Cezalandırılmış Eğrisel Çizgi Regresyonunda Karışık Doğrusal Model Yaklaşımı

Semra Türkan[1,*], Öniz Toktamış[1]

[1] *Hacettepe Üniversitesi, İstatistik Bölümü, Beytepe/ANKARA*

Özet

Bu çalışmada cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı incelenmiştir. Bu amaçla İstanbul Sanayi Odası tarafından yürütülen "Türkiye'nin 500 Büyük Sanayi Kuruluşu" çalışmasının 2009 yılına ilişkin verileri kullanılmıştır. İşletmelerin işgücü verimliliği ve aktif karlılığı arasındaki parametrik olmayan ilişki cezalandırılmış eğrisel çizgi regresyonu ve cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı kullanılarak tahmin edilmiştir. R paket programında karışık doğrusal model için geliştirilen "nlme" paketi kullanılmıştır. Karışık doğrusal model yaklaşımı kullanılarak tahmin edilen modelin hata kareler ortalaması daha küçük bulunmuştur.

Anahtar Kelimeler: Karışık doğrusal model, cezalandırılmış eğrisel çizgi regresyonu, parametrik olmayan regresyon, iş gücü verimliliği, aktif karlılığı

Linear Mixed Model Approach in Penalized Spline Regression

Abstract

In this study, it was examined the linear mixed model approach in the penalized spline regression. For this purpose, the data for the year 2009 which is included in the "Turkey's Top 500 Industrial Enterprises" study conducted by the İstanbul Chamber of Industry was used. The nonparametric relationship between labor productivity and return on assets of the enterprises was estimated by using penalized spline regression and linear mixed model approach in penalized spline regression. "nlme" package which is developed for linear mixed model in R was used. The mean square error of the model which was estimated by using linear mixed model approach is found smaller.

Keywords: Linear mixed model, penalized spline regression, nonparametric regression, labor productivity, the return on assets

1. Giriş

Parametrik olmayan regresyon, (x_i, y_i) veri çiftleri arasındaki ilişkinin şekli bilinmediğinde yaygın olarak kullanılan bir modeldir. Bu model kullanılarak düzgün (smooth) fonksiyonlar oluşturulabilir. Bu fonksiyonlar, veri kümesini temsil eden düzgün bir eğri oluşturarak bağımlı değişken için daha iyi tahminler yapılmasını sağlar. Bu nedenle bu işleme düzleştirme (smoothing) adı verilir [1]. (Redpath, 2008). Literatürde çoğu kez düzleştirme yöntemleri olarak tanımlanan çok sayıda parametrik olmayan regresyon yöntemi önerilmiştir. Cezalandırılmış eğrisel çizgi (penalized splines ya da P-Splines) regresyon yöntemi de bunlardan birisidir.

Cezalandırılmış eğrisel çizgi regresyon yöntemi ile karışık doğrusal modeller arasında yakın bir ilişki vardır. Cezalandırılmış eğrisel çizgi regresyon modeli, karışık doğrusal modele benzetilerek ifade edilebilir. Literatürde karışık doğrusal model ile cezalandırılmış eğrisel çizgi regresyon modeli arasındaki

[*]e-mail: sturkan@hacettepe.edu.tr

ilişkiyi inceleyen çalışmalar vardır. Cezalandırılmış eğrisel çizgilerin karışık doğrusal model gösterimini tanıtılmıştır [2]. Toplamsal modellerde cezalandırılmış eğrisel çizgileri kullanılmıştır [3]. Parise vd. (2001) ve Aerts vd. (2002) karışık model yaklaşımını genelleştirilmiş toplamsal modellere uygulamışlardır [4, 5]. Ruppert vd. (2003), cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımını incelemişlerdir [6].

Karışık doğrusal modelde amaç, hem sabit etkilere hem de rasgele etkilere ilişkin parametre kestirimlerinin elde edilmesidir. Bu amaçla önerilen yöntemlerden birisi, hem sabit hem de rasgele etkilere ilişkin parametrelerin birlikte kestirilebildiği en iyi doğrusal yansız kestirim (best linear unbiased predictor, BLUP) yöntemidir. Bu çalışmada cezalandırılmış eğrisel çizgi regresyon modeli karışık doğrusal modele benzetilip BLUP yaklaşımı kullanılarak tahminler edilmiştir.

2. Materyal ve Yöntem

2.1. Cezalandırılmış eğrisel çizgi regresyonu

Verilen x değerleri için y'nin beklenen değeri ve y ile x arasındaki ilişki genel olarak

$$y = m(x) + \varepsilon \tag{1}$$

biçiminde modellenebilir. Burada $m(x) = E(y/X = x)$ olup regresyon fonksiyonu adını alır. Parametrik olmayan regresyon fonksiyonu m(x)'in belli bir biçiminin olmadığı varsayılır. Parametrik olmayan regresyonun amacı, parametreleri tahmin etmekten ziyade doğrudan m(.) fonksiyonunu tahmin etmektir. $\kappa_1 < ... < \kappa_K$, $(\min(x_i) \le \kappa_1 < ... < \kappa_K \le \max(x_i))$, düğüm noktaları kümesi için Eş. (1)'deki m(x) fonksiyonu cezalandırılmış eğrisel çizgi regresyonu olarak

$$m(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - \kappa_k) + \varepsilon \tag{2}$$

biçiminde ifade edilir. Eş.(2)'de

$$(x - \kappa_k)_+ = \begin{cases} 0, & x \le \kappa_k \\ (x - \kappa_k), & x > \kappa_k \end{cases}$$

dır. Cezalandırılmış eğrisel çizgi regresyonunda düğüm noktaları

$$\kappa_k = \left(\frac{k+1}{K+2}\right)$$

formülü ile seçilebilir. Bu formülde K=min(1/4 (tekrar etmeyen x_i'lerin sayısı), 35) dir [7]. Ayrıca literatürde düğüm noktalarının seçimi için bazı algoritmalar da önerilmiştir. Eş.(2)'deki $\beta = (\beta_0, \beta_1, b_1, ..., b_K)^T$ parametreler vektörünün tahmini,

$$S(\beta) = \|y - X\beta\|^2 + \lambda \beta^T D \beta \tag{3}$$

biçimindeki cezalandırılmış hata kareler toplamı minimum yapılarak bulunur (λ 0). Burada **D** matrisi

$$D = \begin{bmatrix} 0_{2x2} & 0_{2xK} \\ 0_{Kx2} & I_{KxK} \end{bmatrix}$$

biçimindedir. Eş. (3)'deki ifadede $\lambda\beta^T D\ \beta$ terimi pürüzlülük cezası (roughness penalty) olarak, λ ise düzleştirme parametresi olarak adlandırılır. Eş. (3)'deki fonksiyonun β 'ya göre türevi alınıp sıfıra eşitlenirse β 'nın tahmini $\hat{\beta}$,

$$\hat{\beta} = (X^T X + \lambda D)^{-1} X^T y \tag{4}$$

olarak bulunur. Eş. (4)'de λ 'nın seçimi için literatürde birçok yöntem geliştirilmiştir. Çapraz geçerlilik ve genelleştirilmiş çapraz geçerlilik yöntemleri bu yöntemlerin içinde en çok kullanılanlarıdır.

2.2. Cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı

Eş (2)'deki cezalandırılmış eğrisel çizgi regresyon modeli karışık doğrusal modele benzetilerek gösterilebilir. Buna göre Eş.(2)'deki model

$$y_i = m(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k\ (x - \kappa_k)_+ + \varepsilon \tag{5}$$

biçiminde ifade edilebilir. Eş.(5)'de $\beta = [\beta_0, \beta_1]^T$, x değerlerine ilişkin regresyon katsayıları vektörünü ve $u = (u_1,\ldots,u_K)^T$ düğüm noktalarına ilişkin regresyon katsayıları vektörünü göstermektedir. Eş. (5)'deki ifade matris formunda

$$y = X\beta + Zu + \varepsilon \tag{6}$$

biçiminde yazılabilir. Eş. (5)'deki model hem sabit etkilerin hem de rasgele etkilerin birlikte yer aldığı karışık doğrusal modeli ifade etmektedir. Eş.(6)'da sabit etkilere ilişkin tasarım matrisi **X** ve rasgele etkilere ilişkin tasarım matrisi **Z**

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad Z = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_k)_+ \\ \vdots & \vdots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - \kappa_k)_+ \end{bmatrix}$$

biçimindedir. Eş. (6)'daki karışık doğrusal model gösteriminde β , sabit etkilere ilişkin parametreler vektörünü, **u** ise, rasgele etkilere ilişkin parametreler vektörünü ve ε rasgele hatalar vektörünü göstermektedir. Eş. (6)'da **u** ve ε 'nun beklenen değerlerinin ve **u** ile ε arasındaki varyans-kovaryans matrisinin,

$$E\begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{ve Cov} \begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_\varepsilon^2 I \end{bmatrix} \tag{7}$$

biçiminde olduğu varsayılsın. Buna göre Eş. (6)'daki karışık doğrusal modelde β ve **u** parametreler vektörlerinin birlikte tahmini en iyi doğrusal yansız kestirici (BLUP) kullanılarak elde edilebilir. Eş. (7)'den **u** bilindiğinde **y**'nin

$$y \mid u \sim N(X\beta + Zu,\ \sigma_\varepsilon^2 I), \qquad u \sim N(0,\ \sigma_u^2 I)$$

şeklinde normal dağıldığı varsayılır [8]. **u** ve **y**'nin bileşik yoğunluk fonksiyonu,

$$f(u\ ;y) = f(y\ |u)f(u)$$

$$\cong \exp\{\ -\frac{1}{2}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu}\ )^{T}(\mathbf{y}-\mathbf{X\beta}-\mathbf{Zu}\ )+\frac{\sigma_{\varepsilon}^{2}}{\sigma_{u}^{2}}\mathbf{u}^{T}\mathbf{u}\} \tag{8}$$

dır. Eş. (8)'den $f(u\ ;y)$ 'nun log-olabilirlik fonksiyonu

$$L = \log\ (f(u;\ y) \approx (y - X\beta - Zu)^{T}(y - X\beta - Zu) + \lambda\, u^{T}u \tag{9}$$

olarak elde edilir. Eş. (9)'dan düzleştirme parametresinin $\lambda = \sigma_{\varepsilon}^{2}/\sigma_{u}^{2}$ olduğu görülmektedir. Eş. (9)'daki ifadenin β ve **u** 'ya göre türevleri bulunup sıfıra eşitlenirse (β , **u**)'nun en iyi doğrusal yansız kestiricisi (BLUP)

$$\begin{bmatrix}\hat{\beta}\\ \hat{u}\end{bmatrix} = (C^{T}C + \lambda D)^{-1}C^{T}y \tag{10}$$

olarak elde edilir. Burada $C \equiv [X \quad Z]$ ve $D = \begin{bmatrix}0_{2x2} & 0_{2xK}\\ 0_{Kx2} & I_{KxK}\end{bmatrix}$ dir. Kestirim değerleri vektörü ise

$$\hat{m} = X\hat{\beta} + Z\hat{u} = C(C^{T}C + \lambda D)^{-1}C^{T}y \tag{11}$$

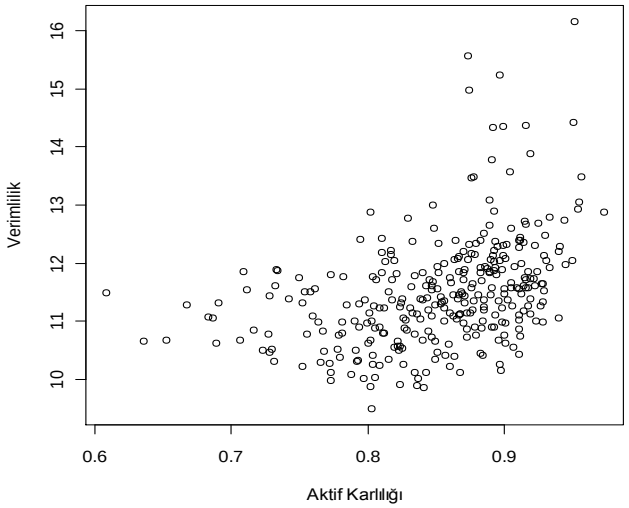
olarak bulunur. Eş. (11)'deki ifade parametrik olmayan m fonksiyonunun en iyi doğrusal yansız kestiricisinin "Ridge Regresyon" gösterimidir [6].

Cezalandırılmış eğrisel çizgi regresyonunun karışık doğrusal model olarak gösterimi oldukça yararlıdır. Çünkü karışık doğrusal model olarak ifade edilebilen cezalandırılmış eğrisel çizgi regresyonunda kestirimler karışık doğrusal model teorisi ve karışık doğrusal model için geliştirilen bilgisayar programları kullanılarak elde edilebilir. Ayrıca karışık doğrusal modelde kestirilen varyans oranı, $\sigma_{\varepsilon}^{2}/\sigma_{u}^{2}$, eğrisel çizgi regresyonundaki düzleştirme parametresine, λ, karşılık gelmektedir. Bu da düzleştirme parametresinin seçiminin önemli olduğu eğrisel çizgi regresyonunda düzleştirme parametresinin belirlenmesinde önemli kolaylık sağlar.

3. Uygulama

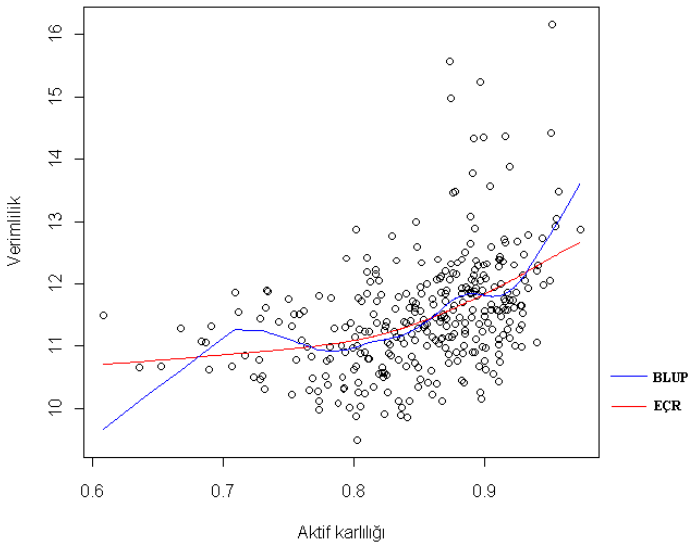
Bu çalışmada, cezalandırılmış eğrisel çizgi regresyonda BLUP yaklaşımı ile parametrik olmayan regresyon fonksiyonunu kestirmek için İstanbul Sanayi Odası tarafından yürütülen "Türkiye'nin 500 Büyük Sanayi Kuruluşu" çalışmasının 2009 yılına ait verileri kullanılmıştır. İşletmelere ait işgücü verimliliği ve aktif karlılığı değişkenleri arasındaki ilişki incelenmiştir. Verimlilik, bir üretim ya da hizmet sisteminin ürettiği çıktı ile bu çıktıyı yaratmak için kullandığı girdi arasında ilişki olarak tanımlanmıştır. Verimlilik bir işletmede kaynakların ne ölçüde etkin kullanıldığını gösteren bir ölçüdür. Bu nedenle verimliliğin ölçülmesi ve verimliliği etkileyen faktörlerin belirlenip kontrol altına alınması işletmeler için oldukça önemlidir. İşgücü verimliliği ise işletmelerin üretim performanslarındaki iyileşmeyi ve gelişmeyi, dolayısıyla sektörlerin rekabet güçlerini yansıtan önemli bir göstergedir [9]. Verimlilik göstergesi olarak işgücü verimliliği alınmıştır. İşgücü verimliliği *İşgücü verimliliği=Brüt Katma değer/Çalışan Sayısı* biçiminde hesaplanmaktadır. Bu çalışmada, işletmelerin *"işgücü verimliliği"* bağımlı değişken, *"aktif karlılığı (Dönem Karı/Net Aktif)"* ise bağımsız değişken olarak ele alınmıştır.

Bazı işletmelere ilişkin verilerin eksik olması nedeniyle bu işletmeler veri kümesinden çıkarılarak 318 tane işletme analize alınmıştır. Şekil 1'deki işgücü verimliliği ve aktif karlılığına ilişkin saçılım grafiğinden bu iki değişken arasında parametrik olmayan bir ilişki olduğu görülmektedir.



Şekil 1. İşgücü verimliliği ve aktif karlılık ilişkin saçılım grafiği

Bu çalışmada işgücü verimliliği ve aktif karlılığı arasındaki ilişki hem eğrisel çizgi regresyonu hem de eğrisel çizgi regresyonunda BLUP yaklaşımı kullanılarak incelenmiştir. Eğrisel çizgi regresyonunda BLUP yaklaşımını kullanmak kolaylık sağlamaktadır. BLUP yaklaşımında eğrisel çizgi regresyonu karışık doğrusal modele benzetilerek tahminler elde edildiği için karışık doğrusal model için geliştirilen paket programları kullanılarak tahminler kolaylıkla elde edilebilir. Bu çalışmada da tahminler R paket programında karışık doğrusal model için geliştirilen "nlme" paketi kullanılarak elde edilmiştir. Ayrıca eğrisel çizgi regresyonunda düzleştirme parametresinin seçimi önemli bir konudur. Düzleştirme parametresinin seçimi için çapraz geçerlik, genelleştirilmiş çapraz geçerlilik vb. gibi birçok yöntem geliştirilmiştir. Ancak eğrisel çizgi regresyonunda BLUP yaklaşımı kullanıldığında düzleştirme parametresinin seçimi için herhangi bir yönteme gerek duyulmaz. Çünkü Eş.(9)'dan BLUP yöntemi ile tahmin edilen varyans oranı ($\sigma_\varepsilon^2 / \sigma_u^2$) düzleştirme parametresi olarak elde edilir. *İlk olarak cezalandırılmış eğrisel çizgi regresyonunda kullanılacak düğüm sayılarına karar* verilmelidir. Bunun için K=min(1/4 (tekrar etmeyen x_i'lerin sayısı), 35) formülünden yararlanılmıştır. Buna göre düğüm noktalarının sayısı K=min(79.2, 35)=35 olarak bulunmuştur. Düğüm noktası sayısı belirlendikten sonra R programında yazılan kodlar ile bu düğüm noktaları için **X** ve **Z** matrisleri yeniden oluşturulmuştur. Karışık doğrusal model için geliştirilen "nlme" paket programı kullanılarak Eş.(6)'daki karışık doğrusal model için parametre kestirimleri elde edilmiştir. Bulunan parametre kestirimleri yerine yazılarak Eş.(11)'deki kestirim değerleri , $\hat{m}_i$, bulunmuştur. Düzleştirme parametresi $\lambda = \sigma_\varepsilon^2 / \sigma_u^2$ = 0.7088 / 120.58 = 0.006 olarak elde edilmiştir. Daha sonra bu veriler için aynı düğüm noktalarında Eş.(2)'deki eğrisel çizgi regresyon modeli kullanılarak $\hat{m}_i$ değerleri elde edilmiştir. Eş.(2)'deki eğrisel çizgi regresyon modeli için düzleştirme parametresi "genelleştirilmiş çapraz geçerlilik" yöntemi ile λ = 0.8875 olarak elde edilmiştir. Eğrisel çizgi regresyonunda BLUP yaklaşımı kullanılarak yapılan düzleştirme ve eğrisel çizgi regresyonu kullanılarak yapılan düzleştirme Şekil 2'deki gibi elde edilmiştir.



Şekil 2. BLUP kullanılarak elde edilen düzleştirme ve Eğrisel Çizgi Regresyonu (EÇR) ile elde edilen düzleştirme

Şekil 2'den görüldüğü gibi eğrisel çizgi regresyonunda BLUP yaklaşımı kullanıldığında daha iyi bir düzleştirme elde edilmiştir. Ayrıca her iki modele ilişkin hata kareler ortalaması

$$\text{HKO}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

formülü kullanılarak hesaplanmış ve Tablo1'deki gibi elde edilmiştir.

Tablo 1. Yöntemlere ilişkin HKO

Kullanılan Düzleştirme Yöntemi	HKO
BLUP Yaklaşımı	0,6926
Eğrisel Çizgi Regresyonu	0,7156

Tablo 1'den görüldüğü gibi eğrisel çizgi regresyonunda BLUP yaklaşımı kullanıldığında elde edilen hata kareler ortalaması daha düşüktür.

4. Sonuç

Eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımının kullanılması tahminlere ilişkin hatayı azaltmıştır. Ayrıca veriyi temsil eden daha iyi bir düzleştirme elde edilmiştir. Cezalandırılmış eğrisel çizgi regresyonun karışık doğrusal model olarak gösterilmesi, karışık doğrusal model teorisinden yararlanarak tahminlerin elde edilmesini sağlamıştır. Bu amaçla karışık doğrusal model için geliştirilen paket programlarından yararlanılabilmiştir. Ayrıca cezalandırılmış eğrisel çizgi regresyonunun karışık doğrusal model olarak gösterilmesi sayesinde düzleştirme parametresinin seçimi için herhangi bir yöntemin kullanılmasına gerek kalmamıştır. Ayrıca incelenen veri kümesi için cezalandırılmış eğrisel çizgi regresyonunda karışık doğrusal model yaklaşımı kullanıldığında elde edilen hata kareler ortalaması, cezalandırılmış eğrisel çizgi regresyonu kullanılarak elde edilen hata kareler ortalamasından daha küçük bulunmuştur. Sonuç olarak cezalandırılmış eğrisel çizgi regresyonunun karışık doğrusal model olarak gösterimi araştırmacılara kolaylık sağlamaktadır.

5. Kaynaklar

[1] Redpath M., Comparing Kernel Smoothing, Spline Smoothing and Smoothing Splines. http://maths.dur.ac.uk/Ug/projects/library/CM3/000463064r.pdf, 2008.

[2] Brumback B. A., Ruppert D., Wand M. P., Comment on Shively, Kohn&Wood, Journal of The American Statistical Association, 94, 794-807, 1999.

[3] Coull B.A., Schwartz J., Wand M.P., Respiratory Health and Air Pollution: Additive Mixed Model Analyses, Biostatistics, 2, 337-350, 2001.

[4] Parise H., Wand M. P., Ruppert D., Ryan L., , Incorporation of Historical Controls Using Semiparametric Mixed Models, Journal of The Royal Statistical Society, Series C, 50, 31-42, 2001.

[5] Aerts M., Claeskens G., Wand M.P., Some Theory for Penalized Spline Generalized Additive Models, Journal of Statistical Planning and Inference, 103, 455-470, 2002.

[6] Ruppert D., Wand M.P., Carroll R.J., Semiparametric Regression, Cambrige University Pres, 2003.

[7] Yao F., Lee T.C.M., On knot placement for penalized spline regression, Journal of the Korean Statistical Society, 37, 259-267, 2008.

[8] Türkan S., Yarı Parametrik Regresyon Modelinde Etkili Gözlem Analizi, Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 106, 2012.

[9] Tezcan N., Verimliliği Etkileyen Faktörlerin Analizi: Türkiye'nin 500 Büyük Sanayi Kuruluşu Üzerinde Bir Uygulama, Verimlilik Dergisi, 3, 2010.