KALINTI ANALİZİ İLE REGRESYON FONKSİYONUNDAKİ İLİŞKİNİN BELİRLENMESİ

Atıf Evren

Abstract: In model specification , residuals or some transformed forms of residuals are employed widely in regression analysis. Besides, to detect the functional form or the nonlinearity in the parameters, some residual plots may be useful.

Çok değişkenli doğrusal regresyon analizinde modelin varsayımları ile gözlem değerleri arasındaki uyum ya da uyumsuzluk; belirli bir noktadan sonra tahminin standart hatası, belirginlik katsayısı R^2 , düzeltilmiş R^2, uygunluk katsayısı gibi "global" istatistikler tarafından yansıtılamaz.[1] Bu durumda modelin varsayımlarından sapmaları sergileme yeteneklerinin yüksek olması nedeniyle kalıntıların analizi önem kazanmaktadır.

Tahmin değerleri ile gözlem değerleri arasındaki farklar olarak tanımlanan kalıntıların incelenmesi ile verilerin modelin varsayımlarına uygunsuzluğu, model dışı bırakılan önemli açıklayıcı değişkenlerin bulunup bulunmadığı, doğrusal olmayan bir bağıntının doğrusal kabul edilmesi ile yapılan bir matematiksel biçimleme hatasının varolup olmadığı, kalıntıların birbirlerinden bağımsız olup olmadıkları, homoskedastisite, heteroskedastisite durumlarının incelenmesi, hata terimlerinin normal dağılıp dağılmadıkları ortaya konabilir.[2]

1.1.H Projeksiyon Matrisi ve Regresyon Analizi

Matris notasyonu kullanılacak olursa çok değişkenli doğrusal regresyon modeli

$$Y_{nx1} = X_{nxp}\beta_{px1} + \varepsilon_{nx1} \qquad (1.1.1)$$

denklemi ile ifade edilir.

---

[1] Bu tür istatistiklerin kullanım sınırlılıkları ile ilgili olarak Genceli,M., "Ekonometride İstatistik İlkeler, Filiz Yayınevi, s96-106 arasına bakılabilir.

[2] Maddala'dan aktaran Genceli,M.,a.g.e.,s453

Burada Y_{nx1} bağımlı değişkenin gözlem değerleri vektörü; X_{nxp} , bağımsız değişkenlerin gözlem değerleri matrisi; β_{px1} değerleri tahmin edilecek olan parametreler vektörü; ve yine ε_{nx1} de modele stokastik olma niteliğini veren anakütle hata payını ifade eden vektördür.

β parametrelerinin en küçük kareler (E.K.K.)tahminleri

$$b = (X'X)^{-1} X'Y \qquad (1.1.2)$$

denklemi ile bulunur.Yine değişkenin tahmin değerleri $\hat{Y}$ vektörü ile ifade edilecek olursa

$$\hat{Y} = Xb \qquad (1.1.3)$$

eşitliği yazılabilir. Yine bu denklemde (1.1.2) no'lu denklem yerine konulacak olursa

$$\hat{Y} = X(X'X)^{-1} X'Y \qquad (1.1.4)$$

tahmin denklemleri elde edilecektir.

H= $X(X'X)^{-1} X'$ olacak biçimde bir **H** matrisi tanımlanırsa

kısaca $\hat{Y} = HY$ (1.1.5)

biçiminde bir formülasyona gitmek de mümkün olacaktır. Son olarak kalıntılar vektörünün

tanımı olan $e = Y - \hat{Y}$ (1.1.6)

denkleminden hareket edilir ve (1.1.5) no'lu eşitlik burada yerine konacak olursa

$$e = Y - HY = (I - H)Y \quad (1.1.7)$$

denklemine ulaşılacaktır.

1.2.H Projeksiyon Matrisinin Bazı Özellikleri[3]

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdot & \cdot & h_{1n} \\ h_{21} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ h_{n1} & \cdot & \cdot & \cdot & h_{nn} \end{bmatrix} \quad \text{ya da } \mathbf{H}=[h_{ij}] \quad (1.2.1)$$

olsun. Burada

H 'ın simetrik; $H' = H$ (1.2.2)

idempotent; $H^2 = H$ (1.2.3)

(aslında p herhangi bir kuvvet olmak üzere $H^p = H$) [4]

$$HX = X \quad (1.2.4)$$

$$(I - H)X = 0 \quad (1.2.5)$$

$$H(I - H) = 0 \quad (1.2.6)$$

---

[3] H matrisinin kimi özellikleri için bkz. Weisberg,S.; Applied Linear Regression, Second Edition; John Wiley; s126-127. ve Montgomery D.C.,Peck,E.A.;Introduction to Linear Regression Analysis; Second Edition; John Wiley; 8. Bölüm

[4] Draper N.,R., Smith H.; Applied Regression Analysis; Third Edition; John Wiley; s205

özelliklerine sahip olduğu kısaca gösterilebilir.
H projeksiyon matrisi **X** matrisinin sütun vektörlerince oluşturulan uzayda bir ortogonal projeksiyon operatörüdür. Eğer (1.2.4) ve (1.2.5) no'lu denklemlerden hareket edilirse; **H** ve **I-H** matrisleri yardımı ile **Y** vektörünün tahmin ve kalıntı uzaylarına projeksiyonu gerçekleştirilmiş olmaktadır. Bu iki alt-uzayın ((1.2.6) no'lu denklem nedeniyle) birbirine ortogonal olduğu da vurgulanmalıdır. Bunlara **H** matrisinin genellikle tersinin olmadığını, ve **X** matrisi ile aynı ranka sahip olduğu da ekleyebiliriz.[5]

H'ın (i,j). elemanı ise şu formülle bulunmaktadır[6]:

$$h_{ij} = x'_i (X'X)^{-1} x_j \qquad (1.2.7)$$

Son olarak **H** matrisinin şu özelliklerinden söz etmek de yerinde olacaktır:

$$\sum_{i=1}^{n} h_{ii} = rank(X) = p \qquad (1.2.8)$$

$$\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1 \qquad (1.2.9)$$

1.3.H Projeksiyon Matrisi ve Kalıntılar

Orijinden geçmeyen bir çok değişkenli doğrusal regresyon modelinde

$\sum_{i=1}^{n} e_i = 0$ eşitliği ve dolayısıyla kalıntı ortalamalarının sıfır olması hali sözkonusudur:

$$E(e) = 0 \qquad (1.3.1)$$

---

[5] Weisberg, S. ;a.g.e.; s110
[6] a.g.y.

Kalıntıların varyans-kovaryans matrisi ise

$$V(e) = (I - H)\sigma^2 \qquad (1.3.2)$$

eşitliği ile hesaplanmaktadır. Yine i. kalıntının varyansının skaler bir biçimde

$$V(e_i) = \sigma^2(1 - h_{ii}) \qquad (1.3.3)$$

ile bulunacağı öngörülebilir. Burada h_{ii} , **H** matrisinin i. köşegen elemanıdır. Öte yandan iki kalıntı arasındaki kovaryans,

$$Cov(e_i, e_j) = -h_{ij}\,\sigma^2 \qquad (1.3.4)$$

eşitliği ile bulunmaktadır. Öte yandan ana kütle varyansı σ^2 genellikle bilinmediği için değeri kalıntı kareleri yardımı tahmin edilebilir:
Eğer kalıntı kareleri toplamı "K.K.T." ifadesi ile gösterilecek olursa

$$\text{K.K.T.} = \sum_{i=1}^{n} e_i^2 \qquad (1.3.5)$$

ve bu toplam uygun serbestlik derecesine bölünecek olursa

$$\text{K.K.O.} = \sum_{i=1}^{n} (e_i - \bar{e})^2 / (n - p) \qquad (1.3.6)$$

olarak bulunur. Yukarıdaki "K.K.O." ifadesi ortalama kalıntı kareleri ifadesinin bir kısaltması olarak kullanılmıştır. Modelin geçerli olması durumunda ise , K.K.O.''nin ana kütle varyansının sistematik hatasız bir tahmincisi olduğu söylenir.[7]

---

[7] Neter J., Wasserman, W.;Applied Linear Statistical Models;Richard D. Irwin Inc., s98

1.4.Kaldıraç Noktaları

H matrisinin köşegende yeralan h_{ii} değerleri kaldıraç noktaları (leverages) olarak adlandırılırlar ve kalıntıların analiz edilmesinde kritik bir rol üstlenirler. Hoaglin ve Welsch(1978)[8] (1.1.5) no'lu denklemin cebirsel bir versiyonunu verdiler:

$$\hat{Y}_i = h_{ii} Y_i + \sum_{j \neq 1} h_{ij} Y_j \qquad (1.4.1)$$

Hoaglin ve Welsh'e göre bu denklemden hareketle, h_{ii} aracılığıyla Y_i 'nin , $\hat{Y}_i$ 'ye yapacağı katkıyı saptamanın mümkündür.

Fakat bu nokta istatistikçiler arasında oldukça tartışmalı bir nokta olmuştur. Çünkü bu analizde Y_j değişkeninin yükleneceği rol tamamen ihmal edilmiştir. Bu yüzden örneğin Cook ve Weisberg bu noktayı "kaldıraç noktası" olarak değil, bir "potansiyel" olarak nitelemektedir.[9]

1.5.Ham Kalıntı Değerlerinin Analizdeki Yetersizliği

h_{ii} değeri i. noktanın X uzayındaki konumunun bir ölçütüdür.[10] e_i 'nin varyansı ile sözkonusu X_i noktasının bulunduğu konum arasında bir ilişkiden sözetmek mümkündür. (1.3.3) no'lu denkleme göz atıldığında X uzayında merkeze (ortalamaların bulunduğu ağırlık merkezine) yakın noktaların varyansı, merkezden uzak noktaların varyansına göre büyük olacağı sonucuna varılacaktır. Oysa modelden sapmaların en büyük olduğu noktalar en uzak noktalardır. Fakat bu noktaların varyanslarının küçüklüğünden dolayı modelden sapmaların bu noktalarda tespit edilmesi güç olmaktadır. Çünkü bu noktalardaki kalıntılar genellikle küçük olmaktadır.[11]

---

[8] Hoaglin,D.C., Welsch; "The hat matrix in regression and ANOVA";American Statist.; 32(1) ;17-22

[9] Weisberg; a.g.e.; s111.

[10] a.g.y.

[11] Montgomery; a.g.e.;s172.

1.6.Dönüştürülmüş Kalıntılar

Ham kalıntılara dayalı analizin yukarıda değinilen sakıncalarını gidermek üzere kalıntıların dönüştürülmesi yoluna gidilmektedir. Temel olarak literatürde, kalıntıların iki türde dönüşümü gerçekleştirilmektedir.

$$d_i = \frac{e_i}{\sqrt{K.K.O.}} \qquad \text{i=1,2,...,n} \qquad (1.6.1)$$

dönüşümü gerçekleştirilirse standardize edilmiş kalıntılar elde edilir.
Burada d_i i. standardize edilmiş kalıntı değerini ifade etmektedir. Standardize değerlerin varyansının da 1 olacağına göre varyans eşitsizliği problemi standardize edilmiş kalıntılar ile ortadan kaldırılmış olmaktadır. Bunun yanısıra kalıntıların varyansından da yararlanılarak "studentize edilmiş kalıntılar bulunabilir:

$$r_i = \frac{e_i}{\sqrt{Var(e_i)}} i = 1,2,...n \qquad (1.6.2)$$

$$Var(e_i) = \sigma^2 (1 - (\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}})) \qquad (1.6.3) \text{ ve}$$

$$r_i = \frac{e_i}{\sqrt{Var(e_i)}} = \frac{e_i}{\sqrt{K.K.O.(1 - (\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}))}} \qquad (1.6.4) \text{ olarak}$$

bulunur.[12]

[12] a.g.e.; s69

$E(r_i)=0$ olmasına karşılık r_i değerleri az da olsa birbirleri ile korelasyonludurlar. Ama bu durum pratikte bir sorun yaratmaz ve modelin geçerli olması halinde studentize edilmiş kalıntıların sabit varyansa (=1) sahip oldukları kabul edilir. Studentize edilmiş kalıntılar arasındaki kovaryans ise şu formülle bulunur:

$$Cov(r_i, r_j) = \frac{-h_{ij}}{(1-h_{ii})^{1/2}(1-h_{jj})^{1/2}} \quad [13] \qquad (1.6.5)$$

Modelin varsayımlarının ve normallik koşulunun yerine gelmesi halinde ise $\frac{r_i^2}{n-p}$ istatistiğinin 1/2 ve (n-p-1)/2 parametre değerlerine sahip bir beta dağılımına uyduğu belirtilmektedir.[14]

Pek çok durumda, özellikle büyük veri setleri için kalıntıların varyansı stabilize olmaktadır. Bu gibi durumlarda standardize ve studentize edilmiş kalıntılar arasındaki fark az olmaktadır. Dolayısıyla standardize edilmiş ve studentize edilmiş kalıntılar genellikle benzer bilgileri sunmaktadırlar. Ama büyük bir kalıntı değerine ve büyük bir h_{ii} değerine sahip bir nokta potansiyel olarak en azından en küçük kareler tahminleri üzerinde etkili olur. Bu durumda studentize edilmiş kalıntıların kullanılması önerilir.[15]

Uygulama

DİE verilerinden hareketle Türkiye toplam nüfusunun yıllara göre değişimi STATİSTİCA programı ile incelenmeye çalışılmış ve modelin biçimlendirilmesine ilişkin kalıntı değerlerinden yararlanılmıştır. Üzerinde çalışılan veriler şöyledir[27]

---

[13] Weisberg; a.g.e.; s114
[14] a.g.y.
[15] Montgomery; a.g.e.;s172
[27] İstatistik Göstergeler:1923-1991;DİE Yayını;No 1472

YILLAR	TOPLAM NÜFUS (BİN)
1927	13648
1935	16158
1940	17821
1945	18790
1950	20947
1955	24065
1960	27755
1965	31391
1970	35605
1975	40348
1980	44737
1985	50664
1990	56473

İlk olarak iki değişkenli doğrusal regresyon modeli denenmiştir. Yıllar bağımsız değişken; toplam nüfus da bağımlı değişken olarak alınmış ve TOPLAM NÜFUS=a+b*(YILLAR) biçiminde bir model düşünülmüştür. Bu arada hesaplamalarda kolaylık sağlaması bakımından sözgelimi 1927 yılı için bağımsız değişken 27; 1935 için 35,... biçiminde girilmiştir. Bu model için elde edilen özetleyici istatistikler şöyledir:

PARAMETRELER	TAHMİNLERİ	STANDARD HATA	T-DEĞERİ
a	-10264.8	2950.065	-3.4795
b	684.5	47.017	14.558

$$YILLAR_{BETA} = 0.975$$

Varyans Analizi Sonuçları ise şöyledir:

Değişim Kaynağı	Kareler Toplamı	S.D.	Kareler Ortalaması	F Değeri
Regresyon	2219980000	1	2219980000	211.9394
Hata	115221000	11	10474600	
Toplam		12		

p değeri=0.000

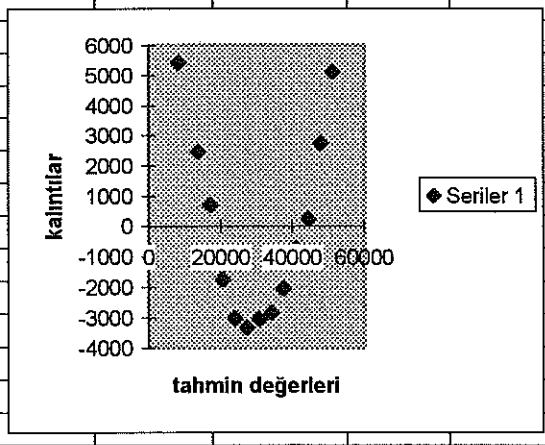
Bunun dışında diğer özetleyici istatistikler de şöyledir:

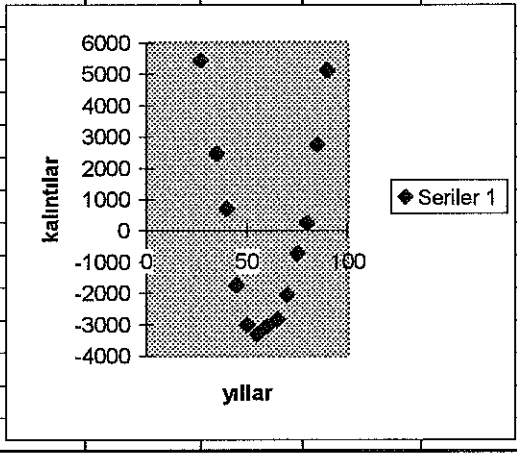
$R^2 = 0.95$ $R_a^2 = 0.946$ Tahminin Standard Hatası(T.S.H.)=3236.5

$DW_D = 0.3037$ ve Serisel Kor.=0.785

Bunun yanısıra kalıntı değerlerine bakıldığında da şunlar görünmektedir. STATİSTİCA bir değeri sapan değer olarak değerlendirirken ortalama artı eksi 2 standard hata kriterini kullanmaktadır. Bu açıdan ele alındığında sapan değer gözükmemektedir. Bunun yanısıra Cook mesafeleri gözlendiğinde 0.01 ile 0.881 arasında değerler ortaya çıkmaktadır. En yüksek değer olan 0.881 1927 yılına aittir. Bu da Cook ölçütü açısından bir sorun yaratmamaktadır. Standardize kalıntılar gibi diğer istatistikler incelendiğinde benzer bir eğilim görülmektedir. Otokorelasyon istatistikleri dışında pek bir problem yok gibi görünse de kalıntıların bağımsız değişkene göre değişiminin incelenmesi (doğrusal olmayan bir trend "heteroskedastisite"ye işarettir) ve yine kalıntıların bağımlı değişkenin tahmin değerlerine göre incelenmesi (böyle bir grafikte doğrusal olmayan bir trend modelin doğrusal olmadığına işarettir) modelin sanıldığı kadar problemsiz olmadığını göstermektedir.

yıllar	nüfus	nüf.tahm	kalıntılar
27	13648	8216,3	5431,7
35	16158	13692,2	2465,8
40	17821	17114,6	706,41
45	18790	20537	-1747
50	20947	23959,4	-3012
55	24065	27381,9	-3317
60	27755	30804,3	-3049
65	31391	34226,7	-2836
70	35605	37649,1	-2044
75	40348	41071,5	-723,5
80	44737	44493,9	243,06
85	50664	47916,4	2747,6
90	56473	51338,8	5134,2





Bu verilerden hareketle, modelin bir kez de doğrusal olmayan regresyon ile incelenmesine gidilmiştir. Bu durumda yine STATİSTİCA yardımı ile $Nüfus = a + \exp(b + c * (yıllar))$ modeli denenmiştir. Bu durumdaki istatistik bulgular ise şöyledir:

n=13	Sabit-a	it-b	c
Tahmin	2083.626	8.62710	0.02541
Standard hata	1525.789	0.14753	0.00139
T-değeri	1.366	58.47817	18.23678
P- değeri	0.202	0.000	0.000

Yine E.K.K. fonksiyonunun değeri 3761557.2342; $R^2 = 0.99$ olarak bulunmuştur. Modelin doğrusal olmaması bazı istatistikler bazında kıyaslama yapmayı engellemektedir. Bununla birlikte hem E.K.K. fonksiyonunun birinci modelden daha iyi sonuç vermesi; parametre tahmin ve standard hatalarına bakıldığında önemli bir probleme rastlanmaması nedeniyle bu model birincisine tercih edilmelidir. Son olarak Türkiye için geçerli olan yıllık ortalama yüzde 2.5 olan nüfus artış oranının "c" parametresinin tahmincisi olarak hesabedilmiş olması da ilginçtir.