SERİ SERIE B CİLT TOME XX SAYI FASCICULE 2 1970

İSTANBUL ÜNİVERSİTESİ

ORMAN FAKÜLTESİ DERGİSİ

REVUE DE LA FACULTÉ DES SCIENCES FORESTIÈRES
DE L'UNIVERSITÉ D'ISTANBUL

BASİT DOĞRUSAL REGRESYON

Dr. MS., Alptekin GÜNEL

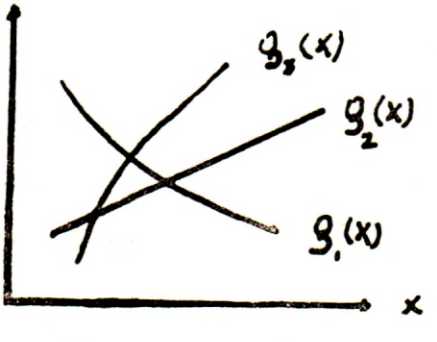
1. GİRİŞ

Bir toplum bir çok karakteristiklere sahiptir ve bu karakteristikler arasındaki ilişkinin mahiyeti tespit edilmek istenir. Tespit edilen bu ilişkiden faydalanarak, daha kolaylıkla elde edilen karakteristikler yardımı ile diğerleri tayin yoluna gidilir.

Burada, en basit hal olan, iki karakteristikli bir toplum ele alınacaktır. Problemin izahı bakımından en basit halin seçilmesi konunun öneminden bir şey kaybettirmemektedir. Zira, daha karışık haller için de uygulanacak teori ve kurallar tamamen aynı olup, fark sadece hesap ayrıntılarındadır.

Yukarda sözü edilen toplumun karakteristiklerini X ve Y ile gösterelim. Tespiti istenen husus, X ile Y arasındaki ilişkinin mahiyeti ve belirli bir X değerine tekabül eden Y değerinin ne şekilde bulunabileceğidir. Böyle bir problemin tipik örneği, bir ağaçta çapla hacım arasındaki ilişkidir. Aynı şekilde, çapın yaşa bağlı olarak değişimi, asimlasyonun, belirli sınırlar içinde, sıcaklığa bağlı olarak artması iki karakteristikli bir topluma ait örneklerdir.

Çapla hacım arasındaki örnekte olduğu gibi, belirli bir X (çap) değerine birden fazla Y (hacım) değeri tekabül edebilir. Bu durumda, belirli bir X değerine tekabül eden Y değerleri bir «toplum» teşkil edeceklerdir; bu toplumun bir ortalaması vardır. Bu ortalamayı g(X) ile gösterelim. Böylece, farklı X değerleri için, Y'lerin ortalamaları da farklı olacaktır. İşte, belirli X değerine tekabül eden Y değerlerinin ortalamasını veren g(X) fonksiyonuna X'in Y'e ait «Regresyon fonksiyonu» adı verilir. Bu fonksiyon çok değişik form-



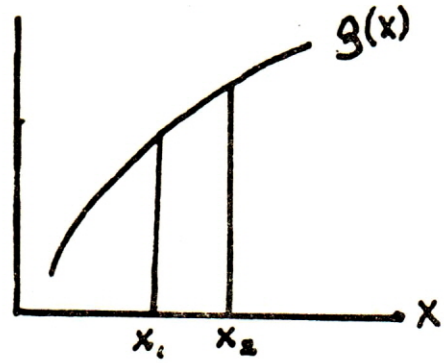
Şekil — 1. g(X)'in alabileceği

Bu fonksiyonun formu, genellikle, deneylerle elde edilen verilere dayanılarak tespit edilmek istenir. Söz konusu fonksiyonun çok çeşitli olması, bu amaçla baş vurulan teori ve prensiplerde her hangi bir değişikliği gerektirmemekte, her hal için, gene aynı esaslardan hareket edilmektedir. Bu nedenle, burada

$$g(X) = C + D\,X \qquad (1)$$

formu üzerinde durulacaktır.

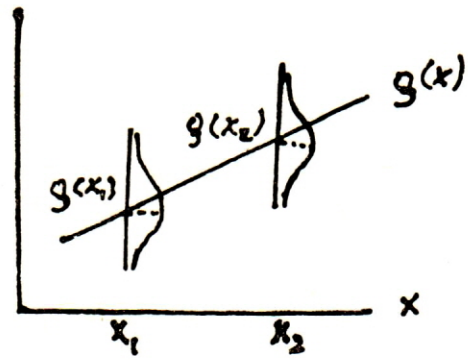
Denklem (1), bilindiği gibi, bir doğru denklemidir. Böyle bir formun seçimini haklı gösteren sebepler vardır: Daha yüksek dereceden eğrilerde bile, belirli sınırlar içinde, X ile Y arasındaki ilişki bir doğru olarak kabul edilebilir (Şekil-2). Ayrıca, birçok fonksiyonlar, logaritmik kâğıt üzerinde doğrusal bir seyir gösterirler. Nihayet, yapılacak değişken değişikliği ile, fonksiyona doğrusal bir karakter verilebilir.



Şekil — 2

Farz edelim ki, N tane X değerine tekabül eden Y değerleri tespit edilmiş olsun. Yukardaki kabule uygun olarak, X ile Y arasındaki ilişkinin doğrusal olduğunu farz edeceğiz.

Belirli bir X değeri için bir çok Y değeri tespit edildiğini ve bu Y değerlerinin bir toplum teşkil ettiğini belirtmiştik. Böyle bir toplumun g(X) gibi bir ortalaması olduğu gibi, bir de varyansı olacaktır. Bu varyansın, ortalamaların aksine, bütün Y toplumları için aynı olduğunu ve Y toplumlarının normal bir dağılım gösterdiğini diğer bir kabul olarak ileri süreceğiz (Şekil-3). Söz konusu varyansı $\sigma^2_{y.x}$ ile gösterelim. (1) nolu denklemi



Şekil — 3. X'e tekabül eden Y toplumları normal dağılımlı ve eşit varyanslıdırlar. Ortalamaları ise farklıdır.

$$g(X) = \mu_{y.x} = A + B(X - \overline{X}) \qquad (2)$$

şeklinde yazmak, bazı kolaylık-

lar sağlaması nedeni ile, daha çok tercih edilmektedir. (*) Denklemde,

A ve B = Denklemin katsayıları

$\overline{X}=\Sigma X/N=X$'lerin ortalamasıdır.

Buna göre, herhangi bir münferit X_i değerine tekabül eden Y_i gerçek değeri

$$Y_i=A+B(X_i-\overline{X})+\varepsilon_i \; ; \qquad i=1, 2, \cdots, N \qquad (3)$$

denkleminden bulunabilir. (3) nolu denklemde ε_i ler normal dağılımlı, ortalamaları sıfır, birbirlerinden bağımsız ve varyansları ise $\sigma^2_{y.x}$ olan tesadüfi değişkenlerdir. (işaretle $\varepsilon_i : N(0\sigma^2_{y\cdot x})$) ε_i'ler hakkındaki bu varsayımlar, özellikle, ε_i'ler birçok hatların bir sonucu ise gerçeğe daha çok uymaktadır. Bu durum ormancılık problemlerinde belirgin bir niteliktedir.

(2) nolu denklemi yazmakla, X ile Y arasındaki ilişkiyi belirten gerçek bir regresyondoğrusunun varlığını kabul etmiş oluyoruz. Amacımız bu doğruyu, deneyle elde ettiğimiz verilere dayanarak, en doğru şekilde tayin etmektir.

Söz konusu doğrunun tayininde en makul yol, bulunacak doğrunun (X_i, Y_i) değer çiftlerine göre elde edilen noktalara mümkün olduğu kadar «yakın» olmasıdır. Böyle bir yakınlık iki şekilde sağlanabilir:

1 — Noktalardan doğruya inilen diklerin (noktaların doğruya olan uzaklıklarının) toplamı minimum olsun, veya

2 — Bulunacak doğrudan alınacak Y değerleri ile tespit edilen Y değerleri arasındaki farkların kareleri toplamı minimum olsun.

İkinci şekilde yapılacak hesap işlemlerinin, birinciye nazaran daha az karışık olması ve ikinci şekilde elde edilen istatistiklerin, aşağıda da gösterileceği gibi, eğilim hatasından ari olması, ikinci yolun daha çok kullanılmasına sebeb olmuştur. İkinci yol, tatbikatta, en küçük kareler metodu olarak tanınır.

2. En Küçük Kareler Metodu:

Yukarda da belirtildiği gibi, en küçük kareler metodunun esası, hesaplanan doğrudan alınan Y değerleri ile, tespit edilen Y değerleri arasındaki farkların kareleri toplamının minimum olması şeklinde ifade edilebilir. Bu şekilde bulunacak doğru, gerçek doğruyu temsil eden birçok doğrudan sadece bir tanesi olup, gerçek regresyon doğrusuna tamamen

* Gerçekte, denklem $g(x)=A+BX$ şeklinde alınır ve (2). bölümde anlatılan esaslar dahilinde A' yı hesaplayıp yerine koyduktan sonra gerekli düzeltmeler yapılırsa denklem 2 elde edilir.

eşit olduğu söylenemez. Bununla beraber, diğer doğrulara üstün tutulması için bir çok sebebler vardır.

En küçük kareler metodu ile hesaplanan regresyon doğrusu denkleminin

$$Y = a + b(X - \overline{X}) \tag{4}$$

şeklinde olduğunu kabul edelim. Bu denklem yardımı ile bulunan ve herhangi bir X_i değerine tekabül eden Y değerini $(\ddot{Y}_i)$ ile gösterelim. (4) nolu denklemin belli olması (a) ve (b) değerlerinin bilinmesine bağlıdır. En küçük kareler metodunun prensibine göre,

$$\Sigma(Y_i - \ddot{Y}_i)^2 = \Sigma[(Y_i - a - b(w_i - \overline{X})]^2 = \text{minimum} \tag{5}$$

olmalıdır, Monoton artan bir fonksiyonun minimumu, türevini sıfır yapan değer olduğundan, (5) nolu eşitliğin bilinmeyenleri olan (a) ve (b) ye göre kısmi türevleri alınır ve $\Sigma(X_i - \overline{X}) = 0$ olduğu göz önünde bulundurulursa,

$$a = \frac{1}{N}\Sigma Y_i = \overline{Y} \tag{6}$$

ve

$$b = \frac{\Sigma Y_i(X_i - \overline{X})}{\Sigma(X_i - \overline{X})^2} \tag{7}$$

elde edilir. (b)'nin payı

$$\Sigma X_i Y_i - \frac{1}{N}(\Sigma X_i)(\Sigma Y_i),$$

payadası ise

$$\Sigma X_i^2 - \frac{1}{N}(\Sigma X_i)^2$$

şekillerinde de yazılabileceğinden; ayrıca,

$$S_x^2 = \frac{\Sigma(X_i - \overline{X})^2}{N-1} = \frac{\Sigma X_i - \frac{1}{N}(\Sigma X_i)^2}{N-1} = \text{X'in varyansı}$$

$$S_y^2 = \frac{\Sigma(Y_i - \overline{Y})^2}{N-1} = \frac{\Sigma Y_i^2 - \frac{1}{N}(\Sigma Y_i)^2}{N-1} = \text{Y'in varyansı}$$

$$S_{xy} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{N-1} = \frac{1}{N-1}\left[(\Sigma X_i Y_i - \frac{1}{N}(\Sigma X_i)(\Sigma Y_i)\right]$$

$$= \text{X ile Y'nin kovaryansı}$$

yazılacak olursa, (b) katsayısı

$$b = \frac{S_{xy}}{S_x^2} \tag{7a}$$

şeklinde de hesaplanabilir.

Yukardaki hesap işlemlerinden de görüleceği gibi, yapılacak bütün iş, ΣX_i, ΣY_i, ΣX_i^2, ΣY_i^2, ve $\Sigma X_i Y_i$ değerlerini hesaplamak ve bunları denklemlerde uygun yerlerine koymaktan ibarettir. Söz konusu beş değer bir defa hesaplandıktan sonra, diğer rakamlara artık ihtiyacımız kalmamıştır.

(a) ve (b) katsayılarından sonra bilinmesi gereken diğer bir ifade $\sigma^2_{y.x}$ dir. Bu amaçla

$$S^2_{y\,.\,x} = \frac{1}{N-2} \Sigma (Y_i - \ddot{Y}_i)^2 \tag{8}$$

formülü kullanılabilir. Paydadaki terimin $(N-2)$ olması, (4) nolu denklemde iki parametre (a ve b) bulunması nedeniyledir. Denklemde hesaplanması gereken dört katsayı bulunsa idi, bu ifade $(N-4)$ olacaktı. (8) nolu denklemde $a = \overline{Y}$ ve $\ddot{Y}_i = a + b\,(X_i - \overline{X})$ yazarak

$$S^2_{y\,.\,x} = \frac{1}{N-2} \Sigma (Y_i - a - b\,(X_i - \overline{X}))^2$$

$$= \frac{1}{N-2} \Sigma ((Y_i - \overline{Y}) - b\,(X_i - \overline{X}))^2$$

$$= \frac{1}{N-2} (\Sigma (Y_i - \overline{Y})^2 - 2b \Sigma (X_i - \overline{X})(Y_i - \overline{Y}) + b^2 \Sigma (X_i - \overline{X})^2)$$

$$= \frac{1}{N-2} [(N-1) S_y^2 - 2b (N-1) S_{xy} + b^2 (N-1) S_x^2]$$

$$= \frac{N-1}{N-2} (S_y^2 - 2b\, S_{xy} + b^2 S_x^2)$$

$b = S_{xy}/S_x^2$ konarak, gerekli kısaltmalardan sonra

$$S^2_{y.x} = \frac{N-1}{N-2} \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right) \tag{9}$$

bulunur.

ÖRNEK:

X_i	2	2	3	3,5	4	4,5	4,5	5	3,6	6
Y_i	1,9	2,1	3,5	2,7	3	3	3,3	3,5	3,6	3,7

$N = 10$

$\Sigma X_i = 40{,}0 \qquad \Sigma X_i^2 = 177{,}0 \qquad \overline{X} = 4{,}00 \qquad (\Sigma X_i)^2 = 1600$

$\Sigma Y_i = 29{,}4 \qquad \Sigma Y_i^2 = 89{,}86 \qquad \overline{Y} = 2{,}94 \qquad (\Sigma Y_i)^2 = 864{,}36$

$$\Sigma X_i Y_i = 125{,}1$$

$$S_x^2 = \frac{1}{9}\left(177{,}0 - \frac{1}{10}(1600)\right) = 1{,}889$$

$$S_y^2 = \frac{1}{9}\left(89{,}86 - \frac{1}{10}(864{,}36)\right) = 3{,}804$$

$$S_{xy} = \frac{1}{9}\left(125{,}1 - \frac{1}{10}(40{,}0)(29{,}4)\right) = 0{,}833$$

$$b = 0{,}833/1{,}889 = 0{,}4409$$

$$a = \overline{Y} = 2{,}94$$

$$S_{y.x}^2 = \frac{9}{8}\left(3{,}804 - \frac{(0{,}833)^2}{1{,}889}\right) = 4{,}126$$

Hesaplanan doğru denklemi:

$$\ddot{Y}_i = 2{,}94 + 0{,}4409\,(X_i - 4{,}00) \quad \text{veya}$$

$$\ddot{Y}_i = 1{,}1764 + 0{,}4409\,X_i \quad \text{dir.}$$

3. a, b, $S^2_{y:x}$ ve $\ddot{Y}$ istatistiklerinin özellikleri

A, B, $\sigma^2_{y.x}$ ve $\mu_{y.x}$ parametrelerinin güven sınırlarını bulmak ve söz konusu olabilecek hipotezlerin testi bakımından yukardaki istatistiklerin dağılım fonksiyonlarının bilinmesine ihtiyaç vardır.

3.1 — a'nın dağılım fonksiyonu

$a = \overline{Y}$ olduğundan, (3) nolu denklem yardımı ile,

$$a = \overline{Y} = \frac{1}{N}\Sigma Y_i = \frac{1}{N}\Sigma(A + B(X_i - \overline{X}) + \varepsilon_i)$$

$$= A + B\frac{\Sigma(X_i - \overline{X})}{N} + \frac{\varepsilon_i}{N}$$

$$= A + \frac{\varepsilon_i}{N}$$

$(\Sigma(X_i - X) = 0$ olduğundan)

bulunur. Ancak, ε_i lerin ortalaması (beklenen değeri) sıfır kabul edildiğinden, $(\varepsilon_i)/N$ nin beklenen değeri de sıfır olacaktır, dolayısiyle

$$a = \overline{Y} = A$$

bulunacaktır; yani a'nın beklenen değeri A'ya eşittir. Böylece, (a) eğilim hatasından aridir.

ε_i lerin varyansları eşit ve $\sigma^2_{y\cdot x}$ kabul edildiğinden, (a)'nın varyansı

$$S_a^2 = \sigma^2_{x.y}/N$$

olur. ε_i ler normal dağılımlı olduklarından, (a) da normal dağılımlı olacaktır. İşaretle gösterilirse

$$a: \quad N(A\,;\ \sigma^2_{y.x}/N)$$

dir.

3.2 — (b)'nin dağılım fonksiyonu

(3) ve (7) nolu formüllerden

$$b = \frac{\Sigma Y_i (X_i - \overline{X})}{\Sigma (X_i - \overline{X})^2} = \frac{\Sigma (A + B(X_i - \overline{X}) + \Sigma_i)(X_i - \overline{X})}{\Sigma (X_i - \overline{X})^2}$$

$\Sigma(X_i - X) = 0$ olduğu göz önünde tutulursa,

$$b = B + \frac{\Sigma \varepsilon_i (X_i - \overline{X})}{(X_i - X)^2}$$

bulunur. ε_i lerin birbirlerinden bağımsız, ortamları sıfır ve X_i lerin sabit kabul edildikleri hatırlanarak,

$$b = B$$

elde edilir.

(b)'nin varyansının

$$S_b^2 = \frac{\sigma^2_{y\,.\,x}}{\Sigma (X_i - X)^2} = \frac{\sigma^2_{y\,.\,x}}{(N-1) S^2_x}$$

olduğu yukardaki eşitlikten kolayca gösterilebilir. Aynı şekilde, ε_i lerin normal dağılımlı olması nedeni ile, (b) de normal dağılımlıdır. Yani

$$b: \quad N(B\,;\ \sigma^2_{y\,.\,x}/(N-1)\,S^2_x)$$

dir.

(b)'nin varyans formülünden de görüldüğü gibi, X'lerin varyansı ne kadar fazla ise, yani X_i'ler ne kadar çok birbirlerinden farklı iseler, (b) nin varyansı o kadar azalacaktır. Aynı şekilde, (b)'nin varyansı gözlem sayısı ile ters orantılıdır; gözlem sayısını arttırmak suretiyle, (b)'nin varyansını küçültmek mümkündür. Bu sonuç, arzu edilen bir özelliktir.

3.3 — $S^2_{y\cdot x}$ 'in dağılım fonksiyonu

İspat edilebilir ki

$$S^2_{y\,.\,x}/\sigma^2_{y\,.\,x}$$

istatistiği (X^2/d.f (khi karesi/serbestiyet derecesi) adı verilen bir dağılım göstermektedir. Bu dağılımın ortalaması (1) olduğundan $S^2_{y\cdot x}$'in ortalaması $\sigma^2_{y\cdot x}$ olacaktır.

3.4 — a, b ve $S^2_{y\cdot x}$ arasındaki ilişki

Yukarda elde edilen sonuçlar göz önünde tutulursa, söz konusu üç istatistiğin birbirlerinden bağımsız oldukları görülür. Böylece, meselâ, (a)'nın hesabında yapılacak bir hata, (b)'yi etkilemiyecektir.

3.5 — $\ddot{Y}$'nin dağılım fonksiyonu

(a) ve (b) birbirlerinden bağımsız ve normal dağılımlı olduklarından, bunların doğrusal bir fonksiyonu olan $\ddot{Y}$ de normal dağılımlı olacaktır. (a) ve (b)'ye ait sonuçlardan $\ddot{Y}$ nin ortalamasının

$$\mu_{y\,.\,x}=A+B\,(X_i-\bar{X})$$

olduğu kolaylıkla çıkartılabilir. Aynı şekilde, $\ddot{Y}$'nin varyansı

$$S^2_{\ddot{y}}=\sigma^2_{y\,.\,x}\left(\frac{1}{N}+\frac{(X_i-\bar{X})^2}{(N-1)S^2_x{}^2}\right)$$

yazılabilir.

Parantezin içindeki ikinci terimin payındaki ifadenin de ortaya koyduğu gibi, X_i değeri ortalamadan ne kadar fazla fark ederse, $\ddot{Y}$'nin sıhhati o kadar daha az olacaktır.

4 — Doğrusal Regresyonla İlgili Hipotezler

4.1 — A ile ilgili hipotezler

Üç çeşit hipotez söz konusudur:
— A belirli bir A_0 değerine eşittir,
— A belirli bir A_0 değerinden küçüktür,
— A belirli bir A_0 değerinden büyüktür.
Bu ifadeler formüle edilecek olursa, aşağıdaki şekilde yazılabilirler:

$$\begin{array}{ll} H_0: & A=A_0 \\ H_A: & A\neq A_0 \end{array} \quad \text{veya} \quad \begin{array}{ll} H_0: & A\leq A_0 \\ H_A: & A>A_0 \end{array} \quad \text{veya} \quad \begin{array}{ll} H_0: & A\geq A_0 \\ H_A: & A<A_0 \end{array}$$

Her üç hipotezin de kontrolunda t- testi kullanılabilir.

$$\frac{a-A_0}{S_a} = \frac{a-A_0}{S_{y.x}/\sqrt{N}}$$

ifadesi t- dağılımı gösterir. Dağılımın serbestiyet derecesi (N−2) dir. Birinci hipotez «iki yanlı» hipotez olup diğerleri «tek yanlı» dırlar.

4.2 — B ile ilgili hipotezler

A ile ilgili hipotezlerin tamamen aynı olup hipotezlerin kontrolu gene aynı şekilde yapılır:

$H_0: B=B_0$ $H_A: B\neq B_0$ veya $H_9: B\leq B_0$ $H_A: B>B_0$ veya $H_0: B\geq B_0$ $H_A: B<B_0$

t — değeri

$$\frac{b-B_0}{S_b} = \frac{b-B_0}{S_{y.x}/S_x\sqrt{N-1}}$$

şeklinde hesaplanmaktadır. Dağılımın serbestiyet derecesi (N−2) dir.

B ile ilgili en önemli hipotez B=0 hipotezi, yani X ile Y arasında herhangi bir ilişki bulunmadığıdır. Böyle bir hipotezin kontrolunda üzerinde durulması gereken iki hal mevcuttur: Az sayıda ölçme yapılması halinde, B'nin sıfırdan çok farklı olmasına rağmen, B=0 hipotezini red etme ihtimali çok düşük olacaktır. Diğer taraftan, B sıfıra çok yakın olabilir; fakat, gözlem sayısı oldukça yüksek ise, B=0 hipotezini red etme ihtimali atar. Böyle bir durumda, güvenirlik derecesini küçültmek, yani daha büyük (α) değeri almak yerinde olur. Unutmamak lâzımdır ki, günlük hayattaki güvenirlik ile istatistikteki güvenirlik aynı şey değildir.

B=0 hipotezinin kontrolu varyans analizi yolu ile de yapılabilir. Bu amaçla kullanılacak istatistik

$$F_{1,(N-2)} = \frac{b^2\Sigma(X_i-\bar{X})^2}{S^2_{y.x}} = t^2_{(N-2)}$$

dir.

Bu testle ilgili varyans analizi tablosu şöyledir:

Varyasyon Kaynağı	Kareler Toplamı	Serbestiyet Derecesi	Kareler Ortalaması	Beklenen Değer
Regresyon	$b^2(X_i-\bar{X})^2$	1		$\sigma^2_{y.x}+B^2\Sigma(X_i-\bar{X})^2$
Hata	$(N-2)\,S^2_{y.x}$	N−2		$\sigma^2_{y.x}$

B=0 hipotezi doğru ise, tablonun son sütunundan da görüldüğü gibi, bu takdirde her iki varyasyon kaynağının beklenen değeri $\sigma^2_{y\cdot x}$ olacağından, F=1 bulunacaktır. B ne kadar sıfırdan fark ederse, regresyonun kareler ortalaması hatanın kareler ortalamasından o kadar fark edecek, neticede F değeri 1 den büyük olcaktır. Bu nedenle, hesaplanan F değeri, seçilen (α) değeri için, tablodan alınacak F değerlerinden büyükse, B=0 hipotezi red edilmelidir.

5 — Güven Aralıkları

A ve B parametleri normal dağılımlı ve birbirlerinden bağımsız olduklarından, bu parametlere ait güven aralıklarının tayininde, t- dağılımı kullanılabilir. (α) kadar güvenirlik derecesi için, A'nın güven aralığı:

$$a - t_{1-\alpha/2}\frac{S_{y\cdot x}}{\sqrt{N}} \leq A \leq a + t_{1-\alpha/2}\frac{S_{y\cdot x}}{\sqrt{N}}$$

B'nin güven aralığı:

$$b - t_{1-\alpha/2}\frac{S_{y\cdot x}}{S_x\sqrt{N-1}} \leq B \leq b + t_{1-\alpha/2}\frac{S_{y\cdot x}}{S_x\sqrt{N-1}}$$

$\sigma^2_{y\cdot x}$'in güven aralığı:

$$\frac{S^2_{y\cdot x}}{K^2_{1-\alpha/2}/(N-2)} \leq \sigma^2_{y\cdot x} \leq \frac{S^2_{y\cdot x}}{K^2_{\alpha/2}/(N-2)}$$

$\mu_{x\cdot y}$'in güven aralığı:

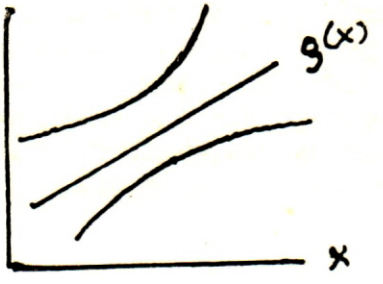
$$\ddot{Y} - t_{1-\alpha/2}\, S^2_{y\cdot x}\sqrt{\frac{1}{N} + \frac{(X_i-\bar{X})^2}{(N-1)\,S_x^2}} \leq \mu_{y\cdot x} \leq \ddot{Y} + t_{1-\alpha/2} S_{y\cdot x}\sqrt{\frac{1}{N} + \frac{(X_i-\bar{X})^2}{(N-1)\,S_x^2}}$$

Münferit bir $\ddot{Y}_i$ değerine ait güven aralığı:

$$\ddot{Y}_i \pm t_{1-\alpha/2} S_{y\cdot x}\sqrt{1+\frac{1}{N} + \frac{(X_i-\bar{X})^2}{(N-1)\,S_x^2}}$$

şeklinde hesaplanırlar.

Son iki denklemden de görüldüğü gibi, X_i $\bar{X}$'den uzaklaştıkça güven aralığı genişlemektedir (Şekil-4).



Şekil — 4. Güven aralıkları

6 — Korelasyon Katsayısı

X ile Y arasındaki doğrusal ilişkinin derecesini tayin amacı ile kullanılabilecek bir kriter, Pearson'ın korelasyon katsayısıdır. (r) ile gösterilen bu katsayı

$$r=\frac{S_{xy}}{S_x S_y} \qquad (10)$$

formülünden hesaplanmaktadır.

Korelasyon katsayısı aşağıdaki özelliklere sahiptir:

1 — $|S_{xy}| = S_x S_y$ olduğundan, r- değeri -1 ile $+1$ değerleri arasındadır veya bunlardan birine eşittir, yani

$$-1 \leq r \leq +1$$

2 — $r=b\frac{S_x}{S_y}$ yapılabileceğinden, $r=0$, ancak ve ancak, $b=0$ için olur. Ayrıca, r ile b'nin işaretleri aynıdır

3 — Kabaca denebilir ki, r, hesaplanan regresyon doğrusunun noktalara çok yakın olması halinde, -1 veya $+1$ değerlerini alır. Doğrunun noktalardan uzak seyretmesi oranında r de sıfıra yaklaşır.

Korelasyon katsayısı r'nin regresyonla ilgi derecesi, Y'lere ait kareler toplamı üzerinde yapılacak ufak bir işlemle daha iyi belirtilebilir. Y'lerin kareler toplamı

$$\begin{aligned}\Sigma(Y_i-\overline{Y})^2 &= \Sigma(\overline{Y}_i-\ddot{Y}_i+\ddot{Y}_i-Y)^2\\ &=\Sigma(Y_i-\ddot{Y}_i)^2+\Sigma(\ddot{Y}_i-\overline{Y})^2+2\,\Sigma(Y_i-\ddot{Y}_i)(\ddot{Y}_i-\overline{Y})\\ &=\Sigma(Y_i-\ddot{Y}_i)^2+\Sigma(\ddot{Y}_i-\overline{Y})^2\,;\quad (\Sigma(Y_i-\ddot{Y}_i)(\ddot{Y}_i-\overline{Y})=0 \text{ dır}).\end{aligned}$$

şeklinde iki kısma ayrılır. Eşitliğin sağ tarafındaki ilk terim hatadan doğan kareler toplamı olup «Hata Kareler toplamı», ikinci terim ise regresyon nedeni ile meydana gelen kareler toplamı olup «Regresyon kareler toplamı» adları ile tanılırlar. (9) nolu formülden, hatalar kareler toplamı

$$(N-1)\left(S_y^2-\frac{S^2_{xy}}{S^2_x}\right)$$

şeklinde yazılabilir. Diğer taraftan, $\Sigma(Y_i-\overline{Y})^2=(N-1)\,S^2_y$ olduğundan, regresyon kareler toplamı

$$\Sigma(\ddot{Y}_i-\overline{Y})^2=(N-1)\frac{S^2_{xy}}{S_x^2}=(N-1)\,S_y^2\,r^2$$

yardımı ile bulunabilir.

Y'nin kareler toplamına ait eşitliğin her iki yanı $\Sigma(Y_i-\overline{Y})^2$ ile bölünürse,

$$1=\frac{(N-2)\left(S_y^2-\dfrac{S^2_{xy}}{S^2_x}\right)}{\Sigma(Y_i-\overline{Y})^2}+\frac{(N-1)+\dfrac{S^2_{xy}}{S_x^2}}{\Sigma(Y_i-\overline{Y})^2}$$

$\Sigma(Y_i-Y)^2=(N-1)\,S_y^2$ yazılarak

$$1=(1-(S^2_{xy}/(S_x^2S_y^2))+(S^2_{xy}/(S_x^2S_y^2)$$

elde edilir.

$$S^2_{xy}/(S_x^2S_y^2))=r^2$$

dir. O halde r^2, Y'lerin kareler toplamının «sebebi bilinen varyasyon» (explained variation) kısmını teşkil etmektedir. Bu durumda, $(1-r^2)$ «Sebebi bilinmeyen varyasyon» (unexplained variation) oranını ifade eder. (r) değeri ne kadar 1'e yakınsa, yani hesaplanan doğru, ölçme ve gözlem sonucunda elde edilen noktalara ne kadar yakınsa, $(1-r^2)$ ifadesi o kadar sıfıra yaklaşacaktır. Bu sonuç, doğrusallık hipotezinin kontrolunda kullanılır.

7 — Regresyonun Doğrusallığının Kontrolu

Yukarıda X ile Y arasındaki ilişkinin doğrusal olduğunu kabul etmiş ve regresyon denklemini

$$g(X)=\mu_{y.x}=A+B(X-\overline{X})$$

şeklinde yazmıştık. En az bir X_i için birden fazla Y değerleri tespit edilmişse (meselâ, aynı çap için birden fazla hacım bulunmuşsa), sözü edilen varsayımın doğruluğunu kontrol etmek mümkündür.

Problemin izahı bakımından aşağıdaki işaretlemeyi kullanacağız:

$Y_{ij}=X_i$ ye tekabül eden j inci ölçme,

$n_i=X_i$ ye takabül eden Y_{ij} sayısı

$$T_{yi.}=\sum_{j=1}^{n_i}Y_{ij},\quad Y_{y..}=\sum_i\sum_j Y_{ij};$$

$$\overline{Y}_{i.}=(T_{yi.}/n_i)\quad;\quad \overline{Y}_{..}=T_{y..}/\Sigma n_i$$

ÖRNEK:

$X_1=5$	$Y_{11}=3,0$	$X_2=5,5$	$Y_{21}=3,8$	$X_3=6$	$Y_{31}=4,2$
	$Y_{12}=4,0$		$Y_{22}=4,2$		$Y_{32}=5,0$
			$Y_{23}=4,4$		

$$n_1=2;\quad n_2=3;\quad n_3=2$$

$$Ty_1.=7,0;\quad Ty_2.=12,4;\quad Ty_3.=9,2;\quad Ty..=28,6$$

$$Y_1.=3,5\ ;\quad Y_2.=4,13\ ;\quad Y_3.=4,6\ ;\quad \overline{Y}..=4,08$$

Farz edelim ki, k tane farklı X değeri mevcuttur; dolaysiyle, k tane Y değerler topluluğu tespit edilmiş olacaktır. X_i değerine tekabül eden Y değerlerinin gerçek ortalamasını U_i ile gösterelim. Aynı zamanda, yukarıda yaptığımız gibi, Y'lerin normal dağılımlı ve varyanslarının eşit olduğunu da kabul edeceğiz.

Bu takdirde, iki ayrı varsayımın kontrolu söz konusudur:

— Bütün U_i'ler birbirlerine eşittir, yani

$$U_1=U_2=\ \ldots\ldots\ldots\ =U_k$$

— U_i'ler X'lerin doğrusal bir fonksiyonudur.

Bu iki varsayımın kontrolunda varyans analizi kullanılabilir. Aynı X_i değerine tekabül eden Y değerlerine «Grup» diyecek olursak, aşağıdaki eşitlikler yazılabilir:

Gruplar içi kareler toplamı

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2 \qquad (11)$$

$$=\sum\sum Y^2{}_{ij}-\sum\frac{T^2{}_{y_i.}}{n_i}$$

Gruplar arası kareler toplamı

$$=\Sigma n_i\,(\overline{Y}_{i.}-\overline{Y}_{..})^2=\Sigma\frac{T^2{}_{y_i}}{n_i}-\frac{T_y{}^2{}_{..}}{\Sigma n_i} \qquad (12)$$

Regresyon Kareler Toplamı

$$=\sum_i\sum_j(Y_{ij}-\ddot{Y}..)^2=(N-1)\,\frac{S^2{}_{xy}}{S_x{}^2}$$

$$=(N-1)\,S_y{}^2\,r^2 \qquad (13)$$

Regresyon etrafındaki Kareler Toplamı

=Gruplararası Kareler toplamı — Regresyon Kareler toplamı (14)

Doğrusallık kontrolu amacı ile hazırlanacak varyans analizi tablosu şöyledir:

Varyasyon Kaynağı	Kareler Toplamı	Serbestiyet Derecesi	Kareler Ortalaması	Kareler Ortalamasının Beklenen Değeri
1	2	3	4	5
Gruplar içi			$\Sigma n_i - k$	$\sigma^2_{y.x}$
Regresyon			1	$\sigma^2_{y.x} + \Sigma n_i \sigma^2_u R^2_{xu}$
Regresyon etraf			$k-2$	$\sigma^2_{y.x} + \frac{\Sigma n_i}{k-2}\sigma^2_\mu (1 - R^2_{x\mu})$
Gruplar arası			$k-1$	$\sigma^2_{y.x} + \frac{\Sigma n_i}{k-2}\sigma^2_u$

Tablonun son sütunundaki

$$\sigma_u^2 = \frac{1}{N}\Sigma n_i (U_i - \overline{U})^2 = \text{Ortalamalar varyansı}$$

$$\overline{U} = \frac{1}{N}\Sigma n_i U_i = \text{Genel ortalama}$$

$R_{xu}=X_i$ ile buna tekabül eden U_i'ler arasındaki korelasyon katsayısıdır.

Birinci varsayımı kontrol için

$$F_1 = \frac{\text{Gruplararası Kareler ortalaması}}{\text{Gruplar içi kareler ortalaması}}$$

istatistiği kullanılır. Varsayımımız doğru ise, $\sigma^2_u=0$ olacağından, pay ve paydanın beklenen değerleri, tablonun son sütunundan da görüldüğü gibi, $\sigma^2_{y\cdot x}$ olacak ve $F_1=1$ bulunacaktır. Hiç değilse, ortalamalardan biri diğerlerine eşit değilse, $F_1 \neq 1$ olacaktır. F_1'in serbestiyet derecesi $(k-1)$ ve $(\Sigma n_i - k)$'dır. Hesaplanan değer, seçilen (α) güvenirlik derecesi ve yukardaki serbestiyet dereceleri için, tablodan alınan değerden büyükse ortalamaların eşitliği varsayımı red edilecektir. **İkinci varsayım ancak birinci varsayım red edildikten sonra kontrol edilir.** İkinci varsayımı kontrol için kullanılacak istatistik

$$F_2 = \frac{\text{Regresyon etrafında kareler ortalaması}}{\text{Gruplar içi kareler ortalaması}}$$

dir.

İkinci varsayım doğru ise, $R_{xu}=1$ olacaktır. Bu taktirde, pay ve paydanın beklenen değerleri aynı ve $\sigma^2_{y\cdot x}$ olacak, dolayısiyle $F_2=1$ bulunacaktır $R_{xu}\neq 1$ ise, F_2'nin değeri 1'den farklıdır. F_2'nin serbestiyet dereceleri $(k-2)$ ve (Σn_i-k)'dır. Tablodan alınan değer hesaplanan değerden küçükse, varsayım red edilir.

Evvelce de belirtildiği gibi, her iki tez birbirinden bağımsız değildir. Testin kuvveti k ve n_i sayıları yükseldikçe artar.

ÖRNEK

Y_{ij} \ X_i	5,0	5,5	6,0	6,5	7,0	7,5	8,0	8,5	90	95
	3,0	3,8	4,2	5,4	5,6	6,3	7,0	7,1	8,6	8,0
	4,0	4,2	5,0	6,0	6,1		8,0	8,0	9,0	9,4
		4,4		6,2	6,1		8,4	8,5		9,8
					6,6			8,5		
n_i	2	3	2	3	4	1	3	4	2	3
$T_{yi.}$	7,0	12,4	9,2	17,6	24,4	6,3	23,4	32,1	17,6	27,2
$\overline{Y}_{i.}$	3,5	4,13	4,6	5,86	6,1	6,3	7,8	8,02	8,8	9,06

$$\sum_{i=1}^{k} n_i=27;\quad k=10;\quad i=1,2,\ldots,10;\quad J=1,2,\ldots,n_i$$

$$\Sigma n_iX_i=198{,}0;\quad \Sigma n_iX_i^2=1505{,}5;\quad (\Sigma n_iX_i)^2=39204$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}=177{,}2,\quad \sum_i\sum_j Y^2_{ij}=1257{,}54;\quad \Big(\sum_i\sum_j Y_{ij}\Big)^2=31399{,}84$$

$$\sum_i^k (T^2_{yi}/n_i)=1251{,}471;\quad \frac{(\Sigma\Sigma Y_{ij})^2}{n_i}=\frac{T^2_{y..}}{n_i}=\frac{31399{,}84}{27}=1162{,}957$$

$$\sum_i\sum_j X_iY_{ij}=1367{,}7;\quad \Big(\sum_i n_iX_i\Big)\Big(\sum_j\sum_i Y_{ij}\Big)=35085{,}6$$

$$S_x^2=\frac{1}{26}\left(1505.50-\frac{39204}{27}\right)=2{,}0576;\quad S_x=1{,}4344$$

$$S_y^2 = \frac{1}{26}\left(1257{,}54 - \frac{31399{,}84}{27}\right) = 3{,}6378; \quad S_y = 1{,}9073$$

$$S_{xy} = \frac{1}{26}\left(1367{,}7 - \frac{35085{,}6}{27}\right) = 2{,}6244$$

$$r = \frac{2{,}6244}{(1{,}4344)\,(1{,}9073)} = 0{,}959; \quad r^2 = 0{,}919; \quad (1-r^2) = 0{,}081$$

Varyans Analizi Tablosu

Varyans Kaynağı	Kareler Toplamı	Serbestiyet Derecesi	Kareler Ortalaması	
Gruplariçi	1257,52 − 1251,471 = 6,049	17	0,357	
Regresyon	$(N-1)S^2_y r^2$ = 86,9218	1	86,9218	$F_1 = \frac{9{,}8349}{0{,}357}$ = 27,55**
Regresyon Etrafı	1,5922	8	0,199	
Gruplar arası	1251,471 − 1162,957 = 88,514	9	9,8349	

Hesaplanan F_1 değeri, (9) ve (17) serbestiyet dereceleri ve $\alpha = \%1$ için tablodan alınan F değerinden büyüktür. Bu durumda, bütün ortalamaların eşitliği hakkındaki varsayım red edilecektir.

Birinci varsayım red edildiği için, ikinci varsayımı kontrol edebiliriz;

$$F_2 = \frac{\text{Regresyon etrafında kareler ortalaması}}{\text{Gruplar içi kareler ortalaması}}$$

Ancak, gruplar içi kareler ortalaması, regresyon etrafında kareler ortalamasından büyüktür. Bu takdirde F_2

$$F_2 = \frac{0{,}257}{0{,}15} = 1{,}70$$

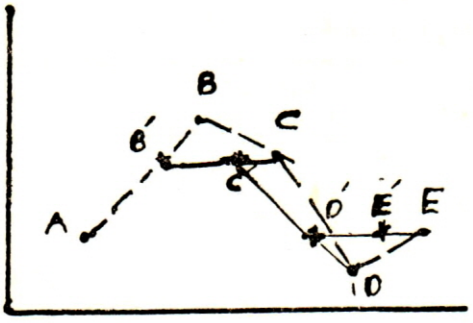
şeklinde hesaplanmalıdır.

Hesaplanan F_2 değeri, (17) ve (8) serbestiyet dereceleri ile $\alpha = \%1$ için bulunan tablo değeri 3,71'den küçüktür. Bu durumda, X ile Y arasındaki ilişkinin doğrusal olmadığı hakkında elimizde yeter delil yoktur. Söz konusu varsayım kabul edilecektir.

8. Regresyon Doğrusunun Geometrik Tayini

Regresyon doğrusunu karakterize eden (a) ve (b) değerlerinin bilinmesine ihtiyaç duyulmadığı hallerde, söz konusu doğruya geometrik olarak tayin mümkündür. Çizim metodu Şekil-5 yardımı ile açıklanmıştır:

Yapılan ölçmelerle A, B, C, D ve E gibi beş nokta elde edildiğini farz edelim. AB doğru parçası üzerinde ve AB uzunluğunun (2/3)'ündeki nokta B′ diyelim. B′ ile C arasında ve gene bu uzaklığın (2/3)'ündeki nokta C′; C′ ile D arasında ve benzer uzaklıktaki nokta D′; D′ ile E arasında, aranan nokta E′ olsun



Şekil — 5

E′ noktası, regresyon doğrusunun geçmesi gereken noktalardan biridir. Doğrunun çizilebilmesi için tayini zorunlu diğer nokta aynı şekilde hareket edilerek, fakat bu sefer E′, D, C, B ve A noktaları için, bulunur. Elde edilecek son nokta B″ ile gösterilirse, regresyon doğrusu B″ E′ doğrusudur.

AB′′ uzaklığı ile E′E uzaklığı birbirine eşitse, çizimde kaba hata yapılmamış demektir.

FAYDALANILAN ESERLER

1 — ANDERSON, R. L., and T. A. BANCROFT, 1952, Statistical Theory in Research, McGraw-Hill, New-York.

2 — DRAPER, N. R., and H. SMITH, 1968, Applied Regression Analysis, John Wiley, New-York.

3 — JEROME, C. R. Li., 1966, Statistical Inference, Edwards Bothers, Ann Arbor, Mich.

4 —PRODAN, M., 1968, Forest Biometrics, Pergamon Press, London.

5 — SELBY, S. M., 1965, Standard Mathematical Tables, The Chemical Rubber Co., Ohio.

6 — SNEDECOR, G. W., 1953, Statistical Methods, State Colege Press, İowa.