REGRESYONDA GİZLENMİŞ DIŞ DEĞER BULMA

n sayıda orijinal $x_{i1}, x_{i2}, ..., x_{ik}$ $(i = 1, 2, ..., n)$ veri noktalarının tümünü içeren en küçük dış bükey kümesi, bağımsız değişken kabuğu (RVH) olarak tanımlansın.

$H = X(X'X)^{-1}X'$ matrisinin köşegen elemanları h_{ii} , **gizlenmiş dış değer bulmanın** saptanmasında oldukça kullanışlıdır.

h_{ii}'nin en büyük değerine sahip olan nokta $h_{\max}$ olmak üzere x noktalar kümesi aşağıdaki eşitsizliği sağlar.

$$x'\left(X'X\right)^{-1}x \leq h_{\max}$$

Bu eşitsizlik, RVH'ın içindeki tüm noktaları kapsayan bir elipsoittir.
Eğer $x_0' = \left[1, x_{01}, x_{02}, ..., x_{0k}\right]$ noktasının kestirimi ya da ön kestirimi ile ilgileniliyorsa bu noktanın RVH'daki yeri aşağıda verilen eşitlikte elde edilen noktadır :

$$h_{00} = x_0'\left(X'X\right)^{-1}x_0$$

- $h_{00} > h_{\max}$ olan noktalar RVH'yi kapsayan noktalar dışındadır ve dış değer bulma noktalarıdır.
- $h_{00} < h_{\max}$ ise bu noktalar elipsoidin ve büyük olasılıkla RVH'nin içindedir.

Genel olarak h_{00}'ın çok küçük değerinde, x_0 noktası x uzayının merkezine yaklaşır.

Örnek 2.12 Gizlenmiş Dış Değer Bulma - Teslim Süresi Verileri
TABLO 2.7 Teslim Süresi Verileri İçin h_{ii} Değerleri

Gözlem, i	Teslim hacmi, x_{i1}	Mesafe, x_{i2}	h_{ii}
1	7	560	0.10180
2	3	220	0.07070
3	3	340	0.09874
4	4	80	0.08538
5	6	150	0.07501
6	7	330	0.04287
7	2	110	0.08180
8	7	210	0.06373
9	30	1460	0.49829 = $h_{\max}$
10	5	605	0.19630
11	16	688	0.08613
12	10	215	0.11366
13	4	255	0.06113
14	6	462	0.07824
15	9	448	0.04111
16	10	776	0.16594
17	6	200	0.05943
18	7	132	0.09626
19	3	36	0.09645
20	17	770	0.10169
21	10	140	0.16528
22	26	810	0.39158
23	9	450	0.04126
24	8	635	0.12061
25	4	150	0.06664

9. gözlem en büyük h_{ii} değerine sahip olduğundan bu gözlemin incelenmesi gerekmektedir. Aşağıdaki dört noktada kestirim ve ön kestirimin ele almak istenildiği varsayılsın:

TABLO 2.8 Dış Değer Bulma Noktalarının Belirlenmesi

Nokta	x_{10}	x_{20}	h_{00}
a	8	275	0.05346
b	20	250	0.58917
c	28	500	0.89874
d	8	1200	0.86736

a noktası için $h_{00} = 0.05346 < h_{\max} = 0.49829$ olduğundan bu nokta bir ara değer bulma noktasıdır. Geriye kalan b, c ve d noktalarının tümü dış değer bulma noktalarıdır.
($h_{00} > h_{\max}$)

STANDARTLAŞTIRILMIŞ REGRESYON KATSAYILARI

Genellikle regresyon katsayılarını doğrudan karşılaştırmak β_j'nin büyüklüğünün x_j bağımsız değişkeninin ölçüm birimini yansıtması sebebiyle zordur.

β_j regresyon katsayısının birimi, "y'nin birimi / x_j'nin birimi"dir. Bu nedenle zaman zaman bağımsız ya da yanıt değişkenleri ölçeklendirerek **"birimsiz regresyon katsayıları"** oluşturulur. Birimsiz regresyon katsayıları çoğunlukla **"standartlaştırılmış regresyon katsayıları"** olarak adlandırılır.

- **Birim Normal Ölçekleme**

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{s_j} \ , \ i = 1,2,...,n \quad , \quad j = 1,2,...,k \qquad \textbf{(2.35)}$$

ve

$$y_i^* = \frac{y_i - \overline{y}}{s_y} \quad , \quad i = 1,2,...,n \qquad \textbf{(2.36)}$$

olmak üzere x_j bağımsız değişkeninin varyansı,

$$s_j{}^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2}{n-1}$$

olup yanıt değişkeninin örneklem varyansı ise

$$s_y{}^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

olarak kullanılır.
Bu yeni değişkenler kullanılarak regresyon modeli,

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + ... + b_k z_{ik} + \varepsilon_i \quad , \quad i = 1,2,....,n \qquad \textbf{(2.37)}$$

olup bağımsız ve yanıt değişkenlerinin $\overline{x}_j$ ve $\overline{y}$'den çıkartılarak merkezileştirilmesi, kesim noktasını modelden kaldırır. (b_0'ın en küçük kareler kestirimi, $\hat{b} = \overline{y}^* = 0$'dır)

$\boldsymbol{b}$ vektörünün en küçük kareler kestiricisi,

$$\hat{\boldsymbol{b}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}^{*} \qquad \textbf{(2.38)}$$

- **Birim Uzunlukta Ölçekleme**

Bu ölçekleme,

$$w_{ij} = \frac{x_{ij} - \overline{x}_j}{\sqrt{s_{jj}}}, \qquad i = 1,2,...,n \;\; , \;\; j = 1,2,...,k \qquad \textbf{(2.39)}$$

ve

$$y_i^{\,0} = \frac{y_i - \overline{y}}{\sqrt{SS_T}} \quad , \quad i = 1,2,...,n \qquad \textbf{(2.40)}$$

olarak verilir.

Burada, $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$ toplamı, x_j bağımsız değişkenleri için düzeltilmiş kareler toplamıdır. Bu ölçeklendirmede, her bir yeni w_j bağımsız değişkeni, $\overline{w}_j = 0$ ortalamaya ve $\sqrt{\sum_{i=1}^{n}\left(w_{ij} - \overline{w}_j\right)^2} = 1$ uzunluğuna sahiptir.

Bu değişkenler ile oluşturulan regresyon modeli,

$$y_i^{\,0} = b_1 w_{i1} + b_2 w_{i2} + ... + b_k w_{ik} + \varepsilon_i \;\; , \;\; i = 1,2,...,n$$

olup en küçük kareler regresyon vektörü,

$$\hat{\boldsymbol{b}} = \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'\boldsymbol{y}^{\boldsymbol{0}} \qquad \textbf{(2.41)}$$

olarak kullanılır.

Birim uzunluk ölçeklemesinde, $\boldsymbol{W}'\boldsymbol{W}$ matrisi bir **korelasyon matrisidir.**

$$W'W = \begin{bmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{12} & 1 & r_{23} & \dots & r_{2k} \\ r_{13} & r_{23} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ r_{1k} & r_{2k} & r_{3k} & \dots & 1 \end{bmatrix}$$

Burada,

$$r_{ij} = \frac{\sum_{i=1}^{n}(x_{ui} - \bar{x}_i)(x_{uj} - \bar{x}_j)}{\sqrt{S_{ii}S_{jj}}} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

x_i ve x_j bağımsız değişkenleri arasındaki **basit korelasyondur**. Benzer şekilde $W'y^0$,

$$W'y^0 = \begin{bmatrix} r_{1y} \\ r_{2y} \\ r_{3y} \\ \dots \\ r_{ky} \end{bmatrix}$$

olup burada,

$$r_{iy} = \frac{\sum_{u=1}^{n}(x_{uj} - \bar{x}_j)(y_u - \bar{y})}{\sqrt{S_{jj}SS_T}} = \frac{S_{jy}}{\sqrt{S_{jj}SS_T}}$$

x_j bağımsız değişkeni ve y yanıt değişkeni arasındaki basit korelasyondur.

*** Eğer birim normal ölçekleme kullanılırsa $Z'Z$ matrisi, $W'W$ matrisiyle,

$$Z'Z = (n-1)W'W$$

biçiminde yakından ilişkilidir.

Her iki yöntem de aynı birimsiz $\hat{b}$ regresyon katsayıları kümesini verir. $\hat{b}$ regresyon katsayıları çoğunlukla **standartlaştırılmış regresyon katsayıları** olarak adlandırılır.

Orijinal ve standartlaştırılmış regresyon katsayıları arasındaki ilişki,

$$\hat{\beta}_j = \hat{b}_j\sqrt{\frac{SS_T}{S_{jj}}} \quad , \quad j = 1,2,...,k \qquad \textbf{(2.42)}$$

olup ayrıca $\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j$ olarak kullanılır.

Birçok bilgisayar programı, $(X'X)^{-1}$ matrisindeki yuvarlama hatalarından kaynaklanan problemleri azaltmak için bu ölçeklendirmeyi kullanır. b_j, diğer x_i, $i \neq j$ bağımsız değişkenleri modelde iken x_j, bağımsız değişkenin etkisini ölçtüğünden bağımsız değişkenin değer aralığından etkilenmektedir.

Sonuç olarak, x_j bağımsız değişkeninin göreli öneminin bir ölçüsü olarak b_j'nin büyüklüğünü kullanmak yanıltıcı olabilir.

Örnek 2.13 Teslim Süresi Verileri

$$SS_T = 5784.5426 \qquad S_{11} = 1136.5600$$
$$S_{1y} = 2473.3440 \qquad S_{22} = 2,537,935.0330$$
$$S_{2y} = 108,038.6019 \qquad S_{12} = 44,266.6800$$

olmak üzere birim uzunlukta ölçekleme kullanılarak,

$$r_{12} = \frac{S_{12}}{\sqrt{S_{11}S_{22}}} = \frac{44,266.6800}{\sqrt{(1136.5600)(2,537,935.0303)}} = 0.824215$$

$$r_{1y} = \frac{S_{1y}}{\sqrt{S_{11}SS_T}} = \frac{2473.3440}{\sqrt{(1136.5600)(5784.53426)}} = 0.964615$$

$$r_{2y} = \frac{S_{2y}}{\sqrt{S_{22}SS_T}} = \frac{108,038.6019}{\sqrt{(2,537,935.0303)(5784.5426)}} = 0.891670$$

elde edilir.

Korelasyon matrisi,

$$W'W = \begin{bmatrix} 1 & 0.824215 \\ 0.824215 & 1 \end{bmatrix}$$

olup standartlaştırılmış regresyon katsayıları,

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.824215 \\ 0.824215 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0.964615 \\ 0.891670 \end{bmatrix} = \begin{bmatrix} 0.716267 \\ 0.301311 \end{bmatrix}$$

olarak bulunur.

Kestirim modeli, $\hat{y}^0 = 0.716267 w_1 + 0.301311 w_2$ olup standartlaştırılmış w_1 bir birim arttığında standartlaştırılmış $\hat{y}^0$ süre değeri 0.716267 birim artar. Standartlaştırılmış w_2, bir birim arttığında ise standartlaştırılmış $\hat{y}^0$ süre değeri 0.301311 birim artar.

Standartlaştırılmış değişkenler yoluyla teslim süresi üzerinde, teslim edilen ürün miktarının mesafeden çok daha önemli bir etkiye sahip olduğu görülmektedir. Eğer teslim miktarı ve mesafe değerlerinin farklı bir aralığı ile başka bir örneklem kullanılsaydı bu bağımsız değişkenlerin önemleri ile ilgili farklı sonuçlar elde edilebilirdi.