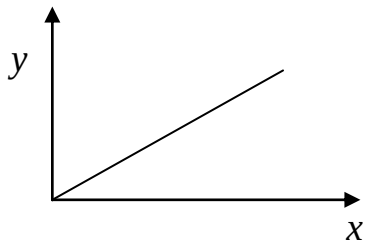
REGRESYON

İki ya da daha çok değişkenin yer aldığı istatistiksel modellerde genellikle sebep-sonuç ilişkisi üzerinde durulur. Yani, değişkenlerden biri ya da bir kaçının, diğer bir ya da birkaç değişkeni ne ölçüde etkilediği incelenir. Eğer değişkenler arsında ilişki varsa, ilişkinin derecesi ve fonksiyonel şekli belirlenmeye çalışılır. İlgilenilen olayı tanımlayan rasgele değişken bağımlı değişken, bu olayla ilgili ya da olayı etkileyen ise bağımsız değişken olarak tanımlanır. Y ile bağımlı değişken, X ile bağımsız değişken gösterilmek üzere, iki yada daha çok değişken arasındaki ilişkinin yapısı regresyon çözümlemesi, ilişkinin yönü ve derecesi ise korelasyon çözümlemesi ile incelenir.

Basit Doğrusal Regresyon Çözümlemesi

$X,(x_1,\ldots,x_n)$ değerlerini alan ve $Y,(y_1,\ldots,y_n)$ değerlerini alan iki rasgele değişken olsun. Bu iki değişken arasındaki ilişki, doğrusal regresyon çözümlemesi ile incelenebilir.

X rasgele değişkeni haftalık çalışma saatini, Y rasgele değişkeni öğrencinin başarısını göstermek üzere n tane öğrencinin haftalık çalışma saatleri ile notları $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ ikilileri ile gözlensin. Bu ikililerin koordinat düzlemi üzerinde serpilme diyagramları çizilerek bu verilere nasıl bir eğrinin uyduğu görülebilir. Eğer haftalık çalışma saati artıkça, başarının da artacağı düşünülürse bu iki değişken arasında doğrusal bir ilişki vardır denir.



X ile Y arasındaki gerçek bağıntı,

$$Y=\beta_0+\beta_1X+\varepsilon$$

doğru denklemi ile gösterilir.

β_0 : doğrunun y eksenini kestiği nokta
β_1 : doğrunun eğimi
ε : gerçek hata

} β_0 ve β_1 bilinmeyen regresyon katsayılarıdır

Kitleden seçilen n birimlik örneklem için doğrusal regresyon denklemi,

$$y_j=b_0+b_1x_j+e_j,\ j=1,2,...,n$$

biçiminde tanımlanır. Bilinen (verilen) bir x_j değeri için y_j değeri tahmin edilir. Tahmini doğrusal regresyon denklemi,

$y_j = b_0 + b_1 x_j,\ j = 1,2,...,n$

ile gösterilir. Genel olarak,

$y = b_0 + b_1 x$ ⟶ Bu denkleme X üzerinde Y nin regresyonu denir.

$y = b_0 + b_1 x$ regresyon doğrusu $E(Y|X) = \beta_0 + \beta_1 x$ kitle regresyon doğrusunun bir tahmindir.

y_j : j. gözleme ilişkin gerçek y değeri

y_j : j. gözleme ilişkin y_j ' nin tahmin değeri

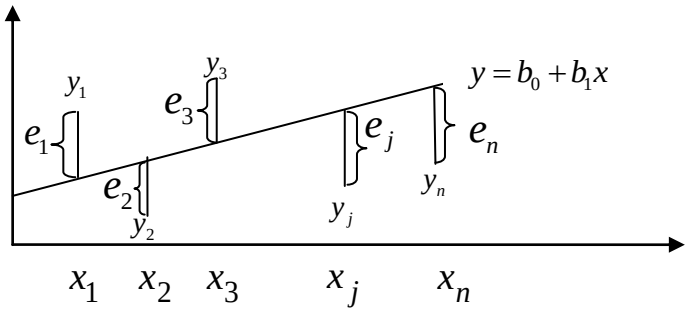
x_j : j. gözleme ilişkin bağımsız değişkenin alacağı değer

$b_0 : \beta_0$ 'ın tahmini (regresyon doğrusunun y eksenini kestiği noktayı gösterir)

$b_1 : \beta_1$ ' in tahmini(regresyon katsayısıdır, doğrunun eğimini gösterir)

e_j : j. gözlemin hata terimidir, $e_j = y_j - y_j$, $e \sim N(0, \sigma^2)$

β_0 ve β_1 Parametreleri için En Küçük Kareler Tahmin Edicileri



$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - y_j)^2 = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$$

$\min \sum_{j=1}^{n} e_j^2$ ⟶ Hata kareleri toplamının minimum olması istenir

$$\frac{\partial \sum_{j=1}^{n} e_j^2}{\partial b_0} = -2 \sum_{j=1}^{n} \left(y_j - b_0 - b_1 x_j \right) = 0$$

$$\sum_{j=1}^{n} y_j = n b_0 + b_1 \sum_{j=1}^{n} x_j \quad (1)$$

$$\frac{\partial \sum_{j=1}^{n} e_j^2}{\partial b_1} = -2 \sum_{j=1}^{n} x_j \left(y_j - b_0 - b_1 x_j \right) = 0$$

$$\sum_{j=1}^{n} x_j y_j = b_0 \sum_{j=1}^{n} x_j + b_1 \sum_{j=1}^{n} x_j^2 \quad (2)$$

(1)'i $-\sum_{j=1}^{n} x_j$, (2) n ile çarpılıp toplanırsa,

$$-\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j = -nb_0 \sum_{j=1}^{n} x_j - b_1 \left(\sum_{j=1}^{n} x_j \right)^2$$

$$n\sum_{j=1}^{n} x_j y_j = nb_0 \sum_{j=1}^{n} x_j + nb_1 \sum_{j=1}^{n} x_j^2$$

$$n\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j = nb_1 \sum_{j=1}^{n} x_j^2 - b_1 \left(\sum_{j=1}^{n} x_j \right)^2$$

$$= b_1 (n\sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2)$$

$$b_1 = \frac{n\sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j}{n\sum_{j=1}^{n} x_j^2 - (\sum_{j=1}^{n} x_j)^2} = \frac{\sum_{j=1}^{n} x_j y_j - \frac{\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j}{n}}{\sum_{j=1}^{n} x_j^2 - \frac{(\sum_{j=1}^{n} x_j)^2}{n}}$$

(1) denkleminden,

$$\sum_{j=1}^{n} y_j - b_1 \sum_{j=1}^{n} x_j = nb_0$$

$$b_0 = \frac{\sum_{j=1}^{n} y_j}{n} - b_1 \frac{\sum_{j=1}^{n} x_j}{n} = \bar{y} - b_1 \bar{x} \Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

olarak bulunur.

$$y = b_0 + b_1 x$$

$b_1 > 0 \Rightarrow$ iki değişken birlikte artıyor yada birlikte azalıyor

$b_1 < 0 \Rightarrow$ değişkenlerden biri artarken diğeri azalıyor

Açıklanan ve Açıklanamayan Değişim

Parametrelerle ilgili sonuç çıkarımına geçmeden önce regresyon analizinin varsayımları gözden geçirilsin,

Varsayım 1. X ve Y arasında doğrusal bir ilişki olduğunda verilen her x_i değeri için ε hata terimi ortalaması 0 olan bir rastgele değişkendir. Yani $E(\varepsilon_i) = 0$'dır.

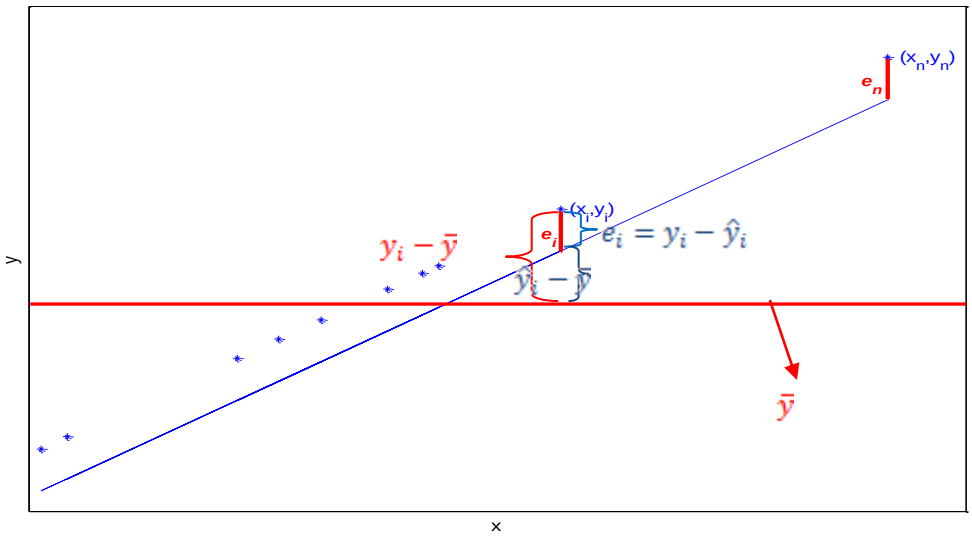
Varsayım 2. Verilen her x_i değeri için ε_i hata terimlerinin varyansı aynıdır. Yani, $Var(\varepsilon_i) = \sigma^2$'dir.

Varsayım 3. Hatalar birbirinden bağımsızdır.

Varsayım 4. Verilen her x_i değeri için ε hata teriminin dağılımı normaldir. Yani $\varepsilon_i \sim N(0, \sigma^2)$'dir.

En küçük kareler yöntemi ile model parametreleri tahmin edildikten sonra artıklar model hatalarının gerçekleşmiş değerleri olarak görüldüğü için bu artık değerler kullanılarak sabit varyans ve ilişkisiz hata varsayımlarının sınanması, artıkların bu özelliklere sahip bir dağılımdan alınan rastgele örneklem olup olmadığının incelenmesi gerekir. Varsayımların açıkça bozulumu, kararsız bir modeli ortaya çıkarabilir. Bundan kaçınmak için artık analizi yapılmalıdır.

Parametrelerle ilgili sonuç çıkarımı yapabilmek için ***"açıklanan değişim"*** ve ***"açıklanamayan değişim"*** kavramlarına bakılsın.



Grafik. Regresyon doğrusu etrafındaki değişim

1) Örneklem ortalaması etrafındaki değerlerin değişimi

$\sum_{j=1}^{n}(y_j - \bar{y})^2$: toplam değişim yada genel kareler toplamı (GnKT)

2) Regresyon doğrusu etrafındaki değişim

$\sum_{j=1}^{n}(y_i - y_j)^2$: açıklanamayan değişim yada hata kareler toplamı (HKT)

3) Ortalama etrafındaki tahmini değerlerin değişimi

$\sum_{j=1}^{n}(y_j-\bar{y})^2$:açıklanan değişim yada regresyon kareler toplamı (RKT)

$$\sum_{j=1}^{n}(y_i-\bar{y})^2=\sum_{j=1}^{n}(y_i-y_j)^2+\sum_{j=1}^{n}(y_i-y_j)^2$$

Toplam değişim=Açıklanamayan değişim+Açıklanan değişim

GnKT=HKT+RKT

$$\text{GnKT}=\sum_{j=1}^{n}y_j{}^2-\frac{(\sum_{J=1}^{N}y_j)^2}{n},\ RKT=\frac{(\sum_{j=1}^{n}x_jy_j-\frac{\sum_{j=1}^{n}x_j\sum_{j=1}^{n}y_j}{n})^2}{\sum_{j=1}^{n}x_j^2-\frac{(\sum_{j=1}^{n}x_j)^2}{n}},\ HKT=GnKT-RKT$$

$$R_{sd}=2-1=2, Gn_{sd}=n-1, H_{sd}=Gn_{sd}-R_{sd}=n-1-1=n-2$$

Her bir kare toplamının kendi serbestlik derecesine bölümü ile kareler ortalamaları bulunur.

$$RKO=\frac{RKT}{R_{sd}}=\frac{RKT}{1}=RKT$$

$$HKO=\frac{HKT}{H_{sd}}=\frac{HKT}{N-2}=S^2=\sigma^2 \longrightarrow$$ X üzerinde Y nin regresyon doğrusunun varyansının yansız tahmin edicisi

Belirtme Katsayısı

Belirtme katsayısı açıklanan değişimin toplam değişime oranıdır. Bağımsız değişkeninin bağımlı değişkendeki değişimin yüzde kaçını açıkladığını gösterir. R^2 ile gösterilir.

$$R^2=\frac{Açıklanan\ Değişim}{Toplam\ Değişim}=\frac{RKT}{GnKT}$$

$1-R^2$:toplam değişimin açıklanamayan yüzdesi

Basit Doğrusal Regresyonda Hipotez Testleri

β_1 için Hipotez Testi

$y = \beta_0 + \beta_1 x$

1) $H_0 : \beta_1 = \beta_{1,0}$
$H_1 : \beta_1 \neq \beta_{1,0}$

$\varepsilon_j \sim N(0, \sigma^2)$ olduğu biliniyor. Buna göre, $y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ olup $y_j \sim N(\beta_0 + \beta_1 x_j, \sigma^2)$ dir.

$$b_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

2) H_0 hipotezinin doğruluğu altında test istatistiği;

$t_h = \frac{b_1 - \beta_{1,0}}{S_{b_1}}$, $S_{b_1} \longrightarrow b_1$ in standart hatası, $S_{b_1} = \sqrt{\frac{HKO}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

3) $|t_h| > t_t$ ise H_0 red edilir ($t_t = t_t(\alpha/2, n-2)$).

β_0 için Hipotez Testi

1) $H_0 : \beta_0 = \beta_{0,0}$
$H_1 : \beta_0 \neq \beta_{0,0}$

2) H_0 hipotezinin doğruluğu altında test istatistiği;

$t_h = \frac{b_0 - \beta_{0,0}}{S_{b_0}}$, $S_{b_0} \longrightarrow b_0$ in standart hatası, $S_{b_0} = \sqrt{HKO(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}$

3) $|t_h| > t_t$ ise H_0 red edilir ($t_t = t_t(\alpha/2, n-2)$).

Regresyon Doğrusunun Anlamlılık Testi

1) $H_0 : \beta_1 = 0$ (Regresyon doğrusu önemsizdir)
$H_1 : \beta_1 \neq 0$ (Regresyon doğrusu önemlidir)

2) $t_h = \frac{b_1}{S_{b_1}}$, $S_{b_1} = \sqrt{\frac{HKO}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

3) $|t_h| > t_t$ ise H_0 red edilir ($t_t = t_t(\alpha/2, n-2)$).

H_0 hipotezi red edildiğinde regresyon doğrusunun anlamlı olduğu söylenebilir. H_0 hipotezi red edilemediğinde regresyon doğrusu anlamsızdır. İki değişken arasında doğrusal bir ilişki olmadığı söylenir. Bu hipotez F testi ile de (varyans çözümlemesi) yapılabilir.

F Testi

" Deneysel noktaların doğrusal regresyona uyumu önemsizdir" ya da "Deneysel noktalar regresyon doğrusu ile gösterilemez" şeklinde yorumlanabilen H_0 hipotezi, $H_0 : \beta_1 = 0$ olarak kurulur.

1) $H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

2) Bu hipotezi test etmek amacıyla varyans analizi tablosu hazırlanır.

Değişim Kaynakları (DK)	Serbestlik Derecesi (Sd)	Kareler Toplamı (KT)	Kareler Ortalaması (KO)	Test
Regresyon	1	RKT	$RKO = RKT/1$	$F_h = \frac{RKO}{HKO}$
Regresyondan ayrılış	$n-2$	$HKT = GnKT - RKT$	$HKO = HKT/n-2$	
Toplam	$n-1$	$GnKT$	-	

3) $F_h > F_t(\alpha, 1, n-2)$ ise H_0 hipotezi red edilir.

Basit Doğrusal Regresyonda Aralık Tahmini

$\varepsilon_j \sim N(0, \sigma^2)$

$$\frac{\beta_1 - \beta_1}{S_{\beta_1}} \sim t_{n-2} \qquad \frac{\beta_0 - \beta_0}{S_{\beta_0}} \sim t_{n-2}$$

1. β_1 için Güven aralığı

$P(b_1 - t_t S_{b_1} \leq \beta_1 \leq b_1 + t_t S_{b_1}) = 1 - \alpha$
$t_t = t_t(\alpha/2, n-2)$

2. β_0 için Güven aralığı

$P(b_0 - t_t S_{b_0} \leq \beta_0 \leq b_0 + t_t S_{b_0}) = 1 - \alpha$
$t_t = t_t(\alpha/2, n-2)$

3. Bilinen bir x_0 değerine karşılık y değerinin ortalaması için güven aralığı verilebilir. x_0 bilindiğinde y değerinin ortalamasını tanımlayan ifade $E(Y|x_0)$ ile gösterilir. Buna göre, $E(Y|x_0)$ için güven aralığı;

$$P(y_0 - t_t S_{y_0} \le E(Y|x_0) \le y_0 + t_t S_{y_0}) = 1-\alpha$$

$y_0 = b_0 + b_1 x_0$ dan hesaplanan değerdir.

$$S_{y_0} = \sqrt{HKO(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}\ ,\ t_t = t_t\left(\alpha/2, n-2\right)$$

4. Bilinen bir x_0 değerine karşılık y nin yeni ya da gelecekteki değerini tahmin etmek regresyon çözümlemesinde önemlidir. X rasgele değişkeninin x_0 gibi bir değeri verildiğinde Y rasgele değişkeninin y_0 gibi özel bir değeri için aralık tahmini verilebilir.

$(y_0 - y_0)$ rasgele değişkeni ortalaması sıfır, standart sapması $S_{y_0 - y_0}$ olan normal dağılıma sahiptir.

$$S_{y_0 - y_0} = \sqrt{HKO(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})}$$

y_0 ' ın güven aralığı;

$$P(y_0 - t_t S_{y_0 - y_0} \le y_0 \le y_0 + t_t S_{y_0 - y_0}) = 1-\alpha$$

$y_0 = b_0 + b_1 x_0$

$t_t = t_t\left(\alpha/2, n-2\right)$