REGRESYON ANALİZİ

Bilgisayar destekli istatistiki yöntemler

Nevzat ADA
0830120007
Doç.Dr. Serdar CARUS

Üzerinde durulan değişkenlerden birinin bağımlı (y), diğerinin (x) bağımsız olması durumunda y’nin x’in bir fonksiyonu olarak ifade edilen ilişkiye **<u>regresyon</u>** denir.

Bağımlı değişken ile bir veya daha çok bağımsız değişken arasındaki ilişkiyi incelemek amacıyla kullanılan bir analiz yöntemine de **<u>regresyon analizi</u>** denir.

Regresyon analizi, değişkenler arasındaki neden-sonuç ilişkisini bulmamıza imkan veren bir analiz yöntemidir.

Regresyon analizi ile bağımlı ve bağımsız değişkenler arasında bir ilişki var mıdır? Eğer bir ilişki varsa bu ilişkinin gücü nedir? Değişkenler arasında ne tür bir ilişki vardır? Bağımlı değişkene ait ileriye dönük değerleri tahmin etmek mümkün müdür ve nasıl tahmin edilmelidir? Belirli koşulların kontrol edilmesi durumunda özel bir değişken veya değişkenler grubunun diğer değişken veya değişkenler üzerindeki etkisi nedir ve nasıl değişir? gibi sorulara cevap aranmaya çalışılır.

Regresyon başlıca 2’ye ayrılır:

1)**Tek Değişkenli Regresyon Analizi;** Tek değişkenli regresyon analizi bir bağımlı değişken ve bir bağımsız değişken arasındaki ilişkiyi inceler. Tek değişkenli regresyon analizi ile bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişkiyi temsil eden bir doğrunun denklemi formüle edilir.

2)Çok Değişkenli Regresyon Analizi; İçinde bir adet bağımlı değişken ve birden fazla bağımsız değişkenin bulunduğu regresyon modelleri çok değişkenli regresyon analizi olarak bilinir.

İki değişken arasında çeşitli ilişkiler

1	$Y = a + bx$	Doğru denklemi
2	$Y = a + bx + cx^2$	Parabolik ilişki
3	$Y = ab^x$ veya $\log y = \log a + x \log b$	Üssel eğri
4	$y = ax^b$ veya $\log y = \log a + b \log x$	Geometrik eğri
5	$y = \frac{1}{a + bx}$ veya $\frac{1}{y} = a + bx$	Hiperbolik ilişki

Regresyon doğrusal regresyon modeli ve doğrusal olmayan regresyon modeli olmak üzere 2 çeşittir.

Doğrusal regresyon modeli:

$y_i=\alpha+\beta x_i$

x_i:bağımsız değişkeni,

α:doğrunun y eksenini kestiği değer olarak başlangıç değerini,

β:x'deki 1 birim değişmeye karşılık y'deki değişim miktarını gösterir.

PARAMETRE TAHMİNİ

$y_i=\alpha+\beta x_i$

β parametresinin tahmin edicisi b;

$$b=\frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}=\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

Benzer şekilde α'nın tahmin edicisi a ise,

$$a=\bar{y}-b\bar{x}$$

REGRESYON KATSAYISININ ÖNEM TESTİ VE GÜVEN SINIRLARI

Regresyon katsayısının önem testi ve güven sınırlarının tahmininde t dağılışı kullanılır ve t dağılışı ile yapılan test tahminlerdeki benzer işlem prosedürü burada da takip edilir.

ÖNEM TESTİ

$$t_h = \frac{b - \beta}{S_b} \quad \text{ve} \quad S_b = \sqrt{S_{yx}^2 / \sum (x_i - \bar{x})^2}$$

$S_b = \sqrt{\dfrac{S_{yx}^2}{\sum x_i^2 - \dfrac{(\sum x_i)^2}{n}}}$ formülü ile hesaplanabilir. Formüldeki S_{yx}^2 ise

$$S_{yx}^2 = \frac{1}{n-2} \sum (y_i - \hat{y})^2 = \frac{1}{2}\left[\sum (y_i - \bar{y})^2 - \frac{(\sum (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum (x_i - \bar{x})^2}\right]$$

olarak hesaplanır.

Kritik cetvel (Tablo) değeri t_c hipotez tek yönlü ise $t_{\alpha\,(n-2)}$, iki yönlü ise $t_{\alpha/2(n-2)}$ olarak belirlenir.

$t_h > t_c$ ise regresyon katsayısı için H_0 hipotezi reddedilir. H_1 hipotezi kabul edilir. $t_h < t_c$ ise H_0 hipotezi kabul edilir.

GÜVEN SINIRLARI

Regresyon katsayısının üst ve alt güven sınırları,

$$\beta_{\text{üs}} = b + t_{\alpha/2,(n-2)} S_b \quad \text{ve} \quad \beta_{as} = b - t_{\alpha/2,(n-2)} S_b$$

formüllerinden hesaplanır. Regresyon katsayısına ait hipotez testi ve güven sınırlarının hesaplanmasında SD = n - 2 olduğuna dikkat edilmelidir.

Regresyon denkleminin uygunluğunu ifade eden belirleme katsayısı,

$$R^2 = \frac{\dfrac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2}}{\sum (y_i - \bar{y})^2}$$

ile ölçülür.

R^2 değerinin 1'e yaklaşması uyumun iyi olduğunu gösterir.

Örnek: Bebekler üzerinde yapılan bir çalışmada, bebeklerin yaş(ay olarak) ve ağırlıkları aşağıdaki gibi bulunmuştur.

a) Ağırlığın yaşa göre ilişkisini ifade eden regresyon denklemini bulunuz?
b) Bebeklerin doğum ve 4. ay ağırlığının bulunuz?
c) Regresyon katsayısı b'nin 0'dan büyüklüğünü tespit ediniz?
d) b'nin güven sınırlarını tahminleyiniz? ($1-\alpha=0.95$)

Yaş(x_i)	1	2	3	4	5	6	7
Ağırlık(y_i)	3.5	4.0	5.0	-	5.2	5.8	6.5

a) Regresyon denkleminin istatistikleri a ve b'yi tahminlemek ve ilgili formüllerdeki terimleri hesaplayabilmek için aşağıdaki Tablo oluşturulur.

no	x_i	y_i	x_i^2	y_i^2	$x_i\ y_i$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(x_i - \bar{x})^2$
1	1	3.5	1	12.25	3.5	3.629	0.017	9
2	2	4.0	4	16.00	8.0	4.086	0.007	4
3	3	5.0	9	25.00	15.0	4.543	0.209	1
4	5	5.2	25	27.04	26.0	5.457	0.066	1
5	6	5.8	36	33.64	34.8	5.914	0.013	4
6	7	6.5	49	42.25	45.5	6.371	0.017	9
Σ	***24***	***30***	***124***	***156.15***	***132.8***		***0.329***	***28***

Regresyon katsayısı **b** formülündeki değerler yerlerine yazılarak

$$b = \frac{132.8 - \dfrac{24 * 30}{6}}{124 - \dfrac{(24)^2}{6}} = \frac{12.8}{28} = 0.457 \text{ olarak bulunur.}$$

$\sum x=24$, $\sum y=30$ ve $\bar{x}=4$, $\bar{y}=5$ olduğundan , $a=\bar{y}-b\bar{x}$ eşitliğinden a $= 5 - 0.457*4 = 3.172$ olarak bulunur ve, Regresyon eşitliği,

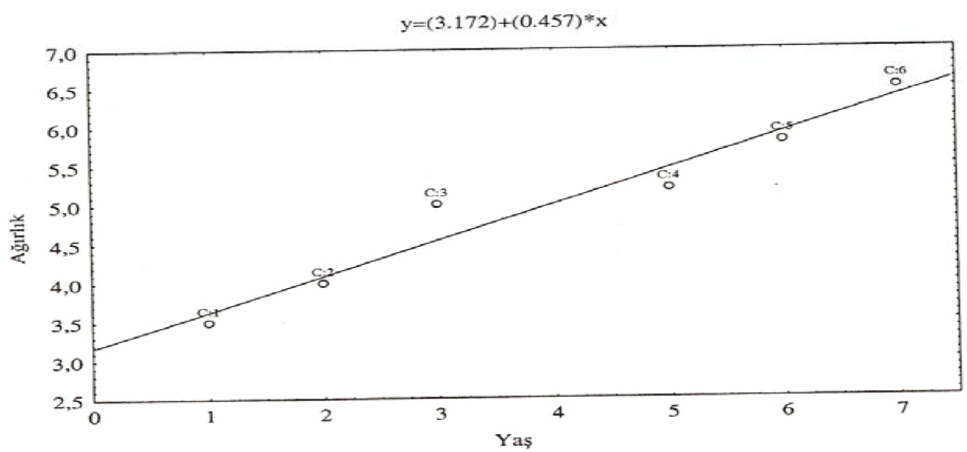
$$y = 3.712 + 0.457x$$

şeklinde kurulur.

b) Doğum ağırlığı regresyon eşitliğinde $x = 0$ alınarak hesaplanan değerdir. Diğer bir ifade ile başlangıç değeri $a = 3.712$ doğum ağırlığıdır.

$y_0 = 3.172 + 0.457*0 = 3.172$

Dördüncü ay ağırlığı ise, $y_4 = 3.172 + 0.457*4 = 5$ kg'dır.



c) Regresyon katsayısı β'nın hipotez testi t dağılımı yardımı ile yapılacaktır.

H_0: Regresyon katsayısı $\beta = 0$

H_1: Regresyon katsayısı $\beta > 0$

$t_h = \dfrac{b-\beta}{S_b}$ olduğuna göre, S_b için önce S^2_{yx} hesaplanırsa,

$$S_{yx}^2 = \frac{1}{n-2}\sum(y_i - \hat{y})^2 = \frac{1}{6-2}0.329 = 0.08225$$ ve buradan,

$$S_b = \sqrt{S_{yx}^2 / \sum(x_i - \overline{x})^2} = \sqrt{0.08225/28} = 0.0542$$ dir.

$$t_h = \frac{0.457 - 0}{0.0542} = 8.43$$

t tablosunda, tek yönlü test de 4 Sd ve 0.01 hata seviyesinde cetvel değeri $t_c = t_{0.01(6-2)} = 3.747$ olarak okunur. Karşılaştırma yapıldığında, $t_h > t_c$ olduğundan (H_0 Ret) β'nın 0'dan büyük olduğuna karar verilir.

d) β'nın güven sınırları ise,

$$\beta \begin{cases} \text{üs} = b + t_{\alpha/2\,(n-2)}S_b = 0.457 + 2.776 * 0.0542 = 0.607 \\ \text{as} = b - t_{\alpha/2\,(n-2)}S_b = 0.457 - 2.776 * 0.0542 = 0.307 \end{cases}$$

olarak tahmin edilir.

Doğrusal olamayan regresyon modelleri

Örnek:

Bir bakteri kültüründe, kültür ekiminden 6 dakika sonra ölçülen kloni sayısı (Y) aşağıda tabloda verilmiştir. a)verilere, $Y=ab^{x}$ formunda, en küçük kareler eşitliğini kurunuz. b)denklemden elde edeceğiniz Y değerlerini gerçek değerler ile karşılaştırınız. c)X=7 iken Y değerini bulunuz?

Dakika(X)	Kloni say.(Y)
0	32
1	47
2	65
3	92
4	132
5	190
6	275

Çözüm:

a) $y = ab^x$ eşitliğinin her iki tarafının logaritması alınırsa,

$\log y = \log a + x \log b$ elde edilir. Sadece bağımlı değişkenin logaritması alındığından, ilişki yarı logaritmiktir. Bu transformasyonla doğrusal regresyon çöxümü için aşağıdaki tablo oluşturularak ***b*** ve ***a*** istatistikleri sırasıyla,

	Dakika (x)	Log(y)	Xy	x^2
	0	1.505	0	0
	1	1.672	1.672	1
	2	1.813	3.626	4
	3	1.964	5.892	9
	4	2.121	8.484	16
	5	2.279	11.395	25
	6	2.439	14.634	36
Toplam	21	13.793	45.703	91
Ortama	3	1.970		

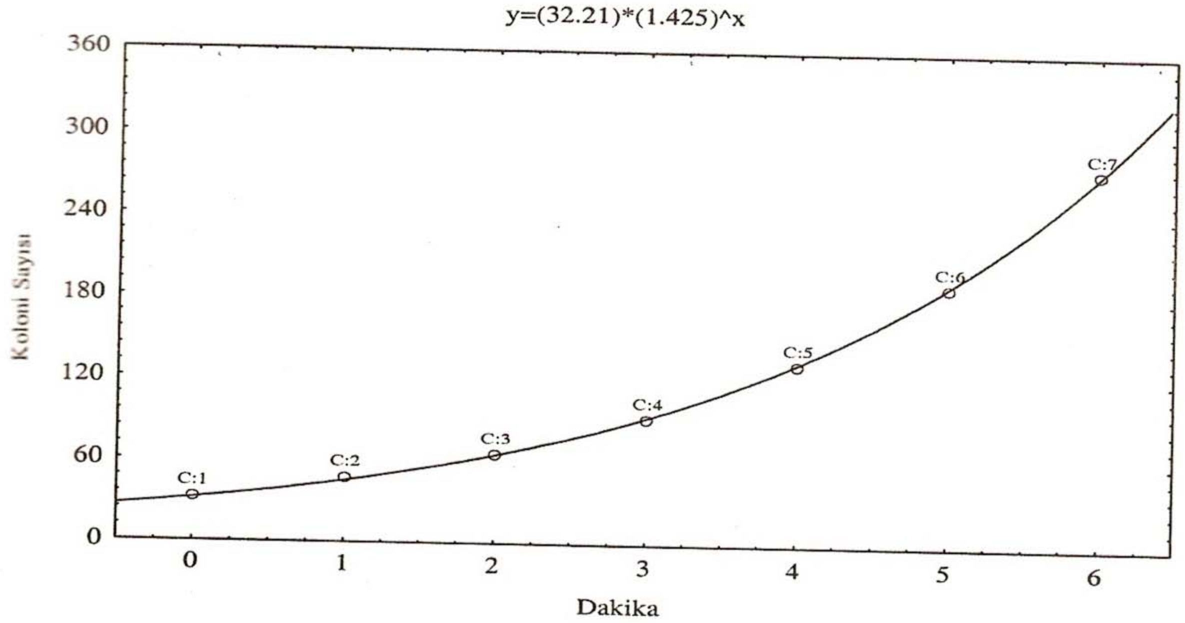
$$b = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{45.703 - \frac{21 * 13.793}{7}}{91 - \frac{21^2}{7}} = \frac{45.703 - 41.379}{91 - 63} = \frac{4.324}{28} = 0.154$$

$a = \bar{y} - b\bar{x} = 1.970 - 0.154*3 = 1.508$ şeklinde tahminlenir.

$\log a = 1.508 \Rightarrow a = 32.21$ ve,

$\log b = 0.154 \Rightarrow b = 1.425$ dir. Bu tahminlere göre $Y = a*b^x$ eşitliği

$y = 32.21*1.425^X$ olarak ifade edilir.



b) $\hat{y}$ değerleri ve y – $\hat{y}$ = farkları aşağıda ki gibi hesaplanır.

x = 0 için	$\hat{y} = 32.21*1.425^0 = 32.21 \quad \Rightarrow y - \hat{y} = 32 - 32.21 = -0.21$
x = 1 için	$\hat{y} = 32.21*1.425^1 = 45.90 \quad \Rightarrow y - \hat{y} = 47 - 45.90 = 1.10$
x = 2 için	$\hat{y} = 32.21*1.425^2 = 65.41 \quad \Rightarrow y - \hat{y} = 65 - 65.41 = -0.41$
x = 3 için	$\hat{y} = 32.21*1.425^3 = 93.20 \quad \Rightarrow y - \hat{y} = 92 - 93.20 = -1.20$
x = 4 için	$\hat{y} = 32.21*1.425^4 = 132.82 \quad \Rightarrow y - \hat{y} = 132 - 132.82 = -0.82$
x = 5 için	$\hat{y} = 32.21*1.425^5 = 189.26 \quad \Rightarrow y - \hat{y} = 190 - 189.26 = 0.74$
x = 6 için	$\hat{y} = 32.21*1.425^6 = 269.70 \quad \Rightarrow y - \hat{y} = 275 - 269.70 = 5.30$

c) X=7 için y=32.21*1.4277=384.32 dir.