HACETTEPE ÜNİVERSİTESİ
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ DERGİSİ
CİLT 8/ SAYI 1 / 1990

DOĞRUSALLIK TESTLERİ(*)

Doğrusal olmayan model parametrelerini tahmin etmede, ana kütle regresyon denkleminin fonksiyonel biçimi bilinmemekte veya a priori olarak varsayılmamakta ise öne sürülen bir hipotezden çok test edilebilir olarak fonksiyonel biçimin spesifikasyonu ele alınabilir. Burada en önemli hipotez doğrusallık hipotezidir. Ana kütle regresyon denkleminin değişkenlerine göre doğrusal olduğu hipotezi alternatif hipotez karşısında test edilebilir.

Aşağıdaki bölümlerde kuvvet fonksiyonunda doğrusallık testi, regresyonun doğrusallığı testi, alt örneklemlerde doğrusallık testi, sapmalara dayalı doğrusallık testi incelenmektedir.

1. KUVVET FONKSİYONUNDA DOĞRUSALLIK TESTİ

En basit doğrusallık testi, regresyon denkleminin belirli bir dereceden kuvvet fonksiyonunu içerdiği alternatif hipotez durumundaki testtir(1). Testin dezavantajı, doğrusal modele alternatif olarak belirli bir dereceden kuvvet fonksiyonunun belirlenmesidir. Bu testin temel düşüncesi, doğrusal bir fonksiyonun bir kuvvet fonksiyonu yani birinci dereceden bir kuvvet fonksiyonu olmasına dayanır. Örneğin,

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \dots\dots\dots\dots + \beta_k X^{k-1}$$

regresyon denkleminde eğer açıklayıcı değişkenlerin yüksek dereceden kuvvetleri ile ilişkili parametrelerin tümü, β_3, β_4,...... β_k, sıfır ise verilen kuvvet fonksiyonu basit bir doğrusal regresyona dönüşür. $H_o : \beta_3 = \beta_4 = \dots = \beta_k = 0$ ve $H_1 : H_0$ doğru değil, hipotezleri kurulur. H_1, alternatif hipotezin reddi halinde doğrusal regresyon kabul edilir.

Bu düşünce, özel bir durum olarak doğrusallığı içeren diğer fonksiyonel biçimler tarafından kullanılabilir. Böyle bir spesifikasyon aşağıda verilmektedir(2).

(*) Dr. Fahamet Akın; H.Ü. İktisadi ve İdari Bilimler Fakültesi Öğretim Görevlisi

(1) KMENTA, Jan; "Elements of Econometrics", McMillan Pub. Co. Inc. ,U.S.A., 1971, s. 466-468.

(2) Bkz. G.E.P. Box- D.R. Cox; "An Analysis of Transformations", Journal of the Royal Statistical Society, Series B, Vol. 26, 1964, s. 211-243.

$$\frac{Y_i^{\lambda}-1}{\lambda}=\alpha+\beta\left(\frac{Y_i^{\lambda}-1}{\lambda}\right)+\varepsilon_i,$$

λ'nın belirli değerleri için bu fonksiyonu inceleyelim.

$\lambda=0$ için $(Y_i^{\lambda}-1)/\lambda$ ve $(X_i^{\lambda}-1)/\lambda$ tanımsız olacaktır. Hospital kuralının uygulanması ile

$$\lim_{\lambda\to 0}\frac{Y^{\lambda}-1}{\lambda}=\log Y_i$$ dir (2). Bununla beraber, sonlu pozitif bir sayı, Z, şöyle olsun;

$$Z=e^{\log Z}.$$

Burada logaritmanın tabanı e dir ve $e^{\log Z}$ şöyle genişletilebilir;

$$e^{\log Z}=1+\log Z=\frac{1}{2!}(\log Z)^2+\frac{1}{3!}(\log Z)^3+\ldots$$

Benzer olarak buradan,

$$\frac{Y_i^{\lambda}-1}{\lambda}=\frac{1}{\lambda}\left[1+\lambda\log Y_i+\frac{1}{2!}(\lambda\log Y_i)^2+\ldots-1\right]$$

$$=\log Y_i+\frac{\lambda}{2!}(\log Y_i)^2+\frac{\lambda^2}{3!}(\log Y_i)^3+\ldots \quad (1.1)$$

ifadesi yazılabilir.

$\lambda=0$ için $\frac{Y^{\lambda}-1}{\lambda}=\log Y_i$ ve

$\frac{X^{\lambda}-1}{\lambda}=\log X_i$ dir.

---

(2) $$\lim_{\lambda\to 0}\frac{Y^{\lambda}-1}{\lambda}=\lim_{\lambda\to 0}\frac{(d/d\lambda)(Y^{\lambda}-1)}{(d/d\lambda)(\lambda)}=\lim_{\lambda\to 0}(Y^{\lambda}.\log Y)=\log Y$$

Bu $\lambda = 0$ için (Pozitif X_i ve Y_i için) regresyon denkleminin

$\log Y_i = \alpha + \beta \log X_i + \varepsilon_i$ denklemine dönüştüğü anlamına gelir.

$\lambda = 1$ için

$(Y_i - 1) = \alpha + \beta (X_i - 1) + \varepsilon_i$ şeklinde basit bir doğrusal regresyon modeli oluşur.

Model bir başka biçimde şöyle de yazılabilir,

$Y_i = \alpha^* + \beta X_i + \varepsilon_i$, buradan $\alpha^* = \alpha - \beta + 1$ dir.

λ'nın farklı değerleri, regresyon denkleminin farklı fonksiyonel spesifikasyonlarını verir. Bu, bizi doğrusal hipoteze karşılık regresyon denkleminin doğrusal olmayan bir fonksiyon olduğunu test etmeye götürür. Biçimsel olarak H_0 ve H_1 hipotezi şöyledir,

$H_0 : \lambda = 1$

$H_1 : \lambda \neq 1$ dir.

Testi uygulamak için λ'nın tahminine ve onun standart hatasına gerek vardır. Daha açık olarak, λ, maksimum benzerlik yöntemi tarafından (1.1) nolu ifadenin diğer parametreleri ile tahmin edilebilir. $Y_1, Y_2, ..., Y_n$ için benzerlik fonksiyonunu düzenlemede normal olduğu varsayılan ε'nin dağılımından Y'nin dağılımını çıkarmak gerekir. Değişken değişimi (=change of variable) teoreminden,

$$f(Y_i) = \left|\frac{d\varepsilon_i}{dY_i}\right| f(\varepsilon_i) \text{ dir.}$$

$$\text{Ancak, } \varepsilon_i = \left(\frac{Y_i^{\lambda} - 1}{\lambda}\right) - \alpha - \beta\left(\frac{X_i^{\lambda} - 1}{\lambda}\right)$$

$$\frac{d\varepsilon_i}{dY_i} = Y_i^{\lambda - 1}$$

$Y_1, Y_2, \ldots, Y_n$ için benzerlik fonksiyonu,

$$L = (\lambda - 1).\Sigma \log Y_i - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}.$$

$$\sum\left[\left(\frac{Y_i^{\lambda} - 1}{\lambda}\right) - \alpha - \beta\left(\frac{X_i^{\lambda} - 1}{\lambda}\right)\right]^2 \text{ olur.}$$

λ, α, β ve σ^2'nin maksimize edilen değerleri bilgisayar yardımı ile bulunabilir ve her birinin standart hataları uygun bilgi matrisine başvurma ile tahmin edilebilir. Büyük örneklemlerde maksimum benzerlik tahminleri normal veya en azından yaklaşık olarak normal dağılmış olacaktır.

2. REGRESYONUN DOĞRUSALLIĞI TESTİ

Bu testte ana kütle korelasyon katsayısı (ρ^2) ve korelasyon oranı (η^2) aracılığı ile regresyonun doğrusal olup olmadığı incelenmektedir[1]. Regresyonun doğrusallığını analiz etmede y üzerine x'in tam doğrusal regresyon durumu ele alınsın. y'ye göre x'in beklenen değerinin varyansı;

$$\mathrm{Var}\{E(x\,|y)\} = E_y\{(E(x\,|y) - E(x))^2\}$$

$$= \beta_1^2\,\sigma_2^2 = \mu_{11}^2/\sigma_2^2$$

y üzerine x'in korelasyon oranı ise,

$$\eta_1^2 = \mathrm{Var}\{E(x\,|y)\}/\sigma_1^2 = \mu_{11}^2/\sigma_1^2 . \sigma_2^2 \text{ dir (2).}$$

Korelasyon katsayısı ρ'ya benzemeyen η_1^2, x ve y'de simetrik değildir, yani

$$\eta_1^2 \neq \eta_2^2 = \mathrm{Var}\{E(y\,(x)\}/\sigma_2^2.$$

Eğer regresyon tam olarak doğrusal ise $\eta_1^2 = \rho^2$ dir.

---

(1) KENDAL, G. Maurice - STUART, Alan; "The Advanced Theory of Statistics", Volume 2, Charles Griffin Comp., London, 1967, s.296-298.

(2) Burada

$$\beta_1 = \frac{n\Sigma xy - \Sigma x.\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\mu 11}{\sigma_1^2}$$

$$\beta_2 = \frac{n\Sigma xy - \Sigma x.\Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{\mu 11}{\sigma_2^2}$$

$\rho = \mu_{11}/\sigma_1.\sigma_2$ ve $\rho^2 = \beta_1.\beta_2$ dir.

Şimdi regresyonun tam olarak doğrusal olmadığı durum ele alınsın.

$$\sigma_1^2 = E[\{x - E(x)\}]^2 = E\left[\left(\left\{x - E(x \mid y)\right\} + \left\{E(x \mid y) - E(x)\right\}\right)^2\right]$$
$$= E\left[\left\{x - E(x \mid y)\right\}^2\right] + E\left[\left\{E(x \mid y) - E(x)\right\}^2\right]$$
$$= E\left[\left\{x - E(x \mid y)\right\}^2\right] + \mathrm{Var}\left\{E(x \mid y)\right\}$$

Burada,

$$2E\left[\left\{x - E(x \mid y)\right\}\left\{E(x \mid y) - E(x)\right\}\right]$$
$$= \underset{y}{E}\left[\left\{E(x \mid y) - E(x)\right\} . \underset{x/y}{E}\left\{x - E(x \mid y)\right\}\right] = 0$$

ifadesinden $0 \leq \eta_1^2 \leq 1$ dir.

Eğer $E\left[\left\{x - E(x \mid y)\right\}^2\right] = 0$ ise $\eta_1^2 = 1$ dir.

Korelasyon katsayısının korelasyon oranına oranı,

$$\frac{\rho^2}{\eta_1^2} = \frac{\mu_{11}^2}{\sigma_2^2 \, \mathrm{Var}\left\{E(x \mid y)\right\}}$$

Cauchy - Schwarz eşitsizliği aracılığıyla,

$$\mu_{11}^2 = \left(\underset{x/y}{E}\left[\{y - E(y)\}\{x - E(x)\}\right]^2\right)$$
$$= \left(\underset{y}{E}\left[\{y - E(y)\}\left\{E(x \mid y) - E(x)\right\}\right]\right)^2$$
$$\leq \underset{y}{E}\left[\{y - E(y)\}^2\right] \underset{y}{E}\left[\left\{E(x \mid y) - E(x)\right\}^2\right] = \sigma_2^2 \, \mathrm{Var}\left\{E(x \mid y)\right\}.$$

$E(x \mid y)$, y'nin doğrusal bir fonksiyonu ise, $\{y - E(y)\}$, $\{E(x \mid y) - E(x)\}$ ifadesine orantılıdır.

Sadece y üzerine x regresyonu doğrusal olduğu zaman

$\dfrac{\rho^2}{\eta_1^2} \le 1$ dir. Buradan, $0 \le \rho^2 \le \eta_1^2 \le 1$ dir.

Bu açıklamalar çerçevesinde aşağıdaki eşitsizlikler yazılabilir;

1) $\rho^2 = 0$ x ve y bağımsız ise,

2) $\rho^2 = \eta_1^2 = 1$ x ve y kesin doğrusal fonksiyonel ilişki içinde ise,

3) $\rho^2 \le \eta_1^2 = 1$ x ve y kesin doğrusal olmayan fonksiyonel ilişki içinde ise,

4) $\rho^2 = \eta_1^2 < 1$ y'nin x üzerine regresyonu doğrusal ancak fonksiyonel ilişki yok,

5) $\rho^2 < \eta_1^2 < 1$ fonksiyonel ilişki yok, ancak doğrusal olmayan regresyon eğrisi düz doğrudan daha iyi uyum sağlar.

Benzer olarak x üzerine y'nin regresyonu için korelasyon oranı şöyledir:

$\eta_1^2 = \mathrm{Var}\{E(y \mid x)\} / \sigma_2^2$ ve

$0 \le \rho^2 \le \eta_2^2 \le 1$ dir.

Fonksiyonel bir ilişki var ise, yani $\eta_1^2 = 1$ ise aynı zamanda $\eta_2^2 = 1$ dir. Genelde korelasyon oranı korelasyon katsayısını aşar. x üzerine y'nin regresyonu doğrusal değil iken y üzerine x'in regresyonu doğrusal ise,

$\eta_1^2 = \rho^2 < \eta_2^2$ dir.

η_1^2, y'nin x üzerine bağımlılığını ölçmez. $\eta_1^2 - \rho^2$ nin pozitif değeri, regresyonun doğrusalsızlığının bir göstergesidir. $\eta_1^2 - \rho^2$ farkı küçük olduğunda regresyonun doğrusal olduğu, buna karşılık bu fark çok büyük olduğunda doğrusal bir ilişkinin bulunmadığı ve doğrusal korelasyon katsayısı ρ'nun bir ilişkiyi ifade edemeyeceği şeklinde kabaca yorum yapılır.

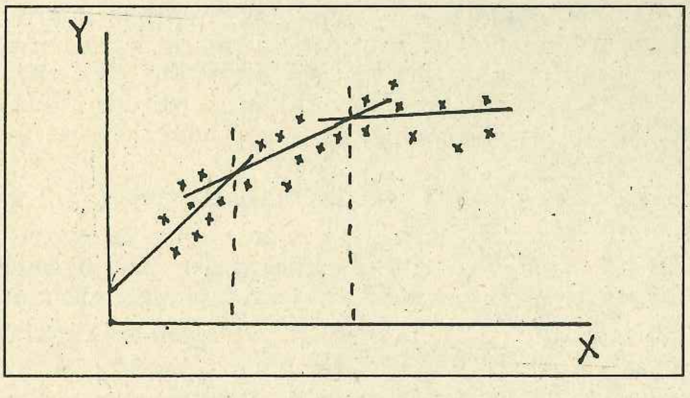
Doğrusal regresyon modellerine ait istatistiksel test yöntemleri, doğrusal olmayan modellere ait parametre tahmininde doğrudan doğruya kullanılmazlar. Örneğin F istatistiği modelin bütününün uygunluğunu belirlemek için kullanılamaz. F testi, açıklayıcı değişkenlerin bağımlı değişken üzerinde gerçekten herhangi bir anlamlı etkiye sahip olup olmadığını saptamayı amaçlamaktadır[1].

F testi için kullanılan korelasyon katsayısı doğrusal regresyon denklemleri içindir. Bu yüzden doğrusal olmayan regresyon modellerinde kullanılamazlar. t testi içinde aynı şey söylenebilir. Buna karşılık korelasyon indeksi veya çoklu determinasyon indeksi kullanılabilir.

3. ALT ÖRNEKLEMLERDE DOĞRUSALLIK TESTİ

Ana kütle regresyon denkleminin değişkenlerine göre doğrusal olduğu hipotezi, alternatif fonksiyonel biçim veya biçimleri belirlemeksizin test edilebilir. Doğrusallığın oldukça belirgin bir başka göstergesi, regresyon denkleminin eğim ve sabit kesmesinin tüm açıklayıcı değişken değerleri üzerinde sabit kalması gerektiğidir[2].

Yapılacak şey, farklı ve birbiri üzerine gelmeyen açıklayıcı değişken değerleri aralığına her bir alt örneklemin karşılık geldiği bir kaç alt örneklem içine gözlem değerlerini bölmektir. Bir alt örneklemden diğerine herhangi bir anlamlı farklılığın olup olmadığı test edilebilir ve her bir alt örneklem için sabit kesme ve eğim tahmin edilebilir. Bu durum aşağıdaki şekilde gösterilmektedir.



Şekil 1

(1) Bkz. KOUTSOYIANNIS, A.; "Theory of Econometrics", Harper Row Pub., U.K., 1978.

(2) KMENTA, Jan; a.g.e., s.469.

Bu testin nasıl uygulandığını açıklayabilmek için $X_1 \leq X_2 \leq X_3 \leq .... \leq X_n$ şeklinde düzenlenmiş olan n gözlemli bir örneklem ele alınsın. Örneklemin aşağıda gösterildiği gibi üç alt örnekleme bölündüğü varsayılsın.

Alt örnneklem 1 : $X_1, X_2, ..., X_k$

Alt örneklem 2 : $X_{k+1}, X_{k+2}, ..., X_m$

Alt örneklem 3 : $X_{m+1}, X_{m+2}, ..., X_n$

Regresyon modeli şöyle düzenlenebilir;

$$Y_i = \beta_1 + \beta_2 X_i + \gamma_i Z_{i1} + \gamma_2 X_i Z_{i1} + \gamma_3 Z_{i2} + \gamma_4 X_i Z_{i2} + \varepsilon_i.$$

Burada $Z_{i1} = 1$ eğer i alt örneklem 1 e ait ise,

$Z_{i1} = 0$ eğer i alt örneklem 1'e ait değil ise,

$Z_{i2} = 1$ eğer i alt örneklem 2'ye ait ise,

$Z_{i2} = 0$ eğer i alt örneklem 2'ye ait değil ise,

Doğrusallık hipotezi, H_0: $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0$

H_1 : H_0 doğru değil, hipotezleri ile test edilebilir. Uygun F testi, örneklemin bölündüğü teste bağlı olabildiğinden dezavantaja sahiptir. Eğer alt örneklemlerin sayısı çok küçükse doğrusallıktan uzaklaşma gizli kalabilir, eğer alt örneklemlerin sayısı büyükse bir çok serbestlik derecesi kaybolur ve böylece testin gücü zayıflar. Uygulamalarda üç veya dört örneklemin yeterli olduğu öne sürülmektedir.

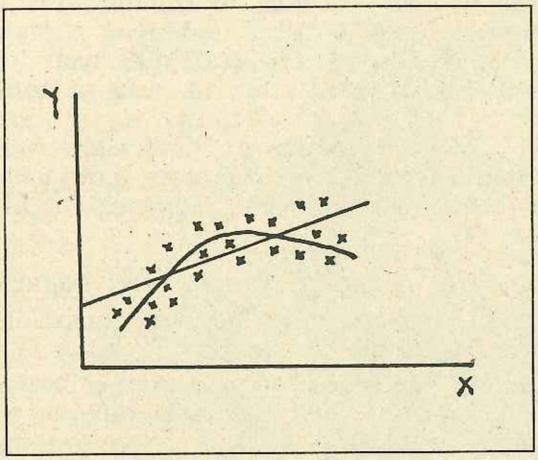
4. SAPMALARA DAYALI DOĞRUSALLIK TESTİ

Bu doğrusallık testi, örneklem regresyon doğrusu etrafındaki sapmaların (residual, e) saçılımına dayanır[1]. Bu düşünce, klasik normal doğrusal regresyon modelinin varsayımları altında, bozukluk teriminin (disturbance, ε_i ana kütle regresyon doğrusundan sapma) ana kütle regresyon doğrusu etrafında tesadüfi olarak dağılmış olmasına güvenmektedir[2]. Eğer ana kütle regresyonu doğrusal değil ise düz bir doğru etrafında ε_i lerin saçılımı pek fazla tesadüfi olmayacaktır. Aşağıdaki Şekil 2'de ilişkinin doğrusalsızlığından düz

(1) KMENTA, Jan; a.g.e., s.469-472.

(2) Bkz. KOUTSIANNIS, a.g.e., s.55.

doğrudan sapmaların önce negatif sonra pozitif ve tekrar negatif eğilimli olduğu görülebilir.



Şekil 2

Bu durumda, regresyon doğrusundan sapmaların sırasının tesadüfi olarak düzenlenip düzenlenmediğini belirleme ile doğrusallığın test edilebileceğini öne sürmektedir.

Ana kütle regresyon doğrusundan sapmaların gözlenememesinden dolayı bu tür testi uygulamada güçlük doğar. Bilinen tek şey, örneklem regresyon doğrusundan sapmalardır. Ortaya çıkan güçlük, ε_i lerin olmasa bile e_i lerin bağımsız olmadığıdır.

$$\mathrm{Cov}(e_i, e_j) = E(y'_i - \hat{\beta}x'_i)(y'_i - \hat{\beta}x'_j)$$

$$= E(\beta x'_i + \varepsilon_i - \hat{\beta}x'_i)(\beta x'_j + \varepsilon_j - \hat{\beta}x'_j)$$

$$= x'_i - x'_j E(\hat{\beta} - \beta)^2 - E(\hat{\beta} - \beta)\varepsilon_i x'_j - E(\hat{\beta} - \beta)\varepsilon_j x'_i + E(\varepsilon_i \varepsilon_j)$$

$$= -x'_i x'_j \left[\frac{\sigma^2}{\Sigma x_i'^2}\right] - \frac{\sigma^2}{n}, \; i \neq j$$

ifadesi sıfır değildir. e'ler ve ε'ler arasındaki bu fark, Durbin Watson testi tarafından dikkate alınmıştır[1].

(1) Bkz. WYNN R.F.- HOLDEN, K.; "An Introduction to Applied Econometric Analysis", McMillian Press Ltd., London, 1974, s.9.

Durbin Watson testi, ana kütle sapmalarının birinci dereceden otoregresif bir yapı izlediği hipotezine karşılık korelasyonlu olmadığı hipotezini test etme için düzenlenmiştir[1]. Eğer örneklem sapmaları zamandan çok artan açıklayıcı değişken değerlerine karşılık düzenlenirse, bu test ana kütle regresyon doğrusundan sapmaların (deviations) tesadüfi olup olmadığını kontrol etmek için kullanılabilir. Eğer zamana karşı sapmaların (residuals) sıralaması, açıklayıcı değişkenin artan değerlerine karşı sapmaların sıralaması ile benzer ise, zamana karşı ana kütle sapmalarının otoregresyonsuz hali yani doğrusallık için aynı test kullanılabilir.

Durbin Watson testinin yapısı, ε_i'ler bağımsız da olsa, en küçük kareler sapmalarının (residuals) bağımsız olmaması ile motive edilmektedir. Bununla beraber, Cov (e_i, e_j) ifadesi örneklem hacmi sonsuza doğru giderken sıfıra yaklaşma eğilimindedir, ki büyük örneklemlerde sapmalar (e_i) bağımsız olmaya yakın olacaktır. Böylece, büyük örneklemlerde çok büyük bir hata riski yapmaksızın genel bir "tesadüfilik testi" kullanılabilir. Böyle bir test, tam bir değişmenin olup olmadığını görmek için bir serinin tepe ve çukurlarını incelemeyi kapsar. Bu test, açıklayıcı değişkenin artan değerlerine göre düzenlenen sapmaların sırasına uygulanabilir.

Eğer,

$e_{i-1} < e_i > e_{i+1}$ ise e_i sapması bir tepe

$e_{i-1} > e_i < e_{i+1}$ ise e_i sapması bir çukur

olarak açıklanır. Eğer iki veya daha fazla ardışık sapmalar aynı değere sahipse ve civarındaki değerleri aşarsa onları bir tepe veya benzer olarak çukur kabul etmek mümkündür. Hem bir tepe hem de bir çukur için genel bir ad, bir dönüm noktasıdır. bağımsız değerlerin bir n serisinde dönüm noktalarının tolam sayısı, diyelim ki p, büyük bir n için yaklaşık olarak normal dağılmakta, ortalaması $E(p) = 2(n-2)/3$ ve variansı $Var(p) = 16n - 29/90$ dır[2].

Doğrusallık tahmini ile ilgili bağımsızlık testi, X'in artan değerlerine karşı düzenlenmiş sapmaların dönüm noktalarının sayısını sayma ve bu sayının anlamlı olarak E(p) den farklı olup olmadığını kontrol etmeyi saptar. Bağımsızlık hipotezi şöyledir:

$H_o : E(p) = 2(n-2)/3$

$H_1 : E(p) \neq 2(n-2)/3$

---

(1) $u_t = \rho u_{t-1} + \varepsilon_t,$

$H_o: \rho = 0$

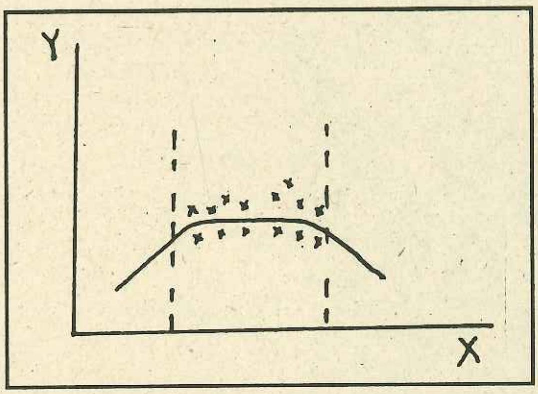
$H_1: \rho \neq 0$

(2) YULE, G. Udny - KENDALL, M.G.; "An Introduction to the Theory of Statistics", Charles Griffin, London, 1950, s. 638.

Büyük örneklemler için uygun test ıstatistiği;

$$\frac{p - 2(n-2)/3}{\sqrt{(16n-29)/90}} \sim N(0,1) \text{ dir.}$$

Bu test, stokastik terimin (ε_i) normal olarak dağıldığı varsayımını gerektirmez. Testin dezavantajı, Durbin Watson testinde olduğu gibi X'in artan değerlerine ve zamana karşılık sapmaların sıralaması benzer ise zaman üzerine ana kütleden sapmaların otoregresyonu ve doğrusalsızlığı arasında ayrım yapılamamasıdır. Bir başka dezavantaj, gözlenen açıklayıcı değişken değerlerinin kapsandığı aralık dışında regresyon doğrusunun biçimi hakkında örneklem verilerinden herhangi bir bilginin sağlanmamasıdır. Eğer ana kütle regresyonu bu aralık içinde yaklaşık olarak doğrusal, aralık dışında doğrusal değil ise, o zaman hiç bir test ana kütle regresyonunun doğrusalsızlığını çıkaramaz. Bu durum aşağıdaki Şekil 3'de verilmektedir.



Şekil 3

Bu örneklem verileri ile yapılacak bir şey yoktur. Bu güçlükten sakınmanın tek mümkün yolu, örneklemi artırmak ve gözlenen açıklayıcı değişken değerlerinin aralığını genişletmektir. Gerçekten, eğer ana kütle regresyonunun doğrusallığı a priori olarak şüpheli ise örneklemi arttırmak ve aralığı genişletmek, öne sürülmüş iyi bir yöntem olurdu. Geçmiş yıllarda uygulamalı ekonometri çalışmalarının çoğu doğrusal modellerle yapılmaktaydı. Bununla beraber, elektronik bilgisayarların yayılması ile bu düşünce güç kaybetmiştir. Teorik olarak ileri sürüldüğünde doğrusal olmayan modelleri kullanmama veya doğrusallığı test etmeme konusunda az bir bahane vardır.

KAYNAKLAR

1. BOX G.E.P- COX D.R.; "An Analysis of Transformations", Journal of the Royal Statistical Society, Series B, Vol. 26, 1964
2. KENDAL, G. Maurice - STUART, Alan; "The Advanced Theory of Statistics", Volume 2, Charles Griffin Comp, London, 1967.
3. KMENTA, Jan; "Elements of Econometrics", McMillian Pub. Co. Inc., U.S.A., 1971.
4. KOUTSOYIANNIS, A., "Theory of Econometrics", Harper Row Pub., U.K., 1978.
5. WYNN R.F.-HOLDEN,K.; "An Introduction to Applied Econometric Analysis", McMillian Press Ltd., London, 1974.
6. YULE, G.Udny-KENDALL, M.G.,; "An Introduction to the Theory of Statistics"', Charles Griffin, London, 1950.

HACETTEPE ÜNİVERSİTESİ
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ DERGİSİ
CİLT 8/ SAYI 1 / 1990

ÇAPRAZ TABLOLARIN ÇÖZÜMLENMESİ VE LOG-LINEAR MODELLER

Doç.Dr. Hamza UYGUN

1- GİRİŞ

Bu çalışmada 1960'lardan sonra, sosyal bilimlerde kullanımı hızlı bir artış gösteren log-linear modeller üzerinde durulacaktır*. Bu model daha önceki çalışmamda (Uygun, 1989), tanıtılmıştı. Bu yazıda ise bu modellerin a) ikiden fazla seçeneği olan değişkenler arasındaki ilişkilerin çözümlenmesinde ve b) ikiden fazla boyutlu tabloların değerlendirilmesinde, kullanımı ele alınacaktır.

İstatistiksel çözümleme, toplanan verilerin oluşturduğu dağlımın seçilen bir modele göre oluşan kuramsal dağılım ile kıyaslanmasına dayalıdır. Bu iki dağılım, kuramsal dağılım ve toplanan verilern tablo halindeki dökümünden oluşan dağılım arasındaki fark değerlendirilir. Bu fark anlamlı ise modelimizin eldeki verilere uymadığından sözedilir. Başka bir anlatımla model olarak ortaya çıkan hipotez reddedilir.

Goodman (1981:; 193-194), nitel değişkenlerden oluşan çapraz tabloların çözümlenmesinde log-linear modellerin kullanımının üç amaca hizmet edeceğinden sözeder. Bunlar; 1) Değişkenlerin oluşturduğu bileşik dağılımı (joint distribution) test etmek, 2) değişkenlerin birbirlerine bağımlı olup olmadıklarını test etmek ve 3) iki değişken arasındaki ilişkiyi, neden-sonuç ilişkisine dayandırmaksızın test etmektir. Bu üç amaç ikiden fazla boyuttaki tabloların değerlendirilmelerinde de geçerlidir. Yani ikiden fazla değişken arasındaki ilişkiye de bu üç amaç doğrultusunda bakılabilir.

Nitel değişkenler arasındaki ilişkiyi test etmeye yönelik log-Linear modellerin kullanımının giderek yaygınlık kazanmasında, bu tekniğin aynı tablo üzerinde Ki-Kare'ye oranla daha değişik hipotezlerin test edilmesine olanak sağlaması başta gelmektedir. Örneğin Goodman (1981; 231) üç boyutlu bir tabloda 18 değişik hipotezin sınanabileceğini belirtmektedir.

İki boyutlu ve her değişkenin iki seçenekten oluştuğu tablolarda log-linear modellerin parametrelerini tahminde basit işlemler yeterli olmaktadır. Değişkenlerin seçenek sayısının yada tablonun boyutunun artması ise parametre tahmin işlemlerinin karmaşıklaşmasına yol açmaktadır. Bu nedenle önce değişkenlerin seçeneklerinin ikiden fazla olduğu iki boyutlu tablolarda log-linear modellerin kullanımı, daha sonra ise ikiden fazla boyutlu tablolar-

(*) Bu konuda daha geniş bilgi için (Fleiss, 1973), (Fienberg, 1977) ve (Andersen, 1980) bakılabilir.

da log-linear modellerin kullanımı üzerinde durulacaktır. Bu modellerin kullanımında önemli olan temel kavram ve işlemler daha önceki makalemde (Uygun, 1989) ele alındığı için burada tekrarlanmayacaktır.

2-Değişkenlerin Seçeneklerinin ikiden fazla olduğu 2 Boyutlu Tabloların Çözümlenmesi

İki ya da daha fazla değişken ilişkisi log-linear modelden yararlanılarak çözümlenmeye çalışıldığında, odds-oranının bulunması gerekmektedir. Odds-oranının hesaplanması ise beklenen değerler üzerinde olmaktadır. Beklenen değerlerin oluşturulması modelin seçimini gerektirmektedir. Belirlenen modele göre ise beklenen değerlerin oluşturduğu kuramsal dağılım bulunur.

Tabloda beklenen değerlerin hesaplanması için Ençok Olabilirlik Tahmininden (Maximum Likelihood Estimation) yararlanılmaktadır. Burada olabilirlik fonksiyonunu en büyük değere ulaştıran değerler parametre tahminleri olarak alınmaktadır*.

Üzerinde çözümlemenin yapıldığı çapraz tabloların boyutları büyüdükce tablonun beklenen değerlerinin hesaplanması da zorlaşmaktadır. Bu hesaplamalar ancak iterasyon yoluyla yapılabilmektedir. İterasyon işlemi ise bilgisayar kullanımı ile kolay hale gelmektedir. Bu nedenle bilgisayar kullanımı bu işlemlerin yapılmasını olanaklı kılmıştır. Iterasyon yoluyla beklenen değerlerin bulunmasında farklı iki algoritma geliştirilmiştir: Deming-Stephan algoritması** ve Newton-Rapson algoritması (Knoke, Burke, 1980;22). Fay ve Goodman'ın ECTA, Bock'ın MULTIQUAL, Baker ve Nelder'in GLIM, Wilkinson'un SYSTAT*** bu algoritmaları kullanan bilgisayar programlarıdır.

Bu anlatılanları bir tablo üzerinde görmeye çalışalım. Tablonun X ve Y değişkenlerinden oluştuğunu ve değişkenlerin herbirinin üç seçeneği olduğunu kabul edelim. Bu durumda gözlenen frekansların (n_{ij}) oluşturduğu tablo (Tablo 1) aşağıdaki gibi olacaktır.

Tablo 1

		Değişken Y			
		Y_1	Y_2	Y_3	Toplam.
	X_1	n_{11}	n_{12}	n_{12}	$n_{1.}$
Değişken X	X_2	n_{21}	n_{22}	n_{23}	$n_{2.}$
	X_3	n_{31}	n_{32}	n_{33}	$n_{3.}$
	Toplam	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

(*) Bu konuda daha geniş bilgi elde edinmek için Mood, Graybil, Boes (1974), Hanushek, Jackson (1977) ve Kendall, Stuart (1973) 'a bakılabilir.

(**) Bu algoritma için Goodman (1979) a bakınız.

(***)Bu programda kullanılan Algoritma için Haberban (1973) e bakınız.

Burada X_i X değişkeninin seçeneklerini, Y_i Y değişkeninin seçeneklerini, n_{ij} i inci satır ve j inci sütundaki gözün frekansını, $n_{i.}$ i inci satırın toplam frekansını, $n_{.j}$ j inci sütunun toplam frakansını, n.. ise genel toplamı göstermektedir.

Beklenen frekansları bulmak için ise öncelikle model seçimi yapılır. Seçilen modele göre de tablonun her gözüne düşen beklenen değerler, Ençok Olabilirlik Tahminiyle bulunur.

Beklenen değerlerin (N_{ij}) oluşturduğu kuramsal dağılım ise Tablo 2 gibi olacaktır.

Tablo 2

		Değişken Y			
		Y_1	Y_2	Y_3	Toplam.
	X_1	N_{11}	N_{12}	N_{12}	$N_{1.}$
Değişken X	X_2	N_{21}	N_{22}	N_{23}	$N_{2.}$
	X_3	N_{31}	N_{32}	N_{33}	$N_{3.}$
	Toplam	$N_{.1}$	$N_{.2}$	$N_{.3}$	$N_{..}$

Burada,

X_i X değişkeninin i inci seçeneğini,

Y_i Y değikeninin i inci seçeneğini

N_{ij} i inci satır ve j. sütundaki gözün beklenen frekansını,

$N_{i.}$ i inci satırdaki beklenen frekanslar toplamını,

$N_{.j}$ j inci sotundaki beklenen frekanslar toplamını,

$N_{..}$ genel toplamı göstermektedir.

Beklenen değerlerin bulunmasından sonra gözlenen frekansların oluşturduğu dağılımla, beklenen frekansların oluşturduğu kuramsal dağılım arasındaki benzerlik test edilir. Dağılımlar benzerlik gösteriyorsa, seçilen model dağılıma uygundur. Benzerlik göstermiyorsa seçilen modelin dağılımı temsil etmediği sonucuna ulaşılır.

Dağılımların benzerlik gösterip göstermediğine karar vermek için Pearson Ki-Kare istatistiğinden yada Olabilirlik Oranı İstatistiğinden yararlanılır. Olabilirlik Oranı İstatistiği,

$$L^2 = 2\,\Sigma\, n_{ij}\ \ln\,(n_{ij}/N_{ij})$$

formülü ile gösterilir.

Formülden de anlaşılacağı gibi aynı satır ve sütundaki göze düşen beklenen ve gözlenen frekanslar birbirlerine yakın değerler içerdikleri oranda n_{ij}/N_{ij} değeri bire yaklaşırken $\ln(n_{ij}/N_{ij})$ ve L^2 değeri sıfıra yaklaşacaktır. Başka bir anlatımla L^2 değeri sıfıra yaklaştıkca seçilen modelin eldeki dağılımı temsil etme olasılığı artacaktır. L^2 değerinin yorumlanmasında seçilen modele göre bağımsız olarak hesaplanabilen göz sayısının (serbestlik derecesi) dikkate alınması gerekmektedir.

Yukarıda anlatılanları baba ve oğulun mesleklerinin statülerine göre dağılımını gösteren bir tablo üzerinde (Tablo 3) daha somut bir şekilde görelim*. Tablo İngilterede dikey hareketliliği ölçmeye yöneliktir. Meslekler statülerine göre üç gruba ayrılmıştır: Üst Orta ve Alt.

Tablo 3

		Oğulun Meslek Statüsü			
		Üst	Orta	Alt	Toplam
Babanın Meslek Statüsü	Üst	588	395	159	1142
	Orta	349	714	447	1510
	Alt	114	320	411	845
	Toplam	1051	1429	1017	3497

Baba ile oğlunu meslek statülerinin birbirlerinden bağımsız olduğu, hipotez olarak ortaya atılmışsa modelimiz

$$N_{ij} = G\ d_i^B\ d_j^0$$

şeklinde olacaktır.

Burada B babanın meslek statüsünü, 0 ise oğulun meslek statüsünü göstermektedir.

Bu modele göre oluşturulacak tabloda bilinenler ve bilinmeyenler Tablo 4'de görülmektedir. Buradaki bilinmeyen değerler Ençok Olabilirlik Tahmini ile bulunmaktadır. Tahmin işlemi ise iterasyon yoluyla yapılmaktadır**.

(*) Bu tablo üzerinde daha önce Glass (1954), White (1963), Goodman (1965), (1969) da çalışmıştır.

(**) Bu amaç için geliştirilmiş algoritmalar için Haberman (1978) ve Goodman (1979)'a bakınız.

Tablo 4:

		Oğulun Meslek Statüsü			
		Üst	Orta	Alt	Toplam
Babanın Meslek Statüsü	Üst	N_{11}	N_{12}	N_{13}	1142
	Orta	N_{21}	N_{22}	N_{23}	1510
	Alt	N_{21}	N_{22}	N_{23}	845
Toplam		1051	1429	1017	3497

Haberman (1973)'ın algoritması ile N_{ij} değerlerini bulan SYSTAT istatistik paket programı ile elde kuramsal dağılım (Tablo 5) ise aşağıdadır.

Tablo 5: Babanın Meslek Statüsü ile Oğulun Meslek Statüsü Arasında Bir İlişkinin Olmaması Durumundaki Dağılım

		Oğulun Meslek Statüsü			
		Üst	Orta	Alt	Toplam
Babanın Meslek Statüsü	Üst	343.22	466.66	332.12	1142
	Orta	453.82	617.04	439.14	1510
	Alt	253.96	345.30	245.74	845
	Toplam	1051	1429	1017	3497

$L^2 = 499.56$ $sd = 4$ $p = .000$

Beklenen frekansların oluşturduğu tablonun (Tablo 5) değerleri ile gözlenen frakansların oluşturduğu tablonun (Tablo 3) değerleri arasındaki farkın büyük olması, seçilen modelin elimizdeki dağılıma uymadığını göstermektedir. Bu nedenle modelimizin temsil ettiği, iki değişkenin bağımsız olduğunu iddia eden hipotezimiz reddedilecektir. İstatistiksel olarak L^2 değerinin serbestlik derecesine göre büyüklüğü bu kararı vermemizde temel oluşturmaktadır.

3. Üç ve Daha Büyük Boyutlu Tabloların Çözümlenmesi

Log-Linear modellerin tabloları yorumlamaktaki esnekliği özellikle 3 ve daha büyük boyutlu tablolarda belirgin olarak görülür. Örneğin 3 değişkenden oluşan bir tablo, değişkenlerin birli, ikili değişik kombinasyonları ile ifade edilebilen farklı modeller altında değerlendirilebilir. Bu değerlendirme biçimi ise eldeki dağılıma en çok uyan modelin seçilmesini kolaylaştırır.

Üç ve daha büyük boyutlu tablolarda modelleri açık formülleriyle yazmak zorlaşacaktır. Çünkü n boyutlu doymuş bir modelde yazılması gereken parametre sayısı 2^n dir. Örneklendirmek için A ve B değişkenlerinden oluşan 2 boyutlu (n=2) bir tablo ile A, B ve C değişkenlerinden oluşan 3 boyutlu (n=3) bir tablo düşünelim.

2 boyutlu tablo için oluşturulan doymuş bir modelde 2^2 yani 4 adet parametre bulunacaktır. Bu nedenle formül.,

$$N_{ij} = G\, d_i^A\, d_j^B\, d_{ij}^{AB}$$

şeklinde olacaktır. Oysa 3 boyutlu bir tabloda parametre sayısı 2 boyutlu modeldeki parametre sayısının iki katına, sekize, (2^3) çıkacaktır. Bu nedenle formül,

$$N_{ij} = Gd_i^A\ d_j^B\ d_k^C\ d_{ij}^{AB}\ d_{jk}^{BC}\ Ac_{d/k}\ d_{ijk}^{ABC}$$

şeklinde olacaktır.

Tablonun boyutu artıkca modelin daha karmaşık bir şekil alması nedeniyle daha basit gösterimi sağlayan semboller kullanmak gerekmektedir. Modellerin daha basit gösterimi ise iki ayrı şekilde olmaktadır. Birincisi parantez kullanmadır. Burada birbiriyle ilişkili olan değişkenler tek bir parantez içerisinde gösterilmektedir. Bağımsız olan değişkenler ise farklı parantezler içinde yer almaktadır. Bu ilkeye göre iki boyut bir tabloda,

$N_{ij} = Gd_i^A\ d_j^B\ d_{ij}^{AB}$ modeli {AB},

$N_{ij} = Gd_i^A\ d_j^B$ modeli {A} {B},

$N_{ij} = Gd_i^A$ modeli {A},

$N_{ij} = Gd_j^B$ modeli {B},

N_{ij} = G modeli ise { } şeklinde gösterilmektedir.

İkinci gösterim şekli ise * ve + sembollerinin kullanımıdır. Bu gösterim şeklinde birbiriyle ilişkili değişkenler arasında *, birbirinden bağımsız değişkenler arasında ise + sembolü bulunmaktadır. Bu ikinci ilkeye göre birinci model A*B, ikinci model A+B, üçüncü model ise A şeklinde gösterilecektir.

Birinci gösterim şekli her değişken için bir alfabetik yada sayısal karakter kullanımını zorunlu kılmaktadır. Çünkü birden fazla karakterin yorumlanması güçlük çıkaracaktır. Örneğin iki boyutlu bir tablonun değişkenleri YAŞ ve BOY olsun. Bu durumda YAŞ değişkeni ile BOY değişkeni arasında ilişkinin varlığını göstermek için kullanılacak {YAŞ BOY} sembolü anlaşılır olmaktan uzak olacaktır. Oysa ikinci gösterim biçiminde, YAŞ*BOY, ise değişkenleri ayıran bir işaretin olması nedeniyle böylesine karmaşıklık olmayacaktır.

Tablo 6

BABANIN MESLEK STATÜSÜ	ÜLKE	OĞULUN MESLEK STATÜSÜ ÜST	ORTA	ALT
ÜST	İNGİLTERE	588	395	159
	DANİMARKA	685	280	118
ORTA	İNGİLTERE	349	714	447
	DANİMARKA	232	348	198
ALT	İNGİLTERE	114	320	411
	DANİMARKA	83	201	246

İngiltere'deki erkek nüfusun dikey toplumsal hareketliliği ile Danimarka'daki erkek nüfusun dikey toplumsal hareketliliğini ölçmek için toplanan verilerden oluşan çapraz tabloların karşılaştırılması (Tablo 6) üç boyutlu modellerin tablo üstünde gösterimi için iyi bir örnek olacaktır.

Tablonun değerlendirilmesi hipoteze göre oluşturulan kuramsal dağılımın tablodaki bu dağılım ile kıyaslanması şeklindedir. Bu karılaştırma sonucunda dağılımların benzerliğine karar verilmesi, hipotezin doğrulanmasıdır.

Ülkeyi Ü, babanın meslek statüsünü B ve oğulun meslek statüsünü ise O değişkeni ile gösterelim. Bu üç değişken arasında bir ilişkinin olmadığını varsayan hipotezimiz, {Ü} { B} {O}

modeli ile ifade edilecektir. Bu modele göre oluşturulan kuramsal dağılım ise Tablo 7'de görülmektedir.

Bu modelde olabilirlik oranı istatistiğinin değerinin büyük çıkması iki dağılımın değerleri arasındaki farkın istatistiksel olarak anlamlı olduğunu göstermektedir. Bu nedenle hipotez kabul edilmeyecektir. Elimizdeki dağılımın (gözlenen frekansların oluşturduğu dağılım) oluşturduğumuz modeldeki kuramsal dağılımdan gelme olasılığı .000 dir.

Log-linear modeller elimizdeki hipotezi test etme yanında, toplanan verilerin oluşturduğu dağılıma en uygun model bulma sorununa da çözüm getirmektedir. Kurulabilecek modellerin hepsi anlamlı olmayacaktır. Bunun için anlamlı olabilecek modellerin seçiminde çok boyutlu kuramsal yönelim önem kazanmaktadır.

Tablo 7

BABANIN MESLEK STATÜSÜ	ÜLKE	OĞULUN MESLEK STATÜSÜ ÜST	ORTA	ALT
ÜST	İNGİLTERE	460.32	506.77	354.38
	DANİMARKA	314.73	346.50	242.30
ORTA	İNGİLTERE	473.35	521.12	364.42
	DANİMARKA	328.64	356.31	249.16
ALT	İNGİLTERE	284.46	313.17	219.00
	DANİMARKA	194.50	214.13	149.73
	L2 = 1170.4	Sd = 12	p = .000	

14 değişik model içinden elimizdeki tabloya en uygun modelin seçimi için gerekli parametreler Tablo 8'de görülmektedir.

Tablodan da anlaşılacağı gibi 14 modelden eldeki dağılıma en uygun model 8. model olmaktadır. Yani ülke ile baba mesleği statüsü, ülke ile oğlu mesleği satüsü ve baba mesleği ile oğul mesleği statüsünün ilişkilerinin varlığı birbirlerinden bağımsız olarak anlamlı olmaktadır.

Tablo 8

MODEL	L2	SD	P
1. {ÜB}	1193.38	12	.000
2. {ÜO}	1365.29	12	.000
3. {BO}	358.53	9	.000
4. {A} {B} {O}	1170.40	12	.000
5. {Ü} {B}	1297.2	14	.000
6. {Ü} {O}	1451.892	14	.000
7. {B} {O}	1379.39	13	.000
8. {ÜB} {ÜO} {BO}	4.11	4	.391
9. {ÜB} {O}	1066.66	10	.000
10. {ÜO} {B}	1083.77	10	.000
11. {BO} {Ü}	149.54	8	.000
12. {ÜB} {ÜO}	980.03	8	.000
13. {ÜB} {BO}	45.80	6	.000
14. {ÜO} {BO}	62.91	6	.000

4. SONUÇ

Log-linear modeller sosyal bilimlerde yoğun olarak kullanılan nitel değişkenler arasındaki ilişkinin test edilmesinde büyük bir duyarlılık ve esneklik getirmektedir. Aynı tablo üzerinde birden fazla hipotezin test edilebilmesi bu modellerle kolaylaşmıştır. Bu modellerin kullanımı ile birlikte sosyal bilimlerdeki ampirik ve kuramsal çalışmalar da yeni bir ivme kazanmıştır.

Ayrıca bu modeller büyük ölçekli verilere gereksinimi de artırmıştır. Çünkü bu modellerin küçük ölçekli verilere uygulanması çözümlemenin güvenirliğini olumsuz yönde etkilemektedir.

Bu modeller bilgisayar kullanımını da zorunlu hale getirmiştir. Çünkü bu modellerle kuramsal dağılımların elle hesaplanması olanaksızdır.

KAYNAKÇA

FIENBERG, S.E. (1977). **The Analysis of Cross Clasified Categorical Data**, Cambridge, Massachusetts: MIT Press.

GLASS, D.V. (1954), **Social Mobility ın Great Britain, Glencoe**, III.: Free Press.

GOODMAN, L.A. (1965). "On The Statistical Analysis of Mobility Data", **American Journal of Sociology**, 70:564-585.

GOODMAN, L.A. (1969), "How to Ransack Social Mobility Tables And Other Kinds of Cross Classification Tables", American Journal of Sociology, 75:1-40.

GOODMAN,L.A. (1979), "Simple Models for The Analysis of Association In Cros-Classifications Having Ordered Categories", **Journal of The American Statistical Association**, 74:573-552.

GOODMAN, L.A. (1981), "Three Elementary Vievs of Log-Linear Models for The Analysis of Cross-Classifications Having Ordered Categories", **Sociological Methodology** 1981, 193-239.

GOODMAN, L.A. (1984), **The Analysis of Cross Classified Data Having Ordered Categories**, Harvard: Harvard Universty Press.

HABERMAN, S.J. (1973), "Log-Linear Fit For Contingency Tables, Algorithm AS 51", **Applied Statistics**, 21:218-224.

HABERMAN, S.J. (1974), **The Analysis of Frequency Data**, Chicago: Universtiy of Chicago Press.

HANUSHEK, E.A. - JACKSON, J.E. (1977), **Statistical Methods For Social Scientists**, New York: Academic Press.

KENDALL, M.G. - STUART, A. (1973), **The Advanced Theory of Statistics: Volume 2, Inference and Relationship**, Third Edition, London: Charles Griffin Company Ltd.

KNOKE, D. - BURKE, P.J. (1980), **Log-Linear Models**, Beverly Hills: Sage Universty Publications.

MOOD, A.M., -GRAYBILL, F.A. - BOES, D.C. (1974), **Introduction to The Theory of Statistics**, Tokyo: McGrawHill-Kagusha.

UYGUN, H. (1989), "Log-Linear Modeller: Sosyal Araştırmalarda Kullanılan Yeni Çözümleme Tekniği", **H.Ü.İktisadi ve İdari Bilimler Fakültesi Dergisi**. 1-2: 241-256

WHITE, H.C. 51963), "Cause and Effect In Social Mobility Tables", **Behavioral Science**, 7:14-27.