

**T.C.
ONDOKUZ MAYIS ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
ZOOTEKNİ ANA BİLİM DALI**



**GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLERİN
ZOOTEKNİDE KULLANIMI**

Yüksek Lisans Tezi

Nursen KURDAL

Danışman

Prof. Dr. Hasan ÖNDER

SAMSUN
2022

TEZ KABUL VE ONAYI

Nursen KURDAL tarafından, Prof. Dr. Hasan ÖNDER danışmanlığında hazırlanan “Genelleştirilmiş Toplamsal Modellerin Zooteknide Kullanımı” başlıklı bu çalışma, jürimiz tarafından 21.1.2022 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Yüksek Lisans Tezi olarak kabul edilmiştir.

	Unvanı Adı Soyadı Üniversitesi Ana Bilim/Ana Sanat Dalı	İmza	Sonuç
Başkan	Prof. Dr. Hasan ÖNDER Ondokuz Mayıs Üniversitesi Zootekni Anabilim Dalı		<input checked="" type="checkbox"/> Kabul <input type="checkbox"/> Ret
Üye	Doç. Dr. Samet Hasan ABACI Ondokuz Mayıs Üniversitesi Zootekni Anabilim Dalı		<input checked="" type="checkbox"/> Kabul <input type="checkbox"/> Ret
Üye	Dr. Öğr. Üyesi Aycan Mutlu YAĞANOĞLU Atatürk Üniversitesi, Ziraat Fakültesi, Zootekni Ana Bilim Dalı		<input checked="" type="checkbox"/> Kabul <input type="checkbox"/> Ret

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

ONAY
... / ... / ...
Prof. Dr. Ali BOLAT
Enstitü Müdürü

BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım Yüksek Lisans tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar'da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

Etik Kurul Gerekli mi ?

Evet (Gerekli ise ekler kısmına ekleyiniz)

Hayır


21 /01/ 2022
Nursen KURDAL


TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

Tez Başlığı: Genelleştirilmiş Toplamsal Modellerin Zooteknide Kullanımı

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 07 Aralık 2021 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 21

Tek kaynak oranı : % 4 çıkmıştır.


21 /01/ 2022
Prof. Dr. Hasan ÖNDER

ÖZET

GENELLEŞTİRİLMİŞ TOPLAMSAL MODELLERİN ZOOTEKNİDE KULLANIMI

Nursen KURDAL

Ondokuz Mayıs Üniversitesi

Lisansüstü Eğitim Enstitüsü

Zootekni Ana Bilim Dalı

Yüksek Lisans, Ocak/2022

Danışman: Prof. Dr. Hasan ÖNDER

Regresyon analizi pek çok bilim dalında olduğu gibi zootekni alanında da en sık kullanılan istatistik yöntemlerden birisidir. Ancak pek çok biyolojik çalışma için doğrusal regresyon varsayımlarının sağlanması mümkün olmamaktadır. Özellikle kesikli verilerin açıklayıcı değişken olarak kullanılması gereken durumlarda doğrusal regresyon kullanılarak elde edilen modellerin istatistiksel olarak hatalı sonuçlar üretebileceği bilinmektedir. Bu gibi durumlar için doğrusal regresyon yerine parametrik olmayan ya da yarı parametrik yöntemlerin kullanılması önerilmektedir. Bu çalışmada, kesikli açıklayıcı değişken varlığında önerilen genelleştirilmiş toplamsal modellerin zootekni alanından elde edilmiş bir veri kümesi kullanılarak tanıtılması amaçlanmıştır. Sonuç değişkeni olarak süt keçilerinden alınan laktasyon süt verimi kullanılırken açıklayıcı değişkenler olarak laktasyon süresi ve ana yaşı kullanılmıştır (n=166). Laktasyon süt verimi, laktasyon süresi ve ana yaşı değişkenlerinin normal dağılım göstermediği belirlenmiştir. Elde edilen doğrusal regresyon modelinin istatistiksel olarak anlamlı olduğu ($P<0,001$), Hata Kareler Ortalamasının (HKO) 5826,039 ve belirtme katsayısının (R^2) 0,16 olduğu belirlenmiştir. Düşük bir belirtme katsayısına rağmen modelin anlamlı olmasının örnek büyüklüğüne bağlı olabileceği yorumlanmıştır. Elde edilen genelleştirilmiş toplamsal model için Ana yaşı değişkeninin laktasyon süresi değişkenine göre daha başarılı düzeltme değerine sahip olduğu belirlenmiştir. Doğrusal regresyon ve genelleştirilmiş toplamsal modelleri karşılaştırabilmek adına Root Mean Square Error (RMSE), Relative Root Mean Square Error (rRMSE) ve ortalama mutlak sapma (MAD) değerleri hesaplanmıştır. Elde edilen bulgulara göre doğrusal regresyon analizinin tahmin başarısının daha yüksek olduğu anlaşılmaktadır. Ancak açıklayıcı değişkenler arasında çoklu bağlantı problemi olmamasına ($VIF=1.000$) ve sapmaların normal dağılıma sahip olmasına ($P>0,05$) rağmen değişkenlerin normal dağılmadığı bilinmektedir. Bu durumda genelleştirilmiş toplamsal modellerin kullanılmasının önerilebileceği söylenebilir. Sonuç olarak zootekni çalışmalarında kesikli değişkenler ile model oluşturmak gerektiğinde doğrusal modeller yerine genelleştirilmiş toplamsal modellerin kullanılmasının elde edilecek modelin güvenilirliğini artıracığı için önerilebileceği belirlenmiştir.

Anahtar Sözcükler: Genelleştirilmiş toplamsal modeller, Regresyon, Kesikli veri, Zootekni

ABSTRACT

USE OF GENERALIZED ADDITIVE MODELS IN ANIMAL SCIENCE

Nursen KURDAL

Ondokuz Mayıs University

Institute of Graduate Studies

Department of Animal Science

Master, January/2022

Supervisor: Prof. Dr. Hasan ÖNDER

Regression analysis is one of the most frequently used statistical methods in the field of animal science, as it is in many branches of science. However, it is not possible to provide linear regression assumptions for many biological studies. It is known that especially in cases where discrete data should be used as explanatory variables, the models obtained using linear regression can produce statistically incorrect results. For such cases, it is recommended to use non-parametric or semi-parametric methods instead of linear regression. In this study, it is aimed to introduce the generalized additive models proposed in the presence of discrete explanatory variables, using a dataset obtained from the field of animal science (n=166). While lactation milk yield was used as the outcome variable, lactation period and maternal age taken from dairy goat were used as explanatory variables. It was determined that the variables of lactation milk yield, lactation period and maternal age did not show normal distribution. It was determined that the obtained linear regression model was statistically significant ($P < 0.001$), the Mean Square Error (MSE) was 5826.039 and the coefficient of determination (R^2) was 0.16. It has been interpreted that the significance of the model, despite a low coefficient of determination, may depend on the sample size. For the generalized additive model obtained, it was determined that the maternal age variable had a more successful smoothing value than the lactation period variable. Root Mean Square Error (RMSE), Relative Root Mean Square Error (rRMSE) and mean absolute deviation (MAD) values were calculated to compare linear regression and generalized additive models. According to the findings obtained, it was understood that the estimation success of the linear regression analysis was higher. However, it was known that although there was no multicollinearity problem between the explanatory variables ($VIF=1.000$) and the deviations had a normal distribution ($P > 0.05$), the variables were not normally distributed. In this case, it can be said to recommend to use of generalized additive models. As a result, it has been determined that when it is necessary to create a model with discrete variables in animal science studies, using generalized additive models instead of linear models can be recommended as it will increase the reliability of the model to be obtained.

Keywords: Generalized additive models, Regression, Discrete data, Animal science

ÖNSÖZ VE TEŞEKKÜR

Bu çalışmamın yürütülmesi sırasında desteğini esirgemeyen değerli danışmanım Prof. Dr. Hasan ÖNDER'e, beni büyütüp bu günlere getiren canım annem Yıldız KURDAL'a ve canım babam Halit KURDAL'a, her zor anımda yanımda olup beni yeniden ayağa kaldırıp yüreklendiren sevgili arkadaşım Ondokuz Mayıs Üniversitesi Coğrafya Bölümü yüksek lisans öğrencisi Kamile AKKOYUN'a, çalışmam boyunca desteğini hep hissettiğim çok sevdiğim canım arkadaşım Ondokuz Mayıs Üniversitesi Bitki Koruma Bölümü lisans öğrencisi Emre YILMAZ'a, büyük küçük destek veren herkese teşekkür ederim.

Nursen KURDAL

İÇİNDEKİLER

ÖZET	iii
ABSTRACT	iv
ÖNSÖZ VE TEŞEKKÜR	v
İÇİNDEKİLER	vi
SİMGELER VE KISALTMALAR	vii
ŞEKİLLER DİZİNİ	viii
TABLolar DİZİNİ	ix
1. GİRİŞ	1
2. MATERYAL VE METOT	6
2.1. Materyal.....	6
2.2. Metot.....	6
2.3. Düzleştirici Fonksiyonlar.....	9
2.3.1. Kübik Splaynlar.....	9
2.3.2. Yerel Regresyon.....	10
2.3.3. Thin-Plane Zincir (TPSPLAYN) Prosedürü.....	10
2.4. Düzleştirme Parametresinin Seçimi.....	11
2.4.1. Çapraz Geçerlilik (CV) ve Genelleştirilmiş Çapraz Geçerlilik (GCV).....	11
2.5. GTM Nasıl Yorumlanmalı?.....	12
2.6. Analiz için Kullanılan SAS Kodları.....	13
3. BULGULAR VE TARTIŞMA	14
4. SONUÇ	17
KAYNAKLAR	18

SİMGELER VE KISALTMALAR

GTM	: Genelleştirilmiş Toplamsal Modeller
GDM	: Genelleştirilmiş Doğrusal Modeller
LSV	: Laktasyon Süt Verimi
LS	: Laktasyon Süt Verimi
AY	: Ana Yaşı
EKK	: En Küçük Kareler
GCV	: Generalized Cross Validation (Genelleştirilmiş Çapraz Geçerlilik)
CV	: Cross Validation (Çapraz Geçerlilik)

ŞEKİLLER DİZİNİ

Şekil 3.1. İncelenen deęişkenlere ait daęılım grafikleri	14
Şekil 3.2. Ana yaşı ve laktasyon süresine ait düzleştirme grafikleri	16

TABLULAR DİZİNİ

Tablo 3.1. Doğrusal regresyon analizi ve parametre tahminleri.....	14
Tablo 3.2. GTM modeli için düzeltirme bileşenleri ve uyum istatistikleri	15
Tablo 3.3. GTM analiz istatistikleri (sapma analizi)	15
Tablo 3.4. Modeller için karşılaştırma istatistikleri.....	16

1. GİRİŞ

Regresyon analizi, aralarında sebep sonuç ilişkisi bulunan değişkenler arasındaki ilişkiyi incelemek, tespit etmek, modellemek ve bu modeli kullanarak ilgili konu üzerinde tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla kullanılan istatistiksel yöntemlerden birisidir (Arı ve Önder, 2013). Regresyon analizi, hemen hemen tüm bilim dallarında en çok kullanılan yöntemlerden birisidir. Regresyon analizinin metodolojisi, sebep değişkenlerin sonuç değişken üzerindeki etkisini ve katkısını değerlendirerek sonuç değişkeni ve açıklayıcı değişkenler arasındaki ilişkiyi ortaya koymak ve modellemektir (Durmuş, 2018). Birçok biyolojik olayda sebep sonuç ilişkisine rastlamak mümkündür. Örneğin; yem miktarının (sebebe), süt miktarına (sonuç) etkisi bir örnek olarak verilebilir. Bu analiz tekniğinde iki (basit doğrusal regresyon) veya daha fazla değişken (çoklu doğrusal regresyon) arasındaki ilişkiyi açıklamak için matematiksel bir model kullanılır. Zootekni alanında da verilerin değerlendirilmesi işleminde değişkenler arası ilişkilerin incelenmesi amacıyla en çok kullanılan istatistiksel yöntemlerden birisi çoklu doğrusal regresyon analizidir.

Çoklu regresyon analizi, bağımlı değişkenle ilişkili olan iki ya da daha fazla açıklayıcı değişkene dayalı olarak, bağımlı değişkenin tahmin edilmesine yönelik bir analiz yöntemidir. Açıklayıcı değişkenler tarafından sonuç değişkenide açıklanan toplam varyansın yorumlanmasına ve açıklayıcı değişkenlerle sonuç değişkeni arasındaki ilişkinin yönüne ve miktarına dair yorum yapma imkanı verir. (Kanıt ve Baykan, 2004).

Doğrusal regresyon modelleri için açıklayıcı değişkenlerin birbirlerinden bağımsız olduğu ve benzer dağılıma sahip oldukları varsayımı kabullenilmektedir. Parametrik regresyon analizinin bir diğer varsayımı ise; regresyon katsayısının test edilmek istenmesi durumunda hata terimlerinin normal dağılıma sahip olmasıdır. Bu varsayımlar altında regresyon katsayısının En Küçük Kareler (EKK) tahmin edicisi $n-k$ serbestlik dereceli t dağılımına sahiptir (burada, n ve k sırasıyla örnek büyüklüğü ve açıklayıcı değişken sayısını göstermektedir). Ayrıca, merkezi limit teoreminden bilindiği gibi regresyon katsayısı optimum örnek büyüklüğüne ulaşıldığında normal dağılıma sahip olmaktadır. Ancak, hata terimlerinin normal dağılıma sahip olmadığı ve örnek büyüklüğünün yeterli düzeye ulaşmadığı durumlarda, EKK tahmin edicisi hatalı sonuçlar üretebilmektedir (Önder, 2007). Yukarıda bahsi geçen özelliklere ek

olarak, çoklu doğrusal regresyon analizi için temel bir varsayım olan, açıklayıcı değişkenle sonuç değişkeni arasındaki ilişkilerin doğrusal, karesel veya küresel gibi tanımlı bir matematiksel fonksiyon olması varsayımının da sağlanması gerekir. Bu varsayımlar sağlandığında, doğrusal regresyon analizi oldukça güçlü bir parametrik yöntemdir. Ancak biyolojik çalışmalarda elde edilen örneklere matematiksel fonksiyon tanımlamak çok zor olmakta hatta bazı durumlarda mümkün olmamaktadır. Bu çerçevede sonuç değişkeni ve açıklayıcı değişkenler arasındaki ilişkinin fonksiyonel şekli biliniyorsa parametrik regresyon modelleri, bilinmiyorsa parametrik olmayan regresyon modelleri kullanılır (Durmuş, 2018). Ancak, veri yapısına uygun olmayan regresyon yöntemlerinin kullanılması yanıltıcı sonuçların elde edilmesine neden olabilmektedir. Regresyon analizinde üzerinde çalışılan değişkenler sürekli ya da kesikli yapıda olabilmektedir. Bu durumda, veri yapısına bağlı olarak farklı regresyon modelleri kullanılabilir (Arı ve Önder, 2013).

Sonuç değişkeninin sürekli olmadığı durumlarda regresyon analizi uygulayabilmek için de geliştirilmiş modellerde bulunmaktadır. Değişkenlerin sürekli olduğu ancak normal dağılım göstermediği durumlarda veri analizi için geliştirilen Genelleştirilmiş Doğrusal Modeller (GDM) kullanılmaktadır (Savaş ve Cengiz, 2009). Şenel, vd. (2009)'nın bildirdiğine göre; GDM, ilk kez Nelder ve Wedderburn (1997) tarafından ileri sürülmüştür. Çok geniş uygulama alanlarında kullanılmıştır. Bu alandaki ilk detaylı kitaplar McCullagh ve Nelder (1989), Aitkin, vd. (1989) ve Dobson(1990) tarafından yazılmıştır. Cengiz ve Percy (2001) genelleştirilmiş doğrusal modellerin genel bir özetini vermektedir.

Doğrusal modeller de her bir tahmin edicide sonuç değişkenini doğrusal kabul ederken toplamsal modeller de ise sadece düz bir şekilde bağımlı değişkenin her bir tahmin edici tarafından etkilediğini varsayılmaktadır. Genelleştirilmiş Toplamsal Modeller (GTM), tahmin ediciler ve bağımlı değişkenler arasındaki esnek toplamsal olmayan ilişkileri modelleyen, doğrusal modeller ve GDM'in genişletilmiş bir ifadesidir (Savaş ve Cengiz, 2009). Doğrusal regresyon modellerinin genelleştirilmiş doğrusal modellerin uzantısı olması gibi genelleştirilmiş toplamsal modeller de toplamsal modellerin bir uzantısıdır. Bu tür modellerin veriye uyumu için kullanılan değişik yaklaşımlardan biri olan düzleştirme, model matrisinde çeşitli düzleştirme fonksiyonlarının kullanıldığı yaklaşımlardandır (Asfha, 2017).

Ağırlıklı en küçük kareler kullanılarak oluşturulan ve kesin bir çözüme sahip

olan doğrusal regresyon modellerinin aksine, bir GTM (veya bir GDM) için tahmin yöntemi, optimal tahminleri bulmak için yinelemeli (iteratif) yaklaşımlar gerektirir. GTM'de tahmin, yerel puanlama (local scoring) algoritması ile geriye uydurma (backfitting) algoritmasının bir bileşimine dayanır. Yerel puanlama algoritması, GDM'de maksimum olabilirlik tahminlerini bulmak için Fisher puanlama prosedürünün bir genellemesidir. Geri uydurma algoritması, herhangi bir toplamsal modeline uydurmak için uygundur ve GTM'de, modele birkaç düzgün fonksiyon dahil edildiğinde yerel puanlama yinelemesinde kullanılır. GTM, mevsimsellik, eğilimler ve hava değişkenlerinin doğrusal olmayan karıştırıcı etkileri için parametrik olmayan düzeltmelere izin verdiği için en yaygın olarak uygulanan yöntem olmuştur ve parametrik alternatiflere göre daha esnek bir yaklaşımdır. GTM, birçok zaman serisi analizinde yaygın olarak kullanılmaktadır (Dominici, vd., 2002).

Literatür incelendiğinde;

Omay (2007)'ın çalışmasında Los Angeles'daki hava kirliliği ile kaydedilen ölüm oranları arasındaki ilişki incelenmiştir. Gözlenen ölüm oranlarına karşılık gelen ozon seviyeleri, karbon monoksit seviyeleri, partikül seviyeleri, ortalama sıcaklık ve zaman ölüm üzerine etkili açıklayıcı değişken olarak belirlenmiştir. Hava kirliliğinin ölüm oranını nasıl etkilediğini test etmek için farklı regresyon modelleri kullanılmış. Modeller sırasıyla doğrusal model, GDM, toplamsal model, GTM ve ince tabakalı splayn ile GTM'dir. Elde edilen sonuçlar incelendiğinde geliştirilmiş toplamsal modelin istatistiksel olarak daha anlamlı sonuçlar verdiği ve semiparametrik toplamsal modele göre daha açıklayıcı olduğu görülmüştür.

Savaş ve Cengiz (2009) çalışmalarında zeka testleri yapılan bir merkezde 118 kişiye yapılan anket çalışmasında kişinin hayat memnuniyet derecesi üzerine etkisini incelemek için yaş, zeka testi, eğitim, aylık gelir, ağırlık gibi değişkenlere ikili lojistik regresyon uygulamışlar ve sonucunda yaşın anlamlı olmadığı görülmüştür. Bu ilişkiyi daha az sınırlı hale getirmek için GTM kullanılmış ve Ki-kare test değerine bakıldığında %5 anlamlılık seviyesinde gelirin etkisi anlamsız çıkmıştır. Yapılan analizlerden elde edilen sonuçta zeka, ağırlık ve yaş etkileri kuadratik alınırken gelirin etkisi doğrusal olarak tekrar test edilmiş ve parametreler %5 anlamlılık seviyesinde anlamlı bulunmuştur.

Ma ve Yan (2014)'ın 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 incelemişlerdir. İnceleme sırasında sürücü yaşı değişkeni anlamlı bir etkiye sahip olup, sürücü yaşı küçüldükçe çarpan (hatalı) taraf olma olasılığının arttığı gözlemlenmiştir.

Kan Kılınç ve Çavuş (2017) yapmış oldukları çalışmada, çoklu bağlantının varlığında, genelleştirilmiş doğrusal model, toplamsal lojistik regresyon model ve sınıflandırma ağaçlarının bir benzetim çalışması ile küçük, orta ve büyük veri kümeleri için modellerin yeterliliğini, bahsedilen yöntemler yardımıyla ortaya koymayı amaçlamışlardır. Sonuç olarak toplamsal modellerden elde edilen sonuçların CART modellerinde düzeltme fonksiyonlarının etkisiyle başarılı sonuçlar elde etmiş olduklarını bildirmişlerdir.

Kovalchik ve Varadhan (2013) çalışmasında toplamsal lojistik modellerin genişletilmiş bir hali olan toplamsal binom modellerini kullanarak mesane kanseri üzerine bir veri seti kullanarak R istatistik programının “blm” paketinin kullanımını göstermişlerdir.

Omay (2014)'in çalışmasında Dünya Bankası, Dünya Kalkınma Göstergeleri'nden alınan verilerle çevresel göstergeler olarak CO₂ ve PM₁₀ emisyonları, ekonomik göstergeler olarak kişi başına düşen GSYH (Gayri safi yıllık hasıla) ve enerji kullanımı, sağlık göstergesi olarak ise kadınlar ve erkeklere ilişkin ölüm oranları alınmıştır. Bu değişkenler dikkate alınarak regresyon modelleri oluşturulmuş ve OECD (Ekonomik Kalkınma ve İşbirliği Örgütü)'ye üye ülkeler kullanılmıştır. Yapılan çalışmada OECD ülkeleri için PM₁₀, CO₂ çevresel göstergelerin ve enerji kullanımı, GSYH ekonomik göstergelerin sağlık üzerindeki etkisi modellenmiş ve farklı varyanslılık problemi gözlenmiş bu ortadan kaldırmak için panel genelleştirilmiş toplamsal model (PGAM) kullanılarak problemin aşıldığı görülmüştür.

Şenel vd. (2009)'ın Samsun il merkezinde hava kirliliğine neden olan SO₂ ölçümlerindeki değişimin nispi nem, ortalama rüzgar hızı, ortalama sıcaklık, düşen yağış miktarı gibi değişkenler ile ilişkisini ortaya koymak için çoklu doğrusal regresyon modeli ve GAM karşılaştırılmıştır. Çoklu doğrusal regresyon modelinde SO₂'yi açıklamada ortalama nispi nem ve aylık yağış miktarı anlamsız olduğu

görülmüştür. GAM kullanılarak açıklayıcı değişkenler arasındaki ilişkinin ortalama rüzgar hızı ve ortalama sıcaklığın doğrusal, ortalama nispi nem ve aylık yağışın kuadratik olduğu ve %5 anlamlılık seviyesinde tüm açıklayıcıların istatistiksel olarak anlamlı olduğu belirlenmiştir.

Dominici vd. (2002) hava kirliliği ve sağlık ilişkisini inceledikleri çalışma sonucunda analiz bulgularının, epidemiyolojik veya diğer araştırmalarda GTM ve diğer parametrik olmayan regresyon tekniklerinin faydasını ortaya koymuştur. Çalışmalarında tanımlanan sorunlara karşı önlem almak için, yakınsama kriterleri, önemli ölçüde daha güvenilir hale getirilmesi gerektiğini bildirmişlerdir. Analizleri gerçekleştirmek için S-Plus yazılımında “gam” fonksiyonunu kullandıklarını belirtmişlerdir.

Bu çalışmada geliştirilmiş toplamsal modellerin tanıtılması ve zootekni alanında bir örnek ile doğrusal regresyon analizi ile karşılaştırılması amaçlanmıştır.

2. MATERYAL VE METOT

2.1. Materyal

Bu çalışmada veri olarak 2014 yılında tamamlanan bir doktora tezi için alınan veriler (Tozlu Çelik, 2014) danışman onayı alınarak kullanılmıştır. Sonuç değişkeni olarak süt keçilerinden alınan laktasyon süt verimi (kg) alınırken açıklayıcı değişkenler olarak laktasyon süresi (gün) ve ana yaşı (yıl) kullanılmıştır. Verilerin analizinde SAS yazılımı kullanılmıştır.

2.2. Metot

Regresyon analizi geniş bir tanım yapıldığında değişkenler arasında sebep sonuç ilişkisinin araştırıldığı bir istatistiksel yöntemdir (Arı ve Önder, 2013; Bağdatlı, 2010). Temel amaçlar ise en genel biçimde, sonuç değişkenindeki değişimi açıklama, herhangi bir gözleme karşılık gelen ortalama y değerini bulma, noktalar için en iyi eğriyi uydurma olarak sıralanabilir (Bağdatlı, 2010).

Geleneksel çoklu doğrusal model;

$$E(Y|X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

olarak gösterilebilir (Arı ve Önder, 2013; Kan Kılınç ve Çavuş, 2017). Bu model oldukça basit olmasına rağmen, varsayımlarının sağlanması durumunda çok güvenilir bir yöntemdir. Doğrusal regresyon modelleri için açıklayıcı değişkenlerin birbirlerinden bağımsız olduğu ve benzer dağılıma sahip oldukları varsayımı kabullenilmektedir. Parametrik regresyon analizinin bir diğer varsayımı ise; regresyon katsayısının test edilmek istenmesi durumunda hata terimlerinin normal dağılıma sahip olmasıdır. Bu varsayımlar altında regresyon katsayısının En Küçük Kareler (EKK) tahmin edicisi $n-k$ serbestlik dereceli t dağılımına sahiptir (burada, n ve k sırasıyla örnek büyüklüğü ve açıklayıcı değişken sayısını göstermektedir). Ayrıca, merkezi limit teoreminden bilindiği gibi regresyon katsayısı optimum örnek büyüklüğüne ulaşıldığında normal dağılıma sahip olmaktadır. Ancak, hata terimlerinin normal dağılıma sahip olmadığı ve örnek büyüklüğünün yeterli düzeye ulaşmadığı durumlarda, EKK tahmin edicisi hatalı sonuçlar üretebilmektedir (Önder, 2007). Biyolojik örneklerin genellikle doğrusal olmaması nedeniyle, daha esnek bir yaklaşım olan istatistiksel yöntemlerden parametrik olmayan yöntemler ve genelleştirilmiş toplamsal modeller alternatif olarak kullanılabilir (Hastie ve Tibshirani, 1990; Kan

Kılınç ve Çavuş, 2017). Parametrik olmayan çoklu regresyon modellerinin tahmininde çok boyutluluğun yarattığı güçlükler (curse of dimensionality) nedeniyle zorluklar yaşanabilmektedir. Bu sorunun çözümü için toplamsal modeller kullanılabilir (Savaş ve Cengiz, 2009).

Genelleştirilmiş toplamsal modeller (GTM) şu şekilde gösterilir (Hastie ve Tibshirani, 1990; Savaş ve Cengiz, 2009);

$$E(Y|X_1, X_2, \dots, X_k) = f_1(X_1) + \dots + f_k(X_k)$$

- X_i : Bağımsız değişken
- Y : Bağımlı değişken
- f_i : Düzleştirici fonksiyonlar (parametrik olmayan) gösterir.

Burada X_i 'nin rastgele seçilen fonksiyon için terimler $f_i(X_i)$, doğrusal denklemdeki $\beta_i X_i$ terimlerine karşılık gelmektedir. Buradaki yaklaşımın farklılığı, her fonksiyona serpilme diyagramı düzleştiricisinin uydurulması ve bu şekilde her f_i fonksiyonunu tahmin edilmesini sağlamaktır (Bağdatlı, 2010). Bu modeller, bağımlı değişkenin ortalamasının, doğrusal olmayan bir bağlantı işlevi aracılığıyla toplamsal bir tahmin ediciye bağlı olduğunu varsayar. Genelleştirilmiş toplamsal modeller GTM'i doğrusal $\sum_j X_j \beta_j$ ifadesini $\sum_j f_j(X_j)$ ile ikame ederek genişletir. GTM'deki lineer regresyon adımı, verilerin uygun düzgün fonksiyonu f belirlemek için kullanıldığı, parametrik olmayan bir eklemeli regresyon adımı ile değiştirilir. Bu, yinelemeli (iteratif) düzleştirme işlemleri aracılığıyla yapılır ve açıklayıcı değişkenlerin çeşitli doğrusal olmayan etkilerine izin verir (Berg, 2007).

Geleneksel bağlantı fonksiyonları ve ilgili genelleştirilmiş toplamsal modeller şu şekildedir.

- $g(\mu) = \mu = \sum_j f_j(x_j)$, geleneksel doğrusal modelin toplamsal genişlemesi
- $g(\mu) = \log(\mu) = \sum_j f_j(x_j)$, poisson için logaritmik toplamsal model
- $\log\left[\frac{\mu}{1-\mu}\right] = \alpha + f_1(x_1) + \dots + f_p(x_p)$ toplamsal logistik regresyon modeli

Yukarıda verilen ifadeler negatif binom ve gamma dağılımını içeren üssel ailesinden örnek modellerinden oluşturulabilir. Bunlar genelleştirilmiş toplamsal modeller ile uyumlu şekilde genişletilmiş doğrusal modeller olarak bilinen sınıfı oluştururlar (McCullagh ve Nelder, 1989).

f_j fonksiyonları bir algoritma için temel yapı bloklarını meydana getiren serpilme

diyagramını kullanarak daha esnek bir şekilde tahmin edilebilir. x_j 'nin etkisinde tahmin edilen $\hat{f}_j(x_j)$ fonksiyonu genellikle doğrusal olmayan durumlar ortaya koyabilir. Burada sürekli ve bazı değişkenlerin kesikli olduğu durumlar için diğer parametrik durumlarda da kolaylıkla kıyaslanabilir olduğunu göstermeyi amaçlamaktadır. Doğrusal olmayan ifadeler ana etkilerle sınıflandırılmamaktadır. İki ya da daha fazla değişken için doğrusal olmayan bileşenler mevcut olabilir ya da kesikli faktörün her seviye düzeyinde eğrilerin kategorize edilebilir özellikte olabilir. Sonuç olarak sınıflandırma amacıyla aşağıda verilenlerden biri kullanılabilir;

- $g(\mu)=X\beta+\alpha_k+f(z)$: Semi-parametrik bir model; burada X doğrusal olarak modellenirken tahmin ediciler için bir faktör olan, α_k , kesikli bir faktörün k 'nci derecesi için etki ve tahmin edici Z 'nin etkisi non-parametrik şekilde modellenir.
- $g(\mu)=f(x)+g_k(z)$: Burada, k bir faktörü için seviye sayısı ve sonuçta k ve Z 'nin bir etkisi için bir etkileşim ifadesini ortaya koyar.
- $g(\mu)=f(x)+g(z,w)$: Burada g iki değişken için non-parametrik bir fonksiyon olarak tanımlanabilir.

Genel doğrusal olmayan modeller şu şekilde gösterilir.

$$Y=(X_1, X_2, \dots, X_p)+\varepsilon$$

- X_1, X_2, \dots, X_p : Açıklayıcı değişken
- ε : rastgele hata
- f : $R^p \rightarrow R$ keyfi seçilen düzgün bir fonksiyon

Ancak $P>2$ olduğu durumlarda doğrusal olmayan model için uyumsuzluk gösterir. Ayrıca açıklayıcı değişken sayısının fazla olması doğrusal olmayan regresyonu yorumlamayı zorlaştırır. Toplamsal modelde; f_j 'yi tahmin etmek için kullanılan X_j 'nin rastgele seçilen fonksiyonuna ait terimler olan $f_j(X_j)$ doğrusal denklemdeki $\beta_j X_j$ 'nin yer değiştirmesidir.

$$Y=\alpha+\sum_{j=1}^p f_j(X_j)+\varepsilon$$

Bağımlı değişken Y , bağımsız değişkenleri X_1, X_2, \dots, X_p 'ye kadar olan toplamsal

model, yukarıdaki forma sahiptir. Hatalar burada X_j 'den bağımsız, $E(\varepsilon)=0$ ve $\text{Var}(\varepsilon)=\sigma^2$ 'dir.

$$E(Y|X_1, X_2, \dots, X_p) = \sum_{j=1}^p f_j(X_j)$$

$E(Y)$ üzerinde X_j 'nin etkisi diğer açıklayıcı değişkenlerin her değeri için benzerdir. Y bir bağımlı değişken olsun ve X_1, X_2, \dots, X_p açıklayıcı değişkenler olarak tanımlandığında standart doğrusal regresyon modeli şu şekilde;

$$E(Y|X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

koşullu beklentisi için doğrusal bir form önerir. $\beta_1, \beta_2, \dots, \beta_p$ 'nin tahminleri EKK yöntemi ile elde edilir. Toplamsal modeller;

$$E(Y|X_1, \dots, X_p) = s_0 + s_1(X_1) + \dots + s_p(X_p)$$

de verilen koşullu beklenen değerlerin modellenmesi ile doğrusal modelleri genelleştirir. Burada $i=1, \dots, p$ olmak üzere $s_i(X_i)$ düzleştirme fonksiyonlarıdır. Düzleştirme fonksiyonları s_i 'lerin $E s_i(X_i) = 0$ standartlaştırılmış koşulları sağlaması gerekir.

2.3. Düzleştirici Fonksiyonlar

Genelleştirilmiş toplamsal modellerde düzleştirici fonksiyonlar, bağımlı değişkendeki değişkenliği bağımsız değişkenin bir fonksiyonuyla açıklar. Düzleştirici fonksiyonlar splayn olarak da bilinir. Splaynlar sonuç değişkeni ile bağımsız değişkenler arasındaki ilişkiyi açıklar. Farklı düzleştiriciler vardır burada yaygın kullanılan düzleştiriciler verilecektir.

2.3.1. Kübik Splaynlar

Basit şekilde farklı kübik polinomlardan oluşan bir eğri olarak tanımlanabilir.

$$\sum_{i=1}^n (n(x_i))^2 + \lambda \int_a^b (n''(t))^2 dt$$

λ bir sabit ve $a \leq x_1 \leq \dots \leq x_n \leq b$ 'dir. İlk terim veriye yakınlığı ölçer, ikinci terim

fonksiyondaki eğriyi düzeltir. $\lambda/(1+\lambda)$ düzleştirme parametrisidir ve λ büyüdükçe düzleştirme parametrisi 1'e yaklaşır ve daha düzgün eğri oluşturur. λ 'nın küçük değeri için düzleştirme parametresi 0'a yaklaşır.

2.3.2. Yerel Regresyon

Yerel regresyon Cleveland vd. (1988) tarafından ortaya konmuştur. Yerel regresyon, bir x açıklayıcı değişkeninde regresyon fonksiyonu $\eta(x)$ 'in bir fonksiyonunun değeriyle hesaplanacağı fikrine dayanır. Böyle bir yerel yaklaşım, x noktasının seçilen bir komşuluğu içinde veri noktalarına bir regresyon yüzeyi düzleştirme ile elde edilir. Ağırlıklı En Küçük Kareler (EKK) algoritması açıklayıcı değişkenlerin doğrusal fonksiyonlarını düzleştirme için kullanılır. Yerel regresyon süreci için düzleştirme parametresi tahmin edilen eğrinin düzgünlüğünü kontrol eder (Savaş ve Cengiz, 2009).

2.3.3. Thin-Plane Zincir (TPSPLAYN) Prosedürü

Thin-Plane Zincir Prosedürü (TPSPLAYN) düzeltilmiş EKK metodunu kullanarak parametrik olmayan regresyon modeli uydurur. Thin-Plane düzleştirme zincirini kullanır. Modelin bir parametrik formu için Genelleştirilmiş Çapraz Geçerlilik (GCV) kullanılır. Regresyon model tanımlanmış ve model parametre sayısı bilindiğinde kolaylıkla kullanıldığı gibi model hakkında bir bilgi yokken, veri sabit sayıda parametrelili bir modelle belirtilmiyorsa veriyi modellemek için de kullanılabilir.

Temel özellikleri;

- Düzeltilmiş en küçük kareler tahmincilerini verir.
- Çok boyutlu veri kullanımını sağlar.
- Çoklu skor ifadelerini verir.
- Hem semiparametrik modelleri hem de non-parametrik modelleri düzleştirir.
- Geniş veri setlerini yönetme de seçenekler sunar.
- Çoklu bağlantılı değişkenleri kullanabilir.
- Model serbestlik derecesini veya düzleştirme parametresinin belirlenmesi ile özel bir modelin seçilmesine imkan sağlar.

Farz edelim ki H_m m'inci dereceden kısmi türevleri $L_2(E_d)$ 'de olan fonksiyonların bir uzayı olsun. Burada E^d x 'in tanım kümesidir. Model,

$$y_i = f(x_1(i), \dots, x_d(i)) + \varepsilon_i, \quad i=1, \dots, n \text{ dir.}$$

burada $f \in H_m$ dir. Sabit bir λ için

$$\frac{1}{n} \sum_i^n (y_i - f(x_i) - z_i \beta) + \lambda J_m(f)$$

ile verilen düzeltilmiş en küçük kareler fonksiyonunun minimum yapılması ile f 'yi tahmin etmek mümkündür. Burada $J_m(f)$ 'yi tanımlamak için birkaç tane yol vardır. TP düzleştirme zinciri için d boyutlu x için $J_m(f)$,

$$J_m(f) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \sum \frac{m!}{a_1! \dots a_d!} \left[\frac{\partial^m f}{\partial x_1^{a_1} \dots \partial x_d^{a_d}} \right]^2 dx_1 \dots dx_d$$

gibi tanımlanır. Burada $\sum_i a_i = m$ 'dir. $d=2$ ve $m=2$ için $J_m(f)$,

$$J_2(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\left[\frac{\partial^2 f}{\partial x_1^2} \right]^2 + 2 \left[\frac{\partial^2 f}{\partial x_1 \partial x_2} \right]^2 + \left[\frac{\partial^2 f}{\partial x_2^2} \right]^2 \right) dx_1 dx_2 \text{ deki gibidir.}$$

Genelde m ve d $2m-d > 0$ koşulunu sağlamalıdır. Kolaylaştırmak için $m=2$ varsayımı kabul edilir (Wahba 1990, Bates vd., 1987).

2.4. Düzleştirme Parametresinin Seçimi

2.4.1. Çapraz Geçerlilik (CV) ve Genelleştirilmiş Çapraz Geçerlilik (GCV)

Düzleştirme parametresinin seçiminde CV kullanılabilir. (x_i, y_i) miktalarından her seferinde biri dışarıda tutularak geriye kalan $n-1$ noktadan x_i 'deki düzleştirme fonksiyonu için artık karelerin tahmini elde edilir ve bu artık karelerin toplamını minimum yapacak düzleştirici seçilir. CV fonksiyonu,

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n (y_i - n_{\lambda}^{-i}(x_i))^2$$

olarak kullanılır. Buradaki $n_{\lambda}^{-i}(x_i)$, x_i deki uyumu gösterir. $nCV(\lambda)$ tahmin kareler toplamı olarak da adlandırılır. GTM prosedürü tarafından uydurulan bütün düzleştiriciler örnek bağımlı değişkenlerin bir doğrusal kombinasyonu olarak formüle edilebilir (Allen, 1974).

$$n(x) = A(\lambda)Y$$

burada $A(\lambda)$, λ' ya bağılı matristir. $A(\lambda)$ ' nın diagonal elemanları a_{ii} ile gösterilirse CV fonksiyonu,

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - n_\lambda(x_i)}{1 - a_{ii}} \right)^2$$

deki gibi ifade edilebilir. Çoğu durumlarda a_{ii} 'yi hesaplamak oldukça zaman alıcıdır. Bu hesaplama problemini çözmek için Wahba (1980, 1990) GCV fonksiyonunu önermiştir. GVC fonksiyonu,

$$GVC(\lambda) = \frac{n \sum_{i=1}^n (y_i - n_\lambda(x_i))^2}{(n - \text{tr}(A(\lambda)))^2}$$

deki gibi ifade edilebilir. GCV, CV' nin ağırlıklı oranıdır. Çoğu durumlarda GCV, CV' ye yakındır. Ancak hesaplaması daha kolaydır (Savaş ve Cengiz, 2009).

Kesikli bağımlı değişkenlerin modellenmesi amacıyla normal dağılım uygun değildir. GTM bahsedilen sorunların aşılması için yardımcı olabilecek bir yöntemdir. Pek çok dağılım için toplamsal modeller uygulanabilir (SAS Institute Inc., 1999a, 1999b, 1999c, 2004).

GTM ve GDM aynı amaçla kullanılabilir. Fakat farklı çözümlenme amaçlarına hitap ederler. GDM model parametreleri için tahmine önem verirken GTM veriyi non-parametrik olarak açıklamayı amaçlar. GTM açıklayıcı değişkenler ile sonuç değişkeni arasındaki ilişkinin ortaya konulmasında ve verinin açıklanmasında daha uygun bir yöntemdir (Savaş ve Cengiz, 2009).

2.5. GTM Nasıl Yorumlanmalı?

GTM için yazılan modelde regresyon parametreleri β_j 'lerin bulunmaması GTM' in önemli özelliklerinden birini yansıtır. GTM kullanıldığında regresyon parametreleri kümesi elde edilemez. Fakat buna rağmen, $\hat{f}_j(X_j)$ olarak yazılan her X_j değeri için $f_j(X_j)$ tahminleri vardır. Her $f_j(X_j)$ için katsayı model oluştururken belirlenir ve dahası $\hat{f}_j(X_j)$, X_j ve sonuç değişkeni arasındaki ilişkiyi gösterir. Çoğu uygulamada standart regresyon uygulamasında olduğu gibi bazı tahmin edicilerin doğrusal etkileri olduğu varsayılır.

Grafiksel yöntemler GTM' in non-parametrik bileşenlerini yorumlamak için

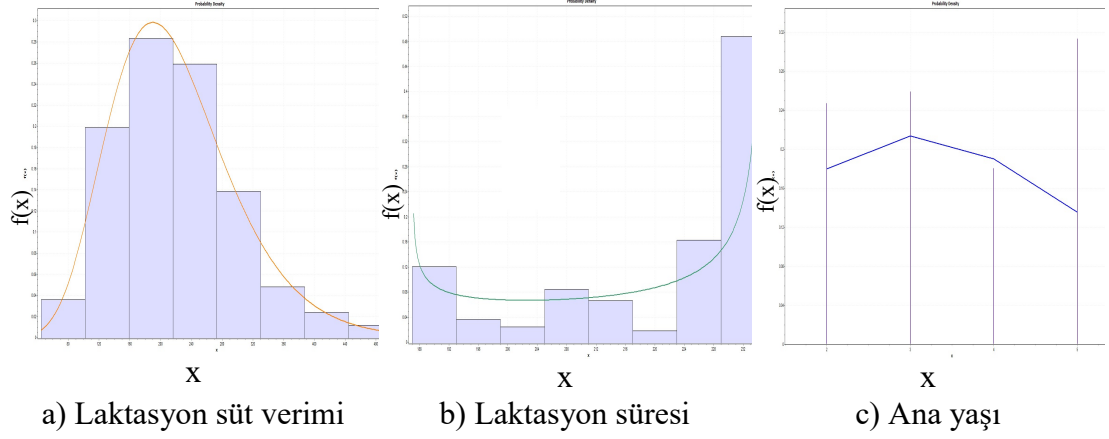
kullanılır. X_j ve $\hat{f}_j(X_j)$ arasındaki grafik X_j ve sonuç değişkeni arasındaki ilişkinin tahmin edilen doğrusal olmayan yapısını gösterir. Standart hatalar ve güven aralıkları $\hat{f}_j(X_j)$ hesaplanabilir ve grafik çizilebilir. Böylece X_j 'nin yükselmesi ve azalması ya da doğrusal uyumundan elde edilen fonksiyonun uyumluluğu değerlendirilebilir. Regresyon katsayıları tablosunun incelenmesi \hat{f}_j grafiğine göre daha basit görünse de bu kullanım kolaylığı muhtemel olarak güvenilir olmayan, kısıtlayıcı ve gereksiz doğrusallık varsayımları karşılığında elde edilir. Diğer yandan eklemelilik GTM'ın tahmin edicilerinin her birinin etkisinin doğrusal regresyonda olduğu gibi diğer tahmin edicilerin etkileriyle oluşan ağırlık yorumlanmasına olanak sağlar (Beck ve Jackman, 1998).

2.6. Analiz için Kullanılan SAS Kodları

```
analiz 'tez';
data tez;
input lsv ay ls;
logCP = log(lsv);
datalines;
272.60 2 235
217.60 2 235
231.00 2 231
      :
301.5 5 201
179.40 5 195
146.60 5 229
;
ods graphics on;
proc gam data=tez;
model logCP = spline(ay) spline(ls);
run;
```

3. BULGULAR VE TARTIŞMA

Sonuç değişkeni olarak laktasyon süt verimi (kg) alınırken açıklayıcı değişkenler olarak laktasyon süresi (gün) ve ana yaşı (yıl) değişkenleri alınmıştır. Bu değişkenlerin dağılışı EasyFit yazılımı ile belirlenmiş olup Şekil 3.1’de verilmiştir.



Şekil 3.1. İncelenen değişkenlere ait dağılım grafikleri

Laktasyon süt veriminin $\sigma=0,24582$, $\mu=5,7345$ ve $\gamma=-99,815$ parametreleri ile lognormal (3P) dağılışı gösterdiği belirlenmiştir. Kolmogorov-Smirnov normallik testi sonucunda LSV değişkeninin normal dağılışa uygun olmadığı ($P=0,038$) belirlenmiştir. laktasyon süresinin $\alpha_1= 0,68651$, $\alpha_2= 0,34817$, $a=187,0$ ve $b=235,0$ parametreleri ile Beta dağılışı gösterdiği belirlenmiştir. Kolmogorov-Smirnov normallik testi sonucunda LSV değişkeninin normal dağılışa uygun olmadığı ($P<0,0001$) belirlenmiştir. Ana yaşının $\lambda=3,5602$ parametresi ile Poisson dağılışı gösterdiği belirlenmiştir. AY değişkeninin normal dağılışa uygun olmadığı ($P<0,0001$) belirlenmiştir (Şekil 3.1).

Doğrusal regresyon analizi ve parametre tahminleri Tablo 3.1’de verilmiştir.

Tablo 3.1. Doğrusal regresyon analizi ve parametre tahminleri

Parametreler	Parametre Tahmini	Standart Hata	t Değeri	P
Sabit	-157,322	84,686	-1,858	0,065
Ana yaşı	21,589	5,068	4,26	<0,0001
Laktasyon süresi	1,364	0,374	3,647	<0,0001

Elde edilen doğrusal regresyon modelinin istatistiksel olarak anlamlı olduğu ($P<0,001$), Hata Kareler Ortalaması (HKO) 5826,039 ve belirtme katsayısının (R^2) 0,16 olduğu belirlenmiştir. Düşük bir belirtme katsayısına rağmen modelin anlamlı

olmasının örnek büyüklüğüne bağlı olabileceği (n=166) düşünülmektedir.

Laktasyon süt verimini(kg) tahmin etmek için laktasyon süresi(gün) ve ana yaşı değişkenleri kullanılarak elde edilen genelleştirilmiş toplamsal model için iterasyon özeti ve uyum istatistikleri; Backfitting Iterasyon sayısı=5, Final Backfitting Kriteri=2,271223E-10 ve sonuç tahminine ait sapma=20,128886164 olarak elde edilmiş olup iterasyon sonucunda uyum sağlandığı (convergence) bilgisine ulaşılmıştır.

GTM modeli için düzleştirici parametre tahminleri ve bunlara karşılık gelen genelleştirilmiş çapraz geçerlilik değerleri Tablo 3.2’de ve GTM analiz istatistikleri Tablo 3.3’de verilmiştir.

Tablo 3.2. GTM modeli için düzleştirme bileşenleri ve uyum istatistikleri

Bileşen	Düzleştirme Parametresi	Serbestlik Derecesi	GCV	Tekil gözlem sayısı
Spline (ana yaşı)	0.009901	1.890859	0.001077	4
Spline (laktasyon süresi)	0.994764	3.000000	0.099842	30

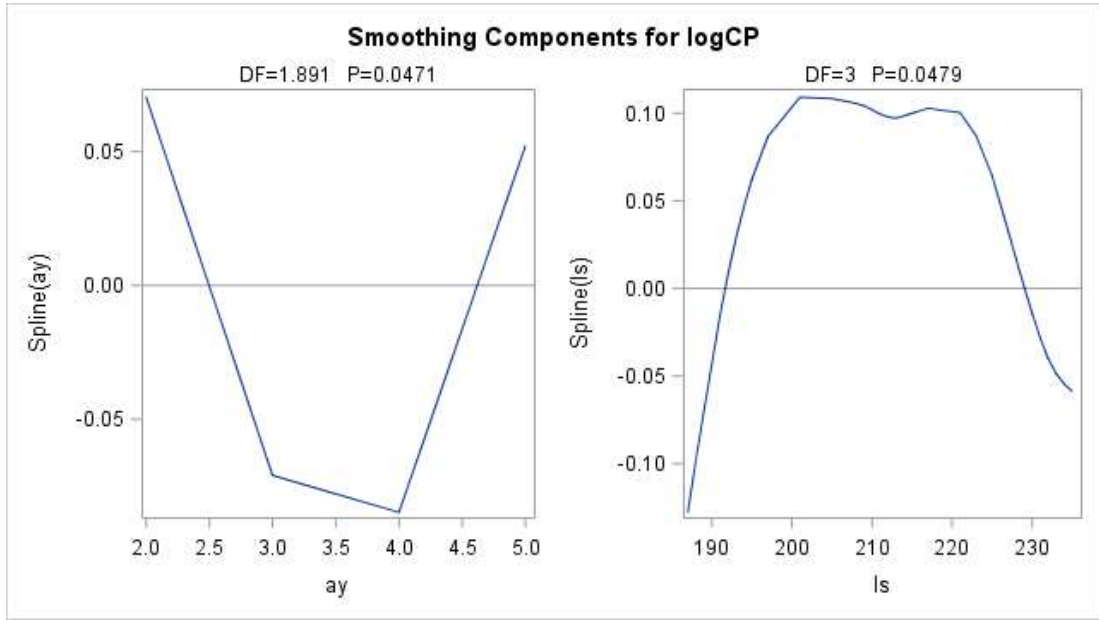
Ana yaşı değişkeninin laktasyon süresi değişkenine göre daha başarılı düzleştirme değerine sahip olduğu söylenebilir çünkü sıfıra yakın değerler düzleştirme parametresinin bir fonksiyonu olarak GCV'nin başarılı bir şekilde en aza indirilmesini sağlamak ve modele uyumu artırmak için oldukça önemlidir (Jansen, 2015).

Tablo 3.3. GTM analiz istatistikleri (sapma analizi)

Bileşen	Serbestlik Derecesi	Kareler Toplamı	Ki-Kare	P
Spline (ana yaşı)	1.89086	0.750597	5.8958	0.0471
Spline (laktasyon süresi)	3.00000	1.006934	7.9093	0.0479

GTM sonuçlarının en önemli kısmı olarak ele alınabilecek olan bilgilere sapma analizi tablosundan ulaşılabilir. Modeldeki düzleştiriciler için tablo tam model ve ilgilenilen değişken yokluğunda oluşturulan model arasındaki sapmayı karşılaştıran bir χ^2 test değeri üretmektedir. Elde edilen sonuçlara göre 0,05 önem düzeyinde ana yaşı ve laktasyon süresi değişkenlerinin model üzerindeki etkisinin istatistiksel olarak anlamlı olduğu anlaşılmıştır.

Ana yaşı ve laktasyon süresine ait düzleştirme grafikleri Şekil 3.2’de verilmiştir.



Şekil 3.2. Ana yaşı ve laktasyon süresine ait düzleştirme grafikleri

Şekil 3.2’de verilen grafikler yorumlandığında ana yaşı ve laktasyon süresi değişkenlerinin genelleştirilmiş toplamsal modellerle değerlendirilmesi gerektiği anlaşılmaktadır.

Doğrusal regresyon ve genelleştirilmiş toplamsal modelleri karşılaştırabilmek adına Root Mean Square Error (RMSE), Relative Root Mean Square Error (rRMSE) ve ortalama mutlak sapma (MAD) değerleri hesaplanmış ve Tablo 3.4’de verilmiştir.

Tablo 3.4. Modeller için karşılaştırma istatistikleri

Model	RMSE	rRMSE	MAD
LR	75,6357	34,4214	58,1024
GTM	79,9291	36,3753	61,9424

Elde edilen bulgulara göre doğrusal regresyon analizinin tahmin başarısının daha yüksek olduğu Tablo 3.4’deki düşük karşılaştırma değerlerinden anlaşılmaktadır. Ancak açıklayıcı değişkenler arasında çoklu bağlantı problemi olmamasına (VIF=1.000) ve sapmaların normal dağılışa sahip olmasına ($P>0,05$) rağmen değişkenlerin normal dağılmadığı bilinmektedir. Doğrusal regresyon analizinin gerekli varsayımları sağlayamadığından sonuçların güvenilir olmadığı bu nedenle Genelleştirilmiş Toplamsal Modellerin kullanılmasının önerilebileceği söylenebilir.

4. SONUÇ

Toplamsal modeller doğrusal regresyon modellerine göre daha kısıtsız modellerdir ancak parametrik olmayan regresyon modellerinden daha kısıtlı modellerdir. Toplamsal modeller açıklayıcı değişkenler arasındaki ilişkiyi dikkate almamaktadır. Bu durum toplamsal modeli parametrik olmayan modellerden daha fazla kısıtlı duruma getirmektedir. Toplamsal modellerin genel gösteriminde açıklayıcı değişkenler arasındaki ilişki dikkate alınmaz ancak bu durum açıklayıcı değişkenler arasında ilişki olmadığı şeklinde yorumlanmamalıdır. Açıklayıcı değişkenler arasında ilişki yok ise, her açıklayıcı değişkenin sonuç değişkenine karşı düzeltilmesiyle elde edilen fonksiyonların toplamı toplamsal modelin tahminini oluşturur. Açıklayıcı değişkenler arasında ilişki olduğu durumda ise modelin tahmininde bu ilişkinin de dikkate alınması gerekir. Bu durumda kısmi regresyonlar yardımıyla hesaplanan kısmi artıklar kullanılabilir (Çağlayan, 2002; Bağdatlı, 2010).

Genelleştirilmiş toplamsal modeller doğrusal regresyon modellerine göre daha kısıtsız modellerdir ancak parametrik olmayan regresyon modellerinden daha kısıtlı modellerdir. Açıklayıcı değişkenlerin normal dağılım göstermediği durumlarda genelleştirilmiş toplamsal modellerin doğrusal modeller yerine kullanılmasının daha doğru ve güvenilir sonuçlar üretebileceği yapılan bu çalışma ile literatürü destekler nitelikte ortaya konulmuştur.

KAYNAKLAR

- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*. 16. 125-127.
- Arı, A. ve H. Önder. (2013). Farklı veri yapılarında kullanılabilir regresyon yöntemleri. *Anadolu J Agr Sci*. 28(3). 168-174.
- Ashfa H. D. (2017). *Performance of spline-based gam in the presence of outliers and mulyicollinearity*. Yüksek Lisans Tezi. Eskişehir Anadolu Üniversitesi Fen Bilimleri Enstitüsü, 73, Eskişehir.
- Bağdatlı, S. (2010). *Semiparametrik regresyon ve bir uygulama*. Yüksek Lisans Tezi. İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü, 81, İstanbul.
- Bates, D. M., M. J. Lindstrom, G. Wahba and B. S. Yandell. (1987). GCVPACK Routines for generalized cross-validation. *Commun. Statist. Simula*.16. 263-97.
- Beck, N. and S. Jackman. (1998). Beyond Linearity by default: generalized additive models. *American Journal of Political Science*. 42(2). 596-627.
- Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Appl. Stochastic Models Bus. Ind.* 23. 129-143.
- Cleveland, W. S., S. J. Devlin and E. H. Grosse. (1988). Regression by local fitting methods, properties and computational algorithms. *Journal of Econometrics*. 37. 87-114.
- Çağlayan, E. (2002). *Yarı parametrik regresyon modelleri ile yaşam boyu sürekli gelir hipotezinin Türkiye uygulaması*. Doktora Tezi, Marmara Üniversitesi Sosyal Bilimler Enstitüsü, 171, İstanbul.
- Dominici, F., A. McDermott, S. L. Zeger and J. M. Samet. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American J. Epidemiology*. 156(3). 193-203.
- Durmuş, S. (2018). *Değişen katsayılı regresyon modeli ile gebelerin ortalama arteriyel kan basıncına etki eden risk faktörlerinin belirlenmesi*. Yüksek Lisans Tezi. Hacettepe Üniversitesi, 62, Ankara.
- Hastie, T. J. ve R. J. Tibshirani. (1990). *Generalized additive models*. New York: Chapman & Hall.
- Jansen, M. (2015). *Generalized Cross Validation in variable selection with and without shrinkage*. Erişim: 14 Kasım 2021, URL: <http://homepages.ulb.ac.be/~majansen/publications/jansen15gcvreprint.pdf>.

- Kan Kılıç, B. ve M. Çavuş. (2017). İkili yanıt değişkenine sahip modellerin yeterliliklerine ilişkin benzetim çalışması – parametrik olmayan yöntemler. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi*. 21(2). 169-177.
- Kanıt R. ve U. N. Baykan. (2004). Bina yaklaşık maliyetinin çoklu doğrusal regresyon ile belirlenmesi. *Politeknik Dergisi*. 7(4). 359-367.
- Kovalchik, S. and R. Varadhan. (2013). Fitting additive binomial regression models with the R package blm. *Journal of Statistical Software*. 54(1). 1-18.
- Ma, L. and X. Yan. (2014). Examining the nonparametric effect of drivers' age in rear-end accidents through an additive logistic regression model. *Accident Analysis and Prevention*. 67. 129-136.
- McCullagh, P. and J. A. Nelder. (1989). *Generalized linear models*. London: Chapman and Hall.
- Omay R. E. (2007). *Regresyonda pürüzlülük ceza yaklaşımı*. Doktora Tezi. Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, 121, Eskişehir.
- Omay R. E. (2014). OECD ülkeleri için hava kirliliği-sağlık ilişkisinin incelenmesi. *Hava Kirliliği Araştırma Dergisi*. 3. 39-46.
- Önder, H. (2007). Permütasyon testlerinin doğrusal regresyonda kullanılabilirliğinin irdelenmesi. *OMÜ Zir. Fak. Dergisi*. 22(2). 157-161.
- SAS Institute Inc., 1999a, SAS language reference: Concepts, Version 8, Cary, NC:SAS Institute Inc.
- SAS Institute Inc., 1999b. SAS language reference: Dictionary, Version 8, Cary, NC:SAS Institute Inc.
- SAS Institute Inc., 1999c. SAS procedures guide, Version 8, Cary, NC: SAS Ins.Inc.
- SAS Institute Inc., 2004. SAS/STAT 9.1 User's guide, Cary, NC: SAS Institute Inc.
- Savaş, N. ve M. A. Cengiz. (2009). Yaşam memnuniyetini etkileyen faktörlerin belirlenmesinde geliştirilmiş toplamsal modellerin kullanımı. *EÜFBED*. 2(1). 105-119.
- Şenel, T., M. A. Cengiz, N. Savaş ve N. Terzi. (2009). Çoklu doğrusal regresyonda model seçiminde geliştirilmiş toplamsal modellerin kullanımı. *EÜFBED*. 2(2). 217-227.
- Tozlu Çelik, H. (2014). *Kıl keçi ve Saanen x Kıl keçi melezlerinin (F1, G1, G2) üretici şartlarında çeşitli verim özellikleri bakımından karşılaştırılması*. Doktora Tezi. Ondokuz Mayıs Üniversitesi Fen Bilimleri Enstitüsü, 103, Samsun.
- Wahba, G. (1980). *Spline bases, regularization and generalized cross-validation for*

solving approximation problems with large quantities of noisy data. New York:
Academic Press.

Wahba, G. (1990). *Spline functions for observational data*. Philadelphia: SIAM.

ÖZ GEÇMİŞ

Nursen KURDAL, Bayburt Anadolu Lisesi'ni bitirdikten sonra Ondokuz Mayıs Üniversitesi Ziraat Fakültesi'nden 2019 yılında mezun oldu. 2019 yılında OMÜ Ziraat Fakültesi Zootekni Bölümü Biyometri ve Genetik alanında Yüksek Lisans programına başladı. Orta derecede İngilizce bilmektedir. Temel ilgi alanları; spor yapmak, sosyal yardım kuruluşlarında gönüllü çalışmak (07 Aralık 2021).

İletişim Bilgileri

Öğrenci no : 19210089

ORCID ID : 0000-0001-7198-2329

Yayımlanmış Çalışmalar:

1. Kurdal N, Önder H. 2020. Estimating the nonparametric confidence interval for correlation coefficient on animal data. International Congress on Animal Breeding, Genetics and Husbandry (ICABGEH-20), Online, 89-92, 12 – 14 August 2020. (Oral, Full Text).

