Eurasian Academy of Sciences
Eurasian Econometrics, Statistics & Emprical Economics Journal
2017 Volume:6 S: 18 - 35
Published Online January 2017 (http://econstat.eurasianacademy.org)
http://dx.doi.org/10.17740/eas.stat.2017-V6-02

SPLINE MODELS WHICH USE IN LONGITUDINAL DATA ANALYSIS

Seda BAĞDATLI KALKAN *
* İstanbul Ticaret Üniversitesi, Uluslararası Ticaret Bölümü
E-mail: sbagdatli@ticaret.edu.tr

Copyright © 2017 Seda BAĞDATLI KALKAN. This is an open access article distributed under the Eurasian Academy of Sciences License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT
Longitudinal data is defined as data obtained by a repeated measurement of variables pertaining to the same units over time. The analysis of longitudinal data cannot be achieved through classical regression models because of the independence and multicollinearity assumptions. For this reason, specific regression models have been developed for such data. Classical parametric models are based on the rationale that the relation between the dependent variable and the independent variable(s) is linear or the relation is expressed through known parametric functions. In such a case, it is not possible to reveal the actual structure of the relation, which will prevent the researcher from achieving reliable and rational outcomes particularly in longitudinal datasets. Non-parametric regression model is utilized in cases where the relation between the dependent variable and the independent variable(s) is more complicated in longitudinal data. In this study spline models in nonparametric regression models which use in longitudinal data are investigated theoretically.

Keywords: Longitudinal data, Nonparametric Regression, Spline Models, Smoothing Splines, penalized Splines, Regression Splines.

JEL Classification: C1, C18, C23

BOYLAMSAL VERİLERİN ANALİZİNDE KULLANILAN SPLAYN MODELLERİ

ÖZET

Boylamsal veriler, aynı birimlere ait özelliklerin zaman içerisinde tekrarlı olarak ölçülmesi ile elde edilen verilerdir. Boylamsal verilerin analizi, bağımsızlık ve sabit varyanslılık varsayımı sağlanamadığından klasik regresyon modelleri ile gerçekleştirilememektedir. Bu nedenden dolayı boylamsal veriler için özel regresyon modelleri geliştirilmiştir. Klasik parametrik modeller, bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkinin doğrusal olması veya ilişkinin bilinen parametrik fonksiyonlarla ifade edilmesi temeline dayanmaktadır. Bu durumda da gerçek ilişki yapısı ortaya çıkarılamamaktadır. Boylamsal veri setlerinde bu durum güvenilir ve mantıklı sonuçlara ulaşılmasını engelleyecektir. Dolayısıyla boylamsal verilerde bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkinin daha karmaşık olduğu durumlarda, parametrik olmayan regresyon modelleri kullanılmaktadır. Bu çalışmada

boylamsal verilerin analizinde kullanılan parametrik olmayan regresyon modellerinden splayn modelleri teorik olarak incelenmiştir.

Anahtar Kelimeler: Boylamsal veriler, Parametrik Olmayan Regresyon, Splayn Modelleri, Düzeltme Splaynları, Cezalı Splaynlar, Regresyon Splaynları

1. GİRİŞ

Boylamsal veriler; aynı birimlere ait özelliklerin zaman içerisinde tekrarlı olarak ölçülmesi ile elde edilirler. Tekrarlı ölçümler, eşit zaman aralıklarında veya eşit olmayan zaman aralıklarında elde edilebilirler. Bu şekilde elde edilen boylamsal verilerde, birimlerin incelenen özellik veya özelliklere göre zaman içerisindeki değişimi incelenebilmektedir. Boylamsal verilerin, klasik regresyon yöntemleri ile analizi bağımsızlık ve sabit varyanslılık varsayımı sağlanamadığından mümkün olmamaktadır. Bu durumda da boylamsal verilerin analizini gerçekleştirmek için yeni yöntemler geliştirilmiş ve hala geliştirilmeye devam edilmektedir. (Bağdatlı Kalkan, 2014: 1). Bu yöntemlerden en sık kullanılanları parametrik olmayan regresyon modelleridir.

Parametrik olmayan regresyon analizi bağımlı değişken ile bağımsız değişken arasındaki ilişkinin fonksiyonel şekli hakkında kesinlik olmaması durumunda kullanılmaktadır. Parametrik olmayan yaklaşımlar ilişkinin fonksiyonel şeklinin belirlenmesinde kolaylık sağlamalarının yanında, düşünülemeyen fonksiyonel şekillerin ortaya çıkarılmasında da yardımcı olmaktadır (Çağlayan, 2002: 8).

İlişkisiz verilerde parametrik olmayan regresyon yöntemleri, Kernel ve Splayn yöntemleri olmak üzere genel olarak ikiye ayrılmaktadır. Kernel yöntemi, lokal olabilirliğe dayalıdır. Splayn yöntemleri ise, düzeltme splaynları *(smoothing splines),* cezalı splaynlar *(penalized splines)* ve regresyon splaynları *(regression splines)* olarak üç grupta incelenmektedir. Düzeltme splaynları ve cezalı splaynlar cezalı olabilirliğe dayalı yöntemlerdir. Parametrik olmayan regresyon analizinde "düzeltme *(smoothing)*" kavramı oldukça önemli bir kavramdır. Bilinmeyen regresyon fonksiyonu için uygun bir tahmin elde etmek amacıyla, gözlemsel hataları azaltarak bağımlı değişkenin bağımsız değişkene göre ortalama bağımlılığının önemli ayrıntılarını vermek, yorumu kolaylaştırır. Böyle bir eğri yaklaştırma (tahmin) işlemi genel olarak "düzeltme" olarak adlandırılmaktadır. (Hardle, 1991: 18).

Silverman (1984: 900) yaptığı çalışmasında ilişkisiz veriler göz önüne alındığında Kernel düzeltmesi ve Splayn düzeltmesinin asimtotik olarak denk olduğunu göstermiştir.

Boylamsal verilerin analizinde en büyük zorluk birim içi korelasyonların dikkate alınmasıdır. Bu korelasyon, Kernel ve Splayn düzeltme yöntemlerinin boylamsal veriler için geliştirilmesinde önemli değişikliklere yol açmaktadır. Bilinen lokal olabilirlik tabanlı Kernel yöntemleri birim içi korelasyonları etkin bir şekilde dikkate alamamaktadır (Lin ve Carroll, 1999: 19). Boylamsal verilerde parametrik olmayan, tutarlı ve etkin tahminciler elde etmek için, lokal olmayan Kernel yöntemlerine ihtiyaç vardır (Shen, 2011: 21).

Parametrik olmayan yöntemlerde, bağımlı değişkenle bağımsız değişken veya değişkenlerle olan ilişkinin şekli veri tarafından belirlenir. Parametrik yöntemlerde ise ilişkinin şekli model tarafından belirlenir. Hem boylamsal hem de ilişkisiz verilerde parametrik olmayan modellerin en büyük avantajı bu şekilde esneklik sağlamasıdır. Boylamsal verilerde

parametrik olmayan regresyon modellerinde splayn düzeltme yöntemi en sık kullanılan yöntemdir. Splayn, bir dizi veri noktalarına polinomiyal bir eğri uydurma ya da bu noktalar arasından pürüzsüz olarak geçen ve bir çok parçadan oluşan esnek bir eğri olarak adlandırılır. Splayn fonksiyonlar da bu fikrin uygulaması olan matematiksel araçlardır. Bu fonksiyonların temel düşüncesi, tanımlanan aralığı bağımsız değişkenler yardımıyla alt aralıklara bölerek, her bir alt aralıkta farklı bir polinomiyal fonksiyon ile bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modellendirerek, istenilen mertebeden türevi olan sürekli bir fonksiyon elde etmektir. Splaynlar parametrik olmayan fonksiyonu tahmin eden yöntemlerden biri olarak sunulur ve genellikle parametrik olmayan ve semiparametrik regresyon ortamında ele alınırlar (Bağdatlı, 2010: 10).

2. DÜZELTME SPLAYNLARI

Düzeltme splaynları yöntemi, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin şekli hakkında önsel bir bilgi olmadığı durumlarda veya şeklin bilinen doğrusal ya da doğrusal olmayan modellerle uyum sağlamadığı durumlarda yararlı olmaktadır. Düzeltme splaynları Doğrusal Karma Etki Modelleri (DKEM) nde kullanıldığında, Sınırlandırılmış En Çok Olabilirlik (SEÇO) aracılığıyla ortalama ve varyansın ortak modellenmesine izin verir (Fitzmaurice ve diğerleri, 2009: 254).

Basit parametrik olmayan regresyon modeli;

$$Y_{ij} = g(X_i) + e_i \qquad (1)$$

olarak gösterilir.

Bu eşitlikte Y_{ij} sürekli bağımlı değişkeni, X_i bağımsız değişkeni ve $g(X_i)$ bilinmeyen pürüzsüz bir fonksiyonu *(smooth function)*, e_i ise 0 ortalamalı, sabit varyanslı, bağımız hata terimini ifade etmektedir $(i = 1,\ldots,N)$

Parametrik olmayan regresyon fonksiyonunun Kernel tahmini, lokal ağırlıklandırılmış ortalamaları (*local weighted averages*), parametrik olmayan regresyon fonksiyonun ($g(X_i)$) düzeltme splaynı tahmini, parçalı polinomiyal fonksiyonları kullanır (Shen, 2011: 24). Splayn modellerini tahmin edebilmek için, doğrusal regresyon doğrularının birleşeceği noktayı belirlemek gerekir, bu noktalara düğüm adı verilir (Bağdatlı, 2010: 11). Burada bütün gözlenen bağımsız değişken değerleri düğüm (*knot*) olarak kullanılır ve düğüm sayısı düzeltmenin miktarını kısıtlar.

Literatürde boylamsal verilerin analizinde en sık kullanılan splaynlar, doğal kübik düzeltme splaynlarıdır. Düzeltme splaynları tahmincisi ile DKEM arasında yakın bir ilişki bulunmaktadır. Ruppert, Wand ve Carroll (2003: 109) karma model çerçevesinde en iyi tahmincinin düzeltme splaynları olduğunu göstermişlerdir. Düzeltme splaynlarının en önemli avantajı düğüm sayısını seçmek zorunda kalınmamasıdır. Çünkü, gözlenen bağımsız değişken değerlerinin hepsi düğüm olarak alınmaktadır. Ancak, büyük örneklem söz konusu olduğunda hesaplamada zorluklar ortaya çıkmaktadır.

Splayn yöntemleri ile ilgili genel bir bilgi vermek adına eşitlik (2) incelenecektir.

$$Y_i = g(t_i) + e_i \qquad (2)$$

Bu eşitlikte, *i*.birime n_i ölçüm yapılmış ve $Y_i = (Y_{i1}, \cdots, Y_{in_i})'$, $t_i = (t_{i1}, \cdots, t_{in_i})'$ zamanlarında gözlenmiştir. Zaman dilimleri, $t_{i1} < t_{i2} < \cdots t_{in_i}$ olarak ifade edilmektedir ve başka bir bağımsız değişkenin bulunmadığı varsayımı söz konusudur. $g(t_i)$ pürüzsüz fakat bilinmeyen fonksiyonu, $e_i = (e_{i1}, \cdots, e_{in_i})'$ ise normal dağılan, 0 ortalamalı ve $V_i = V_i(\phi)$ (ϕ, zaman boyunca birim içi kovaryansı tanımlamaktadır.) kovaryans matrisli hata vektörünü ifade etmektedir. $g(t)$ fonksiyonunun pürüzsüzlüğü için kriter bir cezanın (*penalty)* kullanılmasıdır. Bu ceza, uydurulan fonksiyondaki eğrilik miktarını (düğüm sayısını) kısıtlar. Tahmin edilen fonksiyon $g(t)$, bu pürüzlülük cezası koşulu ile hata kareler toplamını minimize eder.

$$[Y_i - g(t_i)]' V_i^{-1} [Y_i - g(t_i)] + \lambda \int_{t\min}^{t\max} [g''(s)]^2 ds \qquad (3)$$

Eşitlik (3)'de λ düzeltme parametresini ifade eder. Düzeltme parametresi eğrinin pürüzsüzlüğünü ve verilere uyumu dengeler (Aydın, 2005: 40). Düzeltme parametresi negatif olmaz ancak 0 olabilir. λ'nın verilen herhangi bir değeri için cezalı kareler toplamı kriteri, ölçüm zamanları düğüm noktası olarak alınarak doğal kübik splaynlarla minimize edilir. Bu da kübik düzeltme splaynları olarak adlandırılır (Fitzmaurice ve diğerleri, 2009: 255).

Düzeltme splaynı, düzeltme parametresinin değerine bağlıdır. λ arttıkça, cezalı kareler toplamındaki pürüzlülük cezasının etkisi artar ve uydurulan splayn daha düz olur. λ sonsuza yaklaştıkça uydurulan splayn doğrusal hale gelir ve ceza 0 olur. λ arttıkça düşük varyanslı ancak yanlı, λ azaldıkça yüksek varyanslı ancak yansız uyumlar elde edilir (Keele, 2008: 65). Bu nedenden dolayı λ'nın seçimi büyük önem arz etmektedir. λ'nın en uygun değeri, verinin karakteristik özelliklerine bağlı olarak belirlenir. λ nın en uygun seçimi için Craven ve Wahba (1979: 380) tarafından Genelleştirilmiş Çapraz Geçerlilik (GÇG) yöntemi önerilmiştir. Ayrıca Wahba (1985: 1379) GÇG yöntemine alternatif olarak Genel En çok Olabilirlik (GEÇO) metodunu geliştirmiştir.

Kısmi düzeltme splaynı modelleri, semiparametrik veya toplamsal modeller olarak ta bilinmektedirler. Bu modellerde, doğrusal modele düzeltme splaynları yerleştirilmiştir. Eşitlik (4)'de kısmi düzeltme splaynı modeli görülmektedir.

$$Y_i = X_{ri}\beta_r + g(t_i) + e_i \qquad (4)$$

Bu modelde, X_{ri} bağımsız değişkenler için tasarım matrisini ve β_r bu terimler için katsayılar vektörünü ifade etmektedir (*i*.birim için). Bu bilgiler ışığında kismi splayn modelleri karma model çerçevesinde yazılabilmektedir.

$$Y_i = X_{ri}\beta_r + Z_{ri}b_r + g(t_i) + e_i \qquad (5)$$

Eşitlik (5), $g(t_i)$ açık olarak yazıldığında eşitlik (6)'daki gibi ifade edilmektedir.

$$Y_i = X_{ri}\beta_r + Z_{ri}b_r + X_{si}\beta_s + Z_{si}\delta + e_i \qquad (6)$$

Bu eşitlikte düzeltme splaynı $g(t_i)$; sabit katsayı, doğrusal terim $X_{si}\beta_r$ ve tesadüfi etkiler $Z_{si}\delta$'nin toplamı olarak ifade edilmektedir. $X_{ri}\beta_r$, $Z_{ri}b_r$ ve e_i sırasıyla diğer sabit terim, tesadüfi terim ve hata terimini ifade ederler (Fitzmaurice ve diğerleri, 2009: 255). Bu

modeller boylamsal veriler için uygun modellerdir. Çünkü splaynlar hem ortalama hem de kovaryans yapısı için modelde esneklik sağlamaktadırlar.

DKEM'de kübik düzeltme splaynları en iyi doğrusal yansız tahmincidir. Eşitlik (2) ele alındığında, kübik düzeltme splaynı, ölçüm zamanları düğüm noktaları ($r=n_i$ ve $\tau = t_i = (t_{i1},\cdots,t_{in_i})'$) olan doğal kübik splayndır. Cezalı kareler toplamındaki pürüzlülük cezası eşitlik (7)'deki gibi basitleştirilebilir.

$$\int_{t\min}^{t\max} [g''(s)]^2 ds = \delta' G_s^{-1} \delta \tag{7}$$

Bu eşitlikte δ splaynın ikinci türevi, G_s^{-1} t_i cinsinden yazılan matrisi ifade eder ve ölçüm zamanları arasındaki farkı kullanır; $h_j=t_{i(j+1)}-t_{ij}$ $(j=1,\ldots,n_i-1)$. Bu doğal kübik splayn eşitlik (8)'deki gibi ifade edilir.

$$g(t) = \beta_{s,0} + \beta_{s,1} t + \sum_{j=2}^{n_i-1} \delta_j P_j(t) \tag{8}$$

Eşitlik (8)'de fonksiyon $P_j(t)$, $\tau=t_i$ için eşitlik (9)'da gösterildiği gibidir.

$$6P_j(t) = \begin{cases} 0 & t \le \tau_{j-1} \\ h_{j-1}^{-1}(t-\tau_{j-1})^3 & \tau_{j-1} < t \le \tau_j \\ h_{j-1}^{-1}(t-\tau_{j-1})^3 - (h_{j-1}^{-1}+h_j^{-1})(t-\tau_j)^3 & \tau_j < t \le \tau_{j+1} \\ (h_{j-1}+h_j)(3t-\tau_{j-1}-\tau_j-\tau_{j+1}) & \tau_{j+1} < t \end{cases} \tag{9}$$

Eşitlik (8) ve eşitlik (9)'dan yararlanılarak *i*.birim için cezalı kareler toplamı eşitlik (10)'da görüldüğü gibi yazılmaktadır.

$$(Y_i - X_{si}\beta_s - Z_{si}\delta)' V_i^{-1} (Y_i - X_{si}\beta_s - Z_{si}\delta) + \lambda \delta' G_s^{-1} \delta \tag{10}$$

Eşitlik (10)'da X_{si}=[1 t_i] n_ix2 boyutlu matrisi, $\beta_s=(\beta_{s,0}\ \ \beta_{s,1})'$ ve $Z_{si} = \left[P_2(t_i),\cdots,P_{n_i-1}(t_i)\right]$ ise n_i x (n_i-2) boyutlu matrisi ifade etmektedir.

Uydurulan splayna ulaşmak için cezalı kareler toplamının (Eşitlik (10)) β_s ve δ'ya göre minimizasyonu sağlanmalıdır. Bunun için eşitlik (11) çözülmelidir (Fitzmaurice ve diğerleri, 2009: 264).

$$\begin{bmatrix} X_{si}' V_i^{-1} X_{si} & X_{si}' V_i^{-1} Z_{si} \\ Z_{si}' V_i^{-1} X_{si} & Z_{si}' V_i^{-1} Z_{si} + \lambda G_s^{-1} \end{bmatrix} \begin{pmatrix} \hat{\beta}_s \\ \tilde{\delta} \end{pmatrix} = \begin{pmatrix} X_{si}' V_i^{-1} Y_i \\ Z_{si}' V_i^{-1} Y_i \end{pmatrix} \tag{11}$$

Bu eşitlik DKEM olarak eşitlik (12)'deki gibi ifade edilir.

$$Y_i = X_{si}\beta_s + Z_{si}\delta + e_i \tag{12}$$

Bu eşitlikte, β_s sabit etkileri, δ tesadüfi etkileri ve e_i hata terimini ifade etmektedir. Karma model varsayımı, $e_i \sim N(0,V_i)$ burada da geçerlidir. Ayrıca tesadüfi etkiler için $\delta \sim N(0, \sigma_s^2 G_s)$ varsayımı da geçerlidir. Burada, $\lambda=1/\sigma_s^2$'dir. Dolayısıyla σ_s^2, 0 olabilir ancak negatif olamaz. Düzetme splaynı modelleri için bu varsayım istenmeyebilir. Doğrulama işlemi, sadece eşitlik (10) ve (11)'in çözümlerinin uyuşması ile yapılır. Bu nedenle δ'nın G_s ters ceza matrisi ve $\sigma_s^2=1/\lambda$'nın etkiler vektörü olarak alınması uygun olmaktadır. Uydurulan splayn eşitlik (13)'de gösterildiği gibidir.

$$\tilde{g} = \tilde{g}(t_i) = X_{si}\hat{\beta}_s + Z_{si}\tilde{\delta} \tag{13}$$

Bu eşitlikte, $\tilde{g}$ formülasyonu uydurulan splaynın en iyi doğrusal yansız tahminci olduğunu göstermektedir. Eşitlik (13)'te düzeltme parametresinin önceden belirlendiği varsayılmakta ve SEÇO yöntemi ile tahmin edilmektedir.

Özetle, uydurulan splayn eşitlik (14)'de görülmektedir.

$$E(\tilde{g} \mid \beta_s, \delta) = g - V_i P Z_{si}\delta \tag{14}$$

Bu eşitlikte, $P = \sum^{-1} - \sum^{-1} X(X'\sum^{-1} X)^{-1} X'\sum^{-1}$ olarak ifade edilmektedir.

Modeldeki tesadüfi splaynın yani doğrusal olmayan bileşenin istatistiksel anlamlılığının sınanması için SEÇO oran testi kullanılabilmektedir. Bu testte $\sigma^2=0$ dolayısıyla $\delta=0$ hipotezi sınanır (Verbyla ve diğerleri, 1999: 283).

SEÇO oran test istatistiği eşitlik (15)'de görülmektedir.

$$-2\times(RL_o - RL_A) \tag{15}$$

Bu eşitlikte RL_A basit splayn modelinin logaritmik olabilirliğini, RL_0 ise kısıtlanmış modelin logaritmik olabilirliğini ifade etmektedir. Bu test için uygun asimtotik dağılım χ_0^2 ve χ_1^2 dağılımlarının karışımı olarak sunulmaktadır. Karışım olasılığı ise 0.5'tir. Bu teste alternatif olarak, Verbyla ve diğerleri (1999: 283) iki aşamalı bir yöntem önermişlerdir. Bu yöntemde öncelikle $\sigma^2=0$ hipotezi olabilirlik oran testi ile sınanır. Bu aşamadan sonra doğrusal olmayan bileşen hakkında kanıt yok ise, $\beta=0$ hipotezini sınamak için Wald testi kullanılır.

Zhang ve Lin (2003: 61) sıfır hipotezi $\sigma^2=0$ ve alternatif hipotez $\sigma^2\neq0$ için skor testi geliştirmişlerdir. Sıfır hipotezi altında uygun asimtotik dağılım Ki-Kare dağılımının karışımıdır ve dağılım için Satterthwaite yaklaşımını sağlar. Splayn modellerinde hipotez testleri hala üzerinde çalışılan bir konu olup, bu nedenle standart bir yöntem henüz geliştirilmemiştir. Geliştirilen yöntemlerin birçoğu hesaplamalarda yaşanan zorluklar nedeniyle sıkça kullanılamamaktadır.

Splaynlar açıklanırken kolaylık sağlaması adına sadece tek bir birim üzerinde durulmuştur. Ancak, boylamsal verilerde eşzamanlı olarak N birimin incelenmesi gerekmektedir. Birimler arası ölçümlerin, ortak çevresel faktörler, yaşanılan bölge, genetik özellikler gibi değişkenler dışında genellikle bağımsız olduğu, birim içi ölçümlerin ise ilişkili olduğu varsayılmaktadır. Ölçümler arasındaki zaman aralığı arttıkça birim içi korelasyonlar düşmektedir. Aynı birim

için, ölçümler arasındaki korelasyon her birim için hatalara uygulanan varyans modelleri ile (V_i) modellenebilir (Fitzmaurice ve diğerleri, 2009: 268). Aynı işlem, birimler arası korelasyon içinde mümkün olmaktadır. Birim içi korelasyonları dikkate almanın bir diğer yolu da korelasyonu gizli değişkenler aracılığıyla modellemektir. Örneğin, tesadüfi katsayılar (*random-coefficient*) regresyon modellerinde korelasyon bu şekilde modellenmektedir. Boylamsal verilerin karma model kübik düzeltme splaynları ile analizini açıklamak için bir takım modeller incelenecektir.

2.1. Basit Splayn Modeli

Basit splayn modellerini açıklamak adına analiz edilecek birimler, ortak bir çevredeki homojen birimler ve başka bir değişken ya da yapı olmadığı varsayılmıştır. Bu veri için model, zaman boyunca ortak profili içermelidir. Bu ortak profil düzeltme splaynı ($g(t)$) ile modellenebilir. Çünkü, düzeltme splaynları cezalı kareler toplamını minimize eder (Eşitlik 3) Düzeltme splaynı, ortak profildeki trendi takip eden düzgün bir eğri sağlar. Eğrideki düzgünlüğün ya da düzeltmenin miktarı, veriden, SEÇO ile tahmin edilen düzeltme parametresi ile kontrol edilir. Birim içi korelasyon standart kovaryans modeli ile modellenir. $Cov(e_i) = V_i(\phi)$ ve $Cov(e_i, e_j) = 0 \quad i \neq j$ ise (i, j=1,…,N). Bu model eşitlik (16)'daki gibi gösterilmektedir.

$$Y_{ij} = g(t_{ij}) + e_{ij} \qquad i = 1, \cdots, N \qquad j = 1, \cdots, n_i \tag{16}$$

*i.*birim için modelin matris gösterimi eşitlik (2)'deki gibi olmaktadır. Bütün veri seti için model eşitlik (17)'deki gibidir.

$$Y = g(t) + e \tag{17}$$

Bu eşitlikte, $Y = (Y_1', \cdots, Y_N')'$, $t = (t_1', \cdots, t_N')'$ ve $e = (e_1', \cdots, e_N')'$ olarak ifade edilmektedir. Modeldeki düzeltme splaynı doğal kübik splayndır. Doğal kübik splaynın r düğümler kümesi eşitlik (18)'de gösterilmektedir.

$$\tau = (\tau_1, \cdots, \tau_r)' = \{\tau_i \in t : \tau_i < \tau_{i+1}\} \tag{18}$$

Bu bağlamda model, eşitlik (19)'daki gibi yazılmaktadır.

$$Y_{ij} = \beta_{s,0} + \beta_{s,1} t_{ij} + \sum_{j=2}^{r-1} \delta_j P_j(t_{ij}) + e_{ij} \tag{19}$$

Bu eşitliğin *i.* birim için matris gösterimi eşitlik (12)'deki gibi olmaktadır. Bütün veri seti için gösterimi ise eşitlik (20)'de görüldüğü gibidir.

$$Y = X_s \beta_s + Z_s \delta + e \tag{20}$$

Eşitlik (20)'de δ ters ceza matrisi G_s'ye sahip bir vektör, $P_j(t)$ daha önce açıklanan fonksiyon, $X_s = (X_{s1}', \cdots, X_{sn}')'$ N_Tx2 Boyutlu matris, $Z_s = (Z_{s1}', \cdots, Z_{sn}')'$ N_Tx(r-2) boyutlu matrisi ifade etmektedir. Karma model yaklaşımına göre splayn tesadüfi etkiler ortak tutulurlar. Dolayısıyla $b_i = \delta$'dir (i=1,…,N).

Basit splayn modellerinde birim içi değişkenlik, eşitlik (21)'de ifade edildiği gibi gösterilmektedir.

$$Cov(Y_i \mid \beta_s, \delta) = V_i \tag{21}$$

2.2. Birden Fazla Bağımsız Değişken Olma Durumunda Splayn Modeli

Basit splayn modeline birden fazla bağımsız değişken eklenerek model genişletilebilmektedir. Yaş, cinsiyet gibi değişkenler sabit etki olarak, bölge ve çevresel etkiler gibi değişkenler tesadüfi etki olarak alınabilen değişkenlerdir. Sabit ve tesadüfi değişkenler sırasıyla X_{ri} ve Z_{ri} tasarım matrisine eklendiğinde *i*. birim için genişletilmiş model matris notasyonu ile eşitlik (22)'de gösterildiği gibidir.

$$Y_i = X_{ri}\beta_r + Z_{ri}b_r + X_{si}\beta_s + Z_{si}\delta + e_i \tag{22}$$

Cov(b_r)= G_r'dir ve birim içi değişkenlik eşitlik (23)'de ifade edildiği gösterilmektedir.

$$Cov(Y_i \mid \beta_r, \beta_s, \delta) = Z_{ri}G_rZ'_{ri} + V_i \tag{23}$$

2.3. Veride Ayrı Alt Gruplar Olma Durumunda Splayn Modeli

Splayn modeli veride alt grupların olması durumunda genişletilebilmektedir. Örneğin bir tedavi yönteminin yaş ya da cinsiyete göre alt gruplara ayrılarak uygulanmasında bu modeller kullanılabilmektedir. Bu modeller için öncelikle ortak splayn tahmin edilir. Ortak splaynın amacı bütün birimler için ortak trend bileşenini belirlemektir. Bu aşamadan sonra alt gruplar arasındaki farklılıkları belirlemek adına, her bir alt grup için farklı splaynlar tahmin edilir (Fitzmaurice ve diğerleri, 2009: 270).

j ölçüm, *i* birim ve *k* alt grubu göstermek üzere model, eşitlik (24)'deki gibi yazılmaktadır.

$$Y_{ijk} = g(t_{ij}) + h_k(t_{ij}) + e_{ij} \tag{24}$$

Eşitlik (24)'de $h_k(t)$ *k* alt grubu için kübik düzeltme splaynını ifade etmektedir. Bu splayn etkileşim splaynı olarak da adlandırılmaktadır. $g(t)$ ise ortak düzeltme splaynıdır. Alt grup düzeltme splaynı $h_k(t)$ eşitlik (25)'deki gibi ifade edilmektedir.

$$h_k(t) = \xi_{sk,0} + \xi_{sk,1}t + \sum_{j=2}^{r-1} v_{k,j}P_j(t) \tag{25}$$

Bu eşitlikte $P_j(t)$ daha önce açıklanan fonksiyonu (baz fonksiyonu) ifade eder. $v_k = (v_{k,2}, \cdots v_{k,r-1})'$ katsayılarının G_s ters ceza matrisine sahip olduğu varsayılır. Matris formunda *k* alt grubundaki *i* birim için model eşitlik (26)'daki gibi ifade edilmektedir.

$$Y_{ik} = X_{si}(\beta_s + \xi_{sk}) + Z_{si}(\delta + v_k) + e_i \tag{26}$$

Bu eşitlikte, $\xi_k = (\zeta_{sk,0} \quad \zeta_{sk,1})'$ olarak ifade edilir. Birim içi değişkenlik eşitlik (27)'de ifade edildiği gibidir.

$$Cov(Y_{ik} \mid \beta_s, \zeta_k, \delta, v_k) = V_i \tag{27}$$

2.4. Veride Ayrı Birim Profilleri Olma Durumunda Splayn Modeli

Boylamsal verilerde ortalama ve kovaryansın birlikte modellenmesinde, tesadüfi-katsayı regresyon yöntemi en sık kullanılan yöntemlerden biri haline gelmiştir (Fitzmaurice ve diğerleri, 2009: 272). Tesadüfi-katsayı regresyonda birim, sabit katsayı ve eğim katsayısında ufak değişimler ile birimlerin ortak doğrusal bir profil izlemesi beklenir. *i*.birim için doğrusal tesadüfi-katsayı regresyon modeli eşitlik (28)'de gösterilmektedir.

$$Y_{ij} = \beta_{s,0} + \eta_{si,0} + (\beta_{s,1} + \eta_{si,1})t_{ij} + e_i \tag{28}$$

Modelin matris gösterimi ise eşitlik (29)'da gösterildiği gibidir.

$$Y_i = (\beta_s + \eta_{si})X_{si} + e_i \tag{29}$$

Eşitlik (29)'da X_{si}=[1 t_i] ve $\eta_{si} = (\eta_{si,0} \quad \eta_{si,1})'$ olarak ifade edilmektedir. η_{si}'nin normal dağıldığı, 0 ortalama ve G_c kovaryans matrisine sahip olduğu varsayılmaktadır. G_c kovaryans matrisi eşitlik (30)'da gösterilmektedir.

$$G_c = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \tag{30}$$

Bu bağlamda zaman boyunca birim içi kovaryans eşitlik (31)'de gösterildiği gibi olmaktadır.

$$\begin{aligned} Cov(Y_{ij}) &= \sigma_{11} + 2\sigma_{12}t_{ij} + \sigma_{22}t_{ij}^2 + V_i[j,j] \\ Cov(Y_{ij}, Y_{ik}) &= \sigma_{11} + \sigma_{12}(t_{ij} + t_{ik}) + \sigma_{22}t_{ij}t_{ik} + V_i[j,k] \end{aligned} \tag{31}$$

Matris gösteriminde *i*.birim için model eşitlik (32)'de gösterildiği gibidir.

$$Cov(Y_i \mid \beta_s) = X_{si}G_cX'_{si} + V_i \tag{32}$$

Bu regresyon modelinde tesadüfi etkiler birim-özeldir. Dolayısıyla $b_i=\eta_{si}$'dir (i=1,…,N). Ayrıca birim-özel sabit ve eğimin zaman boyunca birim içi kovaryansı hesaba kattığı bu nedenle $V_i=\sigma_e^2 I$ olduğu varsayılır. Birim içi kovaryansın zaman boyunca karesel bir yapı izlediği varsayılmaktadır. Kovaryans parametresi olan σ_{12} varyans yapısındaki esnekliği arttırır.

Tesadüfi-katsayı regresyon modellerinde yüksek dereceli polinomlara izin verilmektedir. Yüksek dereceli polinomlar yerine düzeltme splaynları kullanılarak modeller geliştirilebilmektedir. Bu durumda, modelde ortak splaynla zaman boyunca anakütle ortalaması tahmin edilir. Yapılan işlemler alt grup splayn modellerinde olduğu gibidir. Ancak burada splaynlar her bir birim için ölçümler dizisine uydurulmuştur. Doğrusal tesadüfi-katsayı regresyon modellerinin varsayımı gereğince, her bir birim için sabit ve eğim katsayıları, ortak sabit ve eğim katsayılarından tesadüfi sapmayı gösterir. Model;

$$Y_{ij} = g(t_{ij}) + l_i(t_{ij}) + e_{ij} \tag{33}$$

olarak ifade edilmektedir. Eşitlik (33)'de $l_i(x)$ *i.* birim için alt grup kübik düzeltme splaynını, g(t) ortak düzeltme splaynını göstermektedir. Alt grup düzeltme splaynı eşitlik (34)'deki gibi yazılmaktadır.

$$l_i(t) = \eta_{si,0} + \eta_{si,1}t + \sum_{j=2}^{r-1} \omega_{i,j} P_j(t) \tag{34}$$

Bu eşitlikte, $P_j(t)$ ($j=2,\ldots,r-1$) daha önce açıklanan baz fonksiyonunu ifade eder. Katsayılar $\omega_i = (\omega_{i,2}, \cdots, \omega_{i,r-1})'$ olarak ifade edilir ve G_s ters ceza matrisine sahiptir. Matris gösteriminde *i.* birim için model eşitlik (35)'deki gibi yazılmaktadır.

$$Y_i = X_{si}(\beta_s + \eta_{si}) + Z_{si}(\delta + \omega_i) + e_i \tag{35}$$

Eşitlik (35)'de η_{si} eşitlik (29)'daki gibi tesadüfi etkileri gösterir. Doğrusal tesadüfi –katsayılar regresyonda olduğu gibi birim içi kovaryanslar birim-özel profillerden üretilirler. Bu bağlamda, birim splayn etkilerinin normal anakütleden geldiği varsayılır. Yani $\omega \sim N(0, \sigma^2_\omega G_s)$ olarak ifade edilir (Fitzmaurice ve diğerleri, 2009: 73). Bu bilgiler ışığında birim içi kovaryans eşitlik (36)'daki gibi olmaktadır.

$$Cov(Y_i \mid \beta_s, \delta) = X_{si} G_c X'_{si} + \sigma^2_\omega Z_{si} G_s Z'_{si} + V_i \tag{36}$$

Bu eşitlikte, karma modeller çerçevesinde $Z_i = \begin{bmatrix} Z_{si} & X_{si} & Z_{si} \end{bmatrix}$ olarak ifade edilir. b_i ise $b_i = (\delta' \quad \eta'_{si} \quad \omega'_{si})'$'dir. Burada δ birimler boyunca ortak alınan splayn etkileri, η_{si} ve ω_i birim-özel etkileri ifade eder.

Eşitlik (35) Guo (2002: 123) tarafından yapılan çalışmada kullanılmış ve bu model için ek bir varsayımda bulunmuştur. Bu varsayım $Cov(\eta_{si}, \omega_i)=0$ olduğudur. Bu varsayımdan hareketle, varyans modellerinin veriye uyumunun kullanılan splayn yapısına bağlı olduğu ortaya çıkarılmıştır. Varyans modelinin seçilen splayn yapısına göre değişmemesi için, tesadüfi etkilerin ortak bir kovaryans modeline sahip olması gerekir;

$$Cov\begin{pmatrix} \eta_i \\ \omega_i \end{pmatrix} = \begin{bmatrix} G_c & G'_{sc} \\ G_{sc} & \sigma^2_s G_s \end{bmatrix} \tag{37}$$

Eşitlik (37)'de G_{sc}, (r-2) x r boyutlu kovaryans parametrelerinin matrisini ifade etmektedir. Eşitlik (29)'da görülen tesadüfi–katsayı regresyonda olduğu gibi fazladan parametre daha genel bir kovaryans modeli sağlar. Bununla birlikte varyans modelinin seçilen splayn yapısına göre değişmemesini sağlar ancak tahmini zorlaştırır (Fitzmaurice ve diğerleri, 2009: 273).

Buraya kadar yapılan açıklamalarda, birçok splayn modeli kısaca açıklanmıştır. Splayn modellerini oluştururken en önemli nokta veriye en uygun modelin seçimidir. Varyans parametreleri bağımsız değildir. Bu nedenden dolayı değerleri, modeldeki tesadüfi terimler kümesine göre değişebilir. Ayrıca varyans parametrelerinin değerleri, modelde bulunan sabit terimlerin sistematik trendi tamamen açıklayıp açıklamadığına bağlıdır. Bu bağlamda, öncelikle bütün sabit bağımsız değişkenler modele katılır ve en karmaşık varyans modeli

oluşturulur. Bu aşamadan sonra sabit modeli değiştirmeden varyans modelini adım adım basitleştirmek için olabilirlik oran testi kullanılabilir. Uygun varyans modeli belirlendikten sonra, sabit modeli basitleştirmek için yaklaşık F testi kullanılabilmektedir (Kenward ve Roger, 1997: 986). Birden fazla varyans modelinin uygun bulunması durumunda modeller Akaike Bilgi Kriteri (ABK) ile karşılaştırılabilirler.

3. REGRESYON SPLAYNLAR

Regresyon splaynları, bir dizi belirli düğüm noktalarında birleşen, fonksiyonun tanım bölgesini bölen parçalı polinomlardır. Regresyon splaynları az sayıda düğüm kullanmaktadırlar. Regresyon splayn yöntemi gereksiz düğüm noktalarını silerek fonksiyonu tahmin eder (Aksoy ve Aydın, 2013: 3).

L düğüm sayısını göstermek üzere düğümler kümesi, $\{s_1,\ldots,s_L\}$olarak ifade edilir ve L genellikle küçüktür. Regresyon splaynları, baz fonksiyon (temel ya da taban fonksiyon) olarak genellikle B-splaynlarını kullanırlar (Shen, 2011: 26). Bu bağlamda baz fonksiyonları kümesi, $\{B_1(X),\ldots, B_L(X)\}$. Eşitlik (1)'de gösterilen basit parametrik olmayan bir regresyon modeli için;

$$g(X) \approx \sum_{l=1}^{L} B_1(X)\alpha_l \tag{38}$$

olarak ifade edilir ve $\alpha=(\alpha_1,\ldots, \alpha_l)'$olarak gösterilir.

Parametrik model standart En Küçük Kareler (EKK) ile tahmin edilir ve böylece α'nın tahminine ulaşılır.

$$Y_i = \sum_{l=1}^{L} B_1(X)\alpha_l + e_i \tag{39}$$

Parametrik olmayan regresyon splaynının tahmini (pürüzsüz fonksiyon) eşitlik (40)'da görüldüğü gibidir.

$$\hat{g}(X) = \sum_{l=1}^{L} B_1(X)\hat{\alpha}_l \tag{40}$$

Regresyon splaynlarının en büyük avantajı hesaplamadaki kolaylığıdır. Öncelikle sadece parametrik bir modelin tahmin edilmesine ihtiyaç vardır. Ancak, geleneksel splaynlarda her bir bağımsız gözlem bir düğüm olarak değerlendirilirken regresyon splaynlarda düğümlerin konumu belirlenmelidir (Omay ve Canpolat, 2013: 94). Regresyon splaynlarındaki en önemli nokta düğüm sayısının seçimi ve düğümlerin yerinin belirlenmesidir. Çünkü $g(X)$'in tahmini bu seçimlere duyarlıdır.

4. CEZALI SPLAYNLAR

Cezalı splaynlar (*Penalized Splines-P splines*) regresyon splaynlar ve düzeltme splaynlarının bir karmasıdır. Regresyon splaylarda, pürüzsüz fonksiyonun tahmini düğüm sayısına ve düğüm yerine bağlıdır ve bu durum çok boyutlu verilerde hesaplama sıkıntısına yol açmaktadır. Düzeltme splaynlarında ise düğüm sayısı gözlem sayısına eşittir. Bu durumda da

büyük veri setlerinde problemlerle karşılaşılmaktadır. P-splaynlar her iki splaynın en iyi taraflarını birleştirmiştir. P-splaynlar, düzeltme splaynlarından daha az parametre kullanır ve regresyon splaylarında olduğu gibi düğümlerin seçimi daha esnektir. Ayrıca, tahminde düğüm sayısının ve konumun önemini de azaltır (Güler, 2013: 19).

4.1. Bir Birim İçin Cezalı Splayn Modelleri

Bir birim için P-splayn modeli eşitlik (41)'deki gibi yazılmaktadır.

$$Y_{ij} = s_i(t_{ij}) + e_{ij} \tag{41}$$

Bu eşitlikte $s_i(.)$ pürüzsüz deterministik bir fonksiyonu, e_{ij} bağımsız, normal dağılan ve σ^2 varyanslı hata terimini ifade etmektedir. $s_i(.)$ için düşük boyutlu bir taban belirlenirse $(x_1(.),\ldots,x_p(.))$ örneğin az sayıda düğümlü polinomiyal taban ya da regresyon splayn tabanı gibi. Bu durumda $s_i(t)$ eşitlik (42)'de gösterilen parametrik regresyonla tahmin edilebilir.

$$Y_i = X_i\beta_i + e_i \tag{42}$$

Bu eşitlikte, X_i n_i x p boyutlu matrisi ifade etmektedir. Bu matrisin her bir sütunu $t_1,\cdots t_{n_i}$ zamanlarında tabanın bir boyutunu gösterir. β_i regresyon parametreleri vektörü ve e_i ise hata terimleri vektörü olarak ifade edilir. β_i için EKK tahmincisi eşitlik (43)'de görüldüğü gibidir.

$$\hat{\beta}_i = (X_i'X_i)^{-1}X_i'Y_i \tag{43}$$

Eşitlik (43)'den hareketle $s_i(t)$ için tahminci eşitlik (44)'deki gibi olmaktadır.

$$\hat{s}_i(t) = \sum_{k=1}^{p} x_k(t)\hat{\beta}_{ik} \tag{44}$$

Bu tahminci, bütün yansız doğrusal tahminciler arasında minimum varyansla en iyi doğrusal yansız tahmincidir. Hata terimleri normal dağılıyor ise bu durumda $\hat{s}_i(t)$ En Çok Olabilirlik (EÇO) tahmincisidir. Bu nedenden dolayı da asimtotik olarak etkindir. Hata teriminin normal dağılımdan başka dağılım gösterdiği durumlarda $\hat{s}_i(t)$ asimtotik olarak etkin değildir. Herhangi bir dağılım belirlenmediğinde ise $\hat{s}_i(t)$ diğer asimtotik doğrusal tahminciler arasında semiparametrik etkindir. Merkezi limit teoremine göre büyük örneklemler söz konusu olduğunda asimtotik olarak normal dağılıma yaklaşım söz konusudur (Fitzmaurice ve diğerleri, 2009: 295). Bu durumda ; $\hat{\beta}_i \sim N(\beta_i, \hat{\sigma}^2(X_i'X_i)^{-1})$ olarak gösterilir. Burada varyans tahmincisi eşitlik (45)'de gösterildiği gibi ifade edilmektedir.

$$\hat{\sigma}^2 = \frac{1}{n_i - p}Y_i'(I_{n_i} - H_i)Y_i \tag{45}$$

Bu eşitlikte I_n birim matrisi ifade etmektedir. H_i ise eşitlik (46)'da gösterilmektedir.

$$H_i = X_i(X_i'X_i)^{-1}X_i' \tag{46}$$

$\hat{\beta}_i$'nın bu asimtotik dağılım özelliğinden hareketle $\hat{s}_i(t)$'nin standart sapması ve güven bantlarına ulaşılır.

Cezalı splayn yaklaşımında $s_i(t)$'yi açıklamak için, fazla düğüm içeren yüksek boyutlu regresyon splayn tabanı ya da yüksek dereceli polinomiyal parçalar seçilir. $s_i(t)$'nin tahmininde boyut düşürmek için ceza kullanılır. Yüksek boyutlu splayn tabanını göstermek için X_i'nin daha fazla sütun (daha büyük p) içerdiğini ve $x_k(t)$'nin de t zamanındaki k. taban fonksiyonunu gösterdiği bilindiğinde, hata kareler toplamının minimize edilmiş hali ile cezanın toplanmış eşitlik (47)'deki gibi ifade edilir.

$$\sum_{j=1}^{n_i}(Y_{ij} - X_{ij}\beta_{ij})^2 + \lambda P(\beta_i) \tag{47}$$

Bu ifadeye göre $s_i(t)$'nin cezalı splayn tahmini eşitlik (48)'deki gibi olur.

$$\hat{s}_{cs,i}(t) = \sum_{k=1}^{p} x_k(t)\hat{\beta}_{ik} \tag{48}$$

Bu eşitlikte $P(\beta_i)$ karesel cezayı ifade etmektedir ve eşitlik (49)'da gösterilmektedir. Karesel cezanın tercih nedeni matematiksel hesaplamaların kolay gerçekleştirilebilmesindendir.

$$P(\beta_i) = \beta_i' D \beta_i \tag{49}$$

Eşitlik (49)'da D, simetrik pozitif yarı belirli (karekökü alınabilen) matrisi ifade etmektedir. Ceza fonksiyonu $P(\beta_i)$ pürüzlülük cezasıdır. λ daha öncede açıklanan düzeltme parametresidir. Karesel pürüzlülük cezasından başka pürüzlülük cezaları da bulunmaktadır. Örneğin kübik düzeltme splaynı cezası eşitlik (50)'deki gibidir.

$$P_{kds}(\beta_i) = \int_a^b \left(\sum_{k=1}^{p} x_k''(t)\beta_{ik} \right)^2 dt \tag{50}$$

Eşitlik (50)'de integralin tanım aralığı olan [a, b] veriyi içeren herhangi bir aralıktır.

4.2. Cezalı Splaynlar İçin Karma Etki Model Gösterimi

$s_i(t)$'nin tahmincisi eşitlik (47)'ye dayanır ve eşitlik (51)'de görülen DKEM'nin en iyi doğrusal yansız tahminidir.

$$Y_i = X_i^* \beta_i^* + Z_i b_i + e_i \tag{51}$$

Bu eşitlikte, X_i^*, β_i^* ve Z_i ve b_i tanımlanacaktır. Y_i ve e_i ise önceden açıklanan modellerdeki gibidir. Ancak tek fark çok değişkenli normallik varsayımının eklenmesidir.

Bir tesadüfi etkiler modeli eşitlik (52)'de görülmektedir.

$$Y_i = X_i u_i + e_i \tag{52}$$

Bu eşitlikte u_i p x 1 boyutlu tesadüfi etkiler vektörüdür, e_i'den bağımsızdır ve belirsiz bir dağılımla dağılır; $P(u_i) \propto \exp\left(-\frac{\lambda}{2\sigma^2} u_i' D u_i\right)$. En iyi doğrusal yansız tahmin denklemleri $P(u_i \mid Y_i) \propto p(Y_i \mid u_i)\, p(u_i)$'yi maksimize ederler. D tekil olmayan bir matris olduğunda $p(u_i)$, 0 ortalamalı ve $(\sigma^2/\lambda)D^{-1}$ kovaryanslı çok değişkenli normal dağılıma sahiptir. Ancak D tekil bir matris olduğunda $p(u_i)$ cezalı splayn modellerinde olduğu gibi kısmi belirsiz dağılıma (*partially improper distribution*) sahiptir. D, X_i'de olduğu gibi p sütuna sahiptir ve rankı r ile ifade edilmektedir. D, $Q\Psi Q'$ tayf ayrışımına sahiptir. Burada $Q=(Q_1, Q_2)$ ortonormal matrisi ifade etmektedir. Q_1 sütunları D'nin özvektörlerinden oluşan (0 özdeğerlere karşılık gelir) p x (p-r) boyutlu matrisdir. Q_2 sütunları diğer özvektörleri içeren p x r boyutlu matristir. Ψ özdeğerlerin köşegen matrisini ifade eder ve Ψ_1 ile Ψ_2'nin yani iki köşegen matrisin toplamından oluşur. Ψ_1'de bütün değerler 0'dır. Ψ_2'nin köşegenleri ise 0 olmayan özdeğerlerden oluşmaktadır.

D'nin bu ayrışımını kullanarak eşitlik (52), eşitlik (53)'deki gibi tekrar yazılabilir.

$$Yi = X_i Q_1 \beta_i^* + X_i Q_2 \Psi_2^{-1/2} b_i + e_i \tag{53}$$

Bu eşitliğe göre;

$$Yi = X_i^* \beta_i^* + Z_i b_i + e_i \tag{54}$$

olarak ifade edilir.

Eşitlik (54)'de β_i^* (p-r) boyutlu sabit etkiler vektörü, b_i r boyutlu çok değişkenli normal dağılıma sahip, 0 ortalamalı ve $\sigma^2/\lambda I_r$ kovaryanslı tesadüfi etkiler vektörüdür. Eşitlik (51)'e geri dönüldüğünde X_i^*, X_i Q_1 ve Z_i, $X_i Q_2 \Psi_2^{-1/2}$ olarak tanımlanır. Eşitlik (52) ve (54) λ ve σ^2 bilindiğinde birbirine eşit olmaktadır (Fitzmaurice ve diğerleri, 2009: 298).

Cezalı splaynlar ve karma etkili modeller arasındaki ilişki öncelikle cezalı splaynların Bayes tahmini olarak türetilmesiyle ortaya çıkmıştır (Fitzmaurice ve diğerleri, 2009: 298). Daha sonraki çalışmalarda örneğin, Speed (1991: 43) çalışmasında düzeltme splaynları ile karma etkiler modeli arasındaki ilişkiyi en iyi doğrusal yansız tahmin ve SEÇO ile kurmuştur.

Yüksek boyutlu bir sabit etkiler modeli eşitlik (55)'de görüldüğü gibidir.

$$Y_i = X_i \beta_i + e_i \tag{55}$$

Bu eşitlikte sadece e_i tesadüfidir. Ancak eşitlik (51)'de görüldüğü gibi bir DKEM'de b_i ve e_i tesadüfidir. Modeller arasındaki ilişkiden (uyum) dolayı eşitlik (56) oluşmaktadır.

$$X_i \beta_i = X_i^* \beta_i^* + Z_i b_i \tag{56}$$

β_i^*'ın boyutu b_i ile karşılaştırıldığında daha küçüktür. β_i'nin boyutu ise β_i^* ve b_i'nin boyutlarının toplamına eşit olmaktadır.

Hem cezalı yüksek boyutlu sabit etkiler modelinde hem de DKEM'de, $\{s(t_{i1}),\cdots,s(t_{in_i}\}$'nin tahminiyle aynı $\hat{Y}_i$ tahmin edilir. Daha genel bir gösterimle $\{s(t_a^1),\cdots,s(t_n^a\}$ 'nin tahminiyle aynı $X^a\hat{\beta}_i$ tahmin edilir.

Cezalı splaynlarda diğer splayn modellerinde olduğu gibi düzeltme parametresinin seçimi çok önemlidir. Düzeltme parametresinin seçimi için çapraz geçerlilik ve genelleştirilmiş çapraz geçerlilik yöntemi kullanılmaktadır. Çapraz geçerliliğin esası, veri noktalarından herhangi birini atmak ve kayıp veri noktaları altında geri kalan veriler tarafından en iyi kestirilen λ parametresini seçmektir (Bağdatlı, 2010: 23). Çapraz geçerlilik yöntemi büyük veri setlerine ihtiyaç duyar ancak bu durumda da hesaplamaların karmaşıklığı problemi ortaya çıkmaktadır. Bu problemi ortadan kaldırmak için genelleştirilmiş çapraz geçerlilik yöntemi kullanılır. Özellikle çok küçük veri setlerinde ilişkiler tahmin edilmek isteniyorsa bu durumda görsel metodlar kullanılmalı ve düzeltme parametresinin seçimi elle yapılmalıdır. Ancak veri seti büyüdükçe genelleştirilmiş çapraz geçerlilik gerçek değerine yaklaşmakta ve güvenilir sonuçlara ulaşılmaktadır (Hardle, Hall ve Marron, 1998: 90). Genelleştirilmiş çapraz geçerlilik yöntemi, çapraz geçerlilik yönteminin serbestlik derecesi ile düzeltilmiş biçimidir (Keele, 2008: 87). Genelleştirilmiş çapraz geçerlilik yöntemi istatistiksel yazılımların birçoğunda bulunmaktadır. Dolayısıyla düzeltme parametresinin seçimi kolaylıkla gerçekleştirilebilmektedir.

4.3. Birden Fazla Birim İçin Cezalı Splayn Modelleri

Boylamsal verilerin analizi için oluşturulan P-splayn modelleri eşitlik (57)'de görüldüğü gibi yazılmaktadır.

$$Y_{ij} = s_i(t_{ij}) + e_{ij} \qquad i = 1,\cdots,N \tag{57}$$

Bu modellerde t_{ij}, yani ölçüm zamanları her bir birim için farklı olabilir. Sadece bir birim için veriler söz konusu olduğunda, tahminler gerçekleştirilirken $s_i(.)$'nın deterministik olduğu varsayılır. Çünkü $s_i(.)$'nın sahip olduğu anakütle eğrisi hakkında herhangi bir bilgi bulunmamaktadır. Ancak, birden fazla birim söz konusu olduğunda her bir $s_i(.)$'nın tesadüfi olduğu görülür. Böylelikle boylamsal verilerin eşitlik (57) ile analizinde sadece e_{ij}'ler değil $s_i(t_{ij})$'lerde tesadüfidir. Bu tesadüfilik, ilk nokta tahminini gerçekleştirirken $s_i(.)$ eğrisine bağlı olmanın önüne geçemez. Bu aşamadan sonra nokta tahmini, güven aralığı, güven bantları ve diğer değişim ölçülerini hesaplamak S anakütle eğrisi ile gerçekleşmektedir. Yani tahminler $s_i(.)$ eğrisine bağlı değildir (Fitzmaurice ve diğerleri, 2009: 306). Bu aşamadan sonra kısmi parametrik veya parametrik olmayan bootstrap yöntemi kullanılarak $s_i(.)$'nın tesadüfi olması sağlanır. Bootstrap yeniden örnekleme yöntemidir. Bu yöntem, anakütleden elde edilen örneklemden alt örneklemlerin seçilmesi olarak açıklanır. Bootstrap yöntemi regresyon analizinde, hata terimleri ve bağımsız değişkenlerle ilgili varsayımların gerçekleşmemesi durumunda bir düzeltme işlemi amacıyla da kullanılır. Bootstrap yöntemi, daha küçük tahmin hatalarının elde edilmesi, standart sapmaların küçülmesi ve buna bağlı olarak da daha güvenilir parametre tahmincilerinin elde edilmesi ve güven aralıklarının oluşturulmasını sağlar (Altaş ve Çinko, 2003: 286).

Birden fazla birim, bağımsız ve aynı olasılık dağılımına sahip bir anakütleden seçildiğinde ve uygun bir şekilde modellendiğinde amaç anakütle ortalaması eğrisi olan $E\{s_i(.)\}$'nın tahmini

gerçekleştirmekse bu durumda, $(1/N)\sum_{i=1}^{N}\hat{s}_i(.)$ tahmin edilmelidir. Örneklemde her bir birim için gerçek eğri $s_i(.)$'ler gözlemlendiyse bu durumda $(1/N)\sum_{i=1}^{N}s_i(.)$ $E\{s_i(.)\}$'nın her bir zaman noktasında tutarlı ve etkin tahminidir. Genel olarak bu yöntem; gerçek eğrilerin gözlemlendiğini varsayar ve gerçek eğrilere bağlı olarak tutarlı ve etkin tahminler elde eder.

Bu yöntem için önerilen kısmi parametrik bootsrap yönteminde, tahmin edilen vektörler $\{\hat{s}_i(t_{i1}),\cdots,\hat{s}_i(t_{in_i}\}$ yerine konularak yeniden örneklenir bu durumda bütün tahmin edilen $\hat{e}_{ij}$'ler elde edilir. Önerilen parametrik olmayan bootstrap yönteminde ise Y_i'ler yerine konularak yeniden örneklenir.

Sonuç olarak bu yöntem, kesin ve güvenilir sonuçlara ulaşılması adına araştırmalarda sıkça kullanılan bir yöntemdir.

REFERENCES

- Aksoy, Ö. ve D. Aydın. (2013). Prostat Verisinin Splayn Düzeltme ve Regresyon Splayn Yöntemleri İle Tahmini. *Uluslararası Katılımlı 15. Ulusal Biyoistatistik Kongresi Bildiri Özetleri Kitabı*. Aydın: Adnan Menderes Üniversitesi, 2-4.
- Altaş, D. ve M. Çinko. (2003). Bootstrap Yönteminin Ridge Regresyonda Uygulanması. *Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi.* 22.1, 281-292.
- Aydın, D. (2005). Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı İle Tahmin ve Çıkarsamalar. *Doktora Tezi.* Eskişehir: Anadolu Üniversitesi FBE.
- Bağdatlı Kalkan, S. (2014). Boylamsal Verilerde Semiparametrik Karma Etki Modelleri ve Bir Uygulama . *Doktora Tezi.* İstanbul: Marmara Üniversitesi SBE.
- Bağdatlı, S. (2010). Semiparametrik Regresyon ve Bir Uygulama. *Yüksek Lisans Tezi.* İstanbul: İstanbul Ticaret Üniversitesi FBE.
- Craven, P. ve G. Wahba. (1979). Smoothing Noisy Data With Spline Functions: Estimating The Correct Degree of Smoothing by The Method of Generalized Cross Validation. *Numerische Mathematik.* 31, 377-403.
- Çağlayan, E. (2002). Yarı Parametrik Regresyon Modelleri ile Yaşam Boyu Sürekli Gelir Hipotezinin Türkiye Uygulaması. *Doktora Tezi.* İstanbul: Marmara Üniversitesi SBE.

- Fitzmaurice, G., M. Davidian, G. Verbeke ve G. Molenberghs. (2009). *Longitudinal Data Analysis.* New York: Chapman & Hall/CRC Press.
- Guo, W. (2002). Functional Mixed Effects Models. *Biometrics*. 58.1, 121-128.
- Güler, İ. (2013). P-splines for Longitudinal Data. Application to The Study of Children Growth With Phenylketonuria. *Master's Thesis*. Spain: University of Santiago de Compostela. Department of Statistics.
- Hardle, W. (1991). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.
- Hardle, W., P. Hall ve J.S. Marron. (1998). How Far are Automatically Choosen Regression Smoothing Parameters from their Optimum. *American Journal of Political Science*. 83.401, 86-95.
- Keele, L. (2008). *Semiparametric Regression for The Social Sciences*. England: John Wiley & Sons, Inc.
- Kenward, M.G. ve J.H. Roger. (1997). Small Sample İnference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics.* 53.3, 983-997.
- Lin, X. ve R.J. Carroll. (1999). Nonparametric Function Estimation for Clustured Data when The Predictor is Measured without/with error. Texas A&M University, Department of Statistics. http://www.stat.tamu.edu/~carroll/ftp/2000.papers.directory/jasa. final01.pdf [1 Ekim 2013].
- Omay, R.E. ve E. Canpolat. (2013). Parametrik Olmayan Sabit Etkiler Panel Veri Modelleri: Türkiye İçin Hava Kirliliği ve Gelir İlişkisi. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi-A.* 14.1, 91-103.
- Ruppert D., M.P. Wand ve R.J. Carroll. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.
- Shen, J. (2011). Additive Mixed Modeling Of HIV Patient Outcomes Across Multiple Studies. *Ph. D. Thesis.* Los Angeles: University of California, Department of Statistics.
- Silverman, B.W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *The Annals of Statistics*. 12.3, 898-916.
- Speed, T. (1991). That BLUP is a Good Thing: The Estimation of random effects. *Statistical Science*. 6.1, 42-44.

- Verbyla, A.P., B.R. Cullis, M.G. Kenward ve S.J. Welham. (1999). The Analysis of Design Experiments and Longitudinal Data by Using Smoothing Splines. *Applied Statistics*. 48.3, 269-311.
- Wahba, G. (1985). A Comparision of GCV and GML for Choosing The Smoothing Parameter in The Generalized Spline Problem. *Annals of Statistics*. 13.4, 1378-1402.
- Zhang, D. ve X. Lın. (2003). Hypothesis Testing in Semiparametric Additive Mixed Models. *Biostatistics*. 4.1, 57-74.