1 Gösterge Degisken Kullanımı

Eğer bir kategorik (nitel) değişkenin k tane sınıfı (düzeyi) var ise modelde bu değişken için $k-1$ tane **gösterge (göstermelik, dummy) değişken** tanımlanır.

Örneğin, eğer sigara içme durumunun içmiyor ve içiyor gibi 2 sınıfı var ise modele bu kategorik değişken 2-1=1 değişken olarak tanımlanır.

x	Kategori
0	Sigara içmiyor
1	Sigara içiyor

Benzer olarak iki kategorili bazı diğer değişkenler için de aynı tanımlamayı yapabiliriz

x	Cinsiyet (Kategori)
0	Erkek
1	Kadın

x	Medeni durum (Kategori)
0	Bekar
1	Evli

x	Yaşanılan yer (Kategori)
0	Şehir
1	Kırsal

Bu kodlamalarda kullanılan 0 ve 1 değerlerinin sayısal olarak bir anlamı yoktur. Farklı sayılar da kullanılabilir.

Eğer kategorik değişkenimizde 2 den fazla sınıf var ise,

Örneğin **eğitim düzeyinin sınıfları İlköğretim, Lise, Lisans, Lisansüstü** gibi ise 4 sınıf için 4-1=3 değişken tanımlamamız gerekir.

Gösterge değişken			
1	2	3	Eğitim düzeyi
0	0	0	İlköğretim
1	0	0	Lise
0	1	0	Lisans
0	0	1	Lisansüstü

Bu tanımlamaya göre eğer doğrusal regresyon modelinde 1 tane bağımsız değişken varken yukarıdaki eğitimle ilgili nitel değişkeni modele eklemek istersek

x_2	x_3	x_4	Eğitim düzeyi
0	0	0	İlköğretim
1	0	0	Lise
0	1	0	Lisans
0	0	1	Lisansüstü

olmak üzere modelimiz $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ biçiminde olur. Bu tanımlamaya göre aşağıdaki gibi verilen bir veri setin için gözlemlerin eğitim düzeyleri aşağıdaki gibidir.

Gözlem	y	x_1	x_2	x_3	x_4	
1	10	2	0	0	0	$\rightarrow$ 1. gözlem birimi İlköğretim mezunu
2	15	3.1	0	0	0	$\rightarrow$ 2. gözlem birimi İlköğretim mezunu
3	25	4	0	1	0	$\rightarrow$ 3. gözlem birimi Lisans mezunu
4	12	1.5	1	0	0	$\rightarrow$ 4. gözlem birimi Lise mezunu
5	12	1	1	0	0	$\rightarrow$ 5. gözlem birimi Lise mezunu
6	.	.	0	1	0	$\rightarrow$ 6. gözlem birimi Lisans mezunu
7	.	.	0	1	0	$\rightarrow$ 7. gözlem birimi Lisans mezunu
8	.	.	0	0	1	$\rightarrow$ 8. gözlem birimi Lisansüstü mezunu

Eğer eğitim düzeyinin sınıfları: Lise, Lisans, Lisansüstü gibi ise 3 sınıf için 3-1=2 değişken

1	2	Eğitim düzeyi
0	0	Lise
1	0	Lisans
0	1	Lisansüstü

biçiminde tanımlanır.

İki farklı kategorik değişken için gösterge değişken tanımlama:
Eğitimi durumu (3 sınıf) ve cinsiyet değişkenleri için gösterge değişkenleri oluşturalım.

1	2	Eğitim düzeyi
0	0	Lise
1	0	Lisans
0	1	Lisansüstü

ve

x	Cinsiyet
0	Erkek
1	Kadın

olmak üzere birleştirirsek

1	2	3	Sınıf (2x3=6)
0	0	0	Lise, Erkek
1	0	0	Lisans, Erkek
0	1	0	Lisansüstü, Erkek
0	0	1	Lise,Kadın
1	0	1	Lisans, Kadın
0	1	1	Lisansüstü, Kadın

biçiminde tanımlanır.

Eğer her iki nitel değişken de 3'er sınıf ise

1	2	Eğitim düzeyi
0	0	Lise
1	0	Lisans
0	1	Lisansüstü

ve

1	2	Yaşanılan yer
0	0	Şehir
1	0	İlçe
0	1	Köy

olmak üzere

1	2	3	4	Sınıf (3x3=9)
0	0	0	0	Lise, Şehir
1	0	0	0	Lisans, Şehir
0	1	0	0	Lisansüstü, Şehir
0	0	1	0	Lise, İlçe
1	0	1	0	Lisans, İlçe
0	1	1	0	Lisansüstü, İlçe
0	0	0	1	Lise, Köy
1	0	0	1	Lisans, Köy
0	1	0	1	Lisansüstü, Köy

biçiminde tanımlanır.

Örnek 1: Gebelik haftası ile annenin sigara içip içmemesi doğum ağırlığını etkileyen iki bağımsız değişken olarak düşünülüyor. Bu amaçla 32 anneye ilişkin veriler kullanılarak bu düşüncenin doğruluğu araştırılacaktır.(**Kaynak:** Örnek 7.21, Uygulamalı Çok Değişkenli İstatistiksel Yötntemlere Giriş 1, Reha Alpar, Nobel Yayınevi)

Verideki değişkenlerimiz

$$
\begin{array}{rcl}
y: & \text{doğum ağırlığı} & \\
x_1: & \text{gebelik haftası} & \rightarrow \beta_1 \\
x_2: & \begin{array}{c}\text{sigara içme durumu (Kategorik d.)} \\ \text{İçmiyor: 0 ve İçiyor:1}\end{array} & \rightarrow \beta_2
\end{array}
$$

Kategorik (Nitel) değişken olan annenini sigara içme durumu 2 düzeylidir. Bu nedenle regresyon modeline bu kategorik değişken için **1 (bir) gösterge (göstermelik, dummy) değişken** ekleyeceğiz (yani sadece x_2).

Sigara içmeme durumu **0** ve içme durumu ise **1** olarak tanımlansın. (Tersi de olabilir veya herhangi farklı iki değer de olabilir. Bu değerlerin sayısal olarak bir önemi (anlamı) yoktur.)

Böylece, regresyon modelimiz $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ biçiminde olur.

Eğer $x_2 = 0$ ise regresyon modelimiz $y = \beta_0 + \beta_1 x_1 + \varepsilon$ modeline dönüşür.

Eğer $x_2 = 1$ ise regresyon modelimiz $y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$ modeline dönüşür.

Ayrıca, $x_2 = 0$ ile tanımlanan sigara içmeme kategorisi bu modeldeki diğer durum ile karşılaştırma için **referans kategorisidir.**

Bu veri seti için R programında "lm" fonksiyonunun sonucu aşağıdaki gibi elde edilir.

```
> reg<-lm(y~ x1+x2)  # Tam model: x1 ve x2 kullanilarak elde edilen model
> summary(reg)

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

Residuals:
     Min       1Q   Median       3Q      Max
-223.693  -92.063   -9.365   79.663  197.507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
x1            143.100      9.128  15.677 1.07e-15 ***
x21          -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,    Adjusted R-squared:  0.8892
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15

> anova(reg)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 2895838 2895838 216.962 5.365e-15 ***
x2         1  452881  452881  33.931 2.577e-06 ***
Residuals 29  387070   13347
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bu sonuçlara göre tahmin edlilen regresyon modeli aşağıdaki gibi bulunur.

$$\widehat{y} = -\underbrace{2389.573}_{\widehat{\beta}_0} + \underbrace{143.100}_{\widehat{\beta}_1} x_1 \underbrace{-244.544}_{\widehat{\beta}_2} x_2 \tag{1}$$

Eğer $x_2 = 0$ ise yani **anne sigara içmiyor ise** tahmin edlilen regresyon modeli

$$\widehat{y} = -2389.573 + 143.100 x_1 - 244.544\ (0) = -\underbrace{2389.573}_{\widehat{\beta}_0} + 143.100 x_1 \tag{2}$$

Eğer $x_2 = 1$ ise yani **anne sigara içiyor ise t**ahmin edlilen regresyon modeli

$$\widehat{y} = -2389.573 + 143.100 x_1 - 244.544\ (1) = \underbrace{-2634.117}_{\widehat{\beta}_0 + \widehat{\beta}_2} + 143.100 x_1 \tag{3}$$

biçiminde elde edilir.

(2) ve (3) ile verilen denklemlerin **eğimleri aynıdır** sadece kesim noktaları farklıdır. Bu nedenle bu **regresyon doğruları birbirine paraleldir.**

Ayrıca, $\widehat{\beta}_2$ değeri iki kategorik değişken için bulunan iki regresyon denklemi arasındaki **yükseklik** farkıdır. Açıktır ki (1) denklemi her iki denklemi de içermektedir.

(1) ile verilen regresyon denklemine göre sigara içen annelerin çocuk doğum ağrılığı içmeyenlere göre **244.544 gr daha düşüktür**.

Referans kategorimiz yani 0 olan gösterge değişkeni sigara içmeyen anneler ve $\widehat{\beta}_2 = -244.544$ olduğundan diğer kategori ile arasındaki fark $\widehat{\beta}_2$ kadardır.

Bu model için varyans analizi tablosu aşağıdaki gibi olur.

Değişim kaynağı	Kareler toplamı	Serbestlik derecesi	Kareler ortalaması	F test değeri	$p-$değeri
Regresyon	3348719	2	1674360	$\frac{SSR/2}{SSE/29} = 125.4462$	$5.289x10^{-15}$
Artık	387070	29	13347.24		
Toplam	3735789	31			

$F_h = 125.4462 > F_{2,29,0.05} = 3.33$ veya p-değeri< 0.05 olduğundan $H_0 : \beta_1 = \beta_2 = 0$ **hipotezi reddedilir** yani oluşturulan **regresyon modeli anlamlıdır**.

$R^2 = SSR/SST = 0.8964$ ve $R^2_{adj} = 1 - \dfrac{SSE/(n-p)}{SST/(n-1)} = 0.8892$ olarak bulunur.

Yukarıda verilen R çıktısına göre β_1 ve β_2 regresyon katsayıları için kısmı t-testlerinin değerleri sırasıyla 15.677 ve -5.825 olarak elde edilmiştir. Böylece $\alpha = 0.05$ anlamlılık düzeyinde x_1 ve x_2 bağımsız değişkenlerinin modele anlamlı katkıları olduğu anlaşılır.

Ayrıca, $H_0 : \beta_2 = 0$ ve $H_1 : \beta_2 \neq 0$ için yapılan kısmı t-testinin sonucunun **anlamlı olması sigara içme durumunun doğum ağırlığı üzerindeki etkisinin anlamlı olduğunu** gösterir.

Kategorik x_2 değişkeni modelde olmadığında elde edilen sonuçlar aşağıdaki gibi olur.

```
> summary(lm(y~x1)) #sadece x1 kullanilarak elde edilen model

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-354.03 -115.09   18.07  100.22  263.34

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2037.00     498.11  -4.089 0.000298 ***
x1            130.82      12.86  10.170 3.09e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 167.3 on 30 degrees of freedom
Multiple R-squared:  0.7752,    Adjusted R-squared:  0.7677
F-statistic: 103.4 on 1 and 30 DF,  p-value: 3.085e-11

> anova(lm(y~x1))
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x1         1 2895838 2895838  103.43 3.085e-11 ***
Residuals 30  839951   27998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bu sonuçlara göre tahmin edlilen regresyon modeli $\widehat{y} = \underbrace{-2037}_{\widehat{\beta}_0} + \underbrace{130.82}_{\widehat{\beta}_1} x_1$ biçimindedir ve $R^2_{adj} = 0.7677$ dır.

Örnek 2: Yaş (x_1) ile şiddetli depresyon tedavisinde kullanulan 3 farklı yöntemin (A,B,C), tedavi etkinliği (y) üzerindeki etkisi incelenmek isteniyor. Çalışamanın bir diğer amacı, eğer var ise yaş ile tedavi yöntemi arasındaki etkileşimin etkisini de incelemektir. Bu amaçla, hastalığın tanısına ve şiddetine göre 3 farklı yöntem uygulanan 36 hasta rastgele seçiliyor. (**Kaynak:** Örnek 7.22, Uygulamalı Çok Değişkenli İstatistiksel Yötntemlere Giriş 1, Reha Alpar, Nobel Yayınevi)

```
> data<-print(data.frame(y,x1,x2,x3,x1x2,x1x3))
    y x1 x2 x3 x1x2 x1x3
1  56 21  1  0   21    0
2  41 23  0  1    0   23
3  40 30  0  1    0   30
4  28 19  0  0    0    0
5  55 28  1  0   28    0
6  25 23  0  0    0    0
7  46 33  0  1    0   33
8  71 67  0  0    0    0
9  48 42  0  1    0   42
10 63 33  1  0   33    0
11 52 33  1  0   33    0
12 62 56  0  0    0    0
13 50 45  0  0    0    0
14 45 43  0  1    0   43
15 58 38  1  0   38    0
16 46 37  0  0    0    0
17 58 43  0  1    0   43
18 34 27  0  0    0    0
19 65 43  1  0   43    0
20 55 45  0  1    0   45
21 57 48  0  1    0   48
22 59 47  0  0    0    0
23 64 48  1  0   48    0
24 61 53  1  0   53    0
25 62 58  0  1    0   58
26 36 29  0  0    0    0
27 69 53  1  0   53    0
28 47 29  0  1    0   29
29 73 58  1  0   58    0
30 64 66  0  1    0   66
31 60 67  0  1    0   67
32 62 63  1  0   63    0
33 71 59  0  0    0    0
34 62 51  0  0    0    0
35 70 67  1  0   67    0
36 71 63  0  0    0    0
```

Verideki değişkenlerimiz

y :	tedavi etkinliği	nicel
x_1 :	yaş	nicel
x_2 ve x_3 :	tedavi yöntemi A, B ve C	nitel

olmak üzere kategorik değişkenlerimiz 3-1=2 tane

x_2	x_3	tedavi yöntemi
0	0	C
1	0	A
0	1	B

biçiminde tanımlayalım (Referans yöntem C dir).

2 gösterge değişken eklenerek elde edilen modelimiz

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

biçimindedir. Bu model için R programından elde edilen sonuçlar aşağıdaki gibidir.

Bu sonuçlara göre modelimiz

$$\widehat{y} = 22.29059 + 0.66446 x_1 + 10.25276 x_2 + 0.44518 x_3$$

olarak elde edilir ve $R^2_{adj} = 0.7637$. Bu regresyon denklemine göre;

A yöntemi ile tedavi edilenlerin C yöntemi ile tedavi edilenlere göre tedavi etkinliği puani 10.25276 ($\widehat{\beta}_2$) puan daha fazla iken,

B yöntemi ile tedavi edilenlerin C yöntemi ile tedavi edilenlere göre tedavi etkinliği puani 0.44518 ($\widehat{\beta}_3$) puan daha fazladır. (**Not:** Referans yöntem $x_2 = 0$ ve $x_3 = 0$ durumu olan C yöntemidir.) Böylece,

C yöntemi için regresyon denklemi $x_2 = x_3 = 0 \rightarrow \widehat{y} = 22.29059 + 0.66446 x_1$

A yöntemi için regresyon denklemi $x_2 = 1$ ve $x_3 = 0 \rightarrow \widehat{y} = 32.54335 + 0.66446 x_1$

B yöntemi için regresyon denklemi $x_2 = 0$ ve $x_3 = 1 \rightarrow \widehat{y} = 22.73577 + 0.66446 x_1$

biçiminde bulunur.

Bu durumda 3 yöntem için de elde edilen regresyon denklemleri aynı eğime sahiptir fakat kesim noktaları birbirinden farklıdır. Ancak farklı tedavi yöntemlerinin tedavideki etkilerinin

```
> summary(lm(y~ x1+x2+x3))  # x1,x2,x3 kullanilarak elde edilen model

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5732  -3.3922   0.9829   3.9613   9.5062

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.29059    3.50510   6.359 3.85e-07 ***
x1           0.66446    0.06978   9.522 7.42e-11 ***
x2          10.25276    2.46542   4.159 0.000224 ***
x3           0.44518    2.46399   0.181 0.857763
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.035 on 32 degrees of freedom
Multiple R-squared:  0.784,     Adjusted R-squared:  0.7637
F-statistic: 38.71 on 3 and 32 DF,  p-value: 9.287e-11

> anova(lm(y~ x1+x2+x3))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 3424.4  3424.4 94.0153 4.797e-11 ***
x2         1  803.8   803.8 22.0679 4.775e-05 ***
x3         1    1.2     1.2  0.0326    0.8578
Residuals 32 1165.6    36.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

farklı olması beklenir. Bu nedenle tedavi yöntemlerinin regresyon doğrularının kesim noktası ve eğimde **farklı** olduğunu varsayalım. Bu durumda yaş ile tedavi yöntemleri arasındaki etkileşimi modele ekleyerek bu varsayıma uygun yeni bir model oluşturalım. Etkileşimi ifade eden değişkenler $x_1 * x_2$ ve $x_1 * x_3$ olmak üzere

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + + \beta_4\ x_1 x_2 + \beta_5\ x_1 x_3 + \varepsilon$$

biçiminde yeni bir model oluşturalım.

$\beta_4\ x_1 x_2$ ve $\beta_5\ x_1 x_3$ **etkileşim terimleri** denir. Nitel ve nicel bağımsız değişkenler arasındaki etkileşimi ifade eder.

Bu model için R programından elde edilen sonuçlar aşağıdaki gibidir.

```
> summary(lm(y~ x1+x2+x3+x1x2+x1x3))  # x1,x2,x3,x1*x2 ve x1*x3 kullanilarak elde edilen model

Call:
lm(formula = y ~ x1 + x2 + x3 + x1x2 + x1x3)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4366 -2.7637  0.1887  2.9075  6.5634

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.21138    3.34964   1.854 0.073545 .
x1           1.03339    0.07233  14.288 6.34e-15 ***
x2          41.30421    5.08453   8.124 4.56e-09 ***
x3          22.70682    5.09097   4.460 0.000106 ***
x1x2        -0.70288    0.10896  -6.451 3.98e-07 ***
x1x3        -0.50971    0.11039  -4.617 6.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.925 on 30 degrees of freedom
Multiple R-squared:  0.9143,    Adjusted R-squared:  0.9001
F-statistic: 64.04 on 5 and 30 DF,  p-value: 4.264e-15

> anova(lm(y~ x1+x2+x3+x1x2+x1x3))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
x1         1 3424.4  3424.4 222.2946 2.059e-15 ***
x2         1  803.8   803.8  52.1784 4.857e-08 ***
x3         1    1.2     1.2   0.0772    0.7831
x1x2       1  375.0   375.0  24.3430 2.808e-05 ***
x1x3       1  328.4   328.4  21.3194 6.850e-05 ***
Residuals 30  462.1    15.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sonuçlara göre regresyon denklemi

$$\widehat{y} = 6.21138 + 1.03339x_1 + 41.30421x_2 + 22.70682x_3 - 0.70288\ x_1x_2 - 0.50971\ x_1x_3$$

olarak elde edilir. Bu regresyon denklemine göre;

A yöntemi ile tedavi edilenlerin C yöntemi ile tedavi edilenlere göre tedavi etkinliği puani 41.30421 ($\widehat{\beta}_2$) puan daha fazla iken,

B yöntemi ile tedavi edilenlerin C yöntemi ile tedavi edilenlere göre tedavi etkinliği puani 22.70682 ($\widehat{\beta}_3$) puan daha fazladır.

Böylece,

C yöntemi için regresyon denklemi $x_2 = x_3 = 0 \rightarrow \widehat{y} = 6.21138 + 1.03339x_1$

A yöntemi için regresyon denklemi $x_2 = 1$ ve $x_3 = 0 \rightarrow\ \widehat{y} = 47.51559 + 0.33051x_1$

B yöntemi için regresyon denklemi $x_2 = 0$ ve $x_3 = 1 \rightarrow\ \widehat{y} = 28.9182 + 0.52368x_1$

biçiminde bulunur.

Ayrıca, kısmi t-testlerinin sonucuna göre tüm katsayılar anlamlıdır ve ilk modele göre $R^2_{adj} = 0.7637$ dan $R^2_{adj} = 0.9001$ yükselmiştir.