Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 23, Özel Sayı, 76-83, 2019

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 23, Special Issue, 76-83, 2019

DOI: 10.19113/sdufenbed.436178

Parçacık Sürü Optimizasyonu Yöntemi ile Sayım Modelleri için En Uygun Değişken Kümesinin Belirlenmesi

Haydar KOÇ[1], Tuba KOÇ[*2], Emre DÜNDER[3]

[1,2]Çankırı Karatekin Üniversitesi, Fen Fakültesi, İstatistik Bölümü, Çankırı
[1](ORCID: https://orcid.org/0000-0002-8568-4717)
[2](ORCID: https://orcid.org/0000-0001-5204-0846)
[3]Ondokuz Mayıs Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Samsun
[3](ORCID: https://orcid.org/0000-0003-0230-8968)

(Alınış / Received: 25.06.2018, Kabul / Accepted: 29.01.2019, Online Yayınlanma / Published Online: 11.03.2019)

Anahtar Kelimeler
Değişken seçimi,
Sayım modelleri,
Sezgisel optimizasyon

Özet: Birçok bilimsel çalışmada sayım verisi olarak adlandırılan negatif olmayan tamsayı değerleri alan nicel veriler kullanılmaktadır**.** İstatistiğin en temel analiz yöntemlerinden biri olan regresyon analizi kapsamında da sayım verileri oldukça sık kullanılmaktadır. Bağımlı değişkenin tamsayı ile ifade edilebildiği regresyon modelleri sayım modelleri olarak tanımlanır. Bu çalışmada sayım modelleri kapsamında model seçimi incelendi. Sayım modellerinde model seçimi için klasik seçim yöntemleri ve PSO algoritması kullanıldı. Uygulamalar hem simülasyon hem de gerçek veriler üzerinde yapıldı. Sonuç olarak klasik yöntemlerle kıyaslandığında PSO algoritmasının, modeldeki değişken sayısı arttıkça ve bağımsız değişkenler arasındaki korelasyon değerleri yükseldikçe daha iyi sonuçlar verdiği ve sayım modelleri için PSO algoritmasının değişken seçiminde alternatif bir yöntem olarak kullanılabileceği gösterilmiştir.

Determination of Best Variable Set for Count Models by Particle Swarm Optimization

Keywords
Variable selection,
Count models,
Heuristic optimization

Abstract: In most scientific studies quantitative data are used which take non-negative integer values, called count data. Count data are also used frequently in the context of regression analysis, which is one of the most basic analysis methods of statistical analysis. The regression models in which the dependent variable can be expressed by integers are defined as count models. In this study, the model selection in the context of count models was investigated by using classical selection methods and PSO algorithm. Applications were made on both simulation and real data. As a result, it has been shown that PSO algorithm can be used as an alternative method for PSO algorithm selection for count models when the number of model variables increases and the correlation values between independent variables increases as compared to classical methods.

1. Giriş

Bilimsel çalışmalarda kullanılan nicel verilerin birçoğu belirli bir değer aralığı içerisinde sürekli değerlere sahip değildir. Verilerin bazıları uygulama alanlarına göre negatif olmayan tamsayılar ile ifade edilmektedir. Bu veriler sayım verileri olarak adlandırılır. İstatistiğin en temel analizlerinden biri olan regresyon analizi kapsamında da sayım verileri kullanılmaktadır. Bağımlı değişkenin tamsayı ile ifade edilebildiği regresyon modelleri sayım modelleri olarak tanımlanır. Sayım modellerinin tahmin gücü bağımsız değişkenlerin modeli açıklayabilme gücüne bağlıdır. Bu amaçla farklı regresyon modellerinde olduğu gibi, en uygun değişken kümesi belirlenerek regresyon modeli oluşturulmalıdır. Bağımsız değişkenler arasında yüksek derecede ilişkinin bulunması, gereksiz değişkenlerin modele dahil edilmesi gibi çeşitli hatalar regresyon modelinin tahmin performansını düşürmektedir. Regresyon modeli içerisinde en uygun modelin seçimi için değişken seçim algoritmaları uygulanmaktadır. Değişken seçimi için literatürde en sık kullanılan üç klasik yöntem vardır. Bunlar; geriye doğru, ileriye doğru ve adımsal seçim tekniklerdir. Ancak bu teknikler birçok durumda iyi sonuçlar vermemektedir [1,2]. Değişken sayısı arttıkça klasik seçim yöntemleri çoğu zaman uygun modeli

*İlgili yazar: tubakoc@karatekin.edu.tr

seçememektedir. Ayrıca çok yüksek boyutlu verilerde çoklu bağlantı sorunu sebebiyle klasik seçim yöntemleri ile model belirlemek mümkün değildir. Söz konusu problemleri çözmek amacıyla sezgisel optimizasyon tekniklerinden yararlanılmaktadır [3].

Literatür incelendiğinde model seçiminde yapılan çalışmaların büyük bir kısmının lineer ve lojistik regresyon analizi üzerine olduğu görülmektedir. Lineer regresyon analizi için tabu araştırma [4], genetik algoritma [2], hibrit genetik algoritma-benzetimli tavlama metodu kullanılmıştır [5]. Lojistik regresyon analizinde tabu araştırma algoritması [6] ve parçacık sürü optimizasyonu algoritmasını kullanılarak optimum değişken kümesinin seçimini incelemiştir [7]. Genelleştirilmiş lineer modellerde model seçimi üzerine farklı yöntemler de kullanılmıştır. Poisson regresyon analizinde adımsal bir model seçim yöntemi önermiştir [8]. McLeod ve Xu genelleştirilmiş lineer modeler için değişken seçimi uygulayan bestglm isimli bir paket geliştirmiştir [9]. Calcagno, ve Mazancourt genelleştirilmiş lineer modellerde genetik algoritma kullanarak değişken seçimi uygulayan glmulti isimli bir yazılım paketi oluşturmuştur [10]. Regresyon modellerinde cezalandırma yöntemleri de model seçiminde sıkça kullanılmaktadır. Cezalandırma yöntemleri ile model seçimi konusunda bilinen en genel yöntem LASSO metodudur [11]. Bu metodun ana fikri, modeldeki regresyon katsayılarını sıfıra yaklaştıracak şekilde daraltmaktır. LASSO tekniğinde cezalandırma için elastik net [12], adaptif LASSO [13] yöntemleri geliştirilmiştir.

Cezalandırma yöntemleri kullanarak değişen yayılım parametreli beta regresyon modeli için değişken seçim tekniği uygulanmıştır [14]. Beta regresyon analizinde model seçimi için bootstrap tekniği kullanılmıştır [15].

Model seçimi üzerinde son derece etkili olan bilgi kriterleri için de çok sayıda çalışma mevcuttur. Regresyon analizinde en çok kullanılan kriterler Akaike ve Bayesci bilgi kriterleridir. Akaike bilgi kriteri yanlılık probleminin çözümü ve küçük örneklemlerde de iyi performans sağlaması için alternatif cezalandırmalar ile düzeltilmiş ve yeni Akaike türü kriterler önerilmiştir [16,17] . Bayesci bilgi kriteri de regresyon katsayılarına ilişkin kovaryans matrisi de cezalandırma terimine dahil olacak şekilde düzenlenmiş ve alternatif iki kriter önerilmiştir [18]. Kovaryans matrisini cezalandırma terimine dahil ederek bilgi kriteri hesaplayan yaklaşım ilk olarak Bozdoğan tarafından önerilmiştir [19]. Bu yaklaşım, cezalandırma terimindeki değişiklikler ile düzenlenerek alternatif kriterler türetilmiştir [20, 21, 22].

Model seçimi, istatistiksel modelleme içerisinde son derece önemli bir konudur. Klasik model seçim algoritmaları, sayım modelleri kapsamında modelin performansını optimize etme konusunda zayıf olduğundan, sezgisel optimizasyona dayalı yaklaşımlar daha başarılı sonuçlar sağlama potansiyeline sahiptir. Bu sebeple, çalışmanın temel motivasyon kaynağı, sayım modellerinde model seçimi sürecini parçacık sürü optimizasyonu algoritması kullanarak gerçekleştirmektir.

2. Materyal ve Metot

2.1. Poisson regresyon modeli

Poisson regresyon modeli sayımla elde edilen verilerin modellenmesinde kullanılan temel bir yaklaşımdır. Y_i ($i = 0,1,2, \dots$), ortalaması μ_i olan belli bir zamanda meydana gelen olayların sayısı olsun. Bu durumda Y_i ler Poisson rasgele değişkenidir ve olasılık fonksiyonu

$$f(y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad y_i = 0,1,2, \dots \tag{1}$$

ile verilir. Burada, ortalama ve varyans $E(Y_i) = Var(Y_i) = \mu_i'$ dir. Log-olabilirlik fonksiyonu eşitlik (2)'de verildiği gibidir.

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n}\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\} \tag{2}$$

X: $n \times (p + 1)$ boyutlu açıklayıcı değişken matrisi olsun. Y_i ve X'in i-nci satır vektörü arasındaki ilişki eşitlik (3)'de verilmiştir.

$$\ln(\mu_i) = \eta_i = x_i^T\beta \tag{3}$$

Bu model poisson regresyon veya log-linear model olarak bilinir. β' nın en çok olabilirlik tahmin edicisi $\hat{\beta}$, Fisher scoring yöntemi ile elde edilebilir.

Poisson regresyon modelinde iki temel varsayım vardır. Birincisi, olaylar zamana göre birbirinden bağımsız gerçekleşir. İkincisi, şartlı beklenen değer ve varyans birbirine eşittir. İkinci varsayımda bahsedilen eşitlik durumu, eşit yayılım olarak bilinir ve uygulamada çoğunlukla eşit yayılım durumu sağlanmamaktadır ve genelde varyans ortalamadan büyük çıkmaktadır.

2.2. Negatif binomial modeli

Poisson regresyon modelinde, bağımlı değişken için varyansın ortalamadan büyük olması, bağımlı değişken değerleri arasında pozitif korelasyonun bulunması veya verilerin olasılıkları arasındaki değişimin yüksek çıkması gibi durumlar aşırı yayılıma neden olabilir. Ayrıca verinin dağılımsal varsayımlarında bir ihlal söz konusu olduğu durumlarda da aşırı yayılım ortaya çıkabilir.

Aşırı yayılım durumunda en sık başvurulan yaklaşım Negatif Binomial Regresyon Modelidir (NBRM). NBRM, modele gözlemlenemeyen heterojenlik terimi

dahil edilerek geliştirilebileceği gibi bir çok farklı yöntem ile de elde edilebilir [23].

Negatif binomial dağılımın olasılık fonksiyonu aşağıda verildiği gibidir:

$$\Pr(y_i|x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \quad (4)$$

Burada $\Gamma(.)$ ile gamma fonksiyonu ifade edilmektedir ve α sabit bir parametredir. Negatif binomial dağılımı için y'nin beklenen değeri Poisson dağılımı ile aynıdır. Yani, $\mu_i = E(y_i|x_i) = e^{x_i'\beta}$ 'dır.

Ancak varyansı, eşitlik (5)' de verildiği gibidir:

$$\mathrm{Var}(y_i|x_i) = \mu_i(1+\alpha\mu_i) = e^{x_i'\beta}(1+\alpha e^{x_i'\beta}) \quad (5)$$

NBRM'de, y – nin varyansı, beklenen değerinden büyüktür. Çünkü, μ ve α 'nın her ikisi de pozitiftir. Eşitlik (5), $\alpha = 0$ olduğunda μ_i olur. Yani, $\alpha = 0$ olduğunda negatif binomial dağılımı poissona indirgenir. NBRM parametreleri tahmininde en çok olabilirlik yöntemi kullanılabilir [23].

2.3. Sıfır yoğunluklu poisson model

Sıfır yoğunluklu Poisson (ZIP) model, iki aşamalı bir karışım modelidir. Bu model, sıfır değerli gözlemlerin iki farklı şekilde ortaya çıktığını varsayar. Birinci durum w_i olasılığı ile gerçekleşir ve sadece sıfır üretir. İkinci durum $(1-w_i)$ olasılığı ile meydana gelir ve μ ortalamalı standart bir Poisson sayım verisi elde edilir.

Genelde birinci durumdan elde edilen sıfırlar yapısal sıfırlar olarak ve Poisson dağılımından kaynaklanan sıfırlar örneklem sıfırları olarak adlandırılır [24].

Bu iki aşamalı süreç iki bileşenli bir karışım dağılımı verir ve olasılık fonksiyonu aşağıdaki gibidir.

$$(y_i|x_i,\mu_i,w_i) = \begin{cases} w_i + (1-w_i)e^{-\mu_i}, & y_i = 0, \\ (1-w_i)\dfrac{e^{-\mu_i}\mu_i{}^{y_i}}{y_i!} & y_i = 1,2,3,\ldots \end{cases} \quad 0 \le w_i \le 1 \quad (6)$$

Bu olasılık fonksiyonu Y~ZIP(μ, w) ile gösterilir. Y'nin ortalama ve varyansı aşağıdaki gibidir.

$$\begin{aligned} \mathrm{E}(y_i) &= \mu_i(1-w_i) \\ \mathrm{Var}(y_i) &= (1-w_i)(\mu_i + w_i\mu_i{}^2) \end{aligned} \quad (7)$$

Görüldüğü gibi Y' nin marjinal dağılımı $w > 0$ olduğunda aşırı yayılıma izin vermektedir. $w = 0$ olduğunda standart Poisson dağılımına indirgenmektedir.

Log olabilirlik fonksiyonu eşitlik (8)'de verildiği gibidir.

$$\mathcal{L}_{\mathrm{ZIP}}(\mu, w; y) = \sum_{i=1}^{n}\begin{Bmatrix} I(y_i = 0)\ln[w_i + (1-w_i)e^{-\mu_i}] + \\ I(y_i > 0)[\ln(1-w_i) - \mu_i + y\ln\mu_i - \ln(y!)] \end{Bmatrix} \quad (8)$$

ZIP modeli aşağıdaki iki link fonksiyonu eklenerek elde edilebilir.

$$\begin{aligned} \ln(\mu_i) &= X'\beta \\ \mathrm{logit}(w_i) &= \ln\frac{w}{1-w} = X'\gamma \end{aligned} \quad (9)$$

2.4. Parçacık sürü optimizasyonu

Parçacık Sürü Optimizasyonu (PSO), sürü halinde hareket eden balıklar ve böceklerden esinlenerek geliştirilmiş bir optimizasyon yöntemidir [25]. Temel olarak sürü zekâsına dayanan bir algoritmadır. Sürü halinde hareket eden hayvanların yiyecek ve güvenlik gibi durumlarda, çoğu zaman rastgele sergiledikleri hareketlerin, amaçlarına daha kolay ulaşmalarını sağladığı görülmüştür.

Algoritma temel olarak aşağıdaki basamaklardan oluşur [26];

- **I.** Rasgele üretilen başlangıç pozisyonları ve hızları ile başlangıç sürüsü oluşturulur.
- **II.** Sürü içerisindeki tüm parçacıkların uygunluk değerleri hesaplanır.
- **III.** Her bir parçacık için mevcut jenerasyondan yerel en iyi (pbest) bulunur. Sürü içerisinde en iyilerin sayısı parçacık sayısı kadardır.
- **IV.** Mevcut jenerasyondaki yerel eniyiler içerisinden küresel en iyi (gbest) seçilir.
- **V.** Pozisyon ve hızlar aşağıdaki gibi yenilenir.

$$V_{id} = W * V_{id} + c_1 * \mathrm{rand}_1 * (P_{id} - X_{id}) + c_2 * \mathrm{rand}_2 * (P_{gd} - X_{id})$$
$$X_{id} = X_{id} + V_{id}$$

Burada X_{id} pozisyon ve V_{id} hız değerlerini verirken, rand_1ve rand_2 değerleri rasgele üretilmiş sayılardır. W, atalet ağırlık değeri ve c_1 , c_2 ölçeklendirme faktörleridir.

- **VI.** Durdurma kriteri sağlanıncaya kadar II, III, IV ve V adımları tekrar edilir.

2.5. Bilgi Kriterleri

Log–olabilirlik fonksiyonuna dayalı olan bilgi kriteri tabanlı model seçim teknikleri, hem iç içe hem de iç içe olmayan sayı ile ifade edilen modeller için kullanılabilirler. Modele daha fazla parametre eklendiğinde log–olabilirliğin artacağı düşünülür. Log–olabilirlik artışının cezası, gözlem sayısı kadar parametre sayısını da hesaba katar. Bu çalışmada, iki bilgi kriteri ölçüsü ele alınmıştır.

2.5.1. Akaike bilgi kriteri

Akaike bilgi kriteri (AIC) istatistiksel modellerin uyum iyiliği için kullanılan bir kriterdir.

$$\mathrm{AIC} = -2\mathcal{L} + 2p = D(\hat{\theta}) + 2p \quad (10)$$

Burada p, modeldeki parametre sayısını ve $\mathcal{L}$ ise log–olabilirlik fonksiyonunu göstermektedir.

2.5.2. Bayesci bilgi kriteri

Aynı zamanda Schwarz kriteri olarakta bilinen Bayesci bilgi kriteri (BIC) istatistiksel modellemede çok sık kullanılan bir kriterdir.

$$\mathrm{BIC} = -2\mathcal{L} + (\ln n)\, p \qquad (11)$$

Burada n, veri setindeki gözlem sayısını p ise parametre sayısını göstermektedir.

3. Bulgular

Uygulama aşamasında iki farklı simülasyon ve bir gerçek veri seti uygulaması yapıldı Simülasyon tasarımları çoklu bağlantı bulunan ve bulunmayan regresyon modelleri için uygulandı. Model seçim yöntemlerinden ileriye doğru (İ), geriye doğru (G), adımsal (A) ve PSO metotları kullanılarak sayım modelleri için model seçimi sonuçları elde edildi. Uygulama aşamasında seçim yöntemlerinin seçmiş olduğu sayım modellerine ilişkin AIC ve BIC değerleri incelendi. AIC ve BIC değeri daha düşük olan seçim yöntemi daha başarılı olarak kabul edildi. Analizler R yazılımının 3.4.3 versiyonu kullanılarak yapılmıştır. Ayrıca "pso" ve "COUNT" paketlerinden yararlanılmıştır.

Simülasyon çalışması -1

Çalışmanın bu bölümünde çoklu bağlantı problemi içeren regresyon modelleri oluşturuldu. Simülasyon tasarımı için toplam on açıklayıcı değişken ile Poisson, negatif binomial ve sıfır yoğunluklu regresyon modelleri için klasik seçim yöntemleri ve PSO algoritmaları uygulandı. Bağımsız değişkenler aşağıdaki şekilde türetildi:

$$x_1 = 10 + \varepsilon_1 \qquad (12)$$

$$x_2 = 10 + 0.3\varepsilon_1 + \alpha\varepsilon_2 \qquad (13)$$

$$x_3 = 10 + 0.3\varepsilon_1 + 0.5604\alpha\varepsilon_2 + 0.8282\alpha\varepsilon_3 \qquad (14)$$

$$x_4 = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon_4 \qquad (15)$$

$$x_5 = = -5 + 0.5x_1 + x_2 + 0.5\varepsilon_5 \qquad (16)$$

$\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5 \sim N(0, 1)$ dağılımlı hata terimleridir. Eşitlik 12-16 için α parametresi çoklu bağlantı düzeyini kontrol etmektedir. Bu çalışmada geçmiş çalışmalar baz alınarak α =0.03 alındı ve bağımlı değişken türetimi aşağıdaki şekilde yapıldı:

$$\lambda = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon \qquad (17)$$

ε_i, i= 1, 2,..., n için $\varepsilon \sim N(0,1)$ için dağılımlı hata terimidir. Eşitlik (17)' den bulunan λ değerleri ile $y \sim$ Pois(λ), NegBin(λ) ve ZINFPois(λ) biçiminde sırasıyla Poisson, negatif binomial ve sıfır yoğunluklu Poisson dağılımlı bağımlı değişkenler türetildi. Bağımsız değişkenler kümesine bağımlı değişkenler ile ilişkisi olmayan beş adet düzgün dağılımlı değişkenler $R_j \sim$ Düzgün(0,1) j=1,2,3,4,5 ilave edildi.

1. Simülasyon tasarımı için örneklem sayıları sırası ile n=30,50,100,200,500 olarak alındı.

Simülasyon çalışması -2

Bu simülasyon çalışmasında çoklu bağlantı problemi içermeyen regresyon modelleri oluşturuldu. Simülasyon tasarımı için sırasıyla p=2,3,...,15 değişken ve n=100 için Poisson, negatif binomial ve sıfır yoğunluklu regresyon modelleri kapsamında klasik seçim yöntemleri ve PSO algoritmaları uygulandı. Bağımsız değişkenler birbirinden bağımsız olarak X_k, k=2,3,…15 için N(0,1) dağılımına uygun olarak türetildi. Bağımsız değişkenler arasındaki korelasyon yapısı, çoklu bağlantı problemi içermeyecek biçimde $\mathrm{cor}(X_i, X_j)=0.10^{|i-j|}$ i, j=2,3,…15 için türetildi. Bağımsız değişkenler ile bağımlı değişkenlere ait λ parametreleri ise şu şekilde türetildi:

$$\lambda = X\beta + \varepsilon \qquad (18)$$

$\varepsilon \sim N(0, 1)$ dağılımlı hata terimidir. β=(0.1,0.3,0.5,...,s) biçiminde artımsal olarak seçildi. Bu denklemdeki s, toplam değişken sayısı doğru orantılı olarak 0.1 birim artımlı β değerlerini temsil etmektedir. Eşitlik (18)' den elde edilen λ değerleri ile $y \sim$ Pois(λ) , NegBin(λ) ve ZINFPois(λ) biçiminde sırasıyla Poisson, negatif binomial ve sıfır yoğunluklu Poisson dağılımlı bağımlı değişkenler türetildi.

Gerçek veri uygulaması

Gerçek veri uygulaması için R programında "COUNT" paketi içerisinde yer alan "affairs" hazır veri seti kullanıldı [27]. Bu veri setinde bağımlı değişken tam sayıdır ve veri seti sayım verilerini modellemek için son derece elverişlidir. Bu veri seti üzerinde klasik değişken seçim yöntemleri ve PSO algoritması ile model seçimi uygulandı ve modellerin AIC-BIC değerleri elde edildi.

Tablo 1. 1. Simülasyon tasarımında Poisson regresyon modeli için AIC kriteri sonuçları

n	Seçim yöntemi			
	AIC-İ	**AIC-G**	**AIC-A**	**AIC-PSO**
30	80.76578	80.76578	80.76578	80.61614
50	132.17770	132.17770	132.17770	132.10880
100	268.25710	268.25710	268.25710	268.21730
200	536.20400	536.20400	536.20400	536.19640
500	1333.94000	1333.94000	1333.94000	1333.94400

Tablo 2. 1. Simülasyon tasarımında Poisson regresyon modeli için BIC kriteri sonuçları

n	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
30	85.28353	85.28353	85.28353	85.04804
50	139.04800	139.04800	139.04800	138.69240
100	277.98830	277.98830	277.98830	277.75140
200	548.67820	548.67820	548.67820	548.62150
500	1351.59000	1351.59000	1351.59000	1351.54300

Tablo 1 ve 2' de Poisson regresyon modeli için elde edilen 1.simülasyon tasarımı sonuçları gösterilmektedir. Bu sonuçlara göre 1. simülasyon tasarımı kapsamında PSO algoritması klasik seçim yöntemlerine göre AIC ve BIC değeri daha düşük olan modelleri seçmiştir. Çoklu bağlantı içeren durumlar için PSO daha doğru sonuçlar içeren Poisson regresyon modellerini elde etmektedir.

Tablo 3. 2.Simülasyon tasarımında Poisson regresyon modeli için AIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	AIC-İ	AIC-G	AIC-A	AIC-PSO
2	267.40070	267.40070	267.40070	267.40070
3	268.60160	268.60160	268.60160	268.60160
4	272.08920	272.08920	272.08920	272.08920
5	272.11820	272.11820	272.11820	272.11820
6	267.45050	267.45050	267.45050	267.45050
7	271.93650	271.93650	271.93650	271.93650
8	271.58820	271.58820	271.58820	271.58640
9	271.71860	271.71860	271.71860	271.71860
10	277.01660	277.01660	277.01660	277.01660
11	273.66320	273.66320	273.66320	273.64400
12	270.79890	270.79890	270.79890	270.79850
13	271.18850	271.18850	271.18850	271.18420
14	270.99200	270.99200	270.99200	270.98880
15	271.99420	271.99420	271.99420	271.96750

Tablo 4. 2.Simülasyon tasarımında Poisson regresyon modeli için BIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
2	273.09360	273.09360	273.09360	273.09360
3	276.73480	276.73480	276.73480	276.73480
4	282.54640	282.54640	282.54640	282.54640
5	284.29230	284.29230	284.29230	284.29230
6	281.29520	281.29520	281.29520	281.27240
7	287.23950	287.23950	287.23950	287.23950
8	288.10370	288.10370	288.10370	288.10370
9	289.49090	289.49090	289.49090	289.45200
10	296.19300	296.19300	296.19300	296.16350
11	293.12780	293.12780	293.12780	293.07570
12	291.82580	291.82580	291.82580	291.69400
13	292.45620	292.45620	292.45620	292.38310
14	293.33440	293.33440	293.33440	293.25460
15	295.22230	295.22230	295.22230	295.11200

2. simülasyon tasarımı sonuçları Tablo 3 ve 4' te verilmiştir. Sonuçlar incelendiğinde PSO algoritmasının klasik yöntemlere göre daha iyi sonuçlar verdiği görülmektedir. Ayrıca çoklu bağlantı problemi içermeyen durumlarda özellikle değişken sayısı yükseldikçe PSO algoritması AIC ve BIC değeri daha düşük olan Poisson regresyon modellerini seçmiştir.

Tablo 5. Affair verisi kapsamında Poisson regresyon modeli için AIC kriteri sonuçları

Seçim Yöntemi	AIC
AIC-İ	2823.17500
AIC-G	2823.17500
AIC-A	2823.17500
AIC-PSO	2822.76500

Tablo 6. Affair verisi kapsamında Poisson regresyon modeli için BIC kriteri sonuçları

Seçim Yöntemi	BIC
BIC-İ	2878.86300
BIC-G	2878.86300
BIC-A	2878.86300
BIC-PSO	2873.51200

Poisson regresyon modeli gerçek veri sonuçları Tablo 5 ve 6'da gösterilmektedir. Bu sonuçlara göre PSO algoritması Affairs veri setinde klasik seçim yöntemlerine göre daha başarılı sonuçlar elde etmektedir. PSO algoritmasının seçtiği Poisson regresyon modellerinde AIC ve BIC değerleri klasik yöntemlerin seçtiği modeller göz önünde bulundurulduğunda daha düşük değere sahiptir.

Negatif binomial regresyon modeli için 1. ve 2. simülasyon tasarım sonuçları ve gerçek veri uygulaması aşağıda verilmiştir.

Tablo 7. 1. Simülasyon tasarımında Negatif binomial regresyon modeli için AIC kriteri sonuçları

n	Seçim yöntemi			
	AIC-İ	AIC-G	AIC-A	AIC-PSO
30	82.2550	82.2550	82.2550	82.1273
50	134.1402	134.1402	134.1402	134.1179
100	270.5371	270.5371	270.5371	270.5230
200	536.2455	536.2455	536.2455	536.2729
500	1330.745	1330.745	1330.745	1330.727

Tablo 8. 1. Simülasyon tasarımında Negatif binomial regresyon modeli için BIC kriteri sonuçları

n	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
30	88.3657	88.3657	88.3657	88.1724
50	142.7844	142.7844	142.7844	142.4788
100	282.5406	282.5406	282.5406	282.5349
200	551.6531	551.6531	551.6531	551.7423
500	1353.393	1353.393	1353.393	1353.117

Tablo 7 ve 8'de negatif binomial regresyon modeli için elde edilen 1. simülasyon tasarımı sonuçları gösterilmektedir. Bu sonuçlara göre 1. simülasyon tasarımı kapsamında PSO algoritması klasik seçim yöntemlerine göre AIC ve BIC değeri daha düşük olan modelleri seçmiştir. Çoklu bağlantı içeren durumlar için PSO daha doğru sonuçlar içeren negatif binomial regresyon modellerini elde etmektedir.

Tablo 9. 2.Simülasyon tasarımında Negatif binomial regresyon modeli için AIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	AIC-İ	AIC-G	AIC-A	AIC-PSO
2	269.6135	269.6135	269.6135	269.6135
3	270.5622	270.5622	270.5622	270.5622
4	270.1624	270.1624	270.1624	270.1624
5	271.8613	271.8613	271.8613	271.8613
6	276.1486	276.1486	276.1486	276.1486
7	272.1477	272.1477	272.1477	272.1477
8	273.6337	273.6337	273.6337	273.6337
9	274.8716	274.8716	274.8716	274.8716
10	275.6479	275.6479	275.6479	275.6479
11	275.8439	275.8439	275.8439	275.8423
12	276.6164	276.6164	276.6164	276.607
13	278.2293	278.2293	278.2293	278.1894
14	278.6634	278.6634	278.6634	278.6287
15	275.8236	275.8236	275.8236	275.8094

Tablo 10. 2.Simülasyon tasarımında Negatif binomial regresyon modeli için BIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
2	277.9403	277.9403	277.9403	277.9403
3	281.3456	281.3456	281.3456	281.3456
4	283.0921	283.0921	283.0921	283.0921
5	286.5748	286.5748	286.5748	286.5748
6	292.5683	292.5683	292.5683	292.5683
7	298.1889	298.1889	298.1889	298.1889
8	292.6785	292.6785	292.6785	292.6673
9	295.3407	295.3407	295.3407	295.3357
10	296.9031	296.9031	296.9031	296.8889
11	298.3031	298.3031	298.3031	298.2632
12	300.1552	300.1552	300.1552	300.1241
13	302.7383	302.7383	302.7383	302.5759
14	302.0067	302.0067	302.0067	301.9964
15	301.2273	301.2273	301.2273	301.1695

Negatif binomial regresyon modeli için elde edilen 2. simülasyon tasarımı sonuçları Tablo 9 ve 10'da verilmiştir. Çoklu bağlantı problemi içermeyen durumlarda özellikle değişken sayısı yükseldikçe PSO algoritmasının AIC ve BIC değeri daha düşük olan negatif binomial regresyon modellerini seçtiği görülmektedir.

Tablo 11. Affair verisi kapsamında Negatif binomial regresyon modeli için AIC kriteri sonuçları

Seçim Yöntemi	AIC
AIC-İ	2299.23355
AIC-G	2299.23355
AIC-A	2299.23355
AIC-PSO	2296.10252

Tablo 12. Affair verisi kapsamında Negatif binomial regresyon modeli için BIC kriteri sonuçları

Seçim Yöntemi	BIC
BIC-İ	2223.51400
BIC-G	2223.51400
BIC-A	2223.51400
BIC-PSO	2221.10300

Tablo 11 ve 12'de negatif binomial regresyon modeli gerçek veri sonuçları gösterilmektedir. Bu sonuçlara göre PSO algoritması Affairs veri setinde klasik seçim yöntemlerine göre daha başarılı sonuçlar elde etmektedir.

Sıfır yoğunluklu Poisson regresyon modeli için 1. ve 2. simülasyon tasarım sonuçları ve gerçek veri uygulaması aşağıda verilmiştir.

Tablo 13. 1. Simülasyon tasarımında Sıfır yoğunluklu Poisson regresyon modeli için AIC kriteri sonuçları

n	Seçim yöntemi			
	AIC-İ	AIC-G	AIC-A	AIC-PSO
30	88.3450	88.3450	88.3450	87.9373
50	146.2702	146.2702	146.2702	144.9679
100	281.7771	281.7771	281.7771	280.6280
200	562.3625	562.3625	562.3625	561.2836
500	1402.643	1402.643	1402.643	1401.8672

Tablo 14. 1. Simülasyon tasarımında Sıfır yoğunluklu Poisson regresyon modeli için BIC kriteri sonuçları

n	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
30	94.2478	94.2478	94.2478	93.8963
50	149.7452	149.7452	149.7452	149.2356
100	301.3602	301.3602	301.3602	300.7895
200	592.3547	592.3547	592.3547	592.1254
500	1412.6325	1412.6325	1412.6325	1412.1362

Tablo 13 ve 14' de sıfır yoğunluklu Poisson regresyon modeli için elde edilen 1. simülasyon tasarımı sonuçları gösterilmektedir. Bu sonuçlara göre 1. simülasyon tasarımı kapsamında PSO algoritması klasik seçim yöntemlerine göre AIC ve BIC değeri daha düşük olan modelleri seçmiştir. Çoklu bağlantı içeren durumlar için PSO daha doğru sonuçlar içeren sıfır yoğunluklu Poisson regresyon modellerini elde etmektedir.

Tablo 15. 2.Simülasyon tasarımında Sıfır yoğunluklu Poisson regresyon modeli için AIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	AIC-İ	AIC-G	AIC-A	AIC-PSO
2	276.2348	276.2348	276.2348	276.2348
3	276.9852	276.9852	276.9852	276.9852
4	277.3247	277.3247	277.3247	277.3247
5	278.1456	278.1456	278.1456	278.1456
6	279.0523	279.0523	279.0523	279.0523
7	280.8563	280.8563	280.8563	280.8563
8	281.3659	281.3659	281.3659	281.3659
9	282.1748	282.1748	282.1748	282.1748
10	283.8741	283.8741	283.8741	283.6235
11	284.6398	284.6398	284.6398	284.5236
12	285.1862	285.1862	285.1862	285.1021
13	286.9963	286.9963	286.9963	286.7253
14	288.6548	288.6548	288.6548	288.6321
15	291.3258	291.3258	291.3258	290.9652

Tablo 16. 2.Simülasyon tasarımında sıfır yoğunluklu Poisson regresyon modeli için BIC kriteri sonuçları

%(p/n)	Seçim yöntemi			
	BIC-İ	BIC-G	BIC-A	BIC-PSO
2	285.3659	285.3659	285.3659	285.3659
3	289.6258	289.6258	289.6258	289.6258
4	291.0785	291.0785	291.0785	291.0785
5	292.8574	292.8574	292.8574	292.8574
6	298.5217	298.5217	298.5217	298.5217
7	300.9852	300.9852	300.9852	300.9327
8	301.2386	301.2386	301.2386	301.2321
9	296.2178	296.2178	296.2178	296.1993
10	299.5982	299.5982	299.5982	299.5821
11	302.2185	302.2185	302.2185	301.9962
12	303.1587	303.1587	303.1587	302.8754
13	303.6726	303.6726	303.6726	303.3627
14	304.1289	304.1289	304.1289	304.0185
15	303.8756	303.8756	303.8756	303.5214

Tablo 15 ve 16' da sıfır yoğunluklu Poisson regresyon modeli için elde edilen 2. simülasyon tasarımı sonuçları gösterilmektedir. Çoklu bağlantı problemi içermeyen durumlarda PSO algoritmasının daha iyi sonuçlar verdiği görülmektedir.

Tablo 17. Affair verisi kapsamında Sıfır yoğunluklu Poisson regresyon modeli için AIC kriteri sonuçları

Seçim Yöntemi	AIC
AIC-İ	1928.21512
AIC-G	1928.21512
AIC-A	1928.21512
AIC-PSO	1928.07231

Tablo 18. Affair verisi kapsamında Sıfır yoğunluklu Poisson regresyon modeli için BIC kriteri sonuçları

Seçim Yöntemi	BIC
BIC-İ	1958.16705
BIC-G	1958.16705
BIC-A	1958.16705
BIC-PSO	1955.52127

Sıfır yoğunluklu Poisson regresyon modeli gerçek veri Tablo 17 ve 18'de verilmektedir. PSO algoritmasının seçtiği sıfır yoğunluklu Poisson regresyon modellerinde AIC ve BIC değerleri klasik yöntemlerin seçtiği modeller göz önünde bulundurulduğunda daha düşük değere sahiptir.

4. Tartışma ve Sonuç

Bu çalışmada sayım modelleri kapsamında model seçimleri incelendi. Sayım modellerinde model seçimi için klasik seçim yöntemleri ve PSO algoritması kullanıldı. Çalışma sonucunda klasik değişken seçim yöntemleri olan ileriye doğru, geriye doğru ve adımsal seçim yöntemleri ve PSO algoritması ile elde edilen sonuçlar karşılaştırıldı. Karşılaştırma için değişken seçimi algoritmaları ile elde edilen regresyon modellerinin bilgi kriteri değerleri (AIC ve BIC) baz alındı. Bilgi kriteri değerleri doğrultusunda; seçilen değişken kümeleri ile oluşturulan regresyon modellerinin performansı değerlendirildi. Klasik seçim yöntemlerinin model seçiminde hem simülasyon hem de gerçek veriler için aynı bilgi kriteri değerlerine sahip olan sayım modellerini seçtiği tespit edildi. Optimizasyon algoritmalarına dayalı olmayan seçim yöntemlerinin çalışma mantığına koşut olarak aynı sayım modellerinin seçilmesi dikkat çekici bir sonuçtur. Simülasyon ve gerçek veri setlerine uygulanan analizler sonucunda her ikisinde de PSO algoritmasının klasik yöntemlere göre daha düşük bilgi kriteri değerine sahip regresyon modelleri seçtiği görülmüştür.

Sonuç olarak klasik yöntemlerle kıyaslandığında PSO algoritmasının, modeldeki değişken sayısı arttıkça ve bağımsız değişkenler arasındaki korelasyon değerleri yükseldikçe daha iyi sonuçlar verdiği farklı model yapıları [28] gibi sayım modellerinde de PSO algoritmasının değişken seçiminde alternatif bir yöntem olarak kullanılabileceği gösterilmiştir.

Teşekkür

Bu makale Çankırı Karatekin Üniversitesi Bilimsel Araştırma Projeleri birimi tarafından desteklenen Araştırma Projesi (Proje No: FF090316B15) kapsamında yapılan çalışmalar sonucunda oluşturulmuştur. Yazarlar desteklerinden dolayı Çankırı Karatekin Üniversitesi Bilimsel Araştırma Projeleri Birimine teşekkür etmektedir.

Kaynakça

[1] George, E. I. 2000. The variable selection problem. Journal of the American Statistical Association, 95(2000), 1304-1308.

[2] Bozdogan, H. 2004. Intelligent statistical data mining with information complexity and genetic algorithms. Statistical data mining and knowledge discovery(2004), 15-16.

[3] Lee, K. Y., & El-Sharkawi, M. A. 2008. Modern heuristic optimization techniques: theory and applications to power systems. John Wiley & Sons.

[4] Drezner, Z., Marcoulides, G. A., & Salhi, S. 1999. Tabu search model selection in multiple regression analysis. Communications in Statistics-Simulation and Computation, 28(1999), 346-367.

[5] Örkcü, H. H. 2013. Subset selection in multiple linear regression models: a hybrid of genetic and simulated annealing algorithms. Applied Mathematics and Computation, 23(2013), 11018-11028.

[6] Pacheco, j., Casado, S., & Nunez, L. A. 2009. Variable selection method based on Tabu search for logistic regression models. European Journal of Operational Research, 199(2009), 506-511.

[7] Unler, A., & Murat, A. 2010. A discrete particle swarm optimization method for feature

selection in binary classification problems. European Journal of Operational Research, 206(2010), 528-539.

[8] Sakate, D. M., Kashid, D. N., & Shirke, D. T. 2011. Subset Selection in Poisson Regression. Journal of Statistical Theory and Practice, 5(2011), 207-219.

[9] McLeod, A. I., & Xu, C. 2010. R-project. org/package= bestglm. http://CRAN (Erişim Tarihi: 19.10.2017)

[10] Calcagno, V., & Mazancourt, C. 2010. glmulti: an R package for easy automated model selection with (generalized) linear models. Journal of Statistical Software, 34(2010), 1-29.

[11] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) (1996), 267-288.

[12] Zou, H., & Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2005), 301-320.

[13] Zou, H. 2006. The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(2006), 1418-1429.

[14] Zhao, W., Zhang, R., Lv, Y., & Liu, J. 2014. Variable selection for varying dispersion beta regression model. Journal of Applied Statistics, 41(2014), 95-108.

[15] Bayer, F. M., & Cribari-Neto, F. 2014. Bootstrap-based model selection criteria for beta regressions. TEST(2014), 1-20.

[16] Bozdogan, H. 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika(1987), 345-370.

[17] Hurvich, C. M., & Tsai, C. L. 1989. Regression and time series model selection in small samples. Biometrika(1989), 297-307.

[18] Bollen, K. A., Ray, S., Zavisca, J., & Harden, J. J. 2012. A comparison of Bayes factor approximation methods including two new methods. Sociological Methods & Research, 41(2012), 294-324.

[19] Bozdogan, H. 2000. Akaike's information criterion and recent developments in information complexity. Journal of mathematical psychology, 44(2000), 62-91.

[20] Bozdogan, H. 2010. A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation. Journal of the School of Business Administration(2010), 370-398.

[21] Deniz, E., Akbilgic, O., & Howe, J. A. (2011). Model selection using information criteria under a new estimation method: least squares ratio. Journal of Applied Statistics, 2043-2050.

[22] Pamukçu, E., Bozdogan, H., & Çalık, S. 2015. A Novel Hybrid Dimension Reduction Technique for Undersized High Dimensional Gene Expression Data Sets Using Information Complexity Criterion for Cancer Classification. Computational and mathematical methods in medicine(2015), Article ID 370640, 14 pages.

[23] Cameron, A. C., & Trivedi, P. K. 1998. Regression Analysis of Count Data. Cambridge University Press.

[24] Jansakul, N., & Hinde, J. P. 2002. Score tests for zero-inflated Poisson models. Computational, 40(2002), 75-96.

[25] Eberhart, R., & Kennedy, J. 1995. A new optimizer using particle swarm theory. In Micro Machine and Human Science. Proceedings of the Sixth International Symposium on IEEE., 39-43.

[26] Özsağlam, M. Y., & Çunkaş, M. 2008. Optimizasyon Problemlerinin Çözümü için Parçaçık Sürü Optimizasyonu Algoritması. Politeknik Dergisi(2008), 11.

[27] Hilbe, J. M. 2016. COUNT: Functions, Data and Code for Count Data. R package version 1.3.4. https://CRAN.R-project.org/package=COUNT

[28] Koç, H., Dünder, E., Gümüştekin, S., Koç, T., & Cengiz, M. A. 2018. Particle swarm optimization-based variable selection in Poisson regression analysis via information complexity-type criteria. Communications in Statistics-Theory and Methods, (2018) 47(21), 5298-5306.