1 Çoklu Dogrusal Regresyon

y bağımlı (yanıt) değişkeni k tane bağımsız değişken ile doğrusal olarak açıklandığı

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

modelini matrisler yardımı ile

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

olmak üzere

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

biçiminde yazabiliriz.

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

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

biçiminde bulunur. Ayrıca, σ^2 için yansız tahmin edici

$$\widehat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{\mathbf{e}'\mathbf{e}}{n-p} = \frac{\mathbf{y}'\mathbf{y} - \widehat{\beta}\mathbf{X}'\mathbf{y}}{n-p} \tag{4}$$

biçiminde elde edilir ($p = k + 1$).

Hataların birbilerinden bağımsız ve 0 ortalamalı sabit σ^2 varyans ile **normal dağıldığı** varsayımı altında **en çok olabilirlik (MLE) yöntemi** ile $\boldsymbol{\beta}$ nın tahmin edicisi EKK tahmin edicisi ile aynıdır. Ayrıca

$$\widehat{\sigma}^2_{MLE} = \frac{SSE}{n} = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

olarak bulunur. $E(\widehat{\sigma}^2_{MLE}) = \left(\frac{n-p}{n}\right)\sigma^2$ olduğundan σ^2 için yansız tahmin edici (4) eşitliğindeki gibi bulunur.

Düzeltilmiş R^2:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

biçiminde tanımlanır. Modele eklenen bağımsız değişken $SSE/(n-p)$ değerini yani artık kareler ortalamasını düşürdüğünde R^2_{adj} değeri artacaktır. Bu nedenle modele eklenecek veya modelden çıkarıalcak bağımsız değişkenlerin belirlenmesinde kullanılabiliriz.

Çoklu regresyon modelinin anlamlılık testi:

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
$H_1 : \beta_j \neq 0$ en az bir j için hipotezlerinin test edilmesidir.Eğer $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ hipotezi rededilirse $x_1, ..., x_k$ bağımsız değişkenlerinden **en az birinin modele anlamlı katkısı** olduğu sonucuna varılır.

H_0 hipotezi altında $\frac{SSR}{\sigma^2} \sim \chi^2_k$ ve $\frac{SSE}{\sigma^2} \sim \chi^2_{n-(k+1)}$ olduğundan test istatistiği

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

olarak kullanılır ve $F_{k,n-k-1}$ dağılımına sahiptir.

Eğer F istatistiğinin veriden hesaplanmış değeri $F_h > F_{k,n-k-1,\alpha}$ olur ise $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ hipotezi rededilir.

Regresyon katsayılarının anlamlılık testi:

Herhangi bir regresyon katsayısı β_j nin $(j = 1, ..., k)$ anlamlığını test etmek için $H_0 : \beta_j = 0$ ve $H_1 : \beta_j \neq 0$ hipotezlerini test ermeliyiz. Bu hipotez için test istatistiği

$$t = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 C_{jj}}}$$

olmak üzere $t_{n-(k+1)}$ dağılımına sahiptir. C_{jj}, $(\mathbf{X}'\mathbf{X})^{-1}$ matrisini jj. köşegen elemanıdır.

Eğer t istatistiğinin verilerden hesaplanan değeri $t_h > t_{n-k-1,\alpha/2}$ veya $t_h < -t_{n-k-1,\alpha/2}$ olurs ise H_0 hipotezi rededilir. Bu durumda, x_j bağımsız değişkeninin modelde diğer bağımsız değişkenler varken **modele anlamlı bir katkı** sağladığı görülür. Bu test modelde diğer bağımsız değişkenler varken x_j nin katkısını test eder. Bu nedenle bu test **kısmi** veya **marjinal** test olarak adlandırılır.

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.

Varyans analizi tablosu aşağıdaki gibi olur.

Değişim kaynağı	Kareler toplamı	Serbestlik derecesi	Kareler ortalaması	F test değeri	$p-$değeri
Regresyon	5550.8	2	2775.4	$\frac{SSR/2}{SSE/22} = 261.2$	$4.7x10^{-16}$
Artık	233.7	22	10.6		
Toplam	5784.5	24			

$F_{2,22,0.05} = 3.443357$ (R da "qf(0.95,2,22)" ile bulunur) olduğundan $F_h = 261.2 > 3.443357$ veya p-değeri$= 4.7x10^{-16} < 0.05$ olduğundan $H_0 : \beta_1 = \beta_2 = 0$ **hipotezi reddedilir** yani oluşturulan **regresyon modeli anlamlıdır**.

Not: Bu örnek için p-değeri$= P(F_{2,22} > 261.2) = 4.7x10^{-16}$ (R da "1-pf(261.2,2,22)" ile bulunur.)

$R^2 = SSR/SST = 0.9596$ ve $R^2_{adj} = 1 - \dfrac{SSE/(n-p)}{SST/(n-1)} = 1 - \dfrac{233.7/(22)}{5784.5/(24)} = 0.9559$ olarak bulunur.

β_1 ve β_2 regresyon katsayılarının ayrı ayrı anlamlığı için **kısmi veya marjinal** t-testi yapalım.

$H_0 : \beta_1 = 0$ ve $H_1 : \beta_1 \neq 0$ hipotezlerini test etmek için t istatistiği $t = \widehat{\beta}_1/se(\widehat{\beta}_1)$ olmak üzere

$$t_h = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2 C_{11}}} = \frac{1.615907}{0.170735} = 9.464$$

olarak hesaplanır.

(Burada $\widehat{\sigma}^2 = SSE/(n-p) = 10.62273, C_{11} = 2.743783x10^{-3}$ ve $\sqrt{\widehat{\sigma}^2 C_{11}} = 0.1707234$)

$t_{n-p,\alpha/2} = t_{22,0.025} = 2.074$ olmak üzere $t_h = 9.464 > 2.074$ olduğundan $H_0 : \beta_1 = 0$ hipotezi rededilir, yani x_1 bağımsız değişkeninin modele anlamlı bir katkıda bulunduğu anlaşılır. Bu kısmi t testi olduğundan, bu sonuç modelde x_2 bağımsız değişkeni varken x_1 bağımsız değişkeninin modele eklenmesinin **anlamlı** bir katkı sağladığını gösterir.

Benzer olarak $H_0 : \beta_2 = 0$ ve $H_1 : \beta_2 \neq 0$ hipotezleri için

$$t_h = \frac{\widehat{\beta}_2}{se(\widehat{\beta}_2)} = \frac{\widehat{\beta}_2}{\sqrt{\widehat{\sigma}^2 C_{22}}} = \frac{0.014385}{0.003612842} = 3.98$$

(Burada $C_{22} = 1.228745x10^{-6}$ ve $\sqrt{\widehat{\sigma}^2 C_{22}} = 0.003612842$)

$t_h = 3.98 > 2.074$ olduğundan $H_0 : \beta_2 = 0$ hipotezi rededilir. Böylece, modelde x_1 bağımsız değişkeni varken x_2 bağımsız değişkeninin modele eklenmesinin **anlamlı** bir katkı sağladığı görüliür.

Eğer Örnek 1'de sadece x_1 bağımsız değişkenini kullanırsak 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 $\widehat{\mathbf{y}} = 3.321 + 2.176\ x_1$ ve varyans analizi tablosu

Değişim kaynağı	Kareler toplamı	Serbestlik derecesi	Kareler ortalaması	F test değeri	$p-$değeri
Regresyon	5382.4	1	5382.4	$\frac{SSR/1}{SSE/23} = 307.85$	$8.22x10^{-15}$
Artık	402.1	23	17.5		
Toplam	5784.5	24			

biçimindedir.

$F_h = 307.85 > 4.279344 = F_{1,23,0.05}$ olduğundan regresyon modeli anlamlıdır.

$R^2 = 0.9305$ ve $R^2_{adj} = 1 - \dfrac{402.1/(23)}{5784.5/(24)} = 0.9275$ bulunur.

Eğer Örnek 1'de sadece x_2 bağımsız değişkenini kullanırsak 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 $\widehat{\mathbf{y}} = 4.961159 + 0.042569\ x_2$ ve varyans analizi tablosu

Değişim kaynağı	Kareler toplamı	Serbestlik derecesi	Kareler ortalaması	F test değeri	$p-$değeri
Regresyon	4599.1	1	4599.1	$\frac{SSR/1}{SSE/23} = 89.237$	$2.2x10^{-9}$
Artık	1185.4	23	51.5		
Toplam	5784.5	24			

biçimindedir.

$F_h = 89.237 > 4.279344 = F_{1,23,0.05}$ olduğundan regresyon modeli anlamlıdır.

$R^2 = 0.7951$ ve $R^2_{adj} = 1 - \dfrac{1185.4/(23)}{5784.5/(24)} = 0.7862$ bulunur.

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 ve R^2_{adj} bakarak yorumlayabiliriz.

Model	**SSE**	**S.d.**	**SSE ortalama=SSE/S.d.**	R^2	R^2_{adj}
x_1 ve x_2 bağımsız değişkenli	233.7	22	**10.62273**	0.9596	0.9559
x_1 bağımsız değişkenli	402.1	23	17.48261	0.9305	0.9275
x_2 bağımsız değişkenli	1185.4	23	51.53913	0.7951	0.7862

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