REGRESYON ANALİZİ

İki değişken arasında bir ilişki olduğunda, bu ilişki dağılım grafiğindeki noktalar arasından geçen uygun bir doğru ile tanımlanabilir. Bu doğruya regresyon doğrusu denir ve matematiksel olarak bir denklem ile gösterilebilir. Bu denkleme de regresyon denklemi denir. Bu denklem yardımı ile bağımsız değişkene verilen herhangi bir değere karşı bağımlı değişkenin alacağı değerler hesaplanabilir. Diğer bir deyişle sonuç- sebep arasındaki bağlantının denklem ifadesidir.

İlişkiler ikiye ayrılabilir:

1. Yapısal (structural) İlişkiler: İstatistiği vardır ama genelde kullanılmaz. Eğer bağımlı (Y) ve bağımsız değişken (X) hata içeriyorsa bu, yapısal ilişkiye girer. Formülü şu şekildedir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i + \delta_i$$

2. Fonksiyonel (functional) İlişkiler: Genelde kullanılan ilişkilerdir. Buradaki denklem tek hatayı içerir. Bağımlı değişken (Y) hata içerirken, bağımsız değişken (X) sabit değerdir. Formülü şu şekildedir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Regresyon analizinde bağımsız “X” değerlerine karşı bağımlı “Y” değerleri bir grafik üzerinde gösterilir. Her “X” üzerinde “Y” nin belli bir dağılışı vardır. Regresyon doğrusu noktaların tam ortasından geçer. Noktaların uzaklıkları minimumdur. İki değişken arasında tam bir ilişki varsa ($r = \pm 1$) dağılımdaki bütün noktalar regresyon doğrusu üzerine düşer. Böyle bir durumda iki değişken arasındaki gerçek ilişki regresyon denklemi ile ifade edilen ilişki ile uygunluk sağlar. Ancak iki değişken arasındaki ilişki nadiren tam bir ilişkidir. Çoğu kez regresyon doğrusu tüm noktaların üzerinden geçemez ve dağılımı yaklaşık olarak temsil edebilir.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

$$\sum_{i=1}^{n} (\varepsilon_i)^2 = \sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

a) $\dfrac{\delta \sum_{i=1}^{n} \varepsilon_i^{\,2}}{\delta \beta_0} = 0$ formülünün türevi alındığında

$\sum Y_i = nb_0 + b_1 \sum X_i$ eşitliği elde edilir.

b) $\dfrac{\delta \sum \varepsilon_i^{\,2}}{\delta \beta_1} = 0$ formülünün türevi alındığında

$\sum Y_i X_i = b_o \sum X_i + b_1 \sum X_i^{\,2}$ eşitliği elde edilir.

__NOT__: $b_0 = \hat{\beta}_0$ **ve** $b_1 = \hat{\beta}_1$ **. Bu iki eşitlik ile** β_0 **ve** β_1 **'in tahminleri elde edilmiş olur. Böylece iki bilinmeyenli bir denklem oluşur**.

$\hat{Y}_i = \hat{E}\left(Y/X = X_i\right) = \hat{\beta}_0 + \hat{\beta}_1 X_i$

$\hat{e}_i = Y_i - \hat{E}\left(Y/X = X_i\right) = Y_i - \hat{Y}i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1\right)$

$e_i = Y_i - \left(\beta_0 + \beta_1 X_i\right)$

$\sum (e_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 =$ Hata Kareler Toplamı

SXX= $\sum X_i^{\,2} - \dfrac{\left(\sum X\right)^2}{n}$

SYY= $\sum Y_i^{\,2} - \dfrac{\left(\sum Y\right)^2}{n}$

SXY= $\sum XY - \dfrac{\sum X \cdot \sum Y}{n}$

$\hat{\beta}_1 = \dfrac{S_{XY}}{S_{XX}}$ ve $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

“y”deki değişkenlik **Genel Kareler Toplamı** ile ifade edilir. Bu değişkenliğin “x” tarafından izah edilen kısmı **Hata Kareler Toplamı**; “x” tarafından izah edilemeyen kısmı **Regresyon Kareler Toplamı** olarak ifade edilir.

Genel Kareler Toplamı (GKT) $= \sum (Y_i - \overline{Y})^2$ **= S_{YY}**

Hata Kareler Toplamı (HKT) $= \sum \left(Y_i - \hat{Y}_i\right)^2$ **= S_{YY} - $(S_{XY})^2 / S_{XX}$**

Regresyon Kareler Toplamı (RKT) $= \sum \left(\hat{Y}_i - \overline{Y}\right)^2$ **= $(S_{XY})^2 / S_{XX}$**

Tablo-1

Varyasyon Kaynağı	Serbestlik derecesi	Kareler Toplamı	Kareler Ortalaması
Regresyon	p=1	$(S_{XY})^2 / S_{XX}$	$(S_{XY})^2 / S_{XX}$ /p=1
Hata **(Rezidüel)**	n-p-1	S_{YY}- $(S_{XY})^2 / S_{XX}$	S_{YY}-$(S_{XY})^2 / S_{XX}$ /n-p-1= $\mathbf{S^2}$
Genel	n-1	S_{YY}	S_{YY} / n-1

HKT = Rezidüel Kareler Toplamı
$\mathbf{S^2}$ = Hata Kareler Ortalaması

Varyans ve “t” Değerleri

- β_1 için

$$V\left(\hat{\beta}_1\right) = S_{\hat{\beta}_1}{}^2 = \frac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2} = \frac{S^2}{S_{XX}}$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S_{\hat{B}_1}{}^2}} \approx t_{n-p-1}$$

Bulunan t değeri parantez içinde verilen (n-p-1=serbestlik derecesi; α/2=olasılık)tablo t değeriyle kıyaslanır.

- β_0 için

$$V(\hat{\beta}_0) = S_{\hat{\beta}o}{}^2 = \frac{S^2}{n} + \frac{(\overline{X})^2 S^2}{S_{XX}}$$

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{S_{\hat{B}_0}{}^2}} \approx t_{n-p-1}$$

Bulunan t değeri parantez içinde verilen (n-p-1=serbestlik derecesi; α/2=olasılık)tablo t değeriyle kıyaslanır.

Tahmini "y" Değeri ve % 95 Güven Aralığı

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ iken X_k için $\hat{Y}_k$ tahmin değeridir.

$$\hat{Y}_k = \overline{Y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 X_k$$

$$\hat{Y}_k = \overline{Y} + \hat{\beta}_1 (X_k - \overline{X})$$

$$V(\hat{y}_k) = V(\overline{Y}) + (X_k - \overline{X})^2 \cup (\hat{\beta}_1)$$

$$V(\hat{y}_k) = \frac{\sigma^2}{n} + \frac{(x_k - \overline{x})^2}{\sum (x_i - \overline{x})^2} \sigma^2$$

$$V(\hat{y}_k) = \frac{S^2}{n} + \frac{(x_k - \overline{x})^2}{SXX} S^2$$

$$V(\hat{Y}_k) = S^2 \left(\frac{1}{n} + \frac{(X_k - \overline{X})^2}{S_{XX}} \right)$$

Her X değeri için $\hat{Y}_i$ güven aralığını hesaplatmak için aşağıdaki formül kullanılır.

$$\hat{Y}_k \pm t \cdot \sqrt{V\left(\hat{Y}_k\right)} \quad (t_{n-1,\, 0{,}025})$$

Standardize B

"X" ve "Y" lerin standardize edilmesiyle elde edilen katsayıdır**.** Bunun için tüm "X" ve "Y" lerden $\overline{X}$ ve $\overline{Y}$ çıkarılır. Bu işlemlerden sonra bir de S_x ve S_y' ye bölünürse daha da standardize edilir.

$$Z_x = \frac{\left(X_i - \overline{X}\right)}{S_x} = \frac{\left(X_i - \overline{X}\right)}{S_{XX} / n-1} \qquad S_x = \frac{S_{XX}}{n-1}$$

$$Z_y = \frac{\left(Y_i - \overline{Y}\right)}{S_y} = \frac{\left(Yi - \overline{Y}\right)}{S_{YY} / n-1} \qquad S_y = \frac{S_{YY}}{n-1}$$

Korelasyon

Korelasyonda bağımlı ve bağımsız değişken yoktur."X" ve "Y"nin birlikte değişimi vardır. Karşılıklı bağımlılık var. Formülü aşağıdaki gibidir.

$$\ell = \frac{Kov(X, Y)}{\sqrt{V(X) \cdot V(Y)}}$$

$$Kov(X, Y) = \frac{S_{XY}}{n-1}, \qquad V(X) = \frac{S_{XX}}{n-1}, \qquad V(Y) = \frac{S_{YY}}{n-1}$$

Kovaryansın tahmini değeri $\hat{\ell} = r$

$$r = \frac{\frac{S_{XY}}{n-1}}{\sqrt{\frac{S_{XX}}{n-1} \cdot \frac{S_{YY}}{n-1}}} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$$

$R^2 = r^2$ = Belirtme Katsayısı (determination coefficient). "y" deki değişkenliğin ne kadarının "x" tarafından izah edilebileceğini gösterir.

r^2: İkili karşılaştırmalarda $\qquad$ R^2: Çoklu karşılaştırmalarda kullanılır

$$R^2 = r^2 = \frac{S_{XY}^{\ 2}}{S_{XX} \cdot S_{YY}} = \frac{\text{Regresyon Kareler Toplamı}}{\text{Genel Kareler Toplamı}}$$

$$t = \frac{r-\ell}{S_r} \approx t_{n-2,\alpha/2}$$

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

$$t = \frac{(r-\ell)\sqrt{n-2}}{\sqrt{1-r^2}}$$

- "n" sayısı büyük olduğunda, t büyük çıkar. Ama asıl önemli olan $R^2 = r^2$'nin büyük olmasıdır.
- Regresyon korelasyon ilişkisine bakıldığında; regresyonun önemli saptandığı durumlarda korelasyonla da bunun ispatlandığı görülür.
- r = korelasyon katsayısı= b_{xy} ile b_{yx}'in geometrik ortalamasıdır.
- $b_{yx=}$ X bağımsız, Y bağımlı değişken iken ($\hat{\beta}_1$)
- $b_{xy=}$ Y bağımsız, X bağımlı değişken iken ($\hat{\beta}_1$)
- $r = \sqrt{b_{yx} \cdot b_{xy}}$ şeklinde de bulunur.

Düzeltilmiş R Kare (Adjusted R Square)

“n” sayısı küçük olduğu durumlarda bu değer önem kazanır. “n” büyük olduğu durumlarda ise düzeltilmiş ve düzeltilmemiş R^2 değerleri birbirine yakındır.

$$R_{adj}^2 = r_{adj}^2 = 1 - \frac{HataKT}{GenelKT} = 1 - \frac{HataKO}{GenelKO}$$

$$R_{adj}^2 = r_{adj}^2 = 1 - \frac{HKT/n-2}{GKT/n-1} = 1 - \frac{HKT(n-1)}{GKT(n-2)}$$

$$R_{adj}^2 = r_{adj}^2 = 1 - \frac{HKT}{GKT} \cdot \frac{n-1}{n-2}$$