SAKARYA ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ
http://dergipark.gov.tr/saufenbilder

İkili yanıt değişkenine sahip modellerin yeterliliklerine ilişkin benzetim çalışması – parametrik olmayan yöntemler

Betül Kan Kılınç[1*], Mustafa Çavuş[2]

22.06.2016 Geliş/Received, 23.11.2016 Kabul/Accepted

doi: https://doi.org/10.16984/saufenbilder.297002

ÖZ

Regresyon modelleri; birçok açıklayıcı değişkenin önemini ortaya koyabilmek için tahmin, sınıflama, ve analitik veri araçlarını kullanarak, veri analizinde etkili bir rol oynamaktadır. Oldukça basit olmasına rağmen klasik doğrusal model, gerçek hayattaki örneklerin doğrusal olmaması nedeniyle sıkça yetersiz kalmaktadır. Bu çalışmada, çoklu doğrusal regresyon analizi varsayımlarından biri olan; bağımlı değişkenin açıklayıcı değişkenler ile arasındaki ilişkinin belli bir matematiksel forma uymasının zorunlu olmadığı parametrik olmayan bir değerlendirme süreci ele alınacaktır. Bu anlamda bağımlı değişkenin iki düzeyli değerler aldığı, daha çok neden-sonuç ilişkilerinin ortaya koyulması amacıyla kullanılan klasik lojistik regresyon modelinin yerine, bağımlı değişken ile açıklayıcı değişkenlerin aralarında var olan ilişki bir benzetim çalışması kapsamında; genelleştirilmiş doğrusal model, toplamsal lojistik regresyon model ve karar ağaçları ile incelenecektir. Benzetim çalışmasında söz konusu olan yöntemler ile küçük, orta ve büyük ölçekli veri kümelerinde çoklu bağlantının etkileri incelenecek ve bu yöntemler birbirleriyle karşılaştırılacaktır.

Anahtar Kelimeler: toplamsal modeller, lojistik regresyon, toplamsal lojistik regresyon

Comparative simulation study for model adequancy with binary response variable under multicollinearity – nonparametric approaches

ABSTRACT

Regression models used to explore the importance of several explanatory variables in estimation, classification and analytical tools play an efficient role for many data analysis. Although the classical linear model is quite easy to use, it is often not sufficient for many real data sets as the relationships between variables do not hold the assumption of the linearity of the relationship between dependent and explanatory variables. Under this study, a nonparametric model fitting that does not require to form a strict mathematical relationship between dependent and explanatory variables will be discussed on the contrary the assumption in multiple linear regression. In this study, the relationship between a binary dependent variable and the explanatory variables will be examined in a conducted simulation study by using generalized linear, the additive logistic regression in case of classical logistic regression model and decision trees to explore the cause and effect relationship. The methods in question and the simulation study will be performed for small, medium and large data sets when multicollinearity problem exists and will be compared with each other.

Keywords: additive models, logistic models, additive logistic models

[*] Sorumlu Yazar / Corresponding Author
[1] Anadolu Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir - bkan@anadolu.edu.tr
[2] Anadolu Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Eskişehir - mustafacavus@anadolu.edu.tr

1. GİRİŞ (INTRODUCTION)

Açıklayıcı değişkenler arasında çoklu bağlantı sorununun varlığı durumunda sıradan en küçük kareler (EKK) yönteminin kullanımının sakıncaları literatürde oldukça yaygın olarak işlenmiştir [1]. Benzer şekilde lojistik regresyon için de bazı sakıncalar söz konusudur. Bu durumu ortadan kaldırmak için kullanılacak yöntemler (modelden değişken çıkarmak, birkaç değişkeni bir indeks olarak ifade etmek, örneklem hacmini büyütmek vs) doğrusal regresyonda olduğu gibidir. Bu çalışmada benzer durumlar için klasik lojistik regresyon kullanmak yerine, bağımlı değişkeni açıklamak üzere toplamsal lojistik regresyon ile daha esnek bir model oluşturulacaktır. Bu modeller karar ağaçları algoritmaları ile karşılaştırılacaktır.

Bağımlı değişkenin ikili değerler aldığı ve değişkenler arasında çoklu bağlantının gözlendiği pek çok uygulama çalışmasına literatürde rastlanılabilir. Bunların arasında örneğin Shen ve ark. (2008) çalışmasında lojistik regresyon modellerinde çoklu bağlantı problemi söz konusu olduğunda en çok olabilirlik tahmininin yanlı hatta bazen sonuç vermemesi nedeniyle genelleştirilmiş modellerde cezalı en çok olabilirlik yöntemi ile ridge regresyon yaklaşımını kullanmışlardır [2]. Çalışmanın uygulama kısmında Alzheimer rahatsızlığını ve bunamayı (demans) ortaya koymada yardımcı olan 33 soruluk bir testten alınan yanıtlar üzerinde bir çalışma yapılmıştır ve lojistik regresyon modellerinde çoklu bağlantı problemi söz konusu olduğunda çift cezalı en çok olabilirlik tahmini yönteminin, klasik en çok olabilirlik yönteminden daha iyi bir yaklaşım olduğu sonucuna ulaşılmıştır. Kaşko (2007), çalışmasında bağımlı değişken ile aralarında farklı düzeylerde ilişki bulunan açıklayıcı değişkenlerin yer aldığı lojistik regresyon modellerinin, Tip I hata olasılığını ve testin gücünü birbirleriyle karşılaştırarak çoklu bağlantı probleminden nasıl etkilendiklerini bir benzetim çalışması yardımıyla göstermiştir. Çalışmanın sonucunda çoklu bağlantı probleminin varlığı durumunun, tüm örneklem genişliklerinde Tip 1 hatasını değiştirmezken, testin gücünü büyük ölçüde düşürdüğü tespit edilmiştir [3]. Kovalchi ve ark. (2013) çalışmasında toplamsal lojistik modellerin genişletilmiş hali olan toplamsal binom modellerini tanıtmış ve mesane kanseri üzerine bir uygulama yapmıştır. Çalışma boyunca R programının "blm" paketinden yararlanılmış ve paket ile dahilindeki kodlar detaylı olarak tanıtılmıştır [4]. Ma ve ark. (2014)' nın çalışmasında, doğrusal model durumunda anlamlı bulunmayan sürücü yaşı değişkeninin trafik kazalarında çarpan ya da çarpılan taraf olma üzerine etkisinin olup olmadığı durumu, kübik splayn düzeltmesi kullanılarak toplamsal lojistik regresyon model ile incelenmiştir. İnceleme sırasında sürücü yaşı değişkeni anlamlı bir etkiye sahip olup, sürücü yaşı küçüldükçe çarpan (hatalı) taraf olma olasılığının arttığı gözlenmiştir [5].

Bu çalışmada, çoklu bağlantının varlığında, genelleştirilmiş doğrusal model, toplamsal lojistik regresyon model ve sınıflandırma ağaçlarının bir benzetim çalışması ile karşılaştırılması söz konusudur. Bu kapsamda çoklu bağlantının var olduğu durumlar söz konusuyken, küçük, orta ve büyük veri kümeleri için modellerin yeterliliğini, bahsedilen yöntemler yardımıyla ortaya koymak amaçlanmaktadır.

2. GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLER (GENERALIZED ADDITIVE MODELS)

Regresyon modelleri; farklı girdilerin önemini ortaya koyabilmek için tahmin ve sınıflama araçlarını kullanarak bir çok veri analizi için önemli bir araç haline gelmiştir. Geleneksel doğrusal model $E(Y \mid X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ oldukça basit olmasına rağmen, gerçek hayattaki örneklerin doğrusal olmaması nedeniyle, daha esnek bir yaklaşım olan istatistiksel yöntemlerden "genelleştirilmiş toplamsal modeller" alternatif olarak kullanılabilir.

Regresyon düzeninde, bir genelleştirilmiş toplamsal model şu şekilde gösterilir;

$$E(Y \mid X_1, X_2, \ldots, X_k) = u + f_1(X_1) + \cdots + f_k(X_k) \quad (1)$$

Bilindiği gibi burada $X_1, X_2, \ldots, X_k$ değişkenleri bağımsız değişkenleri ve Y bağımlı değişkeni, $f_i, i = 1,2, \ldots, k$ olmak üzere f_k düzelticileri fonksiyonları (parametrik olmayan) gösterir. Burada X_i' nin keyfi olarak seçilen fonksiyonunun terimleri $f_i(X_i)$, doğrusal denklemdeki $\beta_i X_i$ terimlerine karşılık gelir. Buradaki yaklaşımın farklılığı, her fonksiyona serpilme diyagramı düzleştiricisinin uydurulması ve bu şekilde her f_i fonksiyonunu tahmin etmektir [6].

Toplamsal modeller cezalandırılmış regresyon splaynları ile gösterilebilir ve cezalandırılmış en küçük kareler yöntemi ile tahmin edilir. Cezalandırılmış regresyon, klasik en küçük kareler yöntemine göre daha esnek olduğundan yanıt değişkeni ve bağımsız değişkenler arasındaki ilişkiyi daha iyi bir şekilde açıklar. Regresyon eğrisinin oluşturulmasında cezalandırılmış regresyon modeli ceza terimi denilen yeni bir terim kullanır:

$$\sum_{i=1}^{n} [Y_i - \vartheta(X_i)]^2 - \lambda \int_a^b [\vartheta''(X)]^2 \, dx \quad (2)$$

burada $\lambda \int_a^b [\vartheta''(X)]^2 \, dx$ ceza terimini gösterirken, λ düzeltme parametresi ise hata ve değişkenlik arasındaki değişim oranını gösterir. Düzeltme parametresinin düşük

değerleri regresyon eğrisinin daha kıvrımlı olmasına neden olurken yüksek değerleri daha az kıvrımlı bir hale gelmesine neden olur.

$\lambda \to \infty$ iken regresyon eğrisi düz bir çizgi halini alırken, $\lambda = 0$ olduğunda ise cezalandırılmamış regresyon eğrisi tahminine dönüşür. Cezalandırılmış regresyon, ϑ fonksiyonunu regresyon eğrisi olarak Eşitlik (2) yardımıyla tahminleme sürecidir [7].

2.1. Düzeltici Fonksiyonlar (Smooth Functions)

Düzeltici fonksiyonlar genelleştirilmiş toplamsal modellerde, bağımsız değişkenin bir fonksiyonu ile bağımlı değişkendeki değişkenliği açıklamaya çalışır. Düzeltici fonksiyonlar (smooth functions) splaynlar olarak da bilinir. Splaynlar, yanıt değişkeni ile bağımsız değişkenler arasındaki ilişkiyi açıklamak için kullanılır. Farklı düzelticiler olmasına rağmen, bu çalışma kübik düzleştirici ile sınırlandırılmıştır.

2.2. Kübik Splaynlar (Cubic Splines)

Kübik splaynlar (KS) en basit şekliyle farklı kübik polinomlardan oluşan bir eğri olarak tanımlanabilir. $[X_i, X_{i+1}]$ aralığındaki polinomlardan oluşan $[a, b]$ aralığını ele alalım. Burada x_i değerleri düğüm noktalarını göstermektedir. $[a, b]$ aralığında tanımlanan f fonksiyonu eğer aşağıdaki iki koşulu sağlıyorsa kübik splayn olarak tanımlanır:

- Her bir $[X_i, X_{i+1}]$ aralığında f kübik polinom olmalıdır.
- Her bir düğüm noktasında f fonksiyonu, birinci türevi ve ikinci türevi $[a, b]$ aralığında sürekli olmalıdır.

Kübik splayn yapısına bir örnek aşağıdaki gibi verilebilir:

$$f(X) = a_1(X - X_i)^3 + a_2(X - X_i)^2 + a_3(X - X_i) + a_4 \quad (3)$$

$X_i \leq X \leq X_{i+1}$, a_1, a_2, a_3, a_4 değerleri splayn fonksiyonun şeklini belirleyen sabitlerdir. Bu tür splaynı kullanmanın dezavantajı tahmin edilmesi gereken çok sayıda parametreye sahip olmasıdır [7].

3. TOPLAMSAL LOJİSTİK REGRESYON MODELİ (ADDITIVE LOGISTIC REGRESSION MODEL)

Yalnızca iki sınıf olarak ifade edilen modelleri, iki sınıfa ayırmak için lojistik regresyon modeli kullanılır [6]. İkili yanıt değişkeninin beklenen değerini $\mu(X) = P((Y = 1)|X)$, tahminleyiciler ile ilişkilendirirken, doğrusal bir model kurar ve logit bağlantı fonksiyonu olarak:

$$log\left(\frac{\mu(X)}{1 - \mu(X)}\right) = u + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (4)$$

alınır. Toplamsal lojistik regresyon modeli ise her doğrusal terim yerine daha genel bir yapı kullanır:

$$log\left(\frac{\mu(X)}{1-\mu(X)}\right) = u + f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k) \quad (5)$$

burada her f fonksiyonu belirlenmemiş düzeltici fonksiyonları ifade eder. Bu parametrik olmayan düzeltici fonksiyonlar, modeli daha esnek hale getirirken, toplamsallık yine aynı yöntemlerle model tahmininin yapılabilmesini sağlar. **Toplamsal lojistik regresyon modeli,** genelleştirilmiş toplamsal modellerin bir özel bir halidir. Genel olarak Y yanıt değişkeninin koşullu ortalaması olan $\mu(X)$, tahmincilerin toplamsal fonksiyonu olan l bağlantı (link) fonksiyonu ile şu şekilde ifade edilir:

$$l\,[\,\mu(X)] = u + f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k) \quad (6)$$

Klasik bağlantı fonksiyonuna ait birkaç örnek verilirse:

- $l(\mu) = \mu$: normal dağılıma sahip yanıt değişkeni söz konusu olduğunda doğrusal ve toplamsal modeller için kullanılır.
- $l(\mu) = logit(\mu)$ ya da $l(\mu) = probit(\mu)$: binomial olasılıkları modellemek için kullanılan probit bağlantı fonksiyonudur. Probit fonksiyonu Normal dağılım fonksiyonun tersi ile ifade edilen bir fonksiyondur: $probit(\mu) = l^{-1}(\mu)$.
- $l(\mu) = log(\mu)$: poisson sayılabilir veri seti için logaritmik doğrusal ya da logaritmik toplamsal modeller için kullanılır.

Bu üçlü üstel dağılım ailesinin örneklem modellerine ek olarak gamma ve negatif binom dağılımlarından ortaya çıkmıştır. Bu dağılım ailelerinden çok iyi bilinen genelleştirilmiş doğrusal modeller ortaya çıktığı gibi aynı yollardan genelleştirilmiş toplamsal modeller de ortaya çıkar [6].

4. KARAR AĞACINA DAYALI YÖNTEMLER (DECISION TREES METHODOLOGY)

Karar ağacına dayalı yöntemler, ilgili değişkenler uzayını, dikdörtgenler seti olarak ifade eder ve sonrasında modeli her biri için tanımlamaya çalışır. Kavramsal olarak basit olmasına rağmen güçlü bir yöntemdir. Bir sonraki bölümde, popüler bir yöntem olan regresyon ve sınıflama ağaçları olarak isimlendirilen CART (Classification and Regression Trees) yöntemi açıklanacaktır [6].

4.1. Regresyon Ağaçları (Regression Trees)

Varsayılsın ki, veri seti p tane açıklayıcı değişken $X_j = \left(X_{j1}, X_{j2}, \dots, X_{jp}\right)$ ve $j = 1, 2, \dots, N$ olmak üzere, her N gözlem için bir tane yanıt değişkeni içersin, Y_j, X_j. Algoritma otomatik olarak değişkenlerin ayrılmasına, düğüm noktalarına ve ağacın şeklinin nasıl olması gerektiğine karar vermelidir. Bunun için öncelikle $R_1, R_2, \dots, R_i$ olmak üzere i tane bölge olsun ve yanıt değişkeni her bölgede c_i sabiti ile ifade edilsin:

$$f(X) = \sum_{i=1}^{R} c_i \, I(X \epsilon \, R_i) \tag{7}$$

Eğer en küçük kareler yönteminden hareket edersek en uygun $\hat{c}_i$ tahmini R_i bölgesinde y_i' lerin ortalamasıdır:

$$\hat{c}_i = \text{ave}\left(Y_j \middle| X_j \epsilon \, R_i \;\right) \tag{8}$$

En küçük kareler yaklaşımı ile en iyi ikili bölüntüyü (partition) hesaplamak mümkündür. Bunun için bir algoritma geliştirilebilir. Tüm veri setini ele alarak başlanırsa bölüntü değişkeninin m ve düğüm noktasının (split point) a olduğu varsayımı altında yarı düzlem çiftleri belirlenir:

$$R_1(m, a) = \{X | X_m \leq a\}, \;\; R_2(m, a) = \{X | X_m > a\} \tag{9}$$

Daha sonra aşağıdaki denklem çözülerek bölüntü değişkeni m ve düğüm noktası a bulunur:

$$\min_{m,a} \left[\min_{c_1} \sum_{X_j \epsilon R_1(m,a)} (Y_j - c_1)^2 + \min_{c_2} \sum_{X_j \epsilon R_2(m,a)} (Y_j - c_2)^2 \right] \tag{10}$$

Herhangi bir m ve a değeri için minimizasyon şu şekilde hesaplanır:

$$\hat{c}_1 = ave\left(y_j \middle| Y_j \epsilon \, R_1(m, a)\right), \hat{c}_2 = ave\left(Y | X_j \epsilon \, R_2(m, a)\right) \tag{11}$$

Her bölüntü değişkeni için düğüm noktası a' nın belirlenmesi kolaydır. Bunun için tüm açıklayıcı değişkenler gözden geçirilir ve en iyi (m, a) ikilisinin belirlenmesine çalışılır.

En uygun bölüntüyü bulmak için, veri seti iki bölgeye ayrılır ve bu bölüntüleme her iki bölge için tekrar edilir. Daha sonra bu işlem her bölge için tekrar edilir.

Bir ağaç ne kadar genişlemelidir sorusu için açıkça söylenebilir ki, çok geniş bir ağaç veri seti için uygun olmayabilir (overfit sorunu), bunun yanında çok küçük bir ağaç da verideki önemli yapıyı ortaya koyamayabilir [6].

Ağaç boyutu, modelin karmaşıklığını belirleyen bir ayarlama parametresidir (tuning parameter α) ve optimum ağaç boyutu veriden seçilmelidir. Bir yaklaşıma göre ağaç düğümleri, kareler toplamı bazı eşik değerini aştığı anda azalma gösteriyorsa ayrılmalıdır. Bu strateji oldukça sığ görünse de, görünüşte önemsiz olan bir bölüntünün altında daha iyi bir bölüntü olabilir.

Öncelikli strateji, en geniş bir ağaç (T_0) geliştirmek ve bölüntü sürecini en küçük düğüm büyüklüğüne (örneğin 5) ulaşıncaya kadar devam ettirmektir. Daha sonra bu geniş ağaç cost-complexity budama yöntemi ile sonlandırılır [6].

Bunun için öncelikle bir alt ağaç belirlenir $(T \subset T_0)$. Bu ağaç belirli sayıda geçiş düğümlerinden (internal node) oluşabilir. Son düğümler i ile gösterilirse, R_i' nci bölgeye ait i düğümünü gösterecektir. $|T|$ son düğümlerin sayısını göstermek üzere,

$$N_i = \# \left\{X_j \epsilon \, R_i\right\}$$
$$\hat{c}_i = \frac{1}{N_i}$$
$$Q_i(T) = \frac{1}{N_i} \sum_{x_j \, \epsilon \, R_i} (Y_j - \hat{c}_i)^2 \tag{12}$$

cost-complexity kriteri;

$$CC_\alpha(T) = \sum_{i=1}^{|T|} N_i Q_i(T) + \alpha |T| \tag{13}$$

olarak elde edilir. Amaç, her α için $CC_\alpha(T)$ ile minimize edilen alt ağacı bulmaktır. $\alpha \geq 0$, ağaç boyutu ve veriye uygunluk arasındaki dengeyi gösterir. Büyük α değerleri için ağaç boyutu küçülürken, tam tersi durumda büyür. $\alpha = 0$ olduğunda ise en geniş ağaç (T_0) durumu oluşur. α' nın seçimi ile ayrıntılı bilgilere ulaşmak için Breiman (1984) ve Ripley (1996) kaynakları önerilir [8,9].

4.2. Sınıflama Ağaçları (Classification Trees)

Araştırmalarda en sık kullanılan bir diğer sınıflandırma yöntemi ise sınıflama ağaçlarıdır. Regresyon ağaçlarına göre farklı bir algoritma kullanarak değişkenleri sınıflandırma yoluna gider.

Eğer, amaç sonuçları $1, 2, \dots, K$ şeklinde sınıflandırmak ise sadece ağaç algoritmasının düğüm ve bitiş kriterleri değiştirilmelidir. Regresyon için hata kareler düğüm uygunsuzluk ölçümü olarak $Q_i(T)$ kullanılır, fakat bu ölçü sınıflandırma için uygun değildir. Bir i düğümünde

R_i bölgesi N_i gözlemleri, I indikatör fonksiyon olmak üzere:

$$\hat{p}_{ik} = \frac{1}{N_i} \sum_{x_j \epsilon R_i} I\left(Y_j = k\right) \tag{14}$$

i.düğümdeki k gözlemin sınıf oranıdır. i.düğümdeki gözlemler ile $k(i) = arg \max_k \hat{p}_{ik}$ i.düğümdeki çoğunluk sınıfına dahil edilir. $Q_i(T)$ uygunsuzluk ölçüsü izleyen farklı ölçüleri içerir,

Yanlış Sınıflama Hatası:

$$\frac{1}{N_i} \sum_{j \,\epsilon\, R_i} I\left(Y_j = k(i)\right) = 1 - \hat{p}_{ik(i)} \tag{15}$$

Gini İndeksi:

$$\sum_{k \neq k'} \hat{p}_{ik} \hat{p}_{ik'} = \sum_{k=1}^{K} \hat{p}_{ik}(1 - \hat{p}_{ik}) \tag{16}$$

Cross-Entropi veya Sapma:

$$-\sum_{k=1}^{K} \hat{p}_{ik} \log \hat{p}_{ik} \tag{17}$$

5. ÇOKLU BAĞLANTI SORUNU (MULTICOLLINEARITY PROBLEM)

Çalışmanın başında da söz edildiği gibi çoklu bağlantı probleminin ele alınan yöntemler üzerindeki etkisi incelenecektir. Bu nedenle çoklu bağlantı problemi, ortaya çıkış nedenleri ve neden olduğu problemler bu bölümde detaylı olarak ele alınmıştır.

Çoklu bağlantı herbir gözlem için açıklayıcı değişkenler arasında bir ya da birden çok doğrusal bağlantının varlığı olarak açıklanır [10]. Çoklu bağlantının etkileri aşağıdaki gibi özetlenebilir [11]:

- Güçlü çoklu bağlantı $(X'X)^{-1}$ köşegen öğelerinin böylece de $\hat{\beta}$' ların standart hatalarının büyük çıkmasına neden olur. Bu da t istatistik değerlerini küçük göstereceğinden değişkenlerin anlamlılığında yanlış bulgular ortaya çıkabilir.
- Çoklu bağlantı regresyon katsayılarını değerce ve işaretçe etkilediğinden gerçektekinden oldukça ayrı kestirimler ortaya çıkabilir.
- Katsayı kestirimleri örneklem verilerine duyarlı olduğundan veri kümesine birkaç gözlem eklenmesi bu kestirimlerde büyük değişliklere yol açar.

Çoklu bağlantının kaynakları şu şekilde özetlenebilir [11]:

- Geniş tanımlı model (değişken sayısının gözlem sayısından büyük olduğu model),
- Örnekleme teknikleri,
- Model ve anakütle üzerindeki fiziksel kısıtlar (anakütlede var olan gerçek ilişkinin örneklemde de korunması durumu).

Çoklu bağlantı sorununun giderilmesi için önerilen çözüm yolları arasında; gözlem sayısının arttırılması, regresyon katsayılarıyla ilgili önbilgilerden önceden kestirimlerin elde edilmesi, değişkenlerin birleştirilerek tek değişken olarak alınması, yanlı kestirimlerin kullanılması ve değişken seçimi gelir [11].

Çoklu doğrusal regresyon çözümlemelerinde verileri en iyi tanımlayacak önemli değişkenlerin modele alınması, modele katkısı gerekli olmayan değişkenlerin modelden çıkartılması "değişken seçimi" ya da "en iyi alt küme denkleminin seçimi" olarak bilinir. Genel regresyon sürecinde, artıkların ve çoklubağlantının incelenmesinden sonra en iyi modeli oluşturabilmede önemli bir aşama değişken seçimidir. Bu aşamaya yalnızca model kurmak için değil çoklu bağlantıya çözüm getirebilmek için de sık sık başvurulur. Değişken seçimi yapmanın iki ana nedeni vardır [12]:

- Pratik nedenler: Modelde az sayıda değişken olması uygulama ve ekonomik açıdan yararlar sağlar. Modeldeki değişkenleri sonradan izlemek gerekebileceğinden bu değişkenlere ilişkin verileri toplamadaki güçlükler ya da bunları elde etme maliyetinin yüksekliği, araştırmacıları az sayıda değişkenle çalışmaya zorlar (parsimony).
- Kuramsal nedenler: Bir model kurulurken kestirimlerin ve önkestirimlerin istenilir istatistik özelliklere sahip olmaları gerekir. Alt küme modeli ele alındığında, gereksiz değişkenlerin modelde bulunmasının veya bazı gerekli değişkenlerin modelden çıkarılmasının ortaya çıkardığı bulgular anlamlı sonuçlar verir.

Farklı nedenlere bağlı olarak ortaya çıkan bu sorun özellikle kestirim konusunda ciddi problemlere neden olabilmektedir. Bu nedenle çoklu bağlantı probleminin istatistiksel yöntemler üzerindeki etkileri incelenmeli, eğer önüne geçilemiyor ise en az etkilenen yöntem kullanılmalıdır. En az etkilenen yöntemin belirlenmesi benzetim çalışmaları ile mümkün olabilmektedir. Çalışmanın izleyen bölümünde buna uygun olarak yöntemler üzerinde çoklu bağlantı durumunun etkisini

gözlemleyebilmek için benzetim çalışmalarına başvurulmuştur.

6. MODEL YAPISI VE BENZETİM ÇALIŞMASI (MODEL STRUCTURE AND SIMULATION STUDY)

Bu aşamada, çalışmanın amacı doğrultusunda geliştirilen benzetim çalışmalarına ilişkin bilgiler verilmiştir. Buna göre birinci aşamada, çoklu bağlantının varlığında genelleştirilmiş doğrusal model ile toplamsal lojistik regresyon modellerin karşılaştırılması yapılmıştır. Son aşamada ise, önceki modeller R^2 kavramı temelinde CART modelleriyle karşılaştırılmıştır.

Benzetim çalışmasında çoklu bağlantının olduğu küçük, orta ve büyük ölçekli veri kümelerini temsil etmek üzere, 50, 100 ve 500 gözlem birimli oluşan rassal örneklemler ele alınmıştır. Korelasyonun tahminleme üzerindeki etkisini gözlemleyebilmek için her veri kümesi için yüksek ve düşük dereceli olmak üzere iki farklı korelasyon düzeyi seçilmiştir. Yazılan algoritma kullanıcının belirleyebileceği herhangi bir korelasyon değeri için toplamsal lojistik regresyon modeli tahmini vermesine rağmen bu çalışmada korelasyon değerleri ∓.90 ve ∓.10 ile sınırlandırmanın yeterli olacağı varsayılmıştır.

İlk olarak link fonksiyonu $l_1 = 0.5 + 3x_1 - 5x_2 + 6x_3$ ele alınmıştır. Burada x_1 ve x_2 açıklayıcı değişkenleri standart normal dağılımdan ve aralarında $r = (0.10, 0.90)$ düzeyde korelasyon olacak şekilde üretilmiştir. İkinci durumda, link fonksiyonu olarak $l_2 = 0.75 + 2x_1 + 0.5x_2 + 1.5x_3$ alınmıştır. Burada ilk modelden farklı olarak sadece x_i değişkeni bağımlı değişkeni pozitif yönde etkileyecek şekilde modelde yer almıştır. Üçüncü durumda ise x_1 ve x_2 arasında $r = (-0.10, -0.90)$ düzeyde negatif korelasyon üretilmiş ve link fonksiyonu olarak $l_3 = 0.75 + 2x_1 + 0.5x_2 + 1.5x_3$ alınmıştır. Son durumda link fonksiyonu $l_4 = 0.75 - 2x_1 + 0.5x_2 + 1.5x_3$ olmak üzere x_1 ve x_2 standart normal dağılımdan ve aralarında $r = (-0.10, -0.90)$ düzeyde negatif korelasyon olacak şekilde üretilmiştir. Tüm durumlar için $x_3 = x_1 + N(0,1)$ olarak alınmıştır. Bunun yanında, yanıt değişkeni $y_i \sim binom(n, z_i)$, $i = 1,2,3,4$ olacak şekilde örneklem hacminin üç farklı değeri için üretilmiştir. Analizler için hazırlanan algoritma R yazılımında yazılmıştır ve 1000 tekrardan elde edilen sonuçlar verilmiştir [13].

6.1. Genelleştirilmiş Doğrusal Model ve Toplamsal Lojistik Regresyon Model Karşılaştırması (Comparison of Generalized Additive Model and Additive Logistic Regression Model)

Yanıt değişkeninin normallik varsayımını sağlamadığı durumlarda Genelleştirilmiş Doğrusal Model (GDM) sıkça kullanılan bir yaklaşımdır. Varsayımsal esneklik bakımından kullanılan bir diğer yaklaşım ise toplamsal lojistik regresyon modelidir. Çoklu bağlantı probleminin varlığında doğrusal modeller ve GDM ile yapılan tahminler güvenilir sonuçlar vermemektedir. Bu yüzden çoklu bağlantı problemi ile karşılaşıldığında toplamsal lojistik regresyon modeli (TLRM) ile daha esnek bir model kurabilmenin daha uygun olacağı düşünülerek bir bilgisayar programı yazılmış, GDM ile toplamsal lojistik regresyon modellerinin performansları karşılaştırılmıştır.

Tablo 1' de, çoklu bağlantının olduğu dört ayrı durum için, örneklem hacmine ve değişkenler arasındaki korelasyon düzeylerine bağlı olarak kurulan genelleştirilmiş doğrusal regresyon modelleri ve toplamsal lojistik regresyon modellerinin 1000 tekrardan elde edilen ortalama AIC değerleri verilmiştir.

Düzeltici fonksiyonların farklılaştığı karşılaştırmalı çalışmalarda, çoğunlukla daha küçük AIC değerleri veren modeller, düzeltici fonksiyon olarak CS' nin kullanıldığı modellerden elde edilir [14]. Bu sonuçtan yararlanarak, toplamsal lojistik regresyon modelleri için düzeltici fonksiyon olarak CS kullanılmıştır. Varsayılan korelasyonlar ve durumlar için yanıt değişkeni ile açıklayıcı değişkenler arasında kurulan modellerin performansları Tablo 1' de özetlenmiştir.

Tablo 1. $n = 50,100,500$ ve $r = \mp 0.90, \mp 0.10$ için GDM ve TLRM AIC sonuçları (AIC for GDM and TLRM in case of $n = 50,100,500$ and $r = \mp 0.90, \mp 0.10$)

Durum	n	r	GDM	TRLM
	50	0.10	13.259	10.510
		0.90	18.394	13.535
1	100	0.10	25.260	18.766
		0.90	35.495	27.773
	500	0.10	116.020	113.914
		0.90	163.443	161.671
	50	0.10	32.449	21.799
		0.90	27.559	18.815
2	100	0.10	62.732	55.491
		0.90	53.475	46.257
	500	0.10	297.230	295.425
		0.90	251.234	249.378
	50	0.10	32.982	22.276
		0.90	38.967	26.937
3	100	0.10	63.895	56.822
		0.90	75.913	70.608
	500	0.10	302.705	300.882
		0.90	362.143	360.453
	50	0.10	51.528	36.379
		0.90	45.541	31.836
4	100	0.10	99.832	95.558
		0.90	88.639	83.312
	500	0.10	484.690	483.064
		0.90	427.597	426.027

Tablo 1' de ki sonuçlar incelendiğinde birinci durumda, örneklem hacmi sabit iken korelasyon parametresinin değeri arttığında, her iki yöntemden elde edilen modellerin AIC değeri yükselmiştir, ancak düşük AIC değeri toplamsal lojistik regresyon modelinden elde edilmiştir. Örneğin örneklem hacmi 50 için, $r = 0.10$ iken en düşük AIC değeri 10.510 olarak toplamsal lojistik regresyon modelinden elde edilmiştir. Benzer şekilde, tüm örneklem büyüklükleri için korelasyon parametresinin değeri arttığında AIC değeri de artmaktadır, ancak toplamsal lojistik regresyon modelinden elde edilen AIC değerleri daha küçüktür.

İkinci durumda, tüm örneklem büyüklükleri için korelasyon parametresinin değeri arttığında, AIC değerinin azaldığı gözlenmiştir. Örneklem hacmi 50 için $r = 0.90$ iken en düşük AIC değeri 18.815 olarak elde etmiştir. Örneklem hacminin arttırılmasıyla AIC değerinin de arttığı gözlenmiştir Örneklem hacmi sabit iken korelasyon parametresinin değeri arttırıldığında, her iki yöntemden elde edilen modellerin AIC değeri azalmıştır ancak düşük AIC değeri toplamsal lojistik regresyon modelinden elde edilmiştir.

Üçüncü durumda ise, ikinci durumda alınan model aynı kalmak üzere, değişkenler arasında negatif korelasyon olacak şekilde düzenlenmiştir. Buna göre, tüm örneklem büyüklükleri için korelasyon parametresinin değeri artırıldığında, AIC değeri de artmaktadır. Örneklem hacmi 50 için, $r = 0.10$ iken en düşük AIC değeri 22.276 olarak toplamsal lojistik regresyon modelinden elde edilmiştir. Örneklem hacmi sabit iken korelasyon parametresinin değeri arttırıldığında, her iki yöntemden elde edilen modellerin AIC değerleri yükselmiştir, ancak en düşük AIC değeri toplamsal lojistik regresyon modelinden elde edilmiştir.

Son durumda, tüm örneklem büyüklükleri için korelasyon parametresinin değeri arttığında, AIC değerinin azaldığı gözlenmiştir. Örneklem hacmi 50 için $r = -0.90$ iken en düşük AIC değeri 31.836 olarak elde etmiştir. Örneklem hacminin arttırılmasıyla AIC değerinin de arttığı gözlenmiştir. Örneklem hacmi sabit iken korelasyon parametresinin değeri artırıldığında, her iki yöntemden elde edilen modellerin AIC değerleri azalmıştır ancak düşük AIC değeri toplamsal lojistik regresyon modelinden elde edilmiştir.

6.2. Genelleştirilmiş Doğrusal Model, Toplamsal Lojistik Regresyon Modeli ve CART Karşılaştırması (Comparison of Generalized Additive Model, Additive Logistic Regression Model and CART)

Bu aşamada, GDM, TLRM ve CART ile uyumu yapılmış modellerin performanslarının karşılaştırılması hedeflenmiştir. Bunun için ortak bir ölçüt kullanılması önemlidir. Genellikle toplamsal lojistik modeller için bu ölçüt GCV kriteri olarak karşımıza çıkmaktadır [15]. Ancak bu kriter diğer yöntemler için kullanılmadığından karşılaştırma ölçütü olarak kullanılabilmesi mümkün olmamıştır. Bunun yerine, bağımsız değişkenlerin modeli açıklamadaki yeterliliklerinden yola çıkarak, R^2 istatistikleri üzerinde durulabilir. Modelin yeterliliği, lojistik regresyon için olabilirlik oran indeksi olarak da bilinen sözde (pseudo) R^2 ile ifade edilebilir [16]. Benzer şekilde ağaçlandırma algoritmalarında ve toplamsal modellerde de belirlilik katsayısı modelin açıklayıcılığına ilişkin olarak yorumlamalarda kullanılabilir [7].

Bu amaçla çoklu bağlantının var olduğu küçük, orta ve büyük ölçekli veri grupları için her üç yöntemin performansları R2 anlamında karşılaştırılmış ve sonuçlar Tablo 2' de verilmiştir.

Tablo 2. $n = 50,100,500$ ve $r = \mp 0.90,\ \mp 0.10$ için GDM, TLRM ve CART için R^2 sonuçları (R^2 for GDM, TLRM and CART in case of $n = 50,100,500$ and $r = \mp 0.90, \mp 0.10$)

Durum	n	r	GDM	TRLM	CART
	50	0.10	1	1	0.789
		0.90	0.846	0.999	0.666
1	100	0.10	0.874	0.966	0.830
		0.90	0.799	0.926	0.779
	500	0.10	0.839	0.855	0.823
		0.90	0.775	0.787	0.763
	50	0.10	0.637	0.961	0.745
		0.90	0.711	0.957	0.785
2	100	0.10	0.596	0.734	0.745
		0.90	0.666	0.788	0.785
	500	0.10	0.576	0.587	0.667
		0.90	0.645	0.655	0.725
	50	0.10	-0.10	0.629	0.955
		0.90	-0.90	0.536	0.914
3	100	0.10	-0.10	0.546	0.702
		0.90	-0.90	0.495	0.618
	500	0.10	-0.10	0.672	0.672
		0.90	-0.90	0.534	0.551
	50	0.10	-0.10	0.180	0.999
		0.90	-0.90	0.430	0.889
4	100	0.10	-0.10	0.386	0.450
		0.90	-0.90	0.394	0.536
	500	0.10	-0.10	0.311	0.314
		0.90	-0.90	0.390	0.394

Tablo 2' de çoklu bağlantının olduğu dört ayrı durum için, örneklem hacmine ve değişkenler arasındaki korelasyonun düşük ve yüksek düzeyine bağlı olarak kurulan genelleştirilmiş doğrusal regresyon modeli, toplamsal lojistik regresyon modeli ve sınıflandırma modelinin (CART) 1000 tekrardan elde edilen ortalama R2 değerleri verilmiştir. Sonuçlar incelendiğinde birinci

durum için, en yüksek R^2 toplamsal lojistik regresyon modelinden elde edilmiştir.. Örneklem değeri 500 iken, R^2 değerinin 50 ve 100 birimlik örneklemlerden elde edilen değerine göre daha azaldığı gözlenmiştir. Ancak 500 birimlik örneklemlerde, her üç yöntemle elde edilen modellerin R^2 oranlarının birbirlerine yaklaştığı gözlenmiştir (0.78; 0.77; 0.76).

İkinci durumda, tüm örneklem hacimleri için korelasyon değerinin artması, R^2 değerini azaltmıştır. Örneklem hacmi 50 iken $r = (0.10{,}0.90)$ için en yüksek R^2 değeri toplamsal lojistik regresyon modeliyle sırasıyla %96 ve %95 olarak elde edilmiştir. Örneklem değeri 500 iken, R^2 değerinin, 50 ve 100 birimlik örneklemlere göre azaldığı gözlenmiştir. Ancak toplamsal lojistik regresyon modeli ve genelleştirilmiş doğrusal regresyon modelinden elde edilen R^2 değerinin 500 birimlik örneklemlerde birbirlerine yaklaşırken, CART modelinden elde edilen R^2 değerinin daha yüksek olduğu gözlenmiştir.

Üçüncü durumda, tüm örneklem hacimleri için korelasyon değerinin artması, her üç yöntemden elde edilen modellerin R^2 değerini azaltmıştır. Örneklem değeri 500 iken, R^2 değerinin, 50 ve 100 birimlik örneklemlere göre daha azaldığı gözlenmiştir. Ancak toplamsal lojistik regresyon modeli ve genelleştirilmiş doğrusal regresyon modelinden elde edilen R^2 değerinin 500 birimlik örneklemlerde birbirine yaklaşırken, CART modelinden elde edilen R^2' nin daha yüksek olduğu gözlenmiştir.

Son durumda, korelasyon değerinin artması ve örneklem hacminin de arttırılması, her üç yöntemden elde edilen modellerin R^2 değerini belirgin ölçüde azaltmıştır. Örneğin örneklem değeri 50 için, $r = -0.10$ iken R^2 değeri %99 olarak toplamsal lojistik regresyon modelinde gözlenirken, örneklem sayısı 500 iken bu oran %31' e düşmüştür. Korelasyon değeri arttığında CART modelinden elde edilen R^2 değeri, diğer iki yöntemden elde edilen R^2 değerlerine göre az değişim göstermiştir.

7. SONUÇ VE TARTIŞMA (RESULT AND DISCUSSION)

Bu çalışmada, çoklu bağlantının varlığında ikili yanıt değişkeni ve açıklayıcı değişkenler arasındaki ilişkiyi açıklamak için genelleştirilmiş doğrusal regresyon, toplamsal lojistik regresyon ve karar ağaçları kullanılmıştır. Bunlardan toplamsal modellerde, değişkenler arasındaki ilişkiyi daha esnek bir hale getirmek için düzeltme fonksiyonları kullanılmıştır. Çoklu bağlantının varlığında, toplamsal lojistik regresyon yöntemi ile farklı örneklem büyüklükleri ve korelasyon düzeylerinde modeller kurarak bir benzetim çalışması düzenlenmiştir. Çalışmada her üç yöntemin karşılaştırılması sonucunda, CART modellerinden elde edilen R^2 değerinde, örneklem büyüklükleri ve korelasyon düzeyleri değiştiğinde önemli ölçüde düşüş veya yükseliş gözlenmemiştir. CART modelleri çoklu bağlantının varlığından daha az etkilenmiştir.

KAYNAKÇA (REFERENCES)

[1] A. Erar, "Çoklu bağlantı varlığında doğrusal regresyon modellerinde değişken seçimi" Ankara, Hacettepe Üniversitesi, İstatistik Bölümü, 1994.

[2] A. Erar, "Bağlanım (Regresyon) Çözümlemesi Ders Notları" İstanbul, Mimar Sinan Güzel Sanatlar Üniversitesi, 2006.

[3] B. Kan Kılınç, "Yanıt Yüzeyi Modellerine MARS Yaklaşımı", Eskişehir, Anadolu Üniversitesi, İstatistik Bölümü, 2010.

[4] Y. Kaşko, "Çoklu Bağlantı Durumunda İkili Lojistik Regresyon Modelinde Gerçekleşen 1.Tip Hata ve Testin Gücü", Ankara, Ankara Üniversitesi, Biyometri ve Genetik Anabilim Dalı, 2007.

[5] G. Wahba and J. Wendelberger, "Some new mathematical methods for variational objective analysis using splines and cross validation", Monthly Weather Review, vol.108, pp. 1122-1145, 1980.

[6] S. Wood, "Generalized Additive Models: An introduction to R", Chapman and Hall/CRC, 2006.

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and Regression Trees", Wadsworth, 1984.

[8] H. Christian, "Smoothing by spline functions", Journal of Numerische Mathematic, vol.10, no.3, pp. 177-183, 1967.

[9] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces", Constructive Theory of Functions of Several Variables, Springer, 1977.

[10] R. De Veaux and L. Ungar, "Multicollinearity: A tail of two nonparametric regressions", Lecture Notes in Statistics: Selecting Models from Data, pp. 393-402, 2007.

[11] M. Hutchinson and R. Bischof, "A new method for estimating the spatial distribution of mean seasonal and annual rainfall applied to the Huner Valley, New South Wales", Australian Meteorological Magazine , vol.31, no.3, pp.179-184, 1983.

[12] T. Hastie, R. Tibshirani and F. Friedman, "The Elements of Statistical Learning", Springer, 2009.

[13] S. Kovalchik and R. Varadhan, "Fitting additive binomial regression models with the R package blm", Journal of Statistical Software, vol.54, no.1, pp.1-18, 2013.

[14] L. Ma and X. Yan, "Examining the nonparametric effect of drivers' age in rear-end accidents through an additive logistic regression model", Accident Analysis and Prevention, vol.67, pp.129-136, 2014.

[15] D. McFadden, "Conditional logit analysis of qualitative choice behavior", Frontiers in Econometrics ,Academic Press, pp.105-142, 1974.

[16] J. Meinguet, "Multivariate interpolation at arbitrary points made simple", Journal of Applied Mathematics and Physics, vol.30, pp.370-384,1979.

[17] C. Montgomery, E. Peck and G. Vining, " Introduction to Linear Regression Analysis", Wiley, 2012.

[18] W. Press, B. Flannery, S. Teukolsky and W. Vetterling, "Cubic Spline Interpolation. The Art of Scientific Computing", Cambridge University Press, 1992.

[19] S. Silvey, "Multicollinearity and imprecise information", Journal of Royal Statistics Society vol.31, pp.539-552, 1969.

[20] J. Shen and S. Gao, "A solution to seperation and multicollinearity in multiple logistic regression", Journal of Data Science, vol.6, no.4, pp.515-531, 2008.

[21] B. Ripley, "Pattern Recognation and Neural Networks", Cambridge University Press, 1996.