Öneri.C.9.S.35.
Ocak 2011.207-213.

SEMİPARAMETRİK REGRESYON

***Münevver TURANLI*[1], *Seda BAĞDATLI*[2]**

[1]*İstanbul Ticaret Üniversitesi, İstatistik Bölümü, Profesör Dr.*
[2]*İstanbul Ticaret Üniversitesi, İstatistik Bölümü, Araştırma Görevlisi*

SEMIPARAMETRIC REGRESSION

Abstract: Classical (parametric) regression techniques are based on the assumption that the independent variable is correlated linearly with the dependent variables and the pattern of this relation is known. When such assumption cannot be verified, parameter estimations fail to be reliable. In cases where the way of correlation is not known or it does not comply with the known parametric mathematical patterns, nonparametric regression techniques are to be applied. One shortcoming concerning this procedure emerges particularly in the interpretation process due to problems brought about by multidimensional aspect of the existence of more than one independent variable. Whenever confronted with a case that includes more than one independent variable, some of the independent variables correlate linearly with the dependent variable; at other times some of the independent variables might correlate nonlinearly. In order to establish a modeling for such relations, semiparametric regression models, comprising the aggregate of parametric and nonparametric regression function, are utilized. İn this study semiparametric regression definitions, estimation (backfitting algorithm), confidence bands, calculating standard errors and hypothesis testing are explained.

Keywords: Additive Models, Semiparametric Regression, Backfitting Algorithm.

SEMİPARAMETRİK REGRESYON

Özet: Klasik (parametrik) regresyon teknikleri, bağımlı değişkenin bağımsız değişkenlerle doğrusal bir ilişki içerisinde olduğunu ve ilişkinin şeklinin biliniyor olduğunu varsayar. Bu varsayımların sağlanamaması durumunda ise parametre tahminleri güvenilir olmamaktadır. İlişkinin şeklinin bilinmediği ya da bilinen parametrik matamatiksel kalıplara uymadığı durumlarda parametrik olmayan regresyon teknikleri kullanılmaktadır. Ancak bu teknikler birden fazla bağımsız değişken olma durumunda çok boyutluluğun yarattığı sıkıntı nedeniyle özellikle yorumlama aşamasında zorluklara neden olmaktadır. Birden fazla bağımsız değişken söz konusu olduğunda, bağımsız değişkenlerin bazıları bağımlı değişkenle doğrusal ilişki içerisinde bulunabilirken, bazıları doğrusal olmayan ilişki içerisinde bulunabilirler. Bu tür ilişkilerin modellenebilmesi için, parametrik ve parametrik olmayan regresyon fonksiyonlarının toplamsal olarak birleşiminden oluşan semiparametrik regresyon modellerinden yararlanılmaktadır. Bu çalışmada semiparametrik regresyon modellerinin tanımı, tahmini (backfitting algoritması), güven bantları, standart hataların hesaplanması ve hipotez testleri açıklanmıştır.

Anahtar Kelimeler: Toplamsal Modeller, Semiparametrik Regresyon, Backfitting Algoritması.

I. GİRİŞ

Parametrik olmayan regresyon modelleri ile x (bağımsız değişken) ve y (bağımlı) değişkenleri arasındaki doğrusal olmayan ilişki incelemektedir ancak, genellikle bağımlı değişken bir çok bağımsız değişkenden eş zamanlı olarak etkilenmektedir. Bu durumda çoklu regresyon analizine ihtiyaç duyulmaktadır. Parametrik olmayan regresyon modellerinde birden fazla değişkenin yer aldığı durum eşitlik (1)'de gösterildiği gibidir. Eşitlik (1)'de k sayıda bağımsız değişken yer almaktadır.

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}) + \varepsilon \qquad (1)$$

Parametrik olmayan çoklu regresyon modellerinin tahmini, çok boyutluluğun yarattığı sıkıntı (curse of dimensionality) nedeniyle zorlaşmaktadır [1]. Bu sorunun çözümü için toplamsal modeller kullanılabilmektedir.

Toplamsal modeller eşitlik (2)'de gösterildiği gibi ifade edilmektedir.

$$y_i = \alpha + f_1(x_1) + \ldots + f_k(x_k) + \varepsilon \qquad (2)$$

Eşitlik (2)'de gösterildiği gibi toplamsal modellerde düzeltme $[\, f_1(x_1), \ldots, f_k(x_k)\,]$ her bir bağımsız değişken için ayrı ayrı yapılmaktadır. Bu şekilde ifade edilen parametrik olmayan modeller toplamsallık varsayımına sahip olduğundan toplamsal modeller olarak ifade edilmektedirler [1]. Toplamsallık varsayımı araştırmacıya yorumlama aşamasında çok büyük kolaylıklar sağlamaktadır.

Toplamsal modellerde bağımlı değişken bazı açıklayıcı değişkenlerle doğrusal, fakat diğer bazı açıklayıcı değişkenlerle doğrusal olmayan ilişki içerisinde bulunabilmektedir. Bu sorunun çözümü için ise semiparametrik regresyon modelleri kullanılmaktadır

II. SEMİPARAMETRİK REGRESYON MODELLERİ

Semiparametrik regresyon modelleri standart regresyon tekniklerini genelleştiren ve her bir değişkenin etkisinin açık bir şekilde yorumlanmasını sağlayan toplamsal modellerin özel bir durumudur [2]. Bu durumda, semiparametrik regresyon modelleri toplamsal modellere parametrik bileşen eklenerek oluşturulan modellerdir. Semiparametrik regresyon modelleri çok boyutluluğun yarattığı sıkıntı nedeniyle parametrik olmayan modellere tercih edilmektedir.

Semiparametrik regresyon modeli,

$$y_i = \alpha + f_1(x_1) + ... + f_j(x_j) + \beta_1 x_{j+1} + ... + \beta_k x_k + \varepsilon \qquad (3)$$

biçiminde ifade edilir. (3) modelindeki j tane değişkenin y bağımlı değişkeni üzerinde doğrusal olmayan etkisi bulunmaktadır ve modelin parametrik olmayan bölümünü oluşturmaktadır. Diğer değişkenlerin ise, y bağımlı değişkeni üzerinde doğrusal etkisi bulunmaktadır ve modelin parametrik bölümünü oluşturmaktadır. Modelin parametrik bölümünde kukla değişkenler gibi kesikli değişkenlere yer verilebilmektedir. Aşağıda görülen (4) modeli incelendiğinde

$$y_i = \alpha + f_1(x_1) + f_2(x_2) + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \qquad (4)$$

x_3 değişkeni kukla değişken, x_1, x_2, x_4 değişkenleri ise sürekli değişkenlerdir. Böyle bir semiparametrik regresyon modelinde bir çok etkileşim modele dahil edilebilir. Örneğin, x_1 ve x_2 değişkenleri arasındaki doğrusal olmayan ilişki tahmin edilebilir. Bu durumda, çoklu parametrik olmayan regresyon modellerinde olduğu gibi üç boyutlu bir grafik elde edilecektir. Araştırmanın konusuna ve içeriğine bağlı olarak x_3 ve x_4 değişkenleri arasındaki ilişkiler de incelenebilir. Ayrıca modelin parametrik ve parametrik olmayan bölümündeki değişkenlerin birbirleriyle etkileşimi de analiz edilebilmektedir. Semiparametrik regresyon modelinin parametrik kısmı doğrusal ilişkiyi, modelin parametrik olmayan kısmı ise doğrusal olmayan ilişkiyi ifade etmektedir. Bu nedenle semiparametrik regresyon modellerine yarı doğrusal modeller adı da verilmektedir [3].

Bir model kurma aşamasında ilk olarak değişkenler belirlenmektedir. Değişkenler belirlendikten sonra ise modelin fonksiyonel şeklinin veya matematiksel yapısının belirlenmesi gerekmektedir. Matematiksel kalıp oluşturken öncelikli olarak yapılması gereken grafiklerin incelenmesidir. Bağımlı değişken ile herbir bağımsız değişkenin ayrı ayrı çizilecek grafikleri incelenerek, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin yapısı hakkında fikir sahibi olunabilir. Matematiksel kalıp ile ilgili tereddütler söz konusu olduğunda, farklı şekillerin denenmesi en uygun sonucu elde etmek için yararlı olacaktır. Herbir değişkenin ilişkisine tek tek bakıldıktan sonra bağımsız değişkenlerin bir veya bir kaçı için parametrik olmayan, diğerleri için parametrik ilişki uygun ise semiparametrik regresyon modeli en uygun model olarak tercih edilecektir [4].

Semiparametrik regresyon modellerinde parametrik kısım, doğrusal olabileceği gibi dönüşüm yöntemleri (Logaritmik, karesel dönüştürme vs), uygulanarak doğrusallaştırılabilinen yapıda da olabilir. Semiparametrik modelin parametrik kısmının belirlenmesinde farklı modeller tahmin edilebilir. Tahmin edilen bu modellerden artık kareler toplamını minimum yapan model semiparametrik modelin parametrik kısmı olarak tahmin edilebilir.

Değişkenler arasındaki ilişkiyi en iyi şekilde açıklayacak model çeşitli denemeler sonucunda da bulunabilir. Özellikle değişkenler arasında şekli tam belirlenemeyen ilişkiler varsa farklı matematiksel kalıplar veya farklı değişkenler için parametrik olmayan ilişkileri kapsayan modellerin denenmesi ve en uygun olanının seçilmesi gerekecektir [4].

II.1. Semiparametrik Regresyon Modellerinin Tahmini: Backfitting Algoritması

Toplamsal modellerin ve semiparametrik regresyon modellerinin tahmininde tekrarlı (iterative) algoritmalara ihtiyaç duyulmaktadır. Bu modellerin tahmini için geliştirilen birçok algoritma bulunmaktadır ve bu algoritmalar birçok değişik bilgisayar programında yer almaktadır. Özellikle, R programı birçok algoritmayı desteklemektedir. Bu algoritmalardan en çok kullanılanlar Newton- Raphson algoritması, backfitting algoritmasıdır. Backfitting algoritması Hastie ve Tibshirani tarafından 1990 yılında tanıtılmıştır. Bu algoritma parametrik olmayan ve parametrik bileşenleri tahmin edebilen en kolay yöntem olarak bilinmektedir.

Model tahmin etme aşamasında, x değişkenleri birbirlerine dik iseler modelin parametrik kısmı iki değişkenli modeller serisi olarak sıradan en küçük kareler yöntemini kullanarak tahmin edilebilir. Parametrik olmayan bileşenlerin tahmininde ise lowess ya da splaynlar kullanılabilir. Bağımsız değişkenler arasında korelasyon bulunmaması durumuna genellikle rastlanmamaktadır. Bu durumda toplamsal modelleri ya da semiparametrik modelleri tahmin ederken bağımsız değişkenler arasındaki ilişkiyi dikkate alacak yöntemlere ihtiyaç duyulmaktadır. Backfitting algoritması parametrik ve parametrik olmayan bileşenleri tahmin ederken bağımsız değişkenler arasındaki korelasyonu dikkate almak üzere tasarlanmıştır.

Backfitting algoritması kısmi regresyon fonksiyonları fikrini önermektedir. Eşitlik (5)'de iki bağımsız değişkenli toplamsal bir model görülmektedir.

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \qquad (5)$$

Bu modelde f_2'nin gerçek fonksiyonel formunun bilindiği ancak f_1'in bilinmediği varsayılsın. Bu durumda f_1'in tahmini için (5) modeli kısmi regresyon fonksiyonu olarak eşitlik (6)'daki gibi yeniden düzenlenmelidir.

$$y - \alpha - f_2(x_2) = f_1(x_1) + \varepsilon \qquad (6)$$

(6) eşitliğinde x_1'e karşı $y - \alpha - f_2(x_2)$'nin düzgünleştirilmesi $f_1(x_1)$'in tahminini elde etmeyi sağlamaktadır. Bu nedenle, bir kısmi regresyon fonksiyonunu bilmek diğer kismi regresyon fonksiyonunu tahmin etmeye olanak sağlamaktadır. Gerçek durumda, hiçbir regresyon fonksiyonunu bilmek mümkün olmamaktadır. Ancak f 'lerden herhangi biri için bir başlangıç değeri belirlenirse, toplamsal modellerin tahmini için kismi regresyon fonksiyonları tekrarlı yöntemler ile çözümlenir. Model (7) tahmin edilmek istensin:

$$y_i = \alpha + f_1(x_1) + ... + f_k(x_k) + \varepsilon \qquad (7)$$

(7) eşitliğinde S_j, sütunları f_k tahminlerinden oluşan bir matrisi ifade etmektedir. X ise, kolonları x değişkenlerinden oluşan model matrisini ifade etmektedir. Toplamsal modellerin tahmini için backfitting algoritması aşağıdaki adımlardan oluşmaktadır [1].

1. Adım: $\alpha = \bar{y}$ ve $S_j = X \ (j = 1,...,m)$ başlangıç değerleri olarak seçilir.

2. Adım: Her x değişkeni için kismi artıklar hesaplanır. x_1 değişkeni için tahmin edilen kismi artıklar eşitlik (8)'de görüldüğü gibidir.

$$\hat{e}_p^j = y_i - \sum_{i=2}^{k} S_j - \alpha \qquad (8)$$

3. Adım: x_1 değişkeni civarında e_p^j düzgünleştirilir. Bu aşama için parametrik olmayan regresyon modeli seçilmelidir (splaynların özelliklerinden dolayı birçok bilgisayar yazılımı backfitting algoritmasının üçüncü aşaması için splaynları kullanmaktadır).

4. Adım: S_j'deki x_1 değişkeni, x_i'nin düzgünleştirilmiş tahminleri ile değiştirilir.

5. Adım: 2'den k'ya kadar olan her x değişkeni için 2-4 adımları tekrarlanır.

6. Adım: Eşitlik (9)'da görülen model de artık kareler toplamı hesaplanır.

$$RSS = \sum_{i=1}^{n}\left[\left(yi - \sum_{i=1}^{k} S_j\right)^2\right] \qquad (9)$$

7. Adım: Artık kareler toplamındaki değişim belirli bir tolerans seviyesinde ise model yakınsar ve algoritma durur. Eğer değilse, bu işlem artık kareler toplamındaki değişim belirli bir tolerans seviyesine gelene kadar devam eder.

Backfitting algoritması durduğunda S_j'nin her sütünü x değişkeninin parametrik olmayan tahminini içerir. Bu tahminler x değişkenleri arasındaki ilişkiyi dikkate alır. Dolayısıyla, üç x değişkenine sahip bir toplamsal model tahmin edildiğinde $\hat{f}_1$ 'in grafiği, x_2 ve x_3 değişkenlerinin etkisi sabit tutulduğunda x_1'in y_i üzerindeki etkisi olarak yorumlanabilir. Backfitting algoritmasının bir çok varyasyonu bulunmaktadır. Bu varyasyonlardan en çok kullanılanlardan biri ise başlangıç değeri olarak sıradan en küçük kareler tahmincilerini kullanmaktır.

Sıradan en küçük kareler tahmincilerini başlangıç değeri olarak kullanan backfitting algoritması aşağıdaki adımlardan oluşmaktadır.

1. Adım: Her bir değişkenin kendi ortalamasından çıkartılmasıyla oluşan doğrusal regresyon modeli (10) eşitliğinde görüldüğü gibi oluşturulur ve tahmin edilir.

$$y_i - \bar{y} = \beta_1(x_1 - \bar{x}_1) + ... + \beta_k(x_k - \bar{x}_k) + \varepsilon \qquad (10)$$

(10) eşitliği kısaca (11) eşitliğinde olduğu gibi gösterilebilir.

$$y^* = \beta_1 x_1^* + ... + \beta_k x_k^* + \varepsilon \qquad (11)$$

Modellerdeki $\beta_1,...,\beta_k$ parametreleri tekrarlı backfitting algoritması için başlangıç değeri olarak görev yapar.

2. Adım: x_1 için kısmi artıklar (12) eşitliğinden tahmin edilir.

$$\hat{e}_{px_1} = y^* - \beta_2 x_{i2}^* + ... + \beta_k x_k^* \qquad (12)$$

Kısmi artıkların tahmini y ile x_2 değişkenleri arasındaki doğrusal bağlılığı ortadan kaldırır. Ancak en küçük kareler artıklarında ($j = 1,...,m$ için) y ile x_1 arasındaki doğrusal ya da doğrusal olmayan ilişki korunur.

3. Adım: Bir sonraki adımda f_1'in tahminini elde etmek için kısmi artıklar ($\hat{e}_{px_1}^{j}$) x_1'e karşı düzgünleştirilir. Bu aşamada kullanılan düzgünleştiricinin etkisi büyük ölçüde bulunmamaktadır.

4. Adım: $\hat{f}_{x_2}$'nin tahmini için eşitlik (13) oluşturulur.

$$e_{px_2} = y^* - \hat{f}_{x_1} x_1^* + ... + \beta_k x_k^* \qquad (13)$$

5. Adım: $\hat{f}_{x_2}$'nin tahmini için (13) eşitliğindeki kısmi artıklar x_2 değişkenine karşı düzgünleştirilir.

6. Adım: $\hat{f}_{x_2}$'nin yeni tahmini, x_3 için hesaplanacak olan yeni kısmi artıkların hesabında kullanılır. Her bir f_k için başlangıç tahminleri yapılır ve süreç tekrarlanır.

7. Adım: Bu tekrarlı süreç, tahmin edilen kısmi regresyon fonksiyonlarının artık kareler toplamındaki değişimin belli bir tolerans seviyesine ulaşmasına kadar tekrarlanır.

Bu süreç tamamlandığında, x değişkenlerinin y_i değişkeni üzerindeki kısmi etkileri tahmin edilmiş olur.

Backfitting algoritması semiparametrik regresyon modelleri için aynı adımları içermektedir. Öncelikle, modeldeki her bir bağımsız değişken için kısmi artıklar oluşturulur. Eğer seçilen değişkenin doğrusal olmayan uyumu söz konusu ise bu değişken için kısmi artıklar, aynı bağımsız değişkene karşı düzeltilebilirler. Eğer seçilen değişkenin doğrusal uyumu söz konusu ise düzgünleştirme yöntemi yerine sıradan en küçük kareler yöntemi kullanılır. Bacfkitting algoritması, algoritmada yapılabilen değişikliklerden dolayı birçok regresyon modelinin tahmininde kullanılmaktadır.

II.2. Semiparametrik Regresyon Modellerinde Çıkarım: Güven Bantları ve Standart Hataların Hesaplanması

Semiparametrik regresyon modellerinde çıkarım, doğrusal modellerde çıkarım ile parametrik olmayan modellerde çıkarımın birleşiminden oluşmaktadır. Modeldeki doğrusal olmayan değişkenler için güven bantları hesaplanır. Modeldeki doğrusal bileşenler için ise güven aralıklarını oluşturmak ve hipotez testlerini uygulamak için standart hatalar hesaplanır.

Parametrik olmayan değişken için oluşturulacak olan güven bantları ve standart hataların hesaplanması için varyans-kovaryans matrisinin tahminine ihtiyaç duyulmaktadır. Semiparametrik regresyon modellerinde varyans-kovaryans matrisinin tahmini parametrik olmayan regresyon modellerindeki tahmin ile çok benzer ancak daha karmaşıktır.

Semiparametrik modellerde parametrik kısımda bulunan ε_i'nin klasik doğrusal regresyon varsayımlarını sağlaması gerekmektedir. Bu durumda semiparametrik regresyon modellerinin hata teriminin (ε_i) klasik doğrusal regresyonun tüm varsayımlarına sahip olması gerekir [5]. Bu varsayımların gerçekleşmesi tutarlı tahminlerin elde edilmesini sağlayacaktır. Literatürde, varsayımların geçerliliğini incelemek için bazı testler bulunmaktadır. Ancak, semiparametrik regresyon modellerinde varsayımlar çoğunlukla artık grafikleri yardımıyla incelenmektedir.

Semiparametrik regresyon modelinde S matrisi düzgünleştirme matrisi olarak ifade edilmektedir. Bu matris doğrusal regresyon modelinde, şapka matrisi olarak ifade edilen H matrisine benzemektedir. Semiparametrik regresyon modellerinde f'in tahmini (14) eşitliğindeki gibi ifade edilmektedir.

$$\hat{f} = Sy \qquad (14)$$

S matrisi elde edildikten sonra, standart hatalar $\hat{\sigma}^2 SS'$ varyans-kovaryans matrisini kullanarak en küçük kareler yöntemindeki gibi tahmin edilir.

Toplamsal modeller eşitlik (15)'deki gibi ifade edilmektedir [1].

$$\begin{bmatrix} \mathrm{I} & S_1 & S_1 \dots S_1 \\ S_2 & \mathrm{I} & S_2 \cdots S_2 \\ \vdots & \vdots & \vdots \ddots \vdots \\ S_k & S_k & S_k \cdots \mathrm{I} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix} = \begin{bmatrix} S_1 y \\ S_2 y \\ \vdots \\ S_k y \end{bmatrix} \qquad (15)$$

Eşitlik (15)'de I matrisi $(n \times n)$ boyutlu birim matrisi ve $S_1, \cdots, S_k$ ise her bir X değişkeni için düzeltme matrisi olarak ifade edilmektedir.

Teorik olarak (15) eşitliği QR analizi gibi tekrarlı olmayan (noniterative) yöntemlerle çözülebilmektedir. Ancak, bu denklem sistemi tekrarlı olmayan yöntemlerle çözülmek için çok büyük olduğundan backfitting algoritmasının kullanımını zorunlu kılmaktadır [1].

Toplamsal ve semiparametrik regresyon modellerinde güven bantlarının oluşturulması için varyans-kovaryans matrisinin elde edilmesi gerekmektedir. Bu durumda, (15)'deki denklem sistemi eşitlik (16)'daki gibi ifade edilebilir.

$$\hat{S}f = \hat{Q}y \tag{16}$$

(16) eşitliği yeniden düzenlendiğinde (17) eşitliği elde edilmektedir.

$$\hat{f} = \hat{S}^{-1}\hat{Q}y \tag{17}$$

(17) eşitliği yeniden düzenlendiğinde ise (18) eşitliğine ulaşılmaktadır.

$$\hat{f} = Ry \tag{18}$$

(18) eşitliğinde $R = \hat{S}^{-1}Q$ olarak ifade edilmektedir. Eğer gözlemler bağımsız ve aynı dağılıma sahipler ise (19) eşitliği oluşturulabilmektedir.

$$V(\hat{f}) = \sigma^2 RR' \tag{19}$$

(19) eşitliğindeki σ^2 (20) eşitliği ile yer değiştirir ve $V(\hat{f})$ değerine ulaşılır.

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{df_{res}} \tag{20}$$

Artıkların serbestlik derecesi (df_{res}) ise eşitlik (21) ile elde edilmektedir.

$$df_{res} = n - tr(2R - RR') \tag{21}$$

Güven bantları, $\sigma^2 RR'$ matrisinin köşegen elemanlarının karekökleri ile ± 2'nin çarpımından elde edilir [6]. Semiparametrik modellerde R matrisinin köşegen elemanları β tahminlerinin varyansını ifade etmektedir. R matrisini tahmin etmek için tekrarlı yöntemlere ihtiyaç duyulmaktadır. Backfitting algoritması R matrisini tahmin etmek için kullanılabilmektedir. Varyans-kovaryans matrisinin bu yolla tahmini $\hat{f}$'daki yanlılığı düzeltememektedir [1]. Bu durumda bayesgil güven bantları gibi alternatif yöntemler kullanılabilmektedir. Ayrıca bir çok bilgisayar programı ve özellikle R programı, toplamsal ve semiparametrik regresyon modelleri için yanlılıktan arındırılmış varyans-kovaryans matrislerini (bias adjusted variance-covariance matrices) hesaplamaktadır.

II.3. Semiparametrik Regresyon Modellerinde Hipotez Testleri

Semiparametrik regresyon modellerinde hipotez testleri herhangi bir karmaşıklık içermemektedir. Modeldeki parametrik bileşenlerin istatistiksel açıdan anlamlı olup olmadığı araştırılmak istendiğinde, R matrisinden β tahminlerinin standart hataları hesaplanır ve bilinen t testleri uygulanır.

t testleri için hipotezler ise:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

şeklinde ifade edilir.

Semiparametrik regresyon modellerinde, parametrik olmayan bileşen için hipotez testleri iki amaçla uygulanmaktadır. Hipotez testinin birinci amacı, x bağımsız değişkenin y bağımlı değişkeni üzerindeki etkisinin istatistiksel olarak anlamlı olup olmadığını ortaya çıkartmaktır. Hipotez testinin ikinci amacı ise, incelenen değişkenin parametrik olmayan bileşen olarak modelde yer almasının parametrik bileşen olarak modelde yer almasından üstün olup olmadığını belirlemektir. Kısaca amaç, model uyumunun hangi durumda en iyi olduğunu belirlemektir. Her iki hipotez testinde de kısmi F testi ve olabilirlik oran testi kullanılabilmektedir.

Her iki hipotez testini incelemek için eşitlik (22)'de görülen iki değişkenli bir toplamsal model oluşturulmuştur.

$$y = \alpha + f_1(x_1) + f_2(x_2) + \varepsilon \tag{22}$$

x_2 değişkeninin istatistiksel olarak anlamlı olup olmadığını test etmek için, eşitlik (22), eşitlik (23)'e karşı test edilmelidir.

$$y = \alpha + f_1(x_1) + \varepsilon \tag{23}$$

f_2 değişkenin doğrusal olup olmadığını test etmek için ise eşitlik (22), eşitlik (24)'e karşı test edilmelidir.

$$y = \alpha + f_1(x_1) + \beta_1 x_2 + \varepsilon \tag{24}$$

F testinin temeli artık kareler toplamından oluşmaktadır. Herhangi bir toplamsal ya da semiparametrik regresyon modelinin artık kareler toplamı eşitlik (25)'de görüldüğü gibi hesaplanmaktadır.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2 \tag{25}$$

RSS_0 kısıtlı modelin artık kareler toplamı, RSS_1 ise toplamsal veya semiparametrik regresyon modelinin artık kareler toplamı ise F test istatistiği eşitlik (26)'daki gibi hesaplanır.

$$F = \frac{RSS_0 - RSS_1 / [tr(R) - 1]}{RSS_1 / df_{res}} \tag{26}$$

Bu test istatistiği F dağılımına yakınsamaktadır.

Tahmin için backffitting algoritması kullanmak yerine, yeniden ağırlıklandırılmış en küçük kareler ya da kısıtlı maksimum olabilirlik yöntemlerini kullanan bilgisayar programları olabilirlik oran ya da sapma fark (diffence of deviance) testlerini kullanmaktadırlar[1]. Toplamsal ya da semiparametrik regresyon modelleri için olabilirlik oran testi eşitlik (27) yardımıyla uygulanmaktadır.

$$LR = -2(Logolabili\ rlik_0 - Logolabili\ rlik_1) \tag{27}$$

Eşitlik (27)'de $Logolabilirlik_0$ değeri kısıtlı modelin logaritmik olabilirlik değerini, $Logolabilirlik_1$ değeri ise kısıtsız model olan toplamsal ya da semiparametrik regresyon modelinin logaritmik olabilirlik değerini ifade etmektedir. Dikkat edilirse bu test istatistiği iki modelin sapmaları arasındaki fark dikkate alınarak hesaplanır. H_0 hipotezi altındaki test istastiği yaklaşık χ^2 dağılımı göstermektedir. Bu dağılım için serbestlik derecesi ise, iki modelin parametre sayıları arasındaki fark olarak hesaplanmaktadır.

Semiparametrik regresyon modellerinde parametrik olmayan bileşenin tahmininin otomatik düzeltme teknikleri ile yapılması durumunda test istatistiğinin dağılımının yaklaşık χ^2 dağılımı gösterdiği düşünülmektedir. Ancak bu yakınlaşmanın ne kadar olduğu örneklem büyüklüğünden ciddi derecede etkilenmektedir [7]. Hipotez testlerinde H_0 hipotezini güvenle reddebilmek için tahmin edilen p olasılıklarının mümkün olduğunca küçük olması gerekmektedir. Ancak, tahmin edilen p olasılığı H_0 hipotezini zorlukla reddedebilecek seviyede ise bu durumda otomatik düzeltme teknikleri yerine diğer düzeltme tekniklerinin (manual smoothing) kullanılması daha kesin sonuçlara ulaşılmasına olanak sağlayacaktır. Yeniden örnekleme tekniklerinden olan bootstrap yöntemi de (parametrik dağılım varsayımı gerektirmeyen bootstrap yöntemi) bu tür bir problemin çözümüne olanak sağlamaktadır [1].

III. SONUÇ

Regresyon modelinde birden fazla bağımsız değişken söz konusu olduğunda, bağımsız değişkenlerin bazıları bağımlı değişkenle doğrusal ilişki içerisinde bulunurken, bazıları doğrusal olmayan ilişki içerisinde bulunabilir. Bu tür ilişkilerin modellenebilmesi için, parametrik ve parametrik olmayan regresyon fonksiyonlarının toplamsal olarak birleşiminden oluşan semiparametrik regresyon modellerinden yararlanılmaktadır. Bu çalışmada semiparametrik regresyon modellerinin backfitting algoritması ile tahmini güven bantları, standart hataların hesaplanması ve hipotez testleri teorik olarak anlatılmıştır.

YARARLANILAN KAYNAKLAR

[1] Hastie, T. & Tibshirani, RJ. (1999). *Generalized Additive Models*. London: Chapman & Hall.

[2] Aydın, D. (2005). Semiparametrik Regresyon Modellemede Splayn Düzeltme Yaklaşımı İle Tahmin ve Çıkarsamalar. *Yayınlanmamış Doktora Tezi*, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

[3] Lee, K.C. (1990). Avoiding Misspecifications and Improving Efficiency in Hedonic and, Consumption Models: Applications of Semiparametric Method. PhD. Thesis London School of Economics and Political Sciences, London.

[4] Çağlayan, E. (2002). Yarı Parametrik Regresyon Modelleri ile Yaşam Boyu Sürekli Gelir Hipotezinin Türkiye Uygulaması. *Yayınlanmamış Doktora Tezi*, Marmara Üniversitesi, Sosyal Bilimler Enstitüsü, İstanbul.

[5] Fox, J. (2000). *Multiple and Generalized Nonparametric Regression*. Thousand Oaks: A Sage University Paper.

[6] Keele, L. (2008). *Semiparametric Regression For The Social Sciences*. Chichester: John Wiley & Sons.

[7] Hardle, W.; Müller, M.; Sperlich, S. & Werwatz, A. (2004). *Nonparametric and Semiparametrik Models*. Berlin: Springer.

Münevver TURANLI

(mturanli@iticu.edu.tr)

She has graduated from IITBA in 1970. Has completed her graduate and PhD. studies in 1971 and 1975 respectively in the same department. Mrs. Turanlı has been a Professor since 1987. She has worked as a chairwomen in Marmara University Statistics Branch between 1988-2000 and in Marmara University IITB Econometrics Branch between 1990-2000. Mrs Turanlı who has worked as the Dean of Science and Literature Faculty of İstanbul Commerce University between 2001-2004, continues her academic career as The Dean of Commercial Sciences Faculty. Mrs. Turanlı who has many publications, has been married and has two children.

Seda BAĞDATLI

(sbagdatli@iticu.edu.tr)

She has entered Science and Literature Department of İstanbul Commerce University in 2004 after graduating from İstek Acıbadem High Scholl. She has graduated from this department in 2008 and has started her graduate studies in the Statistics Department of İstanbul Commerce University and has fınıshed 2010. She has started PhD. in Marmara University Department of Statistics. She has been working as a research assistant in İstanbul Commerce University Faculty of science and Literature Statistics Department.