1 Çoklu Dogrusal Regresyon

y bağımlı (yanıt) değişkeni k tane bağımsız değişken ile açıklanabilir. Eğer bu değişkenler arasındaki ilişki

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{1}$$

modeli ile ifade ediliyorsa, (1) modeli k **bağımsız değişkenli çoklu regresyon modeli** olarak adlandırılır.

$\beta_j, j = 0, 1, ..., k$ parametreleri **kısmi regresyon katsayıları** olarak adlandırılır.

(1) modeli matrisler yardımı ile

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

olmak üzere (2) eşitliğindeki gibi de yazılabilir.

(2) modelinde $\boldsymbol{\beta}$ için EKK tahmin edicisi $\mathbf{X}'\mathbf{X}$ (**pxp** boyutludur) matrisinin tersi mevcut olduğunda

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

biçiminde bulunur. Bu durumda

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{4}$$

biçiminde modelin tahmin denklemi elde edilir, burada $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ **şapka (hat) matrisi**dir. Böylece, artıklar vektörü

$$\mathbf{e} = \mathbf{y} - \widehat{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{5}$$

biçiminde bulunur.

Ayrıca, σ^2 için yansız tahmin edici

$$\widehat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{\mathbf{e}'\mathbf{e}}{n-p} = \frac{\mathbf{y}'\mathbf{y} - \widehat{\beta}\mathbf{X}'\mathbf{y}}{n-p} \tag{6}$$

biçiminde elde edilir ($p = k + 1$).

$E(\boldsymbol{\varepsilon}) = 0$ ve $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ varsayımı altında $\widehat{\boldsymbol{\beta}}$ EKK tahmin edicisinin beklenen değeri $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ ve varyansı $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}$ elde edilir.

Böylece, σ^2 bilinmediğinde σ^2 için yansız tahmin edici $\widehat{\sigma}^2$ kullanarak,

$$cov(\widehat{\beta}_i, \widehat{\beta}_j) = \widehat{\sigma}^2 C_{ij} (i \neq j)$$

ve $\widehat{\beta}_j$ tahmin edicisinin **standart hatası**

$$se(\widehat{\beta}_j) = \sqrt{\widehat{\sigma}^2 C_{jj}} j = 0, 1, ..., k$$

biçiminde elde edilir.

Örnek 1: y teslim süresi, x_1 stoklanmış ürün sayısı, x_2 operatörün katettiği mesafe olmak üzere $n = 25$ olarak verilen veri için çoklu regresyon analizi yapalım. (Bu veriye R programında **"library(MPV)"** paketinde **"data(softdrink)"** den ulaşabilirsiniz.)

Modelimiz $\mathbf{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boldsymbol{\varepsilon}$ biçimindedir. Bu veri seti için R programında "lm" fonksiyonunun sonucu aşağıdadır.

```
> summary(reg)

Call:
lm(formula = y ~ x1 + x2, data = softdrink)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16

> anova(reg)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 5382.4  5382.4 506.619 < 2.2e-16 ***
x2         1  168.4   168.4  15.851 0.0006312 ***
Residuals 22  233.7    10.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bu sonuçlara göre tahmin modelimiz $\widehat{\mathbf{y}} = 2.341231 + 1.615907\, x_1 + 0.014385\, x_2$ olarak elde edilir.

Yorum: x_2 değişkeni sabit tutulduğunda x_1 değişkenindeki 1 birimlik artışda y bağımlı değişkeninin ortalama 1.615907 birim artması beklenir.

Benzer olarak x_1 değişkeni sabit tutulduğunda x_2 değişkenindeki 1 birimlik artışda y bağımlı değişkeninin ortalama 0.014385 birim artması beklenir.

Sadece x_1 bağımsız değişkenini kullanarak elde edilen basit doğrusal regresyon modelinin sonuçları aşağıdaki gibi olur.

```
> summary(reg1)

Call:
lm(formula = y ~ x1, data = softdrink)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5811 -1.8739 -0.3493  2.1807 10.6342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.321      1.371   2.422   0.0237 *
x1             2.176      0.124  17.546 8.22e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.181 on 23 degrees of freedom
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.9275
F-statistic: 307.8 on 1 and 23 DF,  p-value: 8.22e-15

> anova(reg1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x1         1 5382.4  5382.4  307.85 8.22e-15 ***
Residuals 23  402.1    17.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bu sonuçlara göre tahmin modelimiz $\widehat{\mathbf{y}} = 3.321 + 2.176\ x_1$ olarak elde edilir.

Sadece x_2 bağımsız değişkenini kullanarak elde edilen basit doğrusal regresyon modelinin sonuçları aşağıdaki gibi olur.

```
> summary(reg2)

Call:
lm(formula = y ~ x2, data = softdrink)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1628  -4.8783  -0.5966   6.0810  12.8776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.961159   2.337360   2.123   0.0448 *
x2          0.042569   0.004506   9.447 2.21e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.179 on 23 degrees of freedom
Multiple R-squared:  0.7951,    Adjusted R-squared:  0.7862
F-statistic: 89.24 on 1 and 23 DF,  p-value: 2.214e-09

> anova(reg2)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x2         1 4599.1  4599.1  89.237 2.214e-09 ***
Residuals 23 1185.4    51.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Bu sonuçlara göre tahmin modelimiz $\widehat{\mathbf{y}} = 4.961159 + 0.042569\ x_2$ olarak elde edilir.

Böylece, üç farklı model oluşturduk. Hangisi doğrudur veya hangisini tercih etmeliyiz ? Bu üç modelde doğrudur. Ancak hangisinin daha iyi bir model olduğunu artık karelerin (SSE) ortalamasına bakarak yorumlayabiliriz.

Model	**SSE**	**S.d.**	**SSE ortalama=SSE/S.d.**
x_1 ve x_2 bağımsız değişkenli	233.7	22	**10.62273**
x_1 bağımsız değişkenli	402.1	23	17.48261
x_2 bağımsız değişkenli	1185.4	23	51.53913

Bu sonuçlara göre iki bağımsız değişkenle oluşturduğumuz model oluşturduğumuz basit lineer regresyon modellerinden daha iyi bir modeldir.