Regresyon Modellerinde Artıkların İncelenmesinin Önemi

Reha ALPAR (*)
Levent ÖNER (**)

Özet: *Bu çalışmada, regresyon çözümlemesindeki model bozuklukları ile model bozukluklarının incelenmesinde kullanılan standardize edilmiş e_{is} artıklarının dağılımı üzerinde durulmuştur.*

Summary: *The model deficiencies of linear regression analysis and distribution of standardized residuals (e_{is}) that are used for detecting model deficiencies have been studied in this article.*

GİRİŞ

Değişkenler arasındaki ilişkinin kuvvetini ve yönünü belirlemek (korelasyon) ve de bu ilişkiyi bir denklem ile ifade etmek (regresyon), eczacılıkta sık sık başvurma gereği duyulan iki yöntemdir.

Ancak, bilindiği gibi, bu tür çözümlemeler sonucunda bulunan büyük bir r^2 değeri, korelasyon katsayısı için anlamlı bir t istatistiği ya da regresyon için anlamlı bir F istatistiği her zaman uyumun bir göstergesi değildir. Anscombe, 1973 yılında yaptığı ilginç bir çalışma ile konunun önemini açıklamaya çalışmıştır (1). Bu çalışmada, aynı regresyon denklemini ($y = 3+5x$) veren dört aynı veri kümesi vardır ve aşağıda verilen istatistikler dört ayrı veri kümesi için de aynıdır:

Denek Sayısı (n) =11 $\bar{x}$ =9 $\bar{y}$ =7.5

Regresyon Kareler Toplamı (RKT) = 27.5

Regresyondan Ayrılış Kareler Toplamı (RAKT) = 13.75

Belirtme Katsayısı (r^2) = 0.667

Ancak, regresyon denklemleri aynı olan bu dört ayrı veri kümesine ilişkin grafikler çizildiğinde (Şekil 1), ilk grafik dışında doğrusal model ile deneysel noktalar arasında bir uyumdan söz edilemeyeceği görülmektedir.

Amaç

Bu çalışmada, regresyon çözümlemesindeki model bozukluk-

Başvuru Tarihi: 4.4.1989
Kabul Tarihi: 16.1.1990

(*) H.Ü. Tıp Fakültesi Biyoistatistik Bilim Dalı, Ankara.
(**) H.Ü. Eczacılık Fakültesi Farmasötik Teknoloji ABD, Ankara.

larının incelenmesinde kullanılan standardize edilmiş e_{is} artıkları üzerinde durulacak ve konunun önemi bir örnekle açıklanacaktır.

Yöntem

Regresyon çözümlemelerindeki model bozukluklarını ya da modelin geçerliliğini incelemek için kullanılan basit ve etkin bir yöntem, artıklarının incelenmesidir. Bilindiği gibi i. artık; $e_i = y_i - y'_i$ olarak tanımlanır. Burada, y_i ; gözlenen değerler, y'_i ; gözlenen değerlere ilişkin kestirim değerleridir. i. standardize edilmiş artık e_{is} ise, (1)

$$e_{is} = \frac{e_i}{S_{yx}}$$

olarak tanımlanır. Burada S_{yx}; regresyon denkleminin standart hatasıdır.

Standardize edilmiş e_{is} artıklarının, sıfır ortalama civarında ve ± 2 arasında dağılma eğilimi vardır. Ordinat olarak; e_{is} artıklarının, apsis olarak da y_i ya da x_i değerlerinin alınmasıyla oluşacak grafiğin rastgele bir dağılım göstermesi durumunda modelin geçerliliğinden söz edilebilecektir. Bu rastgele dağılımın dışında meydana gelen ve belli biçimler gösteren nokta grafikleri, elde edilen modelin geçersizliğini bize gösterir (Şekil 2).

Şekil 2'yi inceleyecek olursak; Şekil 2a'ya benzer biçimde oluşacak bir nokta dağılımı, modelin uygunluğunu; Şekil 2b, 2c ve 2d ise, elde edilen modelin uygun olmadığını ve eldeki veriye başka modellerin uygulanması gerektiğini bize gösterir. Çünkü, yukarıda değinildiği gibi, Şekil 2b, 2c ve 2d'de verilen e_{is} dağılımları belli biçimler göstermektedir.

Örnek Uygulama

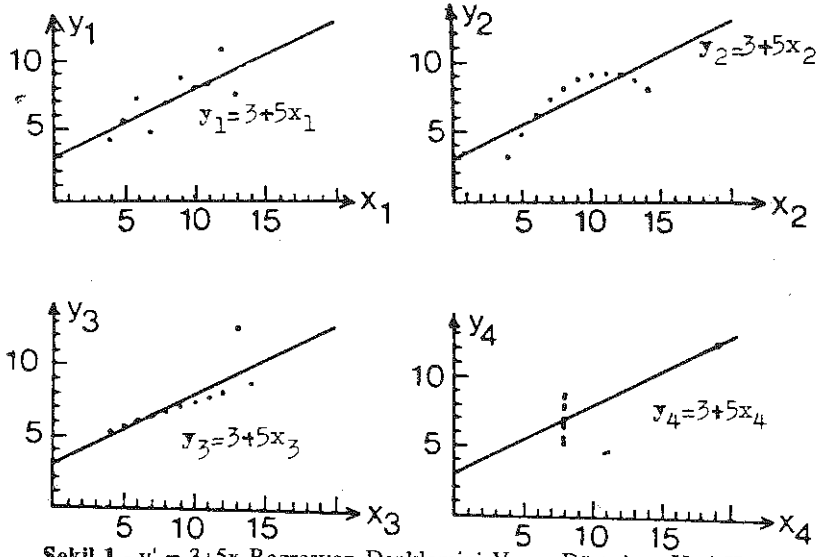
Zamana (x) karşı ilacın çözünme hızı (y) arasında ilişki arayan bir araştırıcı, elde ettiği verilere doğrusal regresyon ve korelasyon çözümlemelerini uygulamış ve Tablo 1'de verilen istatistikleri elde etmiştir.

Tablo 1'deki istatistiklerinden de anlaşılacağı gibi, zaman ile ilacın çözünme hızı arasında çok yüksek bir doğrusal ilişkiden söz edilebileceği düşünülebilir. Bu yüksek ilişkiye bağlı olarak da regresyon denkleminin çok yüksek bir F değerine sahip olduğu, dolayısıyla da zaman-çözünme hızı arasındaki ilişkinin doğrusal olarak gösterilebileceği kanısına rahatlıkla varılabilir.

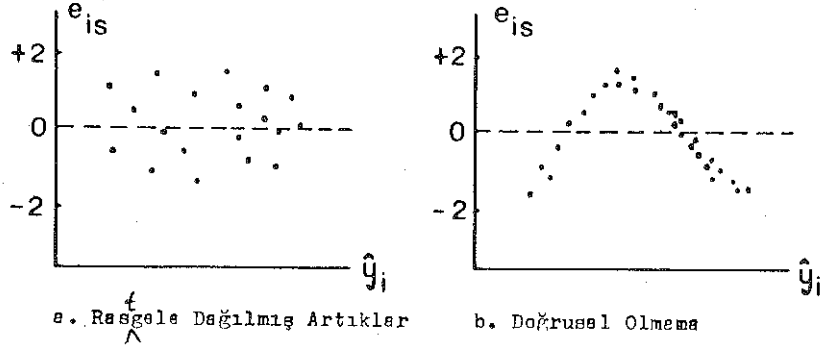
Ancak, daha önce de belirttiğimiz gibi, verilen $y' = 17.6057 + 0.2731x$ denklemine olan uyumun tam olup olmadığına sadece yüksek bir r, r^2 ya da F değeri ile anlayabilmek mümkün olamamakta ve mutlaka bulunan denkleme ilişkin artıkların incelenmesi gerekmektedir. Bu amaçla, $y' = 17.6057 + 0.2731x$ denklemini veren x_i, y_i değerleri ile bunlara ilişkin y'i, e_i ve e_{is} değerleri Tablo 2'de verilmiştir.

Tablo 2'de verilen ve eşitlik (1) yardımı ile bulunan e_{is} artıklarının y_i değerlerine karşılık gelen çizimleri Şekil 3'de verilmiştir.

Şekil 3'de görüldüğü gibi, e_{is} artıklarının y_i değerlerine karşılık gelen çizimlerinin, sıfır civarında rastgele

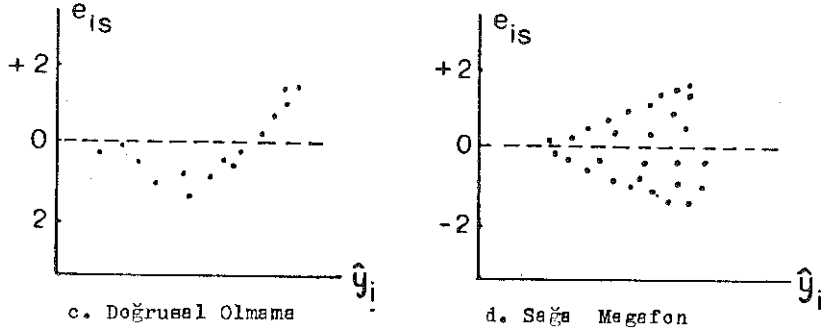


Şekil 1 - y' = 3+5x Regresyon Denklemini Veren Dört Ayrı Veri Kümesinin Grafiği.



a. Rasgele Dağılmış Artıklar

b. Doğrusal Olmama



c. Doğrusal Olmama

d. Sağa Megafon

Şekil 2 - Artıkların Dağılımına İlişkin Dört Örnek.

Tablo 1 - Zaman (x) ile Çözünme Hızı (y) İlişkisine Ait İstatistikler.

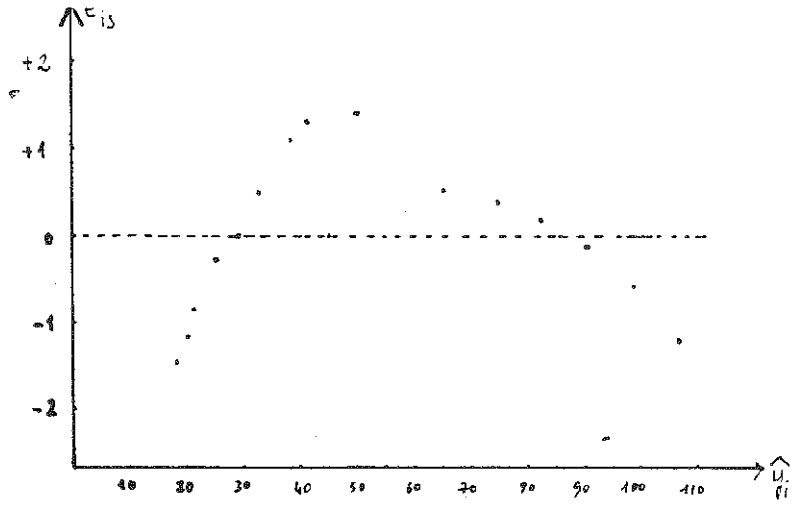
$n = 15$	$y' = 17.6057 + 0.2731x$
$r = 0.9529$ $r^2 = 0.9080$	$F = 128.3$ $p < 0.01$
$S_r = 0.08412$	$RAKO = 104.3191$
$t = 11.327$	$S_{yx} = \sqrt{RAKO} = 10.2137$
$p < 0.01$	

Tablo 2- $y' = 17.6057 + 0.2731x$ Doğrusal Modeline İlişkin x_i, y_i, y'_i, e_i ve e_{is} Değerleri.

x_i	y_i	y'_i	e_i	e_{is}
5	4.42	18.971	- 14.5531	- 1.4247
10	8.10	20.337	- 12.2369	- 1.1981
15	12.12	21.702	- 9.5825	- 0.9382
30	22.14	25.799	- 3.6593	- 0.3583
45	29.92	29.896	0.0240	0.0023
60	40.42	33.993	6.4272	0.6293
75	49.54	38.090	11.4504	1.1210
90	56.52	42.186	14.3336	1.4033
120	66.54	50.380	16.1601	1.5822
180	72.82	66.767	6.0530	0.5926
210	78.92	74.961	3.9594	0.3876
240	84.94	83.154	1.7859	0.1748
270	89.92	91.348	- 1.4277	- 0.1397
300	93.12	99.541	- 6.4212	- 0.6286
330	95.42	107.735	- 12.3148	- 1.2057

dağılmadığı ve dolayısıyla, zaman-çözünme hızı arasıdaki ilişkinin doğrusal modelle ifade edilmesinin hatalı olacağına karar verilir. Bu nedenle, araştırıcı, çözümlemede kullandığı doğrusal model yerine başka bir model (örneğin, doğrusal olmayan) kullanarak sonucu yeniden denetlemelidir.

Diğer taraftan, özellikle son yıllarda istatistiksel çözümlemeler için geliştirilmiş paket programlar yardımıyla, istendiğinde, artıklarının dağılımı grafik olarak elde edilebilmektedir.



Şekil 3 - Tablo 2'de verilen x ve y Değerlerine İlişkin e_{is} Artıklarının y'i Kestirim Değerlerine Karşılık Çizimi.

Kaynaklar

1. Anscombe, F.J., Graphs in Statisticial Analysis, **American Statistician**, 27, 17-20, 1973.

2. Wonnacott, T.H., Introductory Statistics For Bussiness and Economics, New York, John Wiley and Sons, 1977.

3. Chatterjee, Samprit and Bertram, Price, Regression Analysis by Example, New York, Wiley, 1977.

4. Erar, A., Bağlanım (Regresyon) Çözümlemesi (Ders Notları, H.Ü., Fen Fak. İstatistik Bölümü), 1985.

5. Draper, N.R. and Smith H., Applied Regression Analysis, New York, Wiley, 1981.

6. Ertek, T., Ekonometriye Giriş, ODTÜ. Ankara, 1973.

Cesaretli bir adam tek başına çoğunluktur.

Andrew JACKSON