REGRESYON ANALİZİNDE ORTAYA ÇIKABİLECEK HATALAR ve BAZI ÇÖZÜM ÖNERİLERİ

Araş.Gör. Bülent MİRAN*

Temelini matematikten alan istatistik biliminden, tarım da dahil olmak üzere pek çok alanda yaygın olarak yararlanılmaktadır. Bu çerçevede özellikle birbiriyle ilişkili olaylara ait gözlemleri regresyon analiziyle basit denklemler halinde ifade etmek ve geleceğe yönelik tahminlerde bulunmak çok sık karşılaşılan bir uygulamadır. Elle yapılması çok güç olmayan bu uygulama, bilgisayar teknolojisinin getirdiği kolaylıklarla daha kolay bir süreç haline gelmiş ve pek çok sektördeki çalışmaların önemli bir aracı olmuştur. Konuyla ilgili bilgisayar paket programları, kullanıcıya yoğun bir bilgi yükü halinde ayrıntılı sonuçlar sunabilmektedir. Ancak, gerek elle gerekse bilgisayar yardımıyla yapılan regresyon analizlerinde, elde edilen denklemlerin güvenilirlikleri, gerekli tüm bilgilere sahip olunduğu halde genellikle test edilmemektedir. Aynı zamanda, çalışmalardan yararlananların bu testi kendilerinin yapabilmesini sağlayacak bilgilerin verilmesi de çoğunlukla ihmal edilmektedir. Birer başvuru kaynağı olma özelliği taşıyan bu çalışmaların böylesi bir eksikliği, yararlanılmaları sırasında daha dikkatli olunmasını gerektirebilecektir. Gerçekten de elde edilen denklemler, ek bir analize tabi tutulmadığı sürece kolay anlaşılamayacak hayati derecede önemli hataları içerebilecektir.

Yaygın şekilde kullanılan regresyon analizinde karşılaşılabilecek hataların incelenmesi, var olup olmadıklarının belirlenmesi, ortaya çıkış nedenleri ve bunların giderilmesine ilişkin bazi çözüm önerilerinin ortaya konması bu çalışmanın ana amacını oluşturmaktadır. Bu doğrultuda, çalışmanın; daha nitelikli, daha güvenilir bilgilerle yüklü çalışmalar için dikkat çekici olması hedeflenmiştir.

REGRESYON ANALİZİ

Regresyon, bağımlı değişken ile bağımlı değişkeni açıkladığı varsayılan bağımsız değişken veya değişkenler arasındaki matematiksel ilişkinin bir denklemle gösterilmesidir. Regresyon analizinde, açıklanmak istenen bağımlı bir değişkenle bir veya daha fazla bağımsız veya açıklayıcı değişken arasındaki kantitatif ilişki tahmin edilerek istatistiksel açıdan incelenmeye çalışılır. Diğer bir anlatımla iki değişken birbiriyle ilişkiliyse ve değişkenlerden biri değiştiğinde, diğerinde sistematik bir değişme gözleniyorsa regresyon analizine başvurulabilir. Açıklayıcı değişken sayısı tek ise basit regresyon (bivariate), birden fazla ise çoklu (multivariate) regresyondan söz edilir. Regresyon denklemi genel

*E.Ü.Ziraat Fak. Tarım Ekonomisi Böl., Bornova

olarak

$$Y_i = \beta_o + \beta_1 X_i + e_i \quad , \quad i=1,2,..,n$$

şeklinde ifade edilebilir. Burada Y_i bağımlı değişkeni, β_o tahmin edilen denklemin y eksenini kesim noktasını (bağımsız değşken 0 iken Y_i'nin değeri), β_o denklemin eğimini (X'deki bir birim artışın Y'de meydana getirdiği artış), X_i bağımsız değişkeni, e_i hata terimini göstermektedir. e_i, i gözleminin gerçek değeriyle tahmin edilen denklemden elde edilen değeri arasındaki farktır($e_i=Y_i-Y_i$) (5)(9(6).

Regresyon analizinde temel olarak üç varsayım sözkonusudur:

1. Regresyon denkleminin genel ifadesinde yeralan e_i hata terimi, aritmetik ortalaması sıfır, varyansı δ^2 olan bir tesadüf değişkenidir.

2. e_i ve e_{i-i} birbiriyle ilişkili değildir. Bir başka anlatımla, birbirini izleyen verilerin hata terimleri arasında bir korelasyon yoktur.

3. e_i, aritmetik ortalaması sıfır, varyansı δ^2 olan tesadüfi bir normal dağılım değişkenidir. Başka bir deyişle, gözlemler, seri boyunca farklı dağılım göstermez. Bütün hata terimlerinin varyansı sabittir (1)(2)(7).

Tek veya çok açıklayıcılı regresyon denklemlerindeki parametrelerin tahmininde bilindiği gibi "En Küçük Kareler Yöntemi" kullanılmaktadır. Bu şekilde elde edilen regresyon denkleminin güvenilirliği çeşitli açılardan test edilebilir. Çalışmada ele alınacak konuların daha iyi açıklanabilmesi açısından, bu testlere kısaca değinilecektir.

Tahmin edilen parametrelerin tek tek istatistiki açıdan önemli (sıfırdan farklı) olup olmadığını saptamak üzere t oranları hesaplanır

$$t = \frac{\text{Tahmin Edilen Parametre}}{\text{Parametrenin Standart Hatası}}$$

ve önem testine tabi tutulur. Açıklayıcıların, bağımlı değişkeni açıklama dereceleri

$$R^2 = \frac{\text{Y'deki Varyasyonun Açıklanan Kısmı}}{\text{Y'deki Toplam Varyasyon}}$$

ile belirlenir.

Çok açıklayıcılı regresyon denklemlerinde, açıklayıcıların tümünün bağımlı değişkeni açıklama gücünü (sıfırdan farklı olup olmadığını) tesbit etmek amacıyla "Varyans Analizi" yapılır. Bu testte F-testine başvurulur. Öncelikle denkleme ait F değeri

$$F = \frac{\text{Y'deki Varyasyonun Açıklanan Kısmı}/(n-1)}{\text{Y'deki Varyasyonun Açıklanmayan Kısmı}/(n-k)}$$

n:Gözlem sayısı, k:Açıklayıcı değişken sayısı

hesaplanarak açıklayıcıların tümünün istatistiki açıdan önemli olup olmadıkları belirlenir (9)(10).

Regresyon denklemlerindeki parametrelerin tahmin edilmesinde ve testlerin yapılmasında, bilgisayarlardan büyük ölçüde yararlanılmaktadır. Bilgisayar programlarında, regresyon analiziyle ilgili sonuçlar genel olarak aynı formda verilmektedir. Ancak bu programlar tüm testleri kendisi yapıp yorumlarını sunmamaktadır. O nedenle bu formda yeralan unsurların bilinmesi ve testlerin kullanıcı tarafından yapılması gerekmektedir. Varyans Analizine ilişkin genel form ve unsurları:

Regresyon Kaynağı	Serbestlik Derecesi (DF)	Kareler Toplamı (SS) (Varyasyon)	Kareler Ortalaması (MS=SS/DF)
Regresyon (Açıklanan)			
Hata (Açıklanamayan)			
Toplam			
	(3)(8)		

Regresyon Analizinde Karşılaşılabilecek Sorunlar

Regresyon analizi sırasında ortaya çıkabilecek sorunlar, regresyonun üç temel varsayımından en az birinin sağlanamamasına dayalıdır. Bu problemler; çoklu bağlantı, farklı varyans ve otokorelasyondur (3)(4)(9).

Çoklu Bağlantı (Multicollinearity): Çoklu bağlantı, regresyon modelindeki açıklayıcı değişkenlerden ikisinin veya daha fazlasının kendi aralarında sıkı bir ilişki içinde olmasından kaynaklanmaktadır. Bu problem, regresyonun analizinin birinci varsayımının sağlanamadığının bir göstergesidir.

X1	X2	X3
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

Yandaki örnek incelenecek olursa; X2=5X1 dir. Bu nedenle X1 ve X2 arasında tam bir doğrusal ilişki vardır (r_{12}=1.0). X2 ve X3 arasında da oldukça önemli bir doğrusal ilişki bulunmaktadır (r_{23}=0.99). r_{12}=1.0 iken X1 ve X2 değişkenlerinin katsayıları, standart hataları sonsuz olacağından hesaplanamayacaktır. r_{23}=0.99 iken ise değişkenlere ait standart hatalar çok yüksek buna karşılık t oranları çok küçük bulunacaktır.

Çoklu bağlantı, açıklayıcı değişkenler arasındaki doğrusal ilişkiye dayalı bir problemdir. Diğer fonksiyonel ilişkiler çoklu bağlantı hatalarına neden olmaz. Örneğin $X2=X1^2$ veya $X3=X1^3$ ise çoklu bağlantı hatası söz konusu olmayacaktır.

Çoklu bağlantı probleminin sakıncaları:
.Çoklu bağlantı tahmin edilen parametrelerin yanlış olmasını gerektirmez fakat parametrelere ait standart hatalar çok yüksektir
.Parametrelerin t oranları çok küçüktür
.1 ve 2'nci maddelerden dolayı da, R^2 çok yüksek olsa bile parametreler istatistiki açıdan önemsiz bulunur. Diğer bir deyişle, bağımlı ve bağımsız değişkenler arasında çok kuvvetli bir ilişki bulunmasına rağmen, çoklu bağlantı hatası nedeniyle önemsiz gibi görünür.

Çoklu bağlantı probleminin belirtileri:
.Gerek R^2 ve gerekse kısmi korelasyon katsayıları çok yüksektir (0.7 ile 1.0 arasında). Fakat F-testi sonucunda parametreler istatistiki açıdan önemsiz bulunur.
.Kısmi korelasyon katsayılarının yüksek olması tek başına başına yeter şart olmamakla birlikte çoklu doğrusallığın göstergesi olabilir.

Çoklu bağlantı problemini giderme yolları:
.Örnek hacmini artırmak. Diğer bir deyişle daha fazla veri temin etmek
.Daha önceki araştırmaların ortaya koyduğu bilgilerden yararlanarak (örneğin X1 = 2X2 olduğu bilinebilir) çoklu bağlantıya neden olan açıklayıcılardan birini modelden çıkarmak
.Birbiriyle sıkı ilişkili parametrelerden birini veya birkaçını önceden tahmin ederek, bu tahmin değerlerini daha sonra orijinal ilişkideki yerine koymak
.Açıklayıcı değişkenleri transforme etmek (logaritmasını almak gibi)
.Birbiriyle önemli ölçüde ilişkili değişkenlerin birinden vazgeçmek (2)(3)(4)(9).

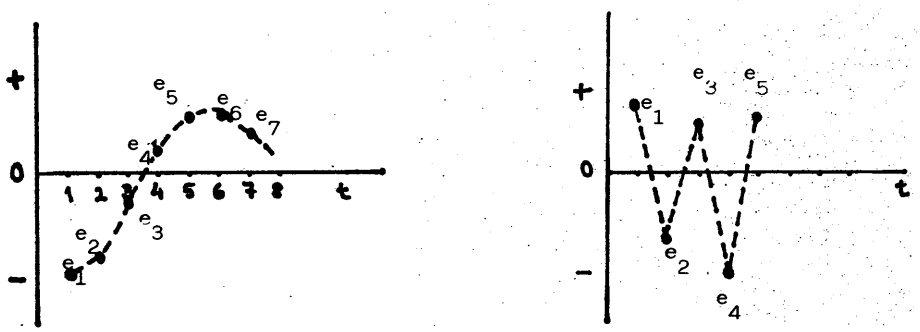
Otokorelasyon (Autocorrelation): Otokorelasyon, birbirini zamana (zaman serileri) veya yere (yatay kesit verileri) göre izleyerek sıralanmış gözlemler arasında ilişki bulunması problemidir. Bu durum regresyonun ikinci varsayımı olan $e_i = e_{i-1}$ varsayımını

bozmaktadır. Diğer bir anlatımla, hata terimleri birbiriyle ilişki halindedir.

Konu örneklerle açıklanacak olursa: Üretimin, işgücü ve sermaye ile ilişkisi 3'er aylık dönemler halinde inceleniyor olsun. Eğer 3 aylık dönemlerden birinde bir grev söz konusu ise bu, sadece o dönemin üretimini etkileyecektir. İzleyen dönemlerin üretiminin bundan etkilenmemesi beklenir. Benzer şekilde, aile tüketim harcamalarıyla aile geliri arasındaki ilişkinin incelendiği bir durumda, ailelerden birinin gelir artışı sadece o ailenin tüketim harcamaları üzerinde etkili olacaktır. Diğer bir ailenin tüketim harcamalarının bu artıştan etkilenmemesi gerekir. İşte her iki örnekteki beklenmeyen durumlar gerçekleştiğinde ya da diğer bir deyişle; grev, yapıldığı dönemi takip eden dönemlerin üretimini de etkiliyorsa veya bir ailenin gelir artışı, diğer ailenin tüketim harcamalarında da değişmeye neden oluyorsa otokorelasyonun varlığından sözedilebilir.

Otokorelasyon, birbirini izleyen hata terimlerinin işaretleri itibariyle pozitif ve negatif otokorelasyon olmak üzere ikiye ayrılır. Birbirini izleyen hata terimlerinin işaretleri aynı ise pozitif otokorelasyon, işaretler sık sık değişiyorsa negatif otokorelasyon söz konusudur. Ekonomide pozitif korelasyon daha yaygındır. Grafik 1a ve Grafik 1b sırasıyla bu iki tür otokorelasyonu göstermektedir.

Grafik 1: Otokorelasyonun Grafik Gösterimi



(a)
Pozitif Otokorelasyon

(b)
Negatif Otokorelasyon

Otokorelasyona neden olan hususlar:
.Gayri safi milli hasıla, fiyat indeksleri, istihdam ve işsizlik gibi, bir döneme ait verilerin diğer döneme ait verileri etkileme şansının yüksek olduğu değişkenlerin regresyon modeline alınması
.Araştırıcının seçtiği açıklayıcılar kendince uygun olmakla

birlikte, daha önemli olanları gözardı etmiş olması ihtimali
.Uygun fonksiyonel formun seçilmemiş olması. Örneğin eğrisel incelenmesi gereken bir ilişki doğrusal olarak ele alınması
.Cobweb teoreminin geçerli olması. Bu, özellikle tarıma dönük çalışmalarda büyük önem taşımaktadır.
.Zaman gecikmeli olarak düşünülmesi gereken değişkenlere dikkat edilmemiş olması. Örneğin tarım ürünleri ekiliş alanlarının genellikle bir önceki senenin ürün fiyatlarından etkilenmesi beklenir. O nedenle gözlemlerin bu özellik göz önüne alınarak zaman gecikmeli olarak düzenlenmiş olması gerekir.
.Nüfus gibi 5 veya 10 yılda bir kez elde edilebilen verilerde, diğer yılların bulunması için interpolasyon veya ekstrapolasyondan yararlanılmış olması

Otokorelasyon probleminin sakıncaları:
.Tahmin edilen parametreler doğru ya da doğruya yakın olmakla birlikte parametrelere ait standart hatalar çok yüksektir
.Parametrelerin t oranları çok küçüktür
.Parametrelere ait güven aralıkları, olması gerekenden farklıdır
.Yapılacak tahminler gerçekçi değildir

Otokorelasyonu saptama yolları:
1.Hata terimini (e_i, i=1,2,..,n) grafik üzerinde gösterilmesi
2.Durbin-Watson veya Von Neumann İstatistiğinden yararlanılması

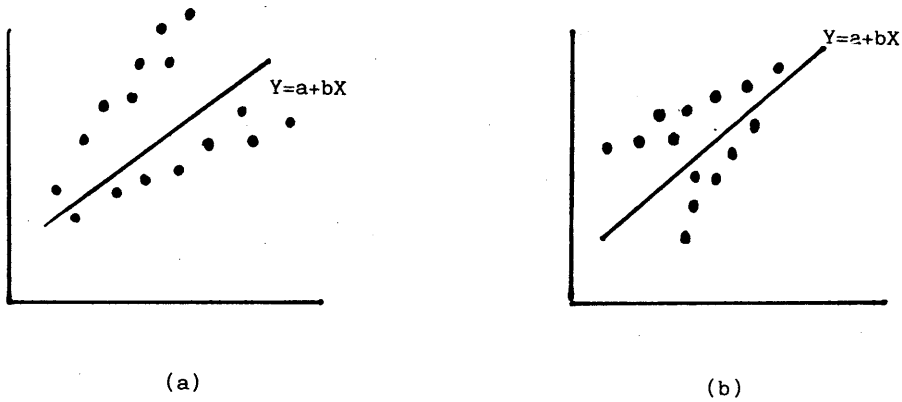
Otokorelasyon problemini giderme yolları:
1.Verilerde mevcut eğilimi saptamak üzere açıklayıcı değişken olarak zamanın modele alınması
2.Önemli olmakla birlikte gözardı edilmiş değişken veya değişkenlerin belirlenerek modele dahil edilmesi
3.Yanlış matematiksel kalıbın seçilmiş olması ihtimaline karşı regresyon denkleminin eğrisel veya başka kalıplarda yeniden belirlenmesi (2)(3)(4)(9).

Farklı varyans (Heteroscedasticity=Heteroskedasticity) : Farklı varyans, regresyon analizinin üçüncü varsayımı olan, hata terimine ait varyansın sabit olduğu koşulunun bozulduğu hallerde ortaya çıkar. Bu problemde, bağımlı değişkene ait gözlemlerin varyansı, bağımsız değişkenin hacmi arttıkça büyür veya azaldıkça küçülür. Özellikle yatay kesit verilerinde rastlanan bir problemdir. Örneğin ailelerin gelir düzeyi ve harcamaların incelendiği bir durumda, düşük gelirli ailelerin harcamalarına ilişkin hata terimi genellikle yüksek gelirli ailelerin harcamalarına ilişkin hata teriminden daha küçük bulunacaktır.

Farklı varyans probleminin sakıncaları:
.Tahmin edilen parametrelere ait standart hatalar doğru değildir
.Tahmin edilen parametrelerin güven aralıkları doğru olarak belirlenemez
.F ve t testlerinin sonuçları istatistiksel olarak gerektiğinden daha önemli bulunur.

Farklı varyans problemini saptama yolları:
1.İncelenen konunun niteliği farklı varyansin göstergesi olabilir. Eğer yatay kesit verileriyle çalışılan bir durumda heterojen birimler söz konusuysa, büyük olasılıkla farklı varyansle karşılaşılabilir
2.Grafik yöntem farklı varyansin varlığının anlaşılması açısından önemli olçüde faydalı olacaktır (Grafik 2a ve Grafik 2b).

Grafik 2: Farklı Varyansın Grafik Gösterimi



(a) (b)

Farklı varyans problemini giderme yolları:
1.Farklı varyansa neden olan açıklayıcı değişkenin logaritmasının alınması
2.'Ağırlıklı En Küçük Kareler' yönteminin kullanılması. Ağırlıklı en küçük kareler yönteminde, bağımlı ve tüm bağımsız değişkenler farklı varyansa yolaçan değişkene bölünerek, yeni regresyon denklemi transforme değişkenlerle elde edilir.
(2)(3)(4)(9)

SONUÇ

Regresyon analizi gerek akademik düzeydeki çalışmalarda gerekse öğrenci seminer ve tezlerinde sıkça başvurulan bir istatistiksel yöntemdir. Özellikle bilgisayar yardımıyla oldukça kolay bir şekilde gerçekleştirilebilir. Bununla birlikte regresyon analizinin dayandırıldığı varsayımlar sağlanamadığı takdirde elde edilen sonuçların kullanılması önemli sakıncalar doğurmaktadır.

Çoklu bağlantı, otokorelasyon ve farklı varyans, regresyon analizinin üç temel varsayımının tam olarak sağlanamamasıyla ortaya çıkan hatalardır. Bu problemlerle karşı karşıya kalınıp kalınmadığı kısa süreçlerle ortaya konabilmektedir. Regresyon analizinden elde edilen sonuçların güvenilir bilgiler halinde sunulabilmesi, bu kısa süreçlerin kullanılmasını ve eğer gerekiyorsa karşılaşılan problemin giderilmesini zorunlu kılmaktadır.

YARARLANILAN KAYNAKLAR

1) Draper, N., Smith, H., Applied Regression Analysis, Second Edition, John Wiley and Sons, USA, 1981.
2) Ertek, T., Ekonometriye Giriş, İstanbul, 1982.
3) Gujarati, D.,Basic Econometrics, Mc Graw-Hill Co., USA, 1979.
4) Judge, G.G., Hill, R.C., Griffits, W.E., Lütkepohl, H., Lee, T.C., Introduction to The Theory and Practice of Econometrics, Second Edition, John Wiley and Sons, USA, 1988.
5) Kazmier, L., Statistical Analysis for Business and Economics, Mc Graw-Hill, USA, 1967.
6) Köksal, B.A., İstatistik Analiz Metodları, İstanbul, 1976.
7) Öztürk, A., Tarım Biyoloji ve Sağlık Bilimlerinde Uygulamalı İstatistik, İzmir, 1979.
8) Ryan, T.A., Joiner, B., Ryan, B.F., Minitab Reference Manual,USA, 1982.
9) Salvatore, D., Managerial Economics, Mc Graw-Hill Co., USA,1989.
10) Zoral, K., Üretim Fonksiyonları, İzmir, 198