~~BÖLÜM 2 :~~ ÇOKLU DOĞRUSAL REGRESYON

ÇOKLU DOĞRUSAL REGRESYON MODELİ

Yanıt değişkeni y , k sayıda **bağımsız değişkenle** ilişkili olabilir. Bu durumda,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \qquad \textbf{(2.1)}$$

modeli k bağımsız değişkenli **çoklu doğrusal regresyon modeli** olarak adlandırılır. β_j, $j = 0,1,...,k$ parametreleri, **regresyon katsayıları** olarak bilinir.

β_j parametresi, $x_i (i \neq j)$ **bağımsız değişkenleri sabit tutulduğu** zaman x_j bağımsız değişkenindeki bir birimlik değişime karşılık gelen y yanıt değişkenindeki beklenen değişimi göstermektedir. Bu nedenle, β_j, $j = 0,1,...,k$ parametreleri genellikle **kısmi regresyon katsayıları** olarak adlandırılır.

MODEL PARAMETRELERİNİN KESTİRİMİ

Regresyon Katsayılarının En Küçük Kareler Kestirimi

y_i, $i.$ gözlenmiş yanıt değişkenini, x_{ij} de x_j bağımsız değişkeninin $i.$ gözlem değerini göstersin. Modelin ε hata terimi için $E(\varepsilon) = 0$ ve $Var(\varepsilon) = \sigma^2$ ve hataların ilişkisiz olduğu varsayılsın.

TABLO 2.1 Çoklu Doğrusal Regresyon İçin Veriler

		Bağımsız Değişkenler			
Gözlem, i	**Yanıt,** y	x_1	x_2	......	x_k
1	y_1	x_{11}	x_{12}	......	x_{1k}
2	y_2	x_{21}	x_{22}	....	x_{2k}
.	.	.	.		.
.	.	.	.		.

.	.	.	.		.
n	y_n	x_{n1}	x_{n2}	....	x_{nk}

Denklem (2.1)'e karşılık gelen **örneklem regresyon modeli**,

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_k x_{ik} + \varepsilon_i \\
&= \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n
\end{aligned}
\tag{2.2}
$$

olarak ifade edilir. En küçük kareler fonksiyonu ise

$$
S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{2.3}
$$

olup bu fonksiyon $\beta_0, \beta_1, \ldots, \beta_k$'ya göre minimize edilerek parametrelerin en küçük kareler kestiricileri elde edilir.

$$
\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{2.4a}
$$

ve

$$
\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \,, \quad j = 1, 2, \ldots, k \tag{2.4b}
$$

olmak üzere en küçük kareler normal denklemleri,

$$
\begin{array}{ccccc}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} & + \ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} & = \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^{\,2} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} & + \ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} & = \sum_{i=1}^{n} x_{i1} y_i \\
\vdots & \vdots & \vdots & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} & + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} & + \ldots\ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^{\,2} & = \sum_{i=1}^{n} x_{ik} y_i
\end{array}
\tag{2.5}
$$

şeklinde elde edilir. Bu denklemler çözülerek $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ **en küçük kareler kestiricileri** bulunur.

Denklem (2.2)'de verilen modelin matris gösterimi,

$$y = X\beta + \varepsilon$$

biçimindedir.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{bmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

$\boldsymbol{y}$, $n \times 1$ boyutlu gözlemler vektörü; $\boldsymbol{X}$, $n \times p$ boyutlu bağımsız değişkenler matrisi; $\boldsymbol{\beta}$, $(p \times 1)$ boyutlu regresyon katsayıları vektörü ve $\boldsymbol{\varepsilon}$, $(n \times 1)$ boyutlu rastgele hatalar vektörüdür.

Aşağıda verilen $S(\beta)$ fonksiyonu minimize edilerek $\hat{\beta}$ en küçük kareler kestiricileri vektörü elde edilir.

$$\begin{aligned} S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon &= (y - X\beta)'(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned}$$

$\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$, 1×1 'lik bir matris ya da skaler olduğundan transpozu da $(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y})' = \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}$ aynı şekilde skalerdir. En küçük kareler kestiricileri,

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

eşitliğini sağlar. $X'X\hat{\beta} = X'y$ olmak üzere, $\boldsymbol{\beta}$ **'nın en küçük kareler kestiricisi,**

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \textbf{(2.6)}$$

olur.

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^{2} & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^{2} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \cdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

Eğer yukarıdaki matris çarpımı gerçekleşirse (2.5) normal denklemlerin skaler formu elde edilir. Burada, $\boldsymbol{X'X}$, $p \times p$ boyutlu simetrik matris ve $\boldsymbol{X'y}$ $p \times 1$ boyutlu sütun vektörüdür.

Bağımsız değişkenlerin tüm $x' = [1, x_1, x_2, ..., x_k]$ düzeylerine karşılık gelen kestirim modeli,

$$\hat{y} = \boldsymbol{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{\boldsymbol{j}=1}^{\boldsymbol{k}} \hat{\boldsymbol{\beta}}_{\boldsymbol{j}} \boldsymbol{x}_{\boldsymbol{j}}$$

olup y_i gözlem değerlerine karşılık gelen $\hat{y}_i$ kestirim değerleri vektörü,

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{\boldsymbol{H}}\boldsymbol{y} = \boldsymbol{Hy} \quad \textbf{(2.7)}$$

olur. $n \times n$ boyutlu $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ matrisi genellikle **şapka matrisi** olarak adlandırılır. Bu matris, gözlenmiş değerler vektörünü $\hat{y}_i$ kestirim değerleri vektörüyle eşleştirir.

y_i, gözlem değerleri ile bunlara karşılık gelen $\hat{y}_i$ kestirim değerleri arasındaki $e_i = y_i - \hat{y}_i$ farkı **artıklardır.** n sayıdaki artıklar, matris biçiminde aşağıdaki gibi yazılabilir:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} \quad \textbf{(2.8a)}$$

Artıklar vektörü $\boldsymbol{e}$, aşağıda gösterildiği gibi birçok farklı biçimde ifade edilebilir.

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{Hy} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \quad \textbf{(2.8b)}$$

Örnek 2.1 Teslim Süresi Verileri

Bir meşrubat şişeleme fabrikası, dağıtım sistemindeki otomatik satış makinelerinin servis güzergahlarını incelemektedir. Güzergahtaki bir operatörün, satış mağazalarından birindeki otomatik satış makinesine bakım yapabilmesi için yaklaşık olarak gereken süre kestirilmek istenmektedir. Bakım hizmetleri, içecek ürünleriyle makineleri stoklama, küçük çaplı bakım ya da temizliği içermektedir. Bu çalışma için Endüstri mühendisi iki önemli değişkenin teslim süresini (y) etkilediğini öne sürmektedir : Stoklanmış ürün sayısı (x_1, teslim hacmi) ve güzergahtaki operatörün yürüyerek katettiği mesafe (x_2). Mühendis teslim süresi için 25 gözlem toplamıştır; bu değerler Tablo 2.2'de görülmektedir.

TABLO 2.2 Örnek 2.1 için Teslim Süresi Verileri

Gözlem Numarası	Teslim Süresi (y)	Teslim Hacmi (x_1)	Mesafe (x_2)
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450

24	19.83	8	635
25	10.75	4	150

Teslim süresinin kestirilebilmesi için aşağıdaki çoklu doğrusal regresyon modeli kullanılmalıdır:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Çoklu regresyon modelini uydurmak için ilk önce $\boldsymbol{X}$ matrisi ve $\boldsymbol{y}$ vektörü oluşturulur.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 7 & 3 & \dots & 4 \\ 560 & 220 & \dots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \dots & \dots & \dots \\ 1 & 4 & 150 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix}$$

$\boldsymbol{\beta}$ 'nın en küçük kareler kestiricisi,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10,232 \\ 219 & 3,055 & 133,899 \\ 10,232 & 133,899 & 6,725,688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7,375.44 \\ 337,072.00 \end{bmatrix} = \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}$$

olup **en küçük kareler uyumu**,

$$\hat{y} = 2.34123 + 1.61591x_1 + 0.01438x_2$$

olarak elde edilir.

TABLO 2.3 Örnek 2.1 İçin Gözlemler, Uyum Değerleri ve Artıklar

Gözlem No	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	16.68	21.7081	- 5.0281
2	11.50	10.3536	1.1464
3	12.03	12.0798	- 0.0498
4	14.88	9.9556	4.9244
5	13.75	14.1944	- 0.4444
6	18.11	18.3996	- 0.2896
7	8.00	7.1554	0.8446
8	17.83	16.6734	1.1566
9	79.24	71.8203	7.4197
10	21.50	19.1236	2.3764
11	40.33	38.0925	2.2375
12	21.00	21.5930	- 0.5930
13	13.50	12.4730	1.0270
14	19.75	18.6825	1.0675
15	24.00	23.3288	0.6712
16	29.00	29.6629	- 0.6629
17	15.35	14.9136	0.4364

18	19.00	15.5514	3.4486
19	9.50	7.7068	1.7932
20	35.10	40.8880	-5.7880
21	17.90	20.5142	-2.6142
22	52.32	56.0065	-3.6865
23	18.75	23.3576	-4.6076
24	19.83	24.4028	-4.5728
25	10.75	10.9626	-0.2126

TABLO 2.4 Meşrubat Şişeleme Fabrikası Teslim Süresi Verileri İçin Minitab Çıktısı

Regression Analysis : Time versus Cases, Distance

The regression equation is
Time = 2.34 + 1.62 cases + 0.0144 Distance

Predictor	Coef	SE Coef	T	P
Constant	2.341	1.097	2.13	0.044
Cases	1.6159	0.1707	9.46	0.000
Distance	0.014385	0.003613	3.98	0.001

S = 3.25947 R - Sq = 96.0 % R - Sq (adj) = 95.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5550.8	2775.4	261.24	0.000
Residual Error	22	233.7	10.6		
Total	24	5784.5			

Source	DF	Seq SS
Cases	1	5382.4
Distance	1	168.4

Ödev

Bir tarlada yetişen bir tahıl türüne gübre ve yağmur miktarlarının etkisi araştırılmak isteniyor. Bu amaçla tarla 7 bölgeye ayrılarak gübre ve yağmur miktarları takip edilip, kilogram cinsinden verimleri ölçülüyor.

Verim (Y)	40	50	50	70	65	65	80
Gübre (X_1)	100	200	300	400	500	600	700
Yağmur (X_2)	10	20	10	30	20	20	30

Verilerin çoklu regresyon modeline uygun olduğu varsayımı altında, verilere göre regresyon parametrelerinin en küçük kareler tahmin değerlerini bulup modeli oluşturunuz.