BÖLÜM 13

REGRESYON

Günlük yaşamda karşılaşılan bazı olaylar birbirlerinden bağımsız olarak düşünülemezler. Örneğin; bir öğrencinin başarısı ile haftalık ders çalışma saatleri arasındaki ilişki, bir ürünün verimi ile gübre arasındaki ilişki, reklamlar ve satışlar arasındaki ilişki incelenmek istenebilir.

İki ya da daha çok değişken arasındaki ilişkinin yapısı 'regresyon analizi' ile ilişkinin yönü ve derecesi ise 'korelasyon analizi' ile incelenir.

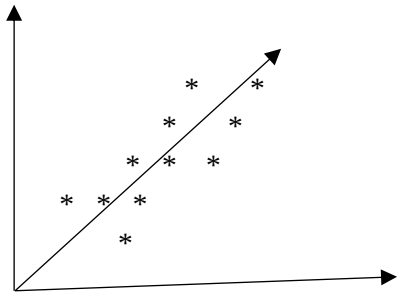
Analizler içinde en çok kullanılan; regresyon analizi, bağımlı ve bağımsız değişkenler arasındaki bağıntının belirlenmesinde ve bu bağıntı yardımıyla çıkarılacak istatistiksel sonuçların elde edilmesinde kullanılan yöntemlerden oluşmaktadır. Burada amaç; bağımlı değişkeni bağımsız değişkenlerin bir fonksiyonu olarak ifade etmek ve bu fonksiyon yardımıyla bağımlı değişkenin değerlerini tahmin etmek, ön görmek; bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerini tahmin etmek; bağımlı veya bağımsız değişkenlerin etkileri ile ilgili öne sürülen hipotezleri test etmektir.

Çoğu zaman iki ya da daha çok değişken arasındaki bir bağıntı olup olmadığını ve bu bağıntının bir denklemle nasıl ifade edilebileceği araştırılmak istenmektedir.

Basit Doğrusal Regresyon Çözümlemesi

X , $(x_1, x_2, ..., x_n)$ değerlerini alan ve Y , $(y_1, y_2, ..., y_n)$ değerlerini alan iki rastgele değişken olsun. Bu iki değişken arasındaki ilişki,doğrusal regresyon çözümlemesi ile incelenebilir.

X rastgele değişkeni haftalık çalışma saatini(bağımsız değişken), Y (bağımlı değişken) rastgele değişkeni öğrencinin başarısını göstersin. n tane öğrencinin haftalık çalışma saatleri ile notları belirlensin. $(x_1, y_1), (x_2, y_2) ..., (x_n, y_n)$ ile gösterilen verilerin koordinat düzlemi üzerinde serpilme diyagramı çizilebilir.Eğer haftalık çalışma saati arttıkça, başarının da artacağı düşünülürse bu iki değişken arasında doğrusal bir ilişki vardır denir.



X ile Y arasındaki gerçek bağıntı,

$Y = \beta_0 + \beta_1 X + \varepsilon \rightarrow$ Kitle için regresyon modeli

doğru denklemi ile ifade edilir.

Y : Bağımlı değişken

X : Bağımsız değişken

β_0 : Regresyon doğrusunun y eksenini kestiği nokta

β_1 : Regresyon katsayısı (Aynı zamanda doğrunun eğimi)

ε : Hata terimi (Bağımlı değişkenin gerçek değeri ile gözlenen değeri arasındaki farkı gösterir.)

β_0 ve β_1 bilinmeyen regresyon katsayılarıdır.

Kitleden n birimlik örneklem için doğrusal regresyon denklemi:

$y_i = b_0 + b_1 x_i + e_i, \quad i = 1,2,3,...,n$

biçiminde tanımlanır. Bilinen bir x_j değeri için y_j değeri tahmin edilir.Tahmini doğrusal regresyon denklemi

$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1,2,3,...,n$

biçimindedir.

b_0: regresyon doğrusunun y eksenini kestiği nokta yı gösterir. Aynı zamanda β_0 'ın tahminidir.

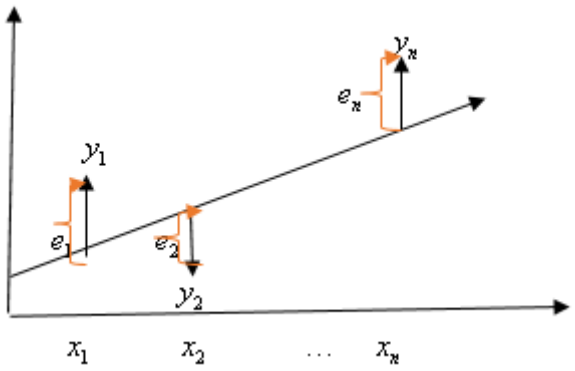
b_1: regresyon katsayısıdır. Doğrunun eğimini gösterir. Bağımsız değişkendeki bir birimlik değişmenin bağımlı değişkende yapacağı değişimi gösterir. β_1'in tahminidir.

e_j: j. Gözlemin hata terimidir. Gözlenen değer ile tahmini değer arasındaki farktır.

$e_j = y_j - \hat{y}_j$ dir. Hata terimleri ortalaması sıfır varyansı σ^2 olan normal dağılıma sahiptir. $e \sim N(0, \sigma^2)$

varsayımlar: $\begin{cases} E(e_i) = 0 \\ Var(e_i) = \sigma^2 \\ e_1, e_2, ..., e_n \text{'ler bağımsız} \end{cases}$

En Küçük Kareler Yöntemi (EKK) İle β_0 Ve β_1 Katsayılarının Tahmini:



$\hat{y} = b_0 + b_1 x,$

$\sum_{j=1}^{n} e_j^2 = \left(y_j - \hat{y}_j\right)^2 = \left(y_j - b_0 - b_1 x_j\right)^2$

b_0 ve b_1

$\min \sum_{j=1}^{n} e_j^2$ olması istenir.

$$\frac{\partial \sum_{i=1}^{n} e_i^{\,2}}{\partial b_0} = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^{\,2}}{\partial b_0} = 2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)(-1) = 0$$

$$\sum_{j=1}^{n} y_j = nb_0 + b_1 \sum_{j=1}^{n} x_j \qquad (1)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^{\,2}}{\partial b_1} = 0$$

$$\frac{\partial \sum_{i=1}^{n} e_i^{\,2}}{\partial b_1} = 2\sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)(-x_i) = 0$$

$$\sum_{j=1}^{n} x_j y_j = b_0 \sum_{j=1}^{n} x_j + b_1 \sum_{j=1}^{n} x_j^2 \qquad (2)$$

(1) Eşitliği $-\sum_{j=1}^{n} x_j$ ile (2) eşitliği n ile çarpılıp taraf taraf toplanırsa

$$-\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j = -nb_0 \sum_{j=1}^{n} x_j - b_1 \left(\sum_{j=1}^{n} x_j\right)^2$$

$$n\sum_{j=1}^{n} x_j y_j = nb_0 \sum_{j=1}^{n} x_j + nb_1 \sum_{j=1}^{n} x_j^2$$

$$n\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j = nb_1 \sum_{j=1}^{n} x_j^2 - b_1 \left(\sum_{j=1}^{n} x_j\right)^2$$

$$b_1 = \frac{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)\left(y_j - \bar{y}\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_j y_j - n\overline{xy}}{\sum_{i=1}^{n} x_j^2 - n\bar{x}^2},$$

(1) denkleminden

$$\sum_{j=1}^{n} y_j - b_1 \sum_{j=1}^{n} x_j = nb_0$$

$$b_0 = \frac{\sum_{j=1}^{n} y_j}{n} - b_1 \frac{\sum_{j=1}^{n} x_j}{n} = \bar{y} - b_1\bar{x} \Rightarrow b_0 = \bar{y} - b_1\bar{x}$$

$b_1 > 0$ iki değişken birlikte artıyor yada birlikte azalıyor.

$b_1 < 0$ değişkenlerden biri artarken diğeri azalacaktır.

Modelin anlamlılığı için F testi

Tanım: Gerçek y değerlerinin kendi ortalaması, $\bar{y}$'dan sapmalarının kareler toplamı SST ile gösterilir.

SST : Kareler toplamı

$$SST = \sum_{i=1}^{n} \left(y_j - \bar{y}\right)^2 = \sum_{j=1}^{n} y_j^2 - n\bar{y}^2$$

Tanım: Tahmin edilmiş y değerlerinin kendi ortalamaları, $\bar{y}$'den sapmalarının kareler toplamı SSR ile gösterilir. Regresyon kareler toplamı denir.

SSR : Regresyon kareler toplamı

$$SSR = \sum_{j=1}^{n} \left(\hat{y}_j - \bar{y}\right)^2 = \frac{\left(\sum_{j=1}^{n} x_j y_j - n\overline{xy}\right)^2}{\sum_{j=1}^{n} x_i^2 - n\bar{x}^2}$$

Tanım: Gerçek y değerlerinin regresyon doğrusu üzerinde karşılık gelen tahmin edilmiş $\hat{y}$ değerlerinden sapmalarının kareler toplamı SSE ile gösterilir.

SSE : Hata kareler toplamı

$$SSE = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

$$SSE = SST - SSR$$

$$SST = SSE + SSR$$

Varyans Analiz Tablosu

Değişimin Kaynağı	**Serbestlik Derecesi**	**Kareler Toplamı**	**Kareler Ortalaması**	**Test İstatistiği**
Regresyon	1	SSR	$MSR = \frac{SSR}{k(1)}$	$F_t = \frac{MSR}{MSE}$
Hata	$= n-2$	SSE	$MSE = \frac{SSE}{n-(k+1)}$	
Toplam	$n-1$	SST		

Hipotez Testi:

Basit doğrusal regresyon modelinin yeterliliğini belirtmek için hipotez testlerine ve güven aralıklarına ihtiyaç vardır. Bu testlerde, hata terimi:

$\varepsilon : \sim N(0, \sigma^2)$ sahip olduğu varsayılır.

1) **Hipotez:**

$H_0 : \beta_1 = 0$ (Regresyon doğrusu önemsizdir.)

$H_1 : \beta_1 \neq 0$ (Regresyon doğrusu önemlidir.)

2) **Test İstatistiği:**

$$F_t = \frac{MSR}{MSE}$$

3) **Karar Aşaması:**

$F_t > F_{\alpha:s_1:s_2}$ *ise* H_0 *red edilir.*

$F_{\alpha:s_1:s_2} \Rightarrow \alpha$ 'ya bağlı s_1 ve s_2 serbestlik dereceli F tablo değeri

Gözlemlerimizin model denklemine uyumu önemlidir.

Belirtme Katsayısı:

Bağımsız değişkenin bağımlı değişkendeki değişimin yüzde kaçını açıkladığını gösterir. 'R^2' ile gösterilir. R^2'nin alabileceği en büyük değer '1'dir.

$$R^2 = \frac{SSR}{SST} \Rightarrow R^2 \text{'nin büyük olması tercih edilir.}$$

$$0 \leq R^2 \leq 1$$

KAYNAKLAR

1. Uygulamalı İstatistik (1994)

   Ayşen APAYDIN , Alaettin KUTSAL, Cemal ATAKAN

2. Olasılık ve İstatistik Problemler ve Çözümleri ile (2008)

   Prof. Dr. Semra ERBAŞ

3. Olasılık ve İstatistik (2006)

   Prof. Dr. Fikri Akdeniz

4. Olasılık ve İstatistiğe Giriş I-II (2011)

   Prof. Dr. Fikri Öztürk

5. Fikri Öztürk web sitesi

   http://80.251.40.59/science.ankara.edu.tr/ozturk/index.html