REGRESYON ANALİZİ-II

•Bu bölümde tahmin edilen anakütle regresyon parametrelerinin istatistiksel olarak anlamlılığını sınayacağız.

1.1 REGRESYON DOĞRUSUNUN ÖZELLİKLERİ VE HATA (ARTIK) KAVRAMI

Regresyon doğrusu denklemi $E(Y_i) = \beta_0 + \beta_1 X_i$ olmak üzere burada β_0, başlangıç terimidir, modelin sabit değeri olarak da bilinir. X=0 iken

$E(Y_i)$'nin değeridir. β_1 doğrunun eğimine karşılık gelir. X'deki bir birim değişmenin $E(Y_i)$ 'deki değişimine karşılık gelir. Burada β_0 ve β_1 parametrelerdir. Bu parametlere veri üzerinden b_0 ve b_1 tahminleriyle ulaşılır.

Y, bağımlı değişkenin elde edilen her bir gözlemi (Y_i) ana kütle ortalaması $E(Y_i)$ olan bir ana kütleden gelen rastlantı değişkeni olduğu varsayılır. Yi gözleminin $E(Y_i)$' den sapması hata terimi olarak isimlendirilir ve ε ile gösterilir.

Belirlenen X değerine karşılık gelen tahmini Y değeri ya da diğer bir ifadeyle X'in belirli değeri için, Y'nin anakütle ortalamasının tahmini $E(Y_i)$ şöyle bulunur:

$$\hat{Y}_i = b_0 + b_1 X_i$$

$$\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X})$$

Gözlenen Y_i değeri ile tahmini değer yani $\hat{Y}_i$ değeri karşılaştırıldığında model ile veri arasındaki uyum için bir büyüklük elde edilir buna **artık** denir.

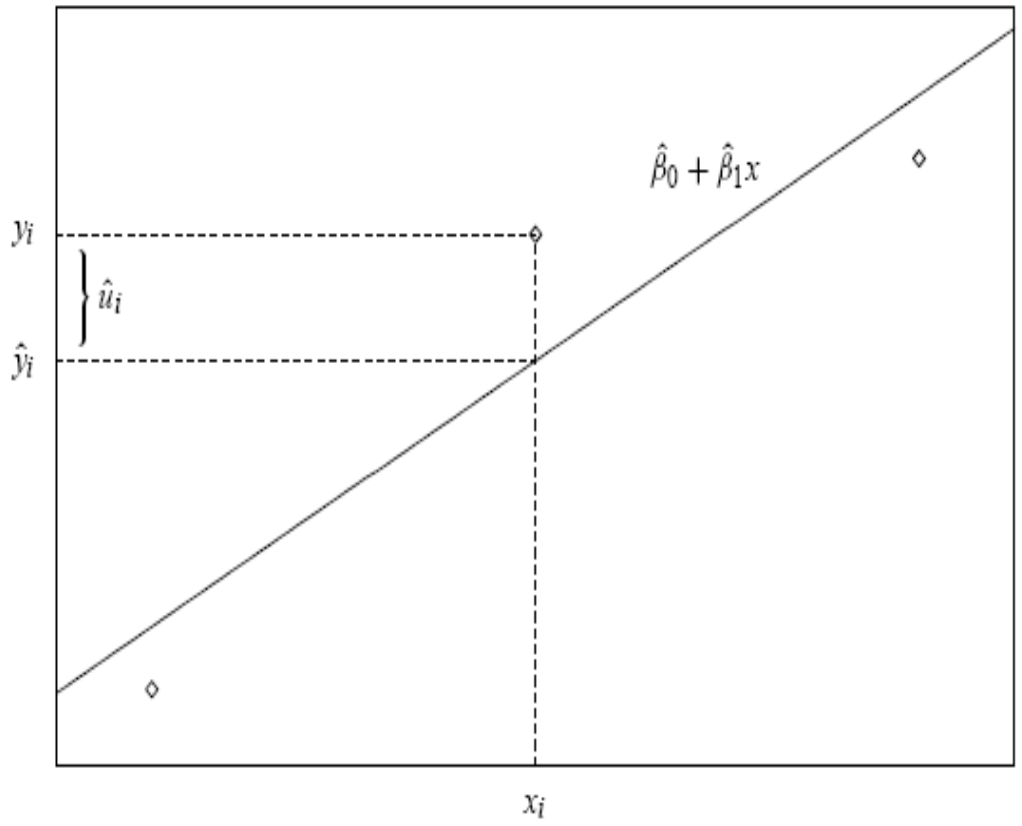
$$e_i = Y_i - \hat{Y}_i$$

Artık, tahmin edilen modelle veri arasındaki farka karşılık gelir eğer modelde sabit terim varsa artıkların toplamı sıfırdır.

$Y_i = \hat{Y}_i + e_i$ eşitliği dikkate alındığında, $\hat{Y}_i$ Y_i gözlemini açıklayabilen fakat e_i Y_i gözlemini açıklayamayan kısımdır.Belirli varsayımlar sağlandığında artıklar tahmini hatalar olarak kabul edilir. Hatırlanacağı üzere hata,

$\varepsilon_i = Y_i - E(Y_i)$ şeklindedir.

Gözlenen Y_i, tahmini Y_i ve hata arasındaki ilişki aşağıdaki temsili çizimde de görülebilmektedir.



Standart Hata

Regresyon denklemini standart hatasının tahmini hata kareler toplamının n-2'ye bölünerek karekökünün alınmasıyla bulunur. Buna tahminin standart hatası da denilir.

$S = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}}$ şeklinde tanımlanır.

Regresyon Denklemi Parametrelenin Hipotez Testi ve Güven Aralığı

EKK yöntemi ile $Y = b_0 + b_1 X$ regresyon denkleminin b_0 ve b_1 parametrelerinin tahmini değerleri $\hat{b}_0$ ve $\hat{b}_1$ örneklemden örnekleme değişen birer rastlantı değişkenidir. Bu rastlantı değişkenlerinin ortalama ve varyansları sırayla aşağıdaki gibidir.

$$E(\hat{b}_0) = b_0 \qquad V(\hat{b}_0) = \frac{\sum X^2}{n\sum (X - \bar{X})^2}\sigma^2$$

$$E(\hat{b}_1) = b_1 \qquad V(\hat{b}_1) = \frac{\sigma^2}{\sum (X - \bar{X})^2}$$

Regresyon analizindeki varsayımlardan ilki hatırlanacağı üzere bağımlı değişkenin Y'nin normal dağıldığı varsayımı idi. Gerek $\hat{b}_0$ ve gerekse $\hat{b}_1$ bağımlı değişkenle doğrusal ilişki içinde olduğundan bu iki rastlantı değişkenin de normal dağıldığı sonucuna varılır. Bu sonuca dayanarak b_0 ve b_1 için $\hat{b}_0$ ve $\hat{b}_1$ üzerinden yola çıkarak hipotez testi ve güven aralıklarını tanımlamak mümkündür. $\hat{b}_0$ ve $\hat{b}_1$ rastlantı değişkenlerinin normal dağılması halinde önceki konulardan hatırlanacağı üzere

$$z = \frac{\hat{b}_0 - b_0}{\sigma_{\hat{b}_0}} \quad \text{ve} \quad z = \frac{\hat{b}_1 - b_1}{\sigma_{\hat{b}_1}}$$

Bu istatistikler n-2 serbestlik dereceli t dağılır. Bu eşitliklerden yola çıkarak sırasıyla sabit ve eğim paramereleri için güven aralıkları şöyledir:

$$P[-t_{n-2,\alpha/2} < \frac{\hat{b}_0 - b_0}{S_{\hat{b}_0}} < t_{n-2,\alpha/2}] = 1-\alpha$$

$$P[\hat{b}_0 - t_{n-2,\alpha/2} S_{\hat{b}_0} \le b_0 \le \hat{b}_0 + t_{n-2,\alpha/2} S_{\hat{b}_0}] = 1-\alpha$$

$$\hat{b}_0 - t_{n-2,\alpha/2} S_{\hat{b}_0} \le b_0 \le \hat{b}_0 + t_{n-2,\alpha/2} S_{\hat{b}_0}$$

Sabit parametre için yukarıdaki eşitsizliklerin düzenlenmesiyle bulunan yaklaşım eğim parametresi için de yapıldığında bulunan güven aralığı şöyledir:

$$\hat{b}_1 - t_{n-2,\alpha/2} S_{\hat{b}_1} \le b_1 \le \hat{b}_1 + t_{n-2,\alpha/2} S_{\hat{b}_1}$$

İlgilinenilen hipotez testi sabit katsayı için aşağıdaki gibi düzenlendiğinde,

$$H_0 : b_0 = b$$
$$H_1 : b_0 \neq b$$

aşağıdaki test istatistiği kullanılır:

$$t_0 = \frac{\hat{b}_0 - b}{S_{\hat{b}_0}}$$

Alternatif hipotez yukarıda çift yönlü düzenlenmiş olmakla beraber aşağıdaki gibi tek yönlü de düzenlenebilir:

$$H_1 : b_0 < b$$

$$H_1 : b_0 > b$$

Alternatif hipotezin çift yönlü olması halinde eğer bulunan kritik değer ve tablo değeri arasında aşağıdaki gibi bir ilişki varsa H_0 reddedilir.

$t_0 > t_{n-2,\alpha/2}$ veya $t_0 < -t_{n-2,\alpha/2}$

Alternatif hipotez eğer aşağıdaki gibi tek yönlü ise,

$H_1 : b_0 < b$

$t_0 < -t_{n-2,\alpha}$ halinde H_0 hipotezi reddedilir.

Ve son olarak da alternatif hipotez

$H_1 : b_0 > b$ şeklindeyse

$t_0 > t_{n-2,\alpha}$ olmalı halinde H_0 hipotezi reddedilir.

Eğim parametresinin hipotez testinde genellikle aşağıdaki şekilde düzenlenir. Katsayının genellikle sıfıra eşitliği sınanır. Çünkü katsayının sıfıra eşitliği bağımlı değişkenin bağımsız değişkene doğrusal bağımlı olmadığı anlamına gelir. Sıfır hipotezi ve olası alternatif hipotezler aşağıdaki gibi düzenlenir;

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

$$H_1 : b_1 > 0$$

$$H_1 : b_1 < 0$$

Burada kullanılan istatistik aşağıdaki gibidir.

$$t_0 = \frac{\hat{b}_1}{S_{\hat{b}_1}}$$

ÖRNEK: Önceki örnekte bir sınıftaki öğrencilerin matematik ve muhasebe derslerindeki başarıları arasındaki ilişki incelenmişti. Regresyon denkleminin paremetrelerinin %99 ve %95 güvenle bulundukları aralıkları belirleyiniz ve katsayların anlamlılığını aynı güven düzeyinde sınayınız.

Öncelikle tahminin standart hatasını bulalım;

$$\hat{y} = -0.625 + 0.9375(2) = 1.25$$

$$\hat{y} = -0.625 + 0.9375(3) = 2.188$$

$$\hat{y} = -0.625 + 0.9375(5) = 4.063$$

$$\hat{y} = -0.625 + 0.9375(6) = 5$$

$$\hat{y} = -0.625 + 0.9375(7) = 5.938$$

$$\hat{y} = -0.625 + 0.9375(10) = 8.75$$

$$\hat{y} = -0.625 + 0.9375(7) = 5.938$$

$$\hat{y} = -0.625 + 0.9375(8) = 6.875$$

Y	$\hat{Y}$	$(Y-\hat{Y})^2$
1	1,25	0,0625
2	2,188	0,035344
3	4,063	1,129969
5	5	0
6	5,938	0,003844
7	8,75	3,0625
8	5,938	4,251844
8	6,875	1,265625
40	**40,002**	**9,811626**

$$S=\sqrt{\frac{9.81}{8-2}}=1.278 \qquad S_{\hat{b}_1}=\frac{S}{\sqrt{\sum(X-\bar{X})^2}}=\frac{1.278}{\sqrt{48}}=0.1846$$

olacağından %99 güvenle eğim parametresinin sınırları şöyledir:

$$0.9375-(0.1846)(3.707)\le b_1\le 0.9375+(0.1846)(3.707)$$

$$0.2532\le b_1\le 1.622$$

$$(t_{0.01;6}=3.707)$$

$$0.9375-(0.1846)(2.447)\le b_1\le 0.9375+(0.1846)(2.447)$$

$$0.486\le b_1\le 1.389$$

($t_{0.05;6}=2.447$)

Sabit parametrenin standart hatası, güven aralığı aşağıdaki gibidir:

$$S_{\hat{b}_0}=\sqrt{\frac{\sum X^2}{n\sum (X-\bar{X})^2}}S=\sqrt{\frac{336}{8(48)}}1.278=1.195$$

$$-0.625-(1.195)(3.707)\le b_0\le -0.625+(1.195)(3.707)$$

$$-5.055\le b_0\le 3.805$$

$$-0.625-(1.195)(2.447)\le b_0\le -0.625+(1.195)(2.447)$$

$$-3.549\le b_0\le 2.299$$

Eğim parametresinin anlamlılığı için kurulan hipotezler ve testi şöyledir:

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

$$t_0 = \frac{0.9375}{0.1846} = 5.079$$

elde edilen değer her iki anlamlılık düzeyindeki kritik değerlerle sınandığında sıfır hipotezi reddedilir. Yani, Y bağımlı değişkeni X'e bağımlıdır.

Sabit katsayının testi benzeri yaklaşımla incelendiğinde,

$$H_0 : b_0 = 0$$
$$H_1 : b_0 \neq 0$$

$$t_0 = \frac{-0.625}{1.195} = -0.523$$

Her iki anlamlılık düzeyinde sıfır hipotezi kabul edilir.

ÖRNEK:

Önceki bölümde ele alınan suç ve işsizlik oranlarına ilişkin örnekteki parametrelerin anlamlılığını % 5 önem düzeyinde sınamak istersek;

X	Y			
işşizlik oranı	**suç oranı**	Ytahmin	(Y-Ytah)	(Y-Ytah)^2
0,8	3	2,592	0,408	0,166464
1,4	6	5,211	0,789	0,622521
2,3	7	9,1395	-2,1395	4,57746
3,5	15	14,3775	0,6225	0,387506
4,5	19	18,7425	0,2575	0,066306
12,5	50	50,0625	-0,0625	5,820258

$$\sum Y = 50 \quad \sum X = 12.5 \quad \sum XY$$

$$\sum X^2 = 40.39 \quad \sum Y^2 = 680$$

$$\sum (X - \bar{X})(Y - \bar{Y}) = 39.9$$

$$\sum (X - \bar{X})^2 = 9.14 \quad \sum (Y - \bar{Y})^2$$

$$Y = -0.9 + 4.365X$$

$$S = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}}\sqrt{\frac{5.82}{5-3}} = 1.39$$

$$S(\hat{b}_1) = \frac{S}{\sqrt{\sum (X - \bar{X})^2}} = \frac{1.39}{\sqrt{9.14}} = 0.455$$

$$S_{\hat{b}_0} = \sqrt{\frac{\sum X^2}{n\sum (X - \bar{X})^2}}S = \sqrt{\frac{40.39}{5(9.14)}}1.39 = 1.306$$

$$4.365 - 3.82 * 0.455 \le b_1 \le 4.365 + 3.82 * 0.455$$

$$2.63 \le b_1 \le 6.1$$

$$t = \frac{4.365}{0.455} = 9.6$$

Eğim parametresi anlamlıdır, güven aralığı sıfırı kapsamadığı gibi hesaplanan t istatistik değeri de tablo değerinden büyüktür (9.6> 3.82)

Sabit parametresi ise anlamsızdır. Aralık sıfırı kapsar ve test istatistiği sonucu kritik değerden küçük çıkmıştır.

$$-0.9 - 3.82 * 1.306 \leq b_0 \leq -0.9 + 3.82 * 1.306$$

$$-5.88 \leq b_0 \leq 4$$

$$t = \frac{-0.9}{1.306} = -0.69$$

Örnek: Aşağıdaki veriden hareketle regresyon denklemi eğim katsayısının anlamlılığını değerlendiriniz.

x	y
20	6
19	8
17	9
16	10
13	12

$$n = 5, \quad \sum X = 85, \quad \sum Y = 45$$

$$\sum XY = 741, \quad \sum X^2 = 1475, \quad \sum Y^2 = 425$$

$$b_1 = \frac{741 - 5(85/5)(45/5)}{1475 - 5(85/5)^2} = -0.8$$

$$b_0 = 45/5 - (-0.8)(85/5) = 22.6$$

Tahmini y değerleri ve hata kareler toplamı şöyledir:

x	y	ytah	e	e^2
20	6	6.6	-0.6	0.36
19	8	7.4	0.6	0.36
17	9	9	0	0
16	10	9.8	0.2	0.04
13	12	12.2	-0.2	0.04
85	45	45	-3.6E-15	0.8

$$s = \sqrt{\frac{0.8}{5-2}} = 0.516$$

$$s_{b1} = \frac{0.516}{\sqrt{30}} = 0.094$$

$$t = \frac{-0.8}{0.094} = -8.9$$

3 sd li t tablo değeri -3.82 olduğundan eğim parametresinin anlamsız olduğunu ifade eden sıfır hipotezi reddedilir.