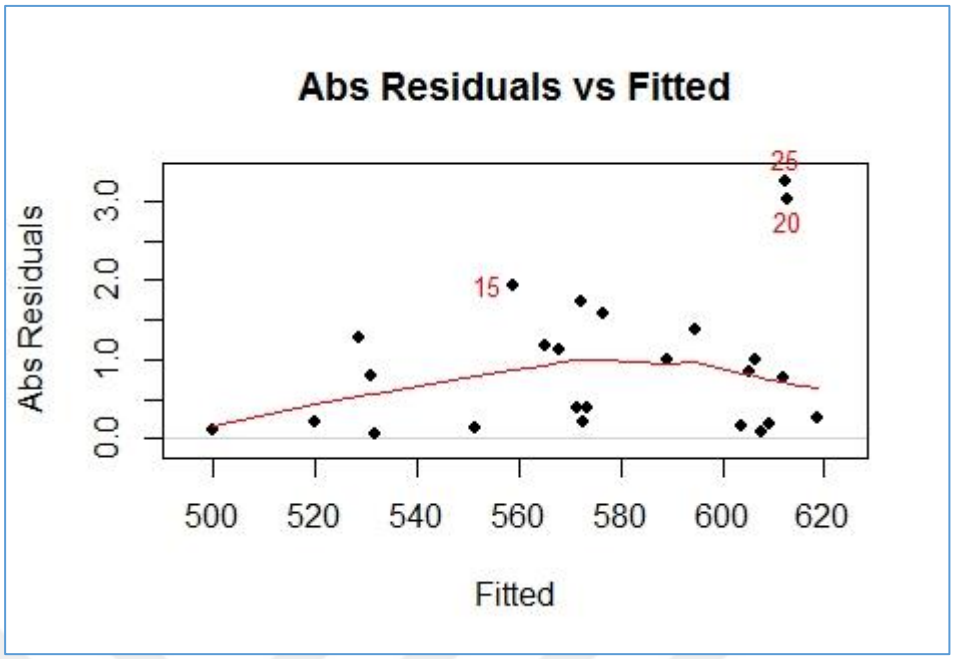
Şekil 4.2. Birikimli dağlım grafiği



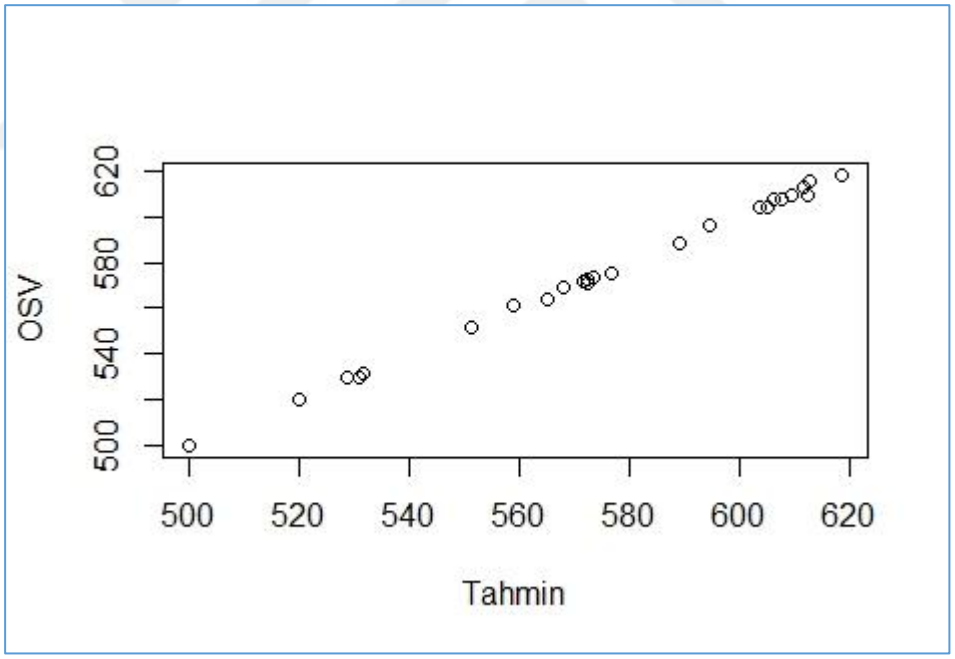
Şekil 4.3. Artıklar ve uyum grafiği



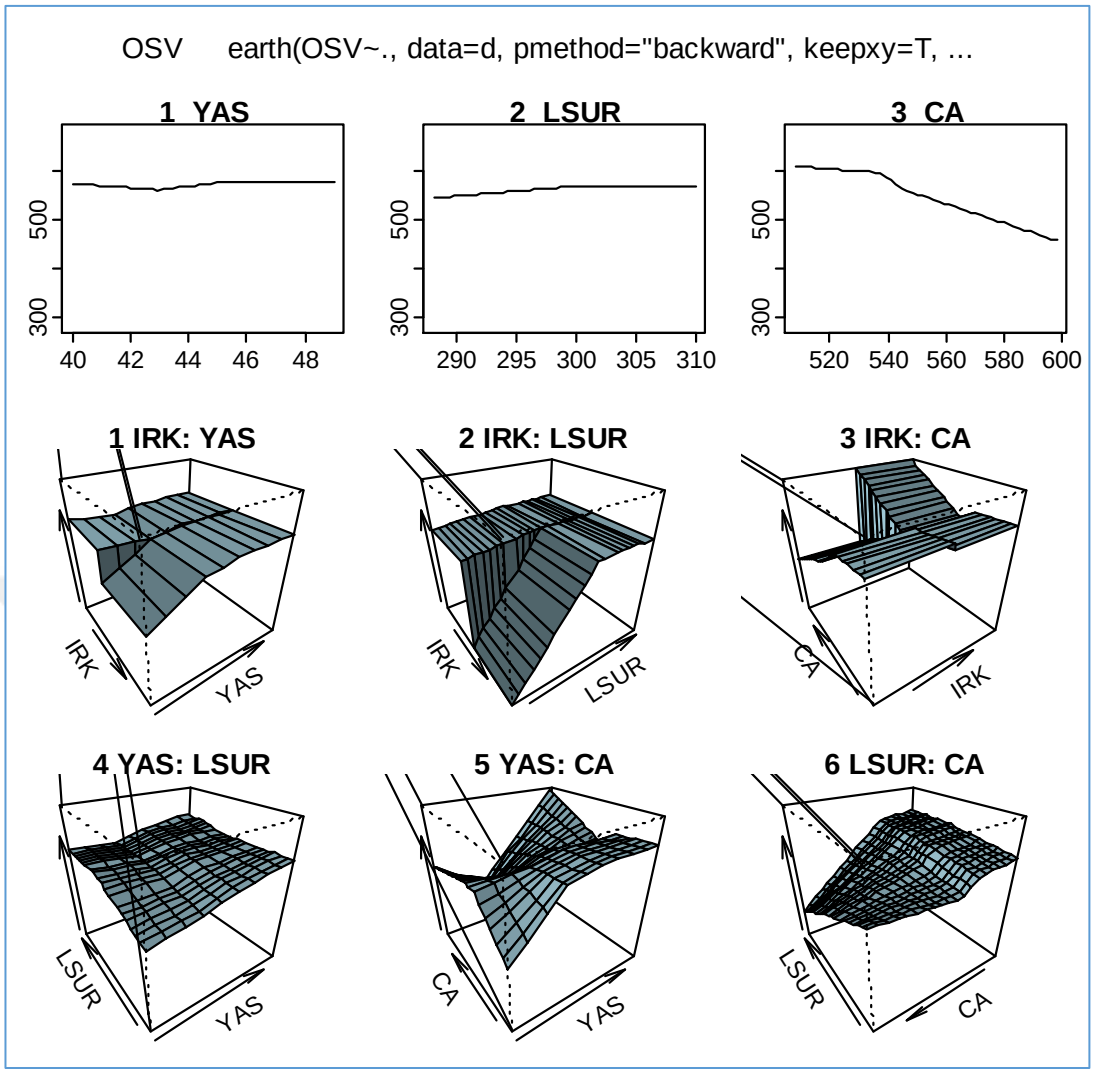
Şekil 4.4. QQ artıkları grafiği



Şekil 4.5. Mutlak değeri alınmış artıklar ve uyum grafiği



Şekil 4.6. Gerçek ve tahmini değerlere ait dağılım grafiği



Şekil 4.7. Bağımlı ve bağımsız değişkenler arasındaki ilişkileri gösteren grafik.

5. SONUÇLAR ve ÖNERİLER

Bu çalışmada, MARS modeli ile ilgili temel kavramlara değinilmiştir. MARS modelinin avantajlarının ortaya konulması amacıyla bir veri setiyle analiz yapılmıştır. Çalışmada bir çiftlikten 25 hayvanın ırk, yaş, laktasyon süresi, canlı ağırlık ve Haziran-Eylül ayları aralığındaki süt verimi verileri alınarak, çok değişkenli uyarlanabilir regresyon eğrileri (Multivariate Adaptive Regression Splines – MARS) modeli algoritmasının kullanılmasıyla ortalama süt verimi (OSV) analiz edilmiştir.

Çalışma, hayvancılık alanından alınarak kullanılan veri setinin analizinde R yazılım programı ile nasıl analiz edilmiştir. Modele giren tüm temel fonksiyonların %99 oranında önem arz ettiği görülmüştür. OSV tahmin etmek için kurulan MARS modellerinde en etkin değişkenin canlı ağırlık (CA) olduğu saptanmıştır. Elde edilen sonuçlara göre MARS modeli uygulanan yöntemlerin OSV tahmininde kullanımının güvenli olabileceği sonucuna ulaşılmıştır.

KAYNAKLAR

ACCİANİ C., FUCİLLİ V. and SARDARO R., 2011. Data mining in real estate appraisal:A model tree and multivariate adaptive regression spline approach. Aestimum 58:27-45s.

AKYOL M, 2011.Yaşam Çözümlemesine Yeni Bir Yaklaşım: MARS, Doktora Tezi, Türkiye Cumhuriyeti Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara

AYTEKİN İ., EYDURAN E., KARADAS K., AKŞAHAN R. ve KESKİN İ., 2018. Prediction of Fattening Final Live Weight from some Body Measurements and Fattening Period in Young Bulls of Crossbred and Exotic Breeds using MARS Data Mining Algorithm. Pakistan Journal of Zoology, 50(1), 189-195s.

ERTÜRK Y.E., 2018. Description of Factors Influencing Final Fattening Weight in Domestic Beef Cattle Breeds through MARS Algorithm. Pakistan Journal of Zoology, 50(5), 1731-1737s.

FRİEDMAN J.H., 1991. Multivariate Adaptive Regression Splines. Annals of Statistics, 19(1):1-67s.

GRZESİAK W., LACROİX R., WÓJCİK J. and BXLASZCZYK P., 2003. A comparison of neural network and multiple regression predictions for 305-day lactation yield using partial lactation records. Can. J. Anim. Sci., 83:307-311s.

GRZESİAK W., BŁASZCZYK P. and LACROİX R., 2006. Methods of predicting milk yield in dairy cowspredictive capabilities of wood's lactation curve and artificial neural networks (ANNs). Computers and Electronic in Agriculture, 54, 69-83s.

HASTİE T. and Stephen MİLBORROW. Derived from mda:mars by, and R. T. U. A. M. F. utilities with T. L. leaps wrapper. (2019). earth: Multivariate Adaptive Regression Splines. https://cran.r-project.org/package=earth

KAKİ B., YEŞİLOVA A. ve ŞEN C., 2004.Yarı Parametrik Regresyon Yönteminin Hayvancılıkta Kullanılması. 4. Ulusal Zootekni Bilim Kongresi Sözlü Bildiriler Programı,Van, 26-32s.

NACAR S., KANKAL M. ve HINIS M.A.,2018. Çok Değişkenli Uyarlanabilir Regresyon Eğrileri (ÇDURE) ile Günlük Akarsu Akımlarının Tahmini-Haldizen Deresi Örneği. Gümüşhane Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 8 (1): 38-47s.

OREKECİ TEMEL G., ÇAMDEVİREN H. and YAZICI A.C., 2005.Regresyon Modellerine Alternatif Bir Yaklaşım: MARS. VIII. Biyoistatistik Kongresi Sözlü Bildiriler. Bursa, 105-123s.

ORHAN H., EYDURAN E., TATLIYER A. ve SAYGICI H., 2016. Prediction of egg weight from egg quality characteristics via ridge regression and regression tree methods. Revista Brasileira de Zootecnia, 45(7), 380-385s.

TEMEL G.O., ANKARALI H. ve YAZICI A.C, 2010. Regresyon Modellerine Alternatif Bir Yaklaşım: MARS. Türkiye Klinikleri Journal of Biostatistics, 2(2): 58-66s.

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

TUNAY B.K., 2010. Bankacılık Krizleri ve Erken Uyarı Sistemleri: Türk Bankacılık Sektörü İçin Bir Model Önerisi. BDDK Bankacılık ve Finansal Piyasalar, 4(1):9-46s.

YERLİKAYA F., 2008. 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, 102s.

EKLER

Salford Predictive modeller 8.0 programının 30 günlük deneme sürümünde yapılan, modele giren bağımsız değişkenlerin grafikleri de ayrıca verilmiştir.

