Cilt: 10 Sayı: 02 Kategori: 02 Sayfa: 56-72 ISSN No: 1303-6319 Volume: 10 Number: 02 Category: 02 Page: 56-72

YENİ DOĞAN BEBEKLERİN DÜŞÜK DOĞUM AĞIRLIĞININ MARS YÖNTEMİNE DAYALI İKİLİ LOJİSTİK REGRESYONLA MODELLENMESİ

Soner ÖZTÜRK* **Volkan SEVİNÇ****

ÖZET

Düşük doğum ağırlıklı bebekler ilerleyen yıllarda sağlık açısından bazı sorunlarla karşılaşmaktadır. Bu yüzden bir bebeğin doğmadan önce düşük doğum ağırlıklı olup olmayacağının tahmini önemlidir. Bu tahminin elde edilebilmesi için ihtiyaç duyulan bir modelin geliştirilmesinde lojistik regresyon modeli uygun bir seçimdir. Lojistik regresyon analizi, bağımlı değişkenin kategorik olduğu durumlarda kullanılan ve kolay yorumlanabilen modelleme tekniklerinden birisidir. Bağımlı değişkenin iki düzeyli olduğu lojistik regresyon analizi ikili lojistik regresyon analizi olarak adlandırılır. Lojistik regresyon analizinin parametrik ve parametrik olmayan çözümleri bulunmaktadır. MARS yöntemi parametrik olmayan ve lojistik regresyon analizinde parametrik çözümlere alternatif olarak kullanılabilecek bir çözüm yöntemidir. Parametrik olmayan modeller, parametrik modellere göre daha az varsayım gerektirir ve daha esnektirler. Uygulama çalışmasında doğacak bebeklerin düşük doğum ağırlıklı olup olmayacağının tahmin edilmesini sağlayacak bir ikili lojistik regresyon modeli oluşturulmuştur. Model MARS yöntemine dayalı olarak tahmin edilmiştir. Analizde 982 bireye ait veri, MARS paket programı kullanılarak incelenmiştir. Sonuç kısmında elde edilen bulgular yorumlanmıştır.

Anahtar Kelimeler: Düşük doğum ağırlığı, Lojistik regresyon, En çok olabilirlik, MARS.

1. GİRİŞ

Eldeki örnek verilerden hareket ederek yorumlama, genelleme ve tahmin yapmak istatistiğin temel konularıdır. İstatistikte bu amaçlara yönelik yöntemlerden biri regresyon analizidir. Regresyon analizi, bir bağımlı değişken ile bir ya da birden fazla bağımsız değişken arasındaki ilişkiyi ifade etmekte kullanılan bir istatistiksel yöntemdir. Regresyon analizi ilişkinin türü, gücü ve yapısını araştırmaktadır. Regresyon analizinin en temel türü, doğrusal regresyon analizidir. Doğrusal regresyon analizinde bağımlı değişken kesikli veya sürekli değerler alabilir. Ancak bazı durumlarda bağımlı değişken kategorik değerler de alabilmektedir. Bağımlı değişkenin kategorik olduğu durumlarda kullanılan regresyon türü lojistik regresyondur. Lojistik regresyonda bağımsız değişkenler sayısal veya kategorik değerler alabilirler. Bağımlı değişkenin sadece iki değer aldığı lojistik regresyon analizi iki düzeyli, ikiden fazla değer alabildiği analiz, çoklu lojistik regresyon analizi olarak adlandırılır.

Lojistik regresyonun parametrik çözümünde kullanılan en yaygın yöntem en çok olabilirlik yöntemidir. Parametrik yöntemlere alternatif olarak parametrik olmayan yöntemler de bulunmaktadır. Parametrik olmayan yöntemler, veri sayısının ve değişken sayısının çok olduğu, kayıp verilerin olduğu durumlarda iyi sonuçlar vermektedir. Bu

*Muğla Sıtkı Koçman Üniversitesi, Fen Bilimleri Enstitüsü, Muğla, e-posta: soner985@hotmail.com
**Yrd. Doç. Dr., Muğla Sıtkı Koçman Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Muğla, e-posta: vsevinc@mu.edu.tr

yöntemlerden biri çok değişkenli uyarlayıcı regresyon uzanımları: MARS (Multivariate Adaptive Splines) yöntemidir.

Düşük doğum ağırlıklı olarak dünyaya gelen bebekler ilerleyen yıllarda sağlık açısından bazı sorunlarla karşılaşmaktadır. Bu nedenle, bir bebeğin doğum ağırlığının düşük değerde olup olmayacağının tahmini için uygun bir modele ihtiyaç vardır. Bu modelin oluşturulmasını içeren uygulama çalışmasında, bebeğin doğum ağırlığını etkileyebilecek faktörler, MARS yöntemine dayalı lojistik regresyon modeli oluşturmak için kullanılmıştır. Çalışmada Danimarka Ulusal Doğum Grubu'nun hamilelikte ateş ve buna bağlı ölü doğumlarla ilgili çalışmasına ait verilerin bir kısmı kullanılmıştır. Bebek doğum ağırlıkları bağımlı değişken olarak seçilmiştir. Çalışmada doğum yapacak kadınlardan elde edilen veriler kullanılarak, bebeğin düşük doğum ağırlıklı olup olmayacağına ilişkin model oluşturulmuştur. Oluşturulan model yardımıyla doğum öncesi bebek doğum ağırlığı için tahminler yapılabilmesi ve gerekli önlemlerin alınabilmesi amaçlanmıştır.

2. MARS YÖNTEMİ

Çok değişkenli uyarlayıcı regresyon uzanımları (MARS), Friedman tarafından 90'lı yılların başında geliştirilmiş, parametrik olmayan bir regresyon yöntemidir. MARS kelimesi açılımı aşağıdaki kavramların baş harflerinden oluşturulmuştur.

Multivariate (çok değişkenli): Çok boyutlu veriler üzerinde işlem yapılabilir, özellikle bağımsız değişken sayısı fazla olduğu durumlarda tercih sebebidir.

Adaptive (uyarlayıcı): Yöntemin basamakları final modeline ulaşana kadar eleme ve seçme aşamalarından oluşur.

Regression (regresyon): Bağımlı ve bağımsız değişkenler arasındaki fonksiyonel ilişkiyi ifade etmektedir.

Splines (uzanımlar): Regresyon eşitliği, düz bir regresyon doğrusu yerine, bükülmüş bir yapıya sahiptir.

MARS yöntemi bankacılık, sigortacılık, ekonominin yanı sıra, yaşam analizi, sosyal bilimler gibi birçok alanda kullanılmaktadır. Literatürde MARS yöntemi ile ilgili olarak, Kim (2000) gençlerin uyuşturucu kullanımı ile ilgili yaptığı çalışma sonucunda, bağımlı değişkenin kategorik olduğu durumlarda da MARS'ın iyi sonuçlar verdiğini göstermiştir. Kuhnert, Do vd. (2000) parametrik olmayan CART ve MARS yöntemlerini, parametrik lojistik regresyonla karşılaştırmıştır. Motor kazalarındaki yaralanma verilerine uygulanmış olan bu çalışma için MARS modelinin diğer ikisine göre daha iyi performans gösterdiği belirtilmiştir. Amerikan Çevre Koruma Kuruluşu (EPA)'ndan 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 yaklaşımlarını MARS ve klasik lojistik regresyonla tahmin etmeye çalışmıştır. Dieterle (2003) zamana bağlı analitik veriler üzerine hazırladığı doktora tezinde yapay sinir ağları, genetik algoritmalar, CART ve MARS'ı karşılaştırmıştır. Lee, Chiu vd. (2004) kredi skorlama ile ilgili çalışmalarında, diskriminant analizi,

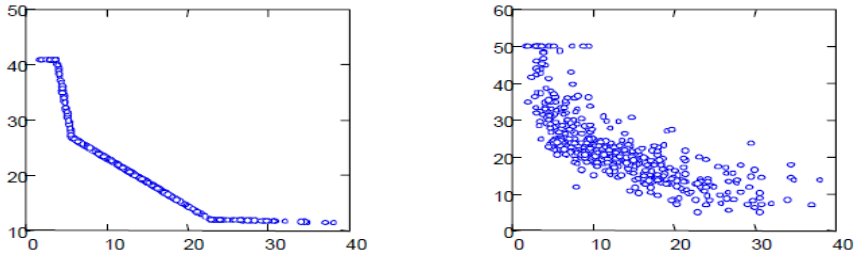
lojistik regresyon, CART ve MARS'ın doğru sınıflama oranlarını ve hatalarını karşılaştırmışlardır. Stokes ve Lattyak (2005) MARS yöntemini ekonometrik bazı sistem ve yazılımlar ile geliştirmiş ve kullanmışlardır. Kriner (2007), yaşam analizini MARS yöntemini kullanarak yapmıştır. Quiros, Felicisimo vd. (2009) MARS yöntemini, arazi örtüsünün uydu görüntülerinden yararlanarak sınıflandırılması için kullanmışlardır. Mina (2009), yoksulluk profilinin belirlenmesinde MARS yöntemini kullanmış ve bazı koşullar altında parametrik lojistik regresyondan daha etkili olduğunu belirtmiştir. Mina (2010) özürlü kişilerin iş seçimi ile ilgili yaptığı çalışmada, parametrik 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. Türkiye'de ise Tunay (2001), Türkiye'de paranın gelir dolaşım hızlarının MARS yöntemiyle tahmini çalışmasını gerçekleştirmiştir. Yerlikaya (2008) MARS üzerinde bir takım düzenlemeler yaparak, oluşturduğu yeni modeli veri madenciliği uygulamaları için kullanmıştır. Kan ve Yazıcı (2010) yakıt tüketimi için, faktöriyel deneyleri, regresyon ağaçları ve MARS yönteminin sonuçlarını karşılaştırmışlardır. Kayri (2010) internet bağımlılığı ölçeğini MARS yöntemini kullanarak analiz etmiştir. Tunay (2010) bankacılık krizlerini ve Türkiye'deki durgunlukları MARS yöntemi ile tahmin etmiştir. Topak (2011) Türkiye'de kurumsal başarısızlığı modellemek için MARS yöntemini kullanmıştır.

2.1 Mars Yöntemi ile Tahmin

Parametrik olmayan regresyon yöntemleri, Kernel tahmini, yerel polinom regresyonu veya düzleştirme uzanımları yöntemlerini kullanır. MARS yöntemi, bağımlı değişken ve bağımsız değişken kümesi arasındaki ilişkiyi düzleştirme uzanımlarını kullanarak belirleyen bir yöntemdir (Friedman, 1991).

MARS yöntemi, her bağımsız değişkenin bağımlı değişkenle olan ilişkisini incelemenin yanı sıra, bağımsız değişkenlerin birbirleri arasındaki etkileşimlerini de belirler ve bu etkileşimlerin bağımlı değişken üzerindeki etkisini ortaya koyar (Tunay, 2001). Bu nedenle gözlem sayısının ve bağımsız değişken sayısının çok olduğu durumlarda MARS yöntemi iyi sonuç vermektedir.

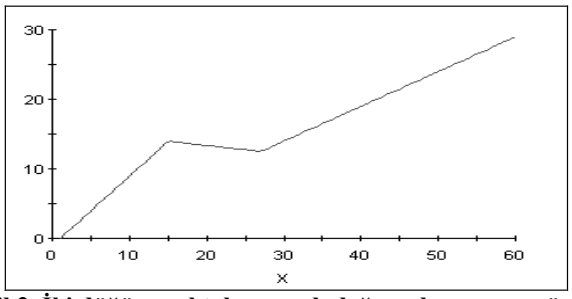
Matematikte iki tür eğri bulunmaktadır. Birinci tür interpolasyon eğrileridir. Bu eğrilerde eğri tüm veri noktalarından geçmektedir. Bu klasik eğri çizimidir. Diğeri ise düzleştirme eğrileridir. Düzleştirme eğrilerinde ise eğri, veri noktalarına yakın olmaktadır. Tam olarak bu noktalardan geçmesi gerekmez. Bu anlamda en basit düzleştirme eğrilerine parçalı doğrusal regresyon eğrisidir.



Şekil 1. Parçalı doğrusal regresyon eğrisi

Şekil 1'de sağdaki resim orijinal veriye aittir. Modelleme yapılırken veriye ait düz bir doğru oluşturmak yerine regresyon doğrusunun bükülmesi sağlanmıştır. Soldaki resimdeki doğru, üç noktadan bükülmüştür ve bir MARS uzanımı haline gelmiştir.

MARS, parçalı doğrusal regresyon uygulayarak esnek modeller oluşturur. Bağımsız değişkenin farklı aralıklarında ayrı regresyon eğim değerleri kullanarak doğrusallığı korur. Regresyon doğrusunun eğiminin değiştiği ve bir aralıktan diğerine geçildiği noktalara düğüm denir. Kullanılacak değişkenler ve her değişken için aralıkların bitiş noktası araştırma sonucu bulunur (Lee, Chiu vd., 2004).



Şekil 2. İki düğüm noktalı, parçalı doğrusal regresyon örneği

Örneğin, Şekil 2'de yer alan ve aralıkları birbirinden ayıran iki tane düğüm noktası olduğu görülmektedir. Regresyon doğrusunun eğimi bir aralıktan diğerine geçtiğinde değişmektedir.

Modeldeki değişkenler, etkileşimleri ve düğüm noktalarının konumu, kaba kuvvet yaklaşımıyla (brute force aproach), katsayılar ise en küçük kareler (EKK) yöntemiyle bulunur. Kaba kuvvet yaklaşımı, tüm olası çözümler bulunduktan sonra en iyi olana karar verilen bir yöntemdir (Dieterle, 2003).

MARS yönteminde diğer bir önemli konu, temel fonksiyonlardır. Temel fonksiyonlar, değişken aralıklarının tanımlandığı bölgesel modellerdir. Temel fonksiyonlar, tek bir eğri fonksiyonu ya da birden fazla değişkenin etkileşim terimi olabilir. Temel fonksiyonlar düğüm noktalarının belirlenmesi açısından önemlidir (Put, Massart vd., 2003). Temel fonksiyonlar bağımlı değişken (Y) ve bağımsız değişkenler $(X_1, X_2, ..., X_m)$ arasındaki ilişkiyi temsil eden fonksiyonlar olarak görev yapar. Bu fonksiyonlar, β_0 sabit parametresi ve diğer temel fonksiyonların ağırlıklandırılmış toplamından oluşur.

$$\hat{Y} = f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) \qquad (1)$$

β_0 : sabit terim

$B_m(X)$: m. temel fonksiyon

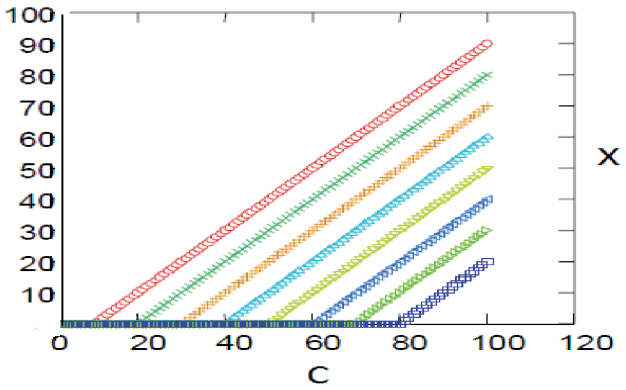
β_m : m. temel fonksiyonun katsayısı

M : Temel fonksiyon sayısı

Hokey sopası temel fonksiyonları - HSTF (hockey stick basis functions - HSBF), MARS modelinde önemli bir unsurdur. HSTF'ler sürekli X değişkeninin dönüşmüş hali olan X^* değerini,

$$X^* = \max(0, X - c), \text{veya} \quad \max(0, c - X) \tag{2}$$

biçiminde tanımlar. X^*, X'in eşik değeri olarak tanımlanan c değerinden küçük tüm değerleri için 0 değerini, c'den büyük X değerleri içinse X değerlerini alır. Örnek olarak X, 0 ve 100 arasında değerler alan bağımsız bir değişken olsun. $c = 10, 20, 30, ..., 70, 80$ için X^*'in aldığı değerler Şekil 3'de gösterilmiştir.



Şekil 3. Farklı eşik değerleri (c) için oluşturulan temel fonksiyonlar

MARS, veri setinin tüm değerleri için, değişik c değerlerine karşılık gelen çok sayıda temel fonksiyon oluşturabilir.

MARS ilk olarak sürekli bağımlı değişkenlerin tahmini için tasarlanmıştır. Daha sonra ikili kategorik bağımlı değişkenler için de kullanılmıştır (Salford Systems, 2001). Tunay (2011) çalışmasında, MARS yönteminin geleneksel regresyon yöntemleri ile parametrik olmayan modellerin üstünlüklerini başarıyla birleştiren ve ekonomik durgunluklar gibi ikili yapıdaki kategorik bağımlı değişkenlere rahatlıkla uygulanabilen yapısının önemli bir avantaj olduğunu belirtmiştir.

Klasik modellemede kategorik değişkenin her bir kategorisi için kukla değişkenler kümesi oluşturulur. Bu küme, regresyon analizinde girdiler olarak kullanılır. Bağımsız değişkenlerin kategorik olduğu durumlarda, MARS da aynı şekilde kukla değişkenler oluşturur. Ancak bu kukla değişkenler, ilgili değişkenin düzeylerinin bütünüdür.

Hastie ve Tibshirani (1990) MARS yöntemini, bağımlı değişkenin ikili olması durumunda, aşağıdaki formül ile logistik regresyon analizine uyarlamıştır.

$$\text{lojit } P(Y = 1) = f(x) + \varepsilon \tag{3}$$

Bu eşitlikte $f(x)$ MARS yöntemi tarafından tahmin edilen temel fonksiyonlar kümesini ifade eder. MARS yöntemi ile formül (3) birlikte ele alınırsa aşağıdaki formül elde edilir.

$$\text{lojit}(P_i) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) \tag{4}$$

Takip eden alt başlıkta MARS yöntemiyle modellemede yer alan basamaklar incelenmiştir. İyi bir model bu üç basamak sonucunda oluşur.

2.2 MARS Yönteminin Basamakları

Put, Massart vd. (2003)'e göre MARS yönteminde en iyi model üç basamaktan oluşan bir süreç sonunda elde edilir. Bu basamaklar yapım, budama ve yumuşatma aşamaları olarak adlandırılır.

2.2.1 Yapım aşaması (constructive phase)

Bu aşamada sabit terimle başlayarak ve sürekli olarak temel fonksiyonlar eklenir. Temel fonksiyonlar eklendikçe karışık ve de esnek bir model oluşur. Bu işlem kullanıcı tarafından sürecin başında belirlenen maksimum terim sayısına ulaşana kadar devam eder.

2.2.2 Budama aşaması (pruning phase)

İkinci aşama geri doğru eleme aşamasıdır. Birinci aşamada fazla sayıda temel fonksiyon eklendiği için oluşturulan model aşırı tahminleme problemiyle karşı karşıya kalacaktır. Aşırı tahminlemeyi ortadan kaldırmak için geri doğru eleme işlemine başlanır. Modelin karmaşıklığının azaltıldığı aşamadır. Modele en az katkısı olan temel fonksiyonlar atılır. Bu aşamada Craven ve Vahba (1979) tarafından geliştirilen Genelleştirilmiş Çapraz Geçerlilik (GÇG) (Generalized Cross-Validation - GCV) ölçütü kullanılır. GÇG aynı zamanda uyum eksikliği kriteridir.

$$\text{GÇG}(M) = \frac{1}{N} \frac{\sum_{i=1}^{N} \left(Y_i - \hat{f}_M(X_i)\right)^2}{\left(1 - C(M)/N\right)^2} \tag{5}$$

Formül (5)'de,
N: örneklem veri sayısı
$C(M)$: modelde uydurulan geçerli parametre sayısı
Eğer modelde M tane doğrusal bağımsız temel fonksiyon varsa,

$$C(M) = M + dc \tag{6}$$

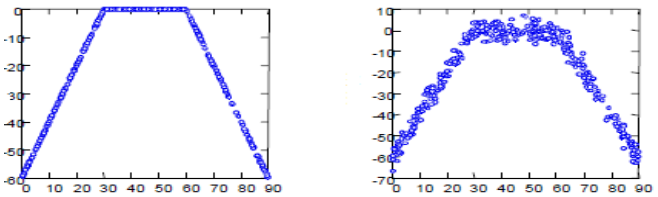
eşitliği sağlanır. Formül (6)'da, c ileri doğru olan süreçte seçilen düğüm sayısı, d ise her bir temel fonksiyon optimizasyonundaki maliyeti ifade eden bir değerdir. MARS

modellerinde genellikle d=3 alınır. Friedman (1991) tüm yapılan çalışmalar sonucunda d için en iyi değerlerin $2 \leq d \leq 4$ aralığında olduğunu belirtmiştir.

2.2.3 Yumuşatma aşaması (smoothing phase)

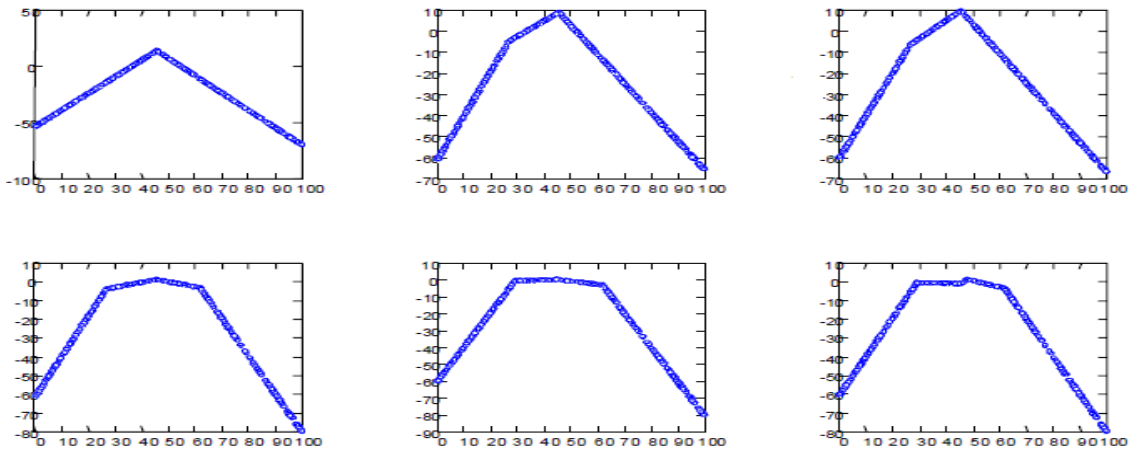
Son olarak bölgesel sınırlar içindeki süreksizliğin giderilmesi ve birincil ve ikincil türevlerin sürekliliğinin sağlanması için yumuşatma gereklidir (Quiros vd., 2009).

MARS'ın belirtilen bu basamaklarının özeti ve modellenmesi Şekil 4 ve Şekil 5'de gösterilmiştir. X ekseni bağımsız değişkeni, Y ekseni bağımlı değişkeni göstermektedir.



Şekil 4. Orijinal verinin dağılım grafiği

Şekil 5'de yer alan soldaki fonksiyon, düz tepeli fonksiyon olarak bilinir ve X=30 ile x=60 düğüm noktalarına sahiptir. Sağdaki fonksiyon ise gözlenen verinin dağılımını göstermektedir.



Şekil 5. MARS modeli ve oluşturduğu düğümler

MARS ilk olarak tek bir düğüm noktasını (X=45) belirlemiştir. İleri aşamada düğüm sayısı arttıkça MARS, Şekil 5'deki yaklaşımla modeli tahmin etmiştir (Salford Systems, 2001).

3. DÜŞÜK DOĞUM AĞIRLIĞINA İLİŞKİN MARS YÖNTEMİNE DAYALI İKİLİ LOJİSTİK REGRESYON MODELİ

Doğum ağırlığı bir bebeğin doğduğu anda ölçülen ağırlığıdır. Doğum ağırlığı yeni doğan ölümleri ve bebeklik dönemi hastalıklarını etkileyen faktörlerden biridir. Bu nedenle, doğum ağırlığı ve bunu etkileyen faktörler her zaman önemli bir klinik araştırma konusu olmuştur.

Düşük doğum ağırlığı, Dünya Sağlık Örgütü tarafından 2500 gramdan daha az olan doğum ağırlığı olarak tanımlanmıştır. Doğum ağırlığı ve gebelik haftası cenin veya yeni doğan ölümlerinin etkenlerinden biridir. Bu nedenle düşük doğum ağırlıkları veya prematüre doğumlar (37 haftadan önce gerçekleşen doğum) bu konudaki araştırmalar için önemli birer veridirler. Bebeğin düşük doğum ağırlıklı olarak doğmasında etkisi olan etkenler, Kramer (1987) tarafından aşağıdaki maddelerle verilmiştir.

- Doğum kusuru da denilen bebeğin doğumsal genetik ve yapısal anomalileri
- Annenin doğum geçmişi: önceki ölü doğum sayısı, düşük sayısı
- Annenin yüksek kan basıncı, şeker, kalp, ciğer ve böbrek hastalıkları
- Annenin alkol uyuşturucu ve sigara alışkanlığı
- Annenin geçirdiği enfeksiyonlar
- Plasentada görülen sorunlar (bebeğe kan ve besin ulaşımında sorunlara sebep olmaktadır).
- Annenin yeterince beslenememesi ve uygun kiloda olmaması
- Sosyoekonomik faktörler (az geliri olan, eğitim düzeyi düşük, 17 yaşından küçük veya 35 yaşından büyük annelerin bu konuda artan riske sahip olduğu saptanmıştır.)

Uygulama çalışması, yeni doğan bebeklerin düşük doğum ağırlığına sahip olma olasılıklarının, ikili lojistik regresyon analizi ile MARS yöntemine dayalı olarak model oluşturulmasını kapsamaktadır.

Bu çalışma sonucunda oluşturulacak model bir annenin bebeğinin doğum ağırlığının düşük olması ya da olmaması ile ilgili tahminde bulunulabilecektir. Böylece düşük doğum ağırlıklı olarak meydana gelme olasılığı bulunan bebek için önlemler alınılabilecektir.

Çalışmada Danimarka Ulusal Doğum Grubu (DUDG)'nun hamilelikte ateş ve buna bağlı ölü doğumlarla ilgili çalışmasına ait verilerin bir kısmı kullanılmıştır. DUDG, 1997 ve 2002 yılları arasında 100.000 kadın ile hamileliğin ilk 12–16 haftasında ve sonunda görüşüp gerekli verileri derlemiştir. Andersen, Vastrup vd. (2002) derlenmiş olan bu verilerden 31 Mart 1999 öncesine ait olanları kullanarak, ölü doğum bağımlı değişkenini etkileyen, bağımsız değişkenleri saptama çalışması gerçekleştirmiştir.

Çalışmamızda Andersen, Vastrup vd. (2002)'nin elde ettiği bağımsız değişkenlerden yararlanılmıştır. Ancak bu çalışmadan farklı biçimde, bağımlı değişken, bebeklerin doğum ağırlığı olarak seçilmiştir. Yeni doğan bir bebeğin ağılığı ortalama olarak 2500–4000 gr. arasındadır. Ağırlığı 2500 gramdan daha az olan bir bebek düşük doğum ağırlıklı olarak nitelendirilmektedir. Bu nedenle 2500 gramdan düşük olan bağımlı değişkenler 1 diğerleri 0 olarak kodlanmıştır. Çalışmamızda yer alan değişkenlerden annede ilk 17 haftada gözlenen yüksek ateş sayısı, annenin yaşı, önceki ölü doğum sayısı, önceki canlı doğum sayısı, gebelik haftası, sigara, alkol ve kahve kullanımı ve bebeğin doğum boyu değişkenlerinin doğum ağırlığını etkileyebileceği düşüncesiyle değerlendirmeye alınmıştır. 982 bireye ait veri kullanılmıştır. Kullanılan aday değişkenler ve kodlama işlemi Tablo 1'de verilmiştir.

Tablo 1. Uygulama verileri ve kodlama işlemi

Bağımlı değişken		Aldığı değerler ve kodları	Kısaltması
Y	Doğum ağırlığı	0=DA $\geq$ 2500 gr 1=DA<2500 gr	DA
Bağımsız değişkenler			
x_1	İlk 17 haftadaki yüksek ateş sayısı	0,1,2,.....	ateş
x_2	Annenin yaşı	0=yas<35 1=yas $\geq$ 35	yer
x_3	Geçmişteki düşük sayısı	0=düşük yapmamışsa 1=diğer d.	düşük
x_4	Önceki canlı doğum sayısı	0=canlı doğum yapmamışsa 1=diğer d.	canlı
x_5	Doğumda gebelik haftası	Sürekli değişkendir	hafta
x_6	Sigara kullanımı	0=Yok 1=Var	sigara
x_7	Alkol tüketimi	0=Yok 1=Var	alkol
x_8	Kahve tüketimi	0=Yok 1=Var	kahve
x_9	Boy	Sürekli değişkendir (cm)	boy

3.1 Verilerin MARS Yöntemi ile Analizi

Verilerin çok değişkenli uyarlayıcı regresyon uzanımları (MARS) yöntemiyle analizi aynı adı taşıyan MARS paket programı ile yapılmıştır. Bu programın MARS 6.6 sürümü, Salford Systems tarafından geliştirilen Salford Predictive Modeler Builder adlı paket program içinde yer almaktadır. Bu program CART, MARS, Combine, Randomforest, Treenet gibi değişik parametrik olmayan analiz yöntemlerini içeren paket programdır. MARS programında veriler girildikten sonra aynı verilerle fazla sayıda model oluşturulabilir. Bu modellerden en iyi olanı en az uyum eksikliğine sahip olandır. Uyum eksikliği en az olan model ise en düşük GÇG değerine sahip olandır. MARS program çıktılarında modele ait GÇG değeri verilmektedir.

Maksimum temel fonksiyon sayısı, maksimum etkileşim, düğümler arasındaki en az gözlem sayısı, düğüm optimizasyonu için serbestlik derecesi gibi değerler oluşturulan modeli belirler (Tablo 2). Bu değerler tamamen uygulayıcıya bırakılmıştır.

Tablo 2. MARS'ta model oluşumunda etkili olan değerler ve kısaltmaları

	MARS'taki ilk değeri	Kısaltması
Maksimum temel fonksiyon sayısı	15	TFmax
Maksimum etkileşim sayısı	1	Emax
Düğümler arasındaki en az gözlem sayısı	0	Omin
Düğüm optimizasyonu için serbestlik derecesi	3	SD

Model tahmininde öncelikle serbestlik derecesi 3 olarak belirlenmiştir. Farklı Tfmax, Emax ve Omin değerleri için GÇG değerleri kaydedilmiştir (Tablo 3-6).

Tablo 3. Emax=1 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04499	0.04506	0.04542
20	0.04499	0.04512	0.04580
30	0.04508	0.04508	0.04553
40	0.04719	0.04734	0.04802
50	0.04719	0.04734	0.04802
100	0.04912	0.04931	0.04966

Tablo 4. Emax=2 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04432
20	0.04522	0.04517	0.04527
30	0.04504	0.04521	0.04531
40	0.04735	0.04734	0.04742
50	0.04735	0.04735	0.04748
100	0.04914	0.04920	0.04938

Tablo 5. Emax=3 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04369
20	0.04522	0.04517	0.04546
30	0.04504	0.04521	0.04513
40	0.04735	0.04734	0.04749
50	0.04735	0.04734	0.04754
100	0.04914	0.04918	0.04922

Tablo 6. Emax=5 için GÇG değerleri

	TFmax		
Omin	15	25	50
0	0.04522	0.04476	0.04320
20	0.04522	0.04517	0.04530
30	0.04504	0.04506	0.04525
40	0.04735	0.04734	0.04753
50	0.04735	0.04734	0.04753
100	0.04914	0.04918	0.04922

Yapılan uygulamalar sonucunda GÇG değerinin, TFmax'ın 50 ve Omin'in 0'a yakın olduğunda en küçük değerlere sahip olduğu görülmüştür. Emax değerinin 5 olduğu durumda ise bu değer en küçük olmaktadır (Tablo 6).

Tfmax=50, Omin=0, Emax=5 olduğunda değişik SD değerleri için hesaplanan GÇG değerleri incelendiğinde SD değeri azaldıkça GÇG değerinin azaldığı ve optimum modele yaklaşıldığı görülmüştür (Tablo7).

Tablo7. Değişik SD değerleri için hesaplanan GÇG değerleri

SD	Genelleştirilmiş Çapraz Geçerlilik (GÇG) Değeri
0	0.04018
1	0.04141
2	0.04234
3	0.04320

Oluşturulan bu modellerden en küçük GÇG değerine sahip olan model uygun model olarak seçilmiş ve bu modele ait çıktılar aşağıda yorumlanmıştır:

Tfmax=50, Omin=0, Emax=5 ve SD=0 değerleri ile MARS uygulanmıştır.

MARS yönteminde sıradan en küçük kareler (EKK) yönteminde olduğu gibi temel fonksiyonlar ile bağımlı değişken arasındaki ilişkinin derecesini gösteren üç tür R-kare değeri kullanılmaktadır:

R-Kare: Bağımlı değişken ve temel fonksiyonlar arasındaki ilişkinin derecesi. Bu değer basit sıradan en küçük kareler regresyon analizi için hesaplanan R^2 ile aynı şekilde hesaplanır.

Düzeltilmiş R-Kare: Temel fonksiyonların sayısı için düzeltilmiştir.

Merkezsel olmayan R-Kare: Uyum iyiliğini test etmek için kullanılmamalıdır (Nash ve Bradford, 2001).

Modele ait R-kare değeri yaklaşık olarak 0.65 bulunmuştur. Düzeltilmiş R-kare 0.64 ve merkezsel olmayan R-kare 0.69 olarak bulunmuştur.

Uygulama sonucunda ileri ve geriye doğru işleyen yöntemle temel fonksiyonlardan 17 tanesinin model oluşumuna katkı sağladığı görülmektedir. Bu fonksiyonların katsayı tahminleri, standart hataları, t ve p değerleri görülmektedir (Tablo 8).

Tablo 8. Modele katkısı olan temel fonksiyonlar ve değerleri

Parametre	Tahmin	S.H	t-Değeri	p-Değeri
Sabit	1.00630	0.03693	27.25161	0.00000
Temel Fonksiyon 5	0.60325	0.08294	7.27300	0.00000
Temel Fonksiyon 9	-0.37637	0.04868	-7.73197	0.00000
Temel Fonksiyon 11	0.42711	0.06110	6.99032	0.00000
Temel Fonksiyon 13	-0.04797	0.01628	-2.94576	0.00330
Temel Fonksiyon 20	-0.03822	0.01477	-2.58753	0.00981
Temel Fonksiyon 21	-0.57432	0.08996	-6.38400	0.00000
Temel Fonksiyon 23	-0.61113	0.08769	-6.96930	0.00000
Temel Fonksiyon 27	-0.92835	0.13938	-6.66081	0.00000
Temel Fonksiyon 29	0.44079	0.09552	4.61449	0.00000
Temel Fonksiyon 31	0.60825	0.09022	6.74164	0.00000
Temel Fonksiyon 33	-0.14866	0.06755	-2.20068	0.02800
Temel Fonksiyon 35	0.04525	0.01951	2.31965	0.02057
Temel Fonksiyon 39	0.02204	0.00688	3.20159	0.00141
Temel Fonksiyon 41	0.04300	0.01453	2.95862	0.00317
Temel Fonksiyon 43	-0.03130	0.01698	-1.84370	0.06553
Temel Fonksiyon 45	-0.02329	0.00655	-3.55531	0.00040
Temel Fonksiyon 49	-0.08965	0.01298	-6.90723	0.00000

Model değerlendirmesi için kullanılan F istatistiği 104.85 ve p değeri 0.00000 olarak hesaplanmıştır.

Tahmin edilen model sonucunda bebeğin doğum ağırlığını, hafta değişkeninin %100 etkilediği, sırasıyla canlı, alkol, yaş ve sigara değişkenlerinin çoktan aza doğru etkisinin olduğu görülmektedir (Tablo 9). Tablodaki son sütun ilgili değişkenin yokluğunda modelin genel uyum iyiliğinde ne kadarlık azalma olacağı yer almaktadır.

Tablo 9. İlişkili değişkenlerin önemi

Değişken	Önemi	Genelleştirilmiş Çapraz Geçerlilik (GÇG) Değeri
Hafta	100.00000	0.11055
Canlı	18.24723	0.04253
Alkol	17.01658	0.04253
Yaş	16.85816	0.04253
Sigara	9.55560	0.04253

Modelin oluşumunda etkisi olan temel fonksiyonlar ve açıklamaları Tablo 10'da verilmiştir.

Tablo 10. Temel fonksiyonlar

BF5 = max(0, HAFTA - 38)
BF7 = (SIGARA in (0))
BF8 = (SIGARA in (1))
BF9 = max(0, HAFTA - 35)
BF11 = max(0, HAFTA - 36)
BF13 = max(0, HAFTA - 37) * BF8
BF15 = (CANLI in (0))
BF17 = (YAS in (0))
BF18 = (YAS in (1))
BF20 = max(0, 35 - HAFTA) * BF17
BF21 = max(0, HAFTA - 38) * BF17
BF23 = max(0, HAFTA - 37) * BF18
BF25 = (ALKOL in (1)) * BF15
BF27 = max(0, HAFTA - 36) * BF25
BF29 = max(0, HAFTA - 38) * BF25
BF31 = max(0, HAFTA - 35) * BF25
BF33 = max(0, HAFTA - 39) * BF25
BF35 = max(0, HAFTA - 40) * BF17
BF36 = max(0, 40 - HAFTA) * BF17
BF39 = (CANLI in (0)) * BF36
BF40 = (CANLI in (1)) * BF36
BF41 = (ALKOL in (1)) * BF40
BF43 = (SIGARA in (0)) * BF41
BF45 = max(0, HAFTA - 32) * BF7
BF49 = max(0, HAFTA - 32) * BF17

Bu temel fonksiyonlar ve katsayılarıyla elde edilen model ise aşağıdaki gibi olacaktır.

$$Y = 1.00629 + 0.60322 * BF5 - 0.376346 * BF9 + 0.42706 * BF11 - 0.0479715 * BF13 - 0.0382146 * BF20 - 0.574262 * BF21 - 0.611076 * BF23 - 0.928152 * BF27 + 0.440698 * BF29 + 0.608121 * BF31 - 0.148641 * BF33 + 0.0452413 * BF35 + 0.0220369 * BF39 + 0.0430005 * BF41 - 0.0313029 * BF43 - 0.023289 * BF45 - 0.0896456 * BF49;$$

Doğru sınıflama oranı ise modelin uygunluğunu ya da verilerin bağımlı değişkenini açıklama yüzdesini gösteren bir değerdir. Modele ait DSO değeri %95.418 olarak hesaplanmıştır. İlgili değişkenler bağımlı değişkenin yaklaşık %95'ini açıklamıştır (Tablo11).

Doğru sınıflama çizelgelerinden modelin duyarlılığı (sensitivity) ve özgüllüğü (specificity) hakkında da yorum yapılabilir.

Tablo 11. MARS için sınıflandırma çizelgesi

Gözlenen		Tahmin edilen		
		DA		toplam
		0.00	1.00	
Adım 1	DA 0.00	850	8	858
	1.00	37	87	124
				95.418

Duyarlılık referans düzeyin doğru tahmin edilme derecesidir. MARS için referans düzey, doğum ağırlığının 2500 gram ve üstünü ifade eden ve 0 olarak kodlanmış düzeydir. Bu durumda duyarlılık=850/858=0.9906 değerine sahiptir.

Özgüllük ise 1 olarak kodlanan doğum ağırlığının 2500 gramdan küçük olduğu değerleri ifade eden düzeyin doğru tahmin edilmesinin derecesidir. MARS yöntemiyle yapılan analiz için özgüllük, 87/124=0.7016 olarak hesaplanmıştır.

4. SONUÇ VE YORUMLAR

Regresyon analizinde amaç bağımlı değişkeni bağımsız değişkenlerin bir fonksiyonu şeklinde yazmaktır. Parametrik regresyon analizinde önceden bilinen modele ait fonksiyonun parametrelerini tahmin etmek amaçlanır. Parametrik olmayan yöntemde ise parametre yerine doğrudan fonksiyonlar tahmin edilir. Model ise bu fonksiyonların bir kombinasyonudur.

Gerçekleştirilen araştırmada verilerin MARS yöntemine dayalı ikili lojistik regresyon ile yapılan analizi sonucunda doğru sınıflandırma oranı %95.418 olarak bulunmuştur. Bu oran oldukça yüksek bir orandır. Oluşturulan modelde, gebelik haftası en fazla etkiye sahip değişken olarak ortaya çıkmıştır (Tablo 9). Aynı şekilde Tablo 10

incelendiği zaman hafta değişkeninin tek başına bir temel fonksiyon oluşturduğu gibi diğer birçok temel fonksiyonun oluşumunda da yer aldığı görülmektedir. Hafta değişkeni sürekli değişken olduğu için birçok düğüm noktası ortaya çıkmıştır. Analiz sonucunda bebeğin doğum ağırlığını, 32-40 arasındaki haftaların etkilediği temel fonksiyonlardan anlaşılmaktadır. Modelin temel fonksiyonlarının katsayıları incelendiğinde en büyük katsayılardan birisinin yine bu değişkene ait olduğu görülmektedir.

Bebeğin doğum ağırlığını etkileyen bir diğer önemli değişken ise, anne adayının daha önce canlı doğum yapıp yapmamasıdır. Daha önce canlı doğum yapmamış bir bireyin aynı zamanda alkol de kullanıyor olmasının bebeğin düşük doğum ağırlıklı olmasına sebep olduğu, 25 numaralı temel fonksiyondan anlaşılmaktadır (Tablo 10).

Yaş değişkeninin de bebeğin doğum ağırlığında etkili olduğu anlaşılmaktadır. 35 yaşına kadar olan yaşlara ait 17 numaralı temel fonksiyonun düşük doğum ağırlığına etki etmediği, 35'in üzerindeki yaşlara ait 18 numaralı temel fonksiyonda ise, yaş değişkeninin temel fonksiyon katsayısı kadar etkili olduğu anlaşılmaktadır (Tablo10).

Araştırmada kullanılan, ilk on yedi haftada ateşli hastalık geçirme ve kahve kullanma durumları çok düşük miktarda etkiye sahip oldukları için için oluşturulan modelde yer almamışlardır. Bu çalışmanın bir devamı niteliğinde, çalışmada kullanılan değişkenlere alternatif yeni değişkenlerin varlığı ve bebek doğum ağırlığına etkileri araştırılabilir.

5. KAYNAKLAR

Andersen A. M. N., Vastrup P., Wohlfahrt J., Amdersen P. K., Olsen J., Melbye M., 2002. Fever in pregnancy and risk of fetal death: a cohort study. Lancet, 360: 1552-1556.

Craven, P., Wahba, G., 1979. Smoothing Noisy Data with Spline Functions. Numer. Math, 31: 377-403.

Dieterle, F. J., 2003. Multianalyte Quantifications by Means of Integration of Artificial Neural Networks, Genetic Algorithms and Chemometrics for Time-Resolved Analytical Data. Ph.D. thesis, Universität Tübingen, Tübingen, 183 s.

Friedman, J. H., 1991. Multivariate Adaptive Regression Splines (with discussion). Ann. of Statistics, 19: 1-141.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Chapman & Hall/CRC, New York, 335s.

Kan, B., Yazıcı, B., 2010. Comparison of the Results of Factorial Experiments, Fractional Factorial Experiments, Regression Trees and MARS for Fuel Consumption Data. WSEAS TRANS. on MAT., 2:110-119.

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

Kim, J. H., 2000. MARS Modeling for Ordinal Categorical Response Data: A Case Study. Soongsil University.

Kolyshkina, I., Brookes, R., 2002. Data Mining Approaches to Modelling Insurance Risk. Electronic Version, Pricewaterhouse Coopers, New York, 20 s.

Kramer, M. S., 1987. Determinants of Low Birth Weight: Methodological Assessment and Meta-analysis. Bull World Health Organ. 1987;65(5):663-737.

Kriner, M., 2007. Survival Analysis with Multivariate Adaptive Regression Splines. Dr. Rer. Nat. Universitat Munchen, 101 s.

Kuhnert, P. M., Do, K. A., Mc, R., 2000. Combining Non-parametric Models with Logistic Regression: an Application to Motor Vehicle Injury Data. Computational Statistics & Data Analysis, 34: 371-386.

Lee, T. S., Chiu, C. C., Chou, Y. C., Lu, C. J., 2004. Mining The Customer Credit Using Classification And Regression Tree and Multivariate Adaptive Regression Splines. Computational Statistics & Data Analysis, Volume 50, Issue 4, 24 February 2006, Pages 1113–1130.

Mina, C., 2009. Profiling Poverty with Multivariate Adaptive Regression Splines. PIDS Discussion Paper Series No. 2009-29, 55 s.

Mina, C., 2010. Employment Choices of Persons with Disability in Metro Manila. PIDS Discussion Paper Series No. 2010-29, 35 s.

Nash, M. S., Bradford, D. F., 2001. Parametric and Nonparametric(MARS; Multivariate AdditiveRegression Splines) Logistic Regressions for Prediction of A Dichotomous Response VariableWith an Example forPresence/Absence of an Amphibian. United States Environmental Protection Agency (EPA), USA, 40s.

Put, R., Massart D. L., Heyden V., 2003. An Application of Multi-variate Adaptive Regression Splines (MARS) in QSRR, IEJMD. BioChemPress.com.

Quiros, E., Felicisimo, A. M., Cuartero, A., 2009. Testing Multivariate Adaptive Regression Splines (MARS) as a Method of Land Cover Classification of TERRA-ASTER Satellite Images. Sensors, 9:9011-9028.

Salford System, 2001. MARS, User Guide. Cal. Stat. SoftWare, Inc., San Diego, California.

Samui, P., Kothari, D.P., 2011. Application of Multivariate Adaptive Regression Splines to Evaporation Losses in Reservoirs, Earthscience, 4:15-20.

Stokes, H. H., Lattyak, W. J., 2005. Multivariate Adaptive Regression Spline (MARS) Modeling Using the B34S ProSeries Econometric System and SCA WorkBench. Scientific Computing Associates Corp., 80 s.

Topak, M. S., 2011. An Empirical Study to Model Corporate Failures In Turkey: A Model Proposal Using Multivariate Adaptive Regression Splines (MARS). Namık Kemal Üniversitesi Sos. Bilim. Metinleri, 15 s.

Tunay, K. B., 2001. Türkiye'de Paranın Gelir Dolaşım Hızlarının MARS Yöntemiyle Tahmini. ODTÜ Geliştirme Dergisi, 28: 431-454.

Tunay, K. B., 2010. Bankacılık Krizleri ve Erken Uyarı Sistemleri: Türk Bankacılık Sektörü İçin Bir Model Önerisi. BDDK Bankacılık ve Finansal Piyasalar Dergisi, 4:9-46.

Tunay, K. B., 2011. Türkiye'de Durgunlukların MARS Yöntemiyle Tahmin ve Kestirimi. Marmara Üniversitesi İİBF Dergisi, XXX:71-91.

Yerlikaya, 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.

MODELLING THE LOW BIRTH WEIGHT OF NEW BORN BABIES WITH BINARY LOGISTIC REGRESSION BASED ON MARS METHOD

ABSTRACT

Babies with low birth weight have some health problems in later years. Therefore, it is important to estimate before the birth whether a new born baby will have a low birth weight or not. In order to obtain this estimation, logistic regression model is a suitable choice. Logistic regression analysis is a modelling technique which is used when the dependent variable is categorical. It is also easily interpreted. When the dependent variable has only two categories, the logistic regression is called binary logistic regression. Logistic regression has parametric and nonparametric solutions. MARS method is a nonparametric method which can be used as an alternative to the parametric solutions in the analysis of logistic regression. The nonparametric models require fewer assumptions compared to the parametric ones and they are also more flexible. In the application, a binary logistic regression model has been fitted to estimate whether a new born baby will have a low birth weight or not. The model has been estimated based on the MARS method. In the analysis, data belonging to 982 subjects have been investigated by applying the MARS software. In the conclusion part, the findings are interpreted.

Keywords: Low birth weight, Logistic regression, Maximum likelihood, MARS.