Volume: 08 Number: 02 Category: 02 Page: 68-75 ISSN No: 1303-6319 Cilt: 08 Sayı: 02 Kategori: 02 Sayfa: 68-75

İZMİR İLİ SICAKLIK VERİLERİNİN REGRESYON EĞRİLERİ İLE MODELLENMESİ

Neslihan DEMİREL*

ÖZET

Regresyon eğrileri parametrik olmayan bir regresyon analizi türüdür. Regresyon çözümlemesine farklı bir yaklaşım getiren bu yöntem, ekonomi, finans, tıp ve politik bilimlerde sıkça kullanılmaktadır. Bu çalışmada; Devlet Meteoroloji İşleri Genel Müdürlüğü'nden alınan İzmir İline ait 34 yıllık sıcaklık verileri, düğüm noktalarının konumlarının bilinmesi ve bilinmemesi durumları için ayrı ayrı incelenmiştir. Her bir veri kümesine regresyon eğrileri uygulanarak modellenmiştir.

Anahtar Kelimeler: Düğüm noktası, Sıcaklık verisi, Regresyon eğrileri modelleri.

1. GİRİŞ

Regresyon eğrileri modelleri bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi belirleyen parametrik olmayan bir regresyon yöntemidir. Marsh (1983); Marsh (1987); Eubank (1999); Lee (2002); Baccini vd. (2007) çalışmaları incelendiğinde regresyon çözümlemesine farklı bir yaklaşım getiren bu yöntemin sosyal bilimlerde, ekonomi ve finans alanında, tıpta, politik bilimlerde ve çeşitli bilimsel çalışmalarda kullanıldığı görülmektedir.

Regresyon eğrileri ile çözümleme yaparken 'böl-fethet' stratejisi kullanılır. Veri kümesi belirli ölçütlere göre parçalara bölünerek, bu parçalara en uygun model uygulanır. Bu durumda her bir parça için ayrı bir fonksiyon elde edilir. Fonksiyonunun eğim değiştirdiği kritik değer düğüm noktası olarak adlandırılır. Buna göre regresyon eğrileri modelleri'nin genel denklemi;

$$E(y) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x) \qquad (1)$$

şeklinde ifade edilir (Hastie vd., 2008). Burada

β_0 : modelin sabit katsayısı,

β_m : her bir parçanın fonksiyonunun katsayısı,

$h_m(x)$: her bir parçanın fonksiyonudur.

*Öğr. Gör. Dr., Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 35160, Buca, İzmir, e-posta: neslihan.ortabas@deu.edu.tr

Regresyon eğrileri modelleri düğüm noktalarının konumu bilindiğinde, düğüm noktalarının konumları bilinmediğinde ve düğüm noktalarının sayısı bilinmediğinde olmak üzere üç ana başlık altında incelenebilir (Marsh ve Cormier, 2002). Bu çalışmada; düğüm noktalarının konumlarının bilinmesi ve bilinmemesi durumları Devlet Meteoroloji İşleri Genel Müdürlüğü'nden alınan İzmir iline ait 34 yıllık (1975-2008) sıcaklık verileri ile incelenmiştir. Yapılan çalışmada doğrusal veya daha yüksek dereceli polinomial regresyon modelleri ile açıklanamayan ya da zamana bağlı olarak sıcaklık verisinin zaman serisi ile çözümlenmesine alternatif bir yöntem olacak şekilde regresyon eğrilerinin uygulanabilir olduğunu göstermek amaçlanmıştır. 2. bölümde yöntemler kısaca anlatılmış, daha çok bulgular üzerinde durulmuştur. Minitab 15 istatistiksel paket programında kod yazılarak elde edilen analiz sonuçları ile tablo ve şekiller verilmiştir. Son olarak 3. bölümde tartışma ve sonuç belirtilmiştir.

2. YÖNTEM VE BULGULAR

2.1 Düğüm Noktalarının Sayısı ve Yeri Bilindiğinde Regresyon Eğrileri Modelleri

Düğüm noktalarının yeri ve sayısının bilindiği durum regresyon eğrilerini modellemede en basit çözümleme yapılan durumdur. Burada aralıklar için hangi model veya modellerin (doğru, kareli, kübik vs.) uygulanacağına araştırmacı karar verir. Karar verme aşamasında serpme diyagramı kullanılabilir. Her aralık için aynı model uygulanabileceği gibi farklı birkaç model de uygulanabilir. Model kurulurken kullanılan gösterge değişkenler;

$$D_{ti} = 0, \quad X_t < t_i$$
$$D_{ti} = 1, \quad X_t \geq t_i \tag{2}$$

şeklinde olmalıdır. Gösterge değişkenler kullanılarak oluşturulan aralıklara doğrusal regresyon uygulandığında regresyon eğrileri modeli

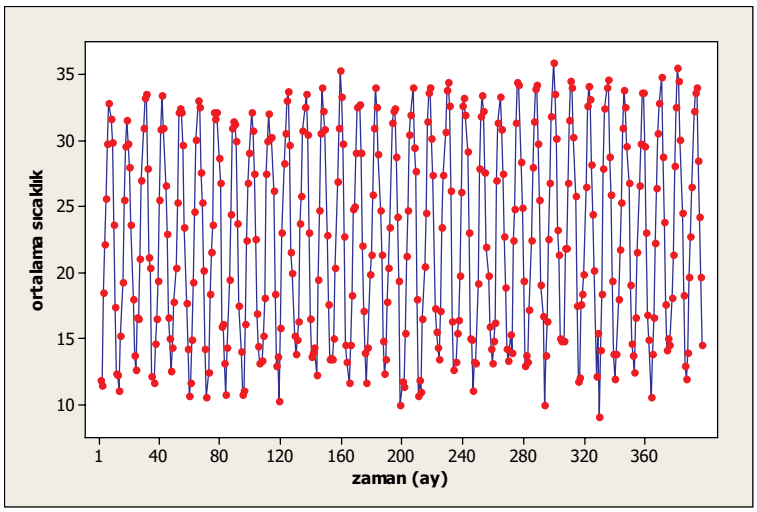
$$E(Y_t) = a_0 + b_0 x_t + b_1 d_{t1}(x_t - t_1) + b_2 d_{t2}(x_t - t_2) + ... + b_i d_{ti}(x_t - t_i) \tag{3}$$

elde edilir. Herhangi bir aralığa yüksek dereceden regresyon uygulanmak istenirse modele

$$c_i d_{ti}(x_t - t_i)^k \tag{4}$$

terimi eklenir. Kareli model için k=2, kübik model için k=3 eşitliği kullanılır (Marsh ve Cormier, 2002). Oluşturulan modellerden hangisinin en iyi olduğuna araştırmacı karar verirken belli ölçütleri göz önünde bulundurmalıdır. Bu amaçla genellikle R-kare, Düzeltilmiş R-Kare, F istatistiği, F istatistiği p-değeri, katsayıların anlamlılığı değerlendirilirken, çoklu doğrusal bağlantı ve otokorelasyon olup olmadığı kontrol edilir. Ayrıca normallik ve varyans homojenliği gibi, kullanılan testlere ait varsayımların sağlanıp sağlanmadığı da incelenmelidir.

Devlet Meteoroloji İşleri Genel Müdürlüğü'nden alınan İzmir iline ait 34 yıllık (1975-2008), ay bazında ortalama sıcaklık verilerinin zamana göre grafiği Şekil 1'de verilmektedir.



Şekil 1. Zamana göre ortalama sıcaklık grafiği

Veri kümesi incelendiğinde, altı aylık periyotlar halinde kırılmaların olduğu görülmektedir. Her altı aya bir düğüm noktası verilerek, oluşacak her parça doğrusal regresyon gibi kabul edilirse; düğüm noktalarının yeri ve sayısı bilindiği regresyon eğrileri modelinin uygun olduğu görülür. Buradaki gösterge değişkenler

$$d_{t1} = 5, d_{t2} = 11, d_{t3} = 17, ..., d_{t67} = 403$$

olarak elde edilir. Tahminlenen model;

$$E(Y_t) = a_0 + b_0 x_t + b_1 d_{t1}(x_t - t_1) + b_2 d_{t2}(x_t - t_2) + ... + b_{67} d_{t67}(x_t - t_{67})$$

şeklinde olacaktır. Tahmin edilen regresyon modeline ait varyans analizi sonuçları Tablo 1'de verilmektedir.

Tablo 1. Ortalama sıcaklık verilerine ait varyans analizi tablosu

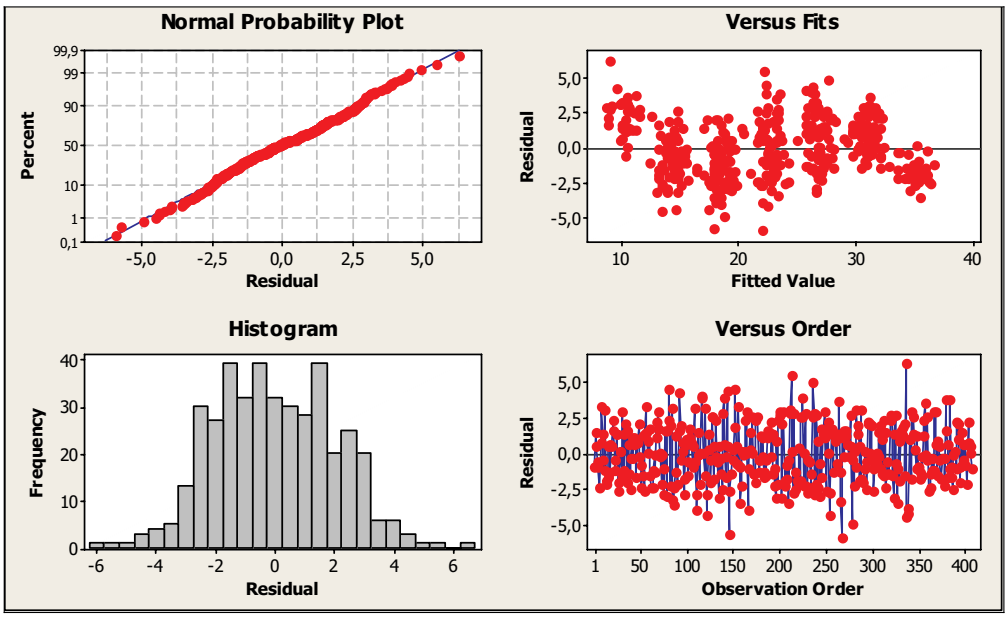
Kaynak	DF	SS	MS	F	P
Regresyon	68	22022,36	323,86	65,60	0,000
Hata	338	1668,64	4,94		
Toplam	406	23691,00			

S = 2,22189 R-Sq = 93,0% R-Sq(adj) = 91,5%

Durbin-Watson istatistiği = 1,75414

Tablo 1 incelendiğinde modelin geçerli olduğu (F=65,60; p=0,000<α=0.05) görülmektedir. Modelde yer alan bağımsız değişkenlerin, bağımlı değişkendeki değişimi açıklama yüzdesi %93 olarak bulunmuştur. Durbin-Watson (DW) test istatistiği 1.75 olarak edilmiştir. DW'nin 2'ye yakın değerleri otokorelasyon olmadığını gösterirken, 0'a yakın değerleri pozitif otokorelasyonu, 4'e yakın değerleri de negatif otokorelasyonu belirtmektedir (Marsh ve Cormier, 2002). Buna göre modelde otokorelasyon yoktur. Regresyon eğrileri modelinin kullanılması için dikkat edilmesi gereken temel ölçütler arasında belirtme katsayısı (R^2), düzeltilmiş R^2, F istatistiği, t

istatistikleri, çoklu doğrusal bağlantı ve otokorelasyon belirtilir (Marsh ve Cormier, 2002). Regresyon eğrileri, parametrik olmayan bir yöntem olduğundan regresyon analizine ait tüm varsayımlar aranmamaktadır. Buna rağmen tahminlenen modele ait hata analizi grafikleri Şekil 2'de verilmektedir. Şekil 2 incelendiğinde hataların sıfır ortalamalı normal dağıldığı (Shapiro-Wilk test istatistiği=0.978, p>0.1), varyansların homojen olduğu (Bartlett test istatistiği=3.69, p=0.297) ve birbirinden bağımsız ve rasgele olduğu görülmektedir.



Şekil 2. Tablo 1'de verilen varyans analiz modeline ait hata analizi grafikleri

2.2 Düğüm Noktalarının Sayılarının Bilinip Konumlarının Bilinmediği Durum Regresyon Eğrileri Modelleri

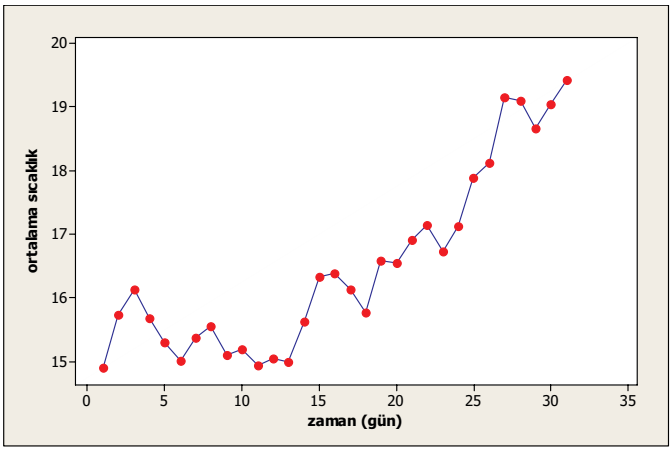
Düğüm noktalarının sayısı ile ilgili bilgi; herhangi bir araştırma sonucu elde edilmiş bir bilgi olabileceği gibi bir tahmin veya varsayımda olabilir. Düğüm sayısına karar verildikten sonra modelin kurulması için aynı sayıda gösterge değişkene ihtiyaç vardır. Düğüm noktalarının yeri ve sayısının bilindiği durumdaki regresyon eğrilerini modellemede kullandığımız eşitlik 2'den yararlanılarak oluşturulan aralıklara doğrusal regresyon uygulandığında elde edilen modeli eşitlik 3 ile göstermek mümkündür. Bu modelde de herhangi bir aralığa yüksek dereceden regresyon uygulanmak istenirse modele eşitlik 4'te yer alan terim eklenir. Kareli model için k=2, kübik model için k=3 eşitliği kullanılır (Marsh ve Cormier, 2002). Bu modelde düğüm noktalarının konumlarının yeri şekil üzerinden belirlenmeye çalışılır. Yakın noktalardaki farklı komşuluklarda değerlendirilerek ortaya çıkan farklı düğüm konumları tek tek denenerek, en uygun olanı regresyon modelinde kullanılan ölçütler dikkate alınarak değerlendirilir.

Bahar mevsiminin başlangıcı olan Mart ayında havalar direkt ısınmaya başlamaz. Belli bir dönem sıcaklık düşer ve sonra tekrar artışa geçer. Bu geçişlerin konumu tam olarak bilinemez Bu nedenle İzmir iline ait Mart ayındaki sıcaklık değişimlerini incelemek için Devlet Meteoroloji İşleri Genel Müdürlüğü'nden elde edilen verilere göre Mart ayının

her günü için 34 yıllık (1975-2008) ortalama değerler Tablo 2'de, günlere ait sıcaklıklar Şekil 3'te verilmektedir.

Tablo 2. Mart ayının her gününe ait 34 yıllık ortalama sıcaklıklar

Gün	°C	Gün	°C	Gün	°C	Gün	°C	Gün	°C
1	14,98	**8**	15,65	**15**	16,28	**22**	17,191	**29**	18,61
2	15,81	**9**	15,15	**16**	16,39	**23**	16,76	**30**	19,20
3	16,30	**10**	15,24	**17**	16,18	**24**	17,141	**31**	19,26
4	15,81	**11**	14,97	**18**	15,95	**25**	17,941		
5	15,36	**12**	15,11	**19**	16,68	**26**	18,141		
6	15,06	**13**	15,46	**20**	16,66	**27**	19,071		
7	15,46	**14**	15,44	**21**	17,11	**28**	19,011		



Şekil 3. Mart ayına ait sıcaklık grafiği

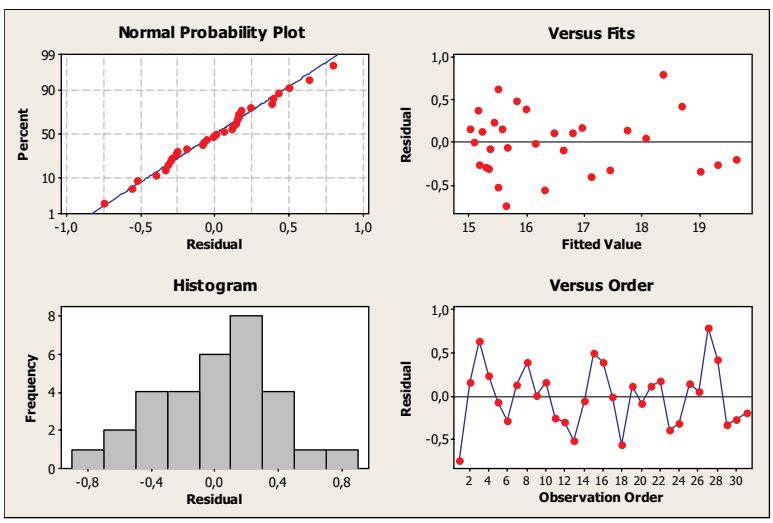
Bu veri kümesine doğru ve kareli modeller ile farklı düğüm noktalarında regresyon eğrileri uygulandığında elde edilen özet sonuçlar Tablo 3'de verilmektedir. Tablo 3 incelendiğinde; en iyi özellikleri gösteren modeller kareli model ile $d_{t10}-d_{t20}, d_{t10}-d_{t21}, d_{t10}-d_{t22}, d_{t10}-d_{t23}$

düğüm noktalarının kullanıldığı regresyon eğrileri modelleridir. Kareli model açıklama oranları ve F istatistiği açısından uygun olmasına rağmen Durbin-Watson istatistiğine bakıldığında modelde negatif otokorelasyon olduğu görülmektedir. Bu durumda en uygun model $d_{t10}-d_{t23}$ düğüm noktaları ile kurulan regresyon eğrileri modelidir.

Tablo 3. Farklı modellere ait özet sonuçlar

Durbin-Watson	Çoklu Doğrusal Bağlantı	Anlamsız katsayı	Anlamlı katsayı	F	R-kare (%) düzeltilmiş	R-kare (%)	Model / Düğüm Noktaları
0,402	YOK	0	2	92,02	75,20	76,00	DOĞRU
1,235	DÜŞÜK	0	3	165,83	91,70	92,20	KARELİ
1,400	DÜŞÜK	0	3	124,28	92,50	93,20	DT10 DT20
1,420	DÜŞÜK	0	3	124,65	92,50	93,30	DT10 DT21
1,434	DÜŞÜK	0	3	126,32	92,60	93,30	DT10 DT22
1,418	DÜŞÜK	0	3	128,62	92,70	93,50	DT10 DT23
1,400	VAR	0	3	123,30	92,40	93,20	DT10 DT19
1,431	DÜŞÜK	0	3	130,65	92,80	93,60	DT11 DT19
1,438	DÜŞÜK	1	2	132,27	92,90	93,60	DT11 DT20
1,484	DÜŞÜK	1	2	135,73	93,10	93,80	DT11 DT22
1,461	VAR	1	2	133,39	93,00	93,70	DT11 DT21

Bu model belirlenirken, en yüksek R-kare, düzeltilmiş R-kare ve F istatistiği değerlerine sahip olması, anlamsız katsayı bulunmaması ve Durbin-Watson istatistiği ile otokorelasyon olmaması ölçütlerinin sağlanmasına dikkat edilmiştir. Durbin Watson için sınırlar d_L=1.299 d_U=1.656. Çoklu doğrusal bağlantı (ÇDB) için kullanılan ölçüt varyans şişirme faktörüdür (VIF). VIF değerinin 4'ten küçük olması ÇDB "Yok", 10'dan büyük olması ÇDB "Var", 4 ile 10 arasında değer alması durumunda "Düşük" olarak belirtilmiştir (Hines ve Montgomery, 1990). $d_{t10} - d_{t23}$ düğüm noktaları ile kurulan regresyon eğrileri modeline ait varsayımların kontrolü Şekil 4'te verilmektedir. Şekil 4 incelendiğinde hataların sıfır ortalamalı normal dağıldığı (Shapiro-Wilk test istatistiği=0.994, p>0.1), varyansların homojen olduğu (Bartlett test istatistiği=0.69, p=0.708) ve birbirinden bağımsız ve rasgele olduğu görülmektedir.



Şekil 4. Tablo 3'de DT10-DT23 düğüm noktaları ile kurulan modelin hata analizi grafikleri

3. TARTIŞMA VE SONUÇ

Regresyon eğrileri modelleri; parametrik yöntemler olan doğru, kareli veya daha yüksek dereceden regresyon modelleri ile zaman serisi analizlerinin uygulanması mümkün olmayan veri kümeleri için kullanışlı bir yöntemdir. Düğüm noktalarının konumu ve sayısı bilindiği durumda regresyon eğrileri modelleri incelenirken İzmir iline ait 34 yıllık aylık ortalama en yüksek sıcaklık verisinden yararlanılmıştır. Veri kümesi doğrusal veya daha yüksek dereceli polinomial regresyon modelleri ile açıklanamamıştır. Aylık sıcaklık verilerinin zaman serisi analizlerinden Box-Jenkins yöntemi ile modellenmesi için de gerekli analizler yapılmış ancak verinin yapısı çok uygun gözükse de doğrusal bir modele uyarlanamamış, varsayımlar sağlanamamıştır. Doğrusal olmayan zaman serisi analizleri ile daha karmaşık modeller üzerinde çalışmak yerine parametrik olmayan bir yöntem olan regresyon eğrileri ile daha uygun sonuçlar elde edilmiştir.

Düğüm noktalarının sayılarının bilinip konumlarının bilinmediği regresyon eğrileri modelleri incelenirken Mart ayının her gününe ait 34 yıllık sıcaklık ortalamalarının dağılımına kareli regresyon modeli uygulandığında yüksek açıklama yüzdesi elde edilmiş ancak varsayımlar sağlanamamıştır. Veri kümesinin yeterli büyüklükte olmaması nedeni ile Box-Jenkins yöntemi uygun görülmemiştir. Veri kümesine regresyon eğrileri modeli uygulanarak varsayımların sağlandığı ve yüksek açıklama yüzdesine sahip bir model elde edilmiştir.

4. KAYNAKLAR

Baccini, M., Biggeri, A., Lagazi, C., Lertxundi, A., Saez, M., 2007. Parametric and Semi-Parametric Approaches in The Analysis of Short-Term Effects of Air Pollution on Health. Computational Statistics & Data Analysis, 51, 4324 – 4336.

Eubank, R. L., 1999. Non-Parametric Regression and Spline Smoothing. 2nd ed., Marcel Dekker, USA.

Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, USA.

Hines, W. W., Montgomery, D. C., 1990. Probability and Statistics in Engineering and Management Science. John Wiley & Sons, Singapore.

Lee, T. C. M., 2002. On Algorithms For Ordinary Least Squares Regression Spline Fitting: A Comparative Study. Journal of Statistical Computation and Simulation, 72(8), 647–663.

Marsh, L.C., 1983. On Estimating Spline Regressions, Proceedings of the Eighth Annual SAS Users Group International Conference. SAS Institute, 723-728.

Marsh, L. C., 1987. Estimating Spline Knots in Time Series Polynomial Regression Models. The Institute of Management Sciences and Operations Research Society of America, St. Louis, 1-15.

Marsh, L. C., Cormier D. R., 2002. Spline Regression Models. Sage Publications, USA.

MODELING IZMIR TEMPERATURE DATA WITH SPLINE REGRESSION

ABSTRACT

Spline regression is a type of non-parametric regression analysis which is frequently applied in economy, finance, medicine and political sciences. In this study, 34 years of temperature data taken from Turkish State Meteorological Service for Izmir is examined under the condition that the locations of knot points are known and unknown. Spline regression method is applied for analyzing each of the data sets.

Keywords: Knot location, Tempature data, Spline regression models.