2. REGRESYON ANALİZİNİN TEMEL KAVRAMLARI

2.1. Tanım

Regresyon analizi, bir değişkenin başka bir veya daha fazla değişkene olan bağımlılığını inceler. Amaç, bağımlı değişkenin kitle ortalamasını, açıklayıcı değişkenlerin sabit (bilinen) değerleri cinsinden tahmin etmek veya öngörebilmektir.

Regresyon analizinde fonksiyonel veya deterministik ilişkilerle değil istatistiki ilişkilerle ilgileniriz. İstatistiki ilişkilerde rassal veya stokastik değişkenler yani olasılık dağılımları olan değişkenlerle çalışırız. Deterministik ilişkilerde yine değişkenler vardır fakat bunlar rassal veya stokastik değildir.

Fakat burada bağımlı ve açıklayıcı değişkenler açısından bir farklılık vardır: Regresyon analizinde bağımlı değişkenin rassal olduğu yani olasılık dağılımları olduğu varsayılır. Fakat açıklayıcı değişkenlerin sabit değerleri olduğu yani her örnekte aynı değerleri aldığı varsayılır.

Regresyon Analizinin Korelasyon Analizinden Farkı:

- Korelasyon analizinde, iki değişken arasındaki doğrusal ilişkinin gücü veya derecesi ile ilgileniriz. Örneğin istatistik dersi ile matematik dersi notları arasındaki ilişki. Ama regresyon analizinde amaç bağımlı değişkenin kitle ortalamasını, açıklayıcı değişkenlerin sabit (bilinen) değerleri cinsinden tahmin etmektir. Örneğin bir öğrencinin matematik notunu biliyorsak istatistik notunu öngörebilir miyiz?

- Korelasyon analizinde bağımlı değişken-açıklayıcı değişken ayrımı yoktur ve değişkenlerle ilgili aynı varsayımlar yapılır: ikisi de rassal değişkenlerdir. Regresyon ise bu ayrım vardır ve analizinde bağımlı değişkenin rassal olduğu, fakat açıklayıcı değişkenlerin sabit yani stokastik olmadıkları varsayımı yapılır.

Regresyon ve nedensellik:

Regresyon analizinde bir değişkenin diğerine bağımlılığından söz edilir. Ancak bu bağımlılık nedensellik anlamında değildir. Yani, istatistiki bir ilişki, ne kadar güçlü olursa olsun, kendiliğinden bir nedensellik göstermez. Bu nedenselliği biz teorik inceleme ile kurarız. Örneğin tüketimin gelire bağlı olduğu ön kabulüyle modeli kurarız.

2.2 Anakütle Regresyon Fonksiyonu (ARF)

Örnek 1: Diyelim ki anakütle 60 aileden oluşmaktadır. Amacımız bu toplulukta haftalık gelir (X) ile haftalık tüketim (Y) arasındaki ilişkiyi belirlemektir. Diğer bir deyişle herhangi bir ailenin gelirini biliyorken haftalık tüketimlerinin ortalama olarak ne kadar olacağını tahmin etmek istiyoruz. Veriler Tablo 2.1'de verilmiştir.

Tablo 2.1: 60 Ailenin Kitle (Y_i, X_i) Değerleri

X_i →	80	100	120	140	160	180	200	220	240	260
Y_i ↓	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	--	88	--	113	125	140	--	160	189	185
	--	--	--	115	--	--	--	162	--	191
Toplam Y_i	**325**	**462**	**445**	**707**	**678**	**750**	**685**	**1043**	**966**	**1211**
Y_i sayısı	**5**	**6**	**5**	**7**	**6**	**6**	**5**	**7**	**6**	**7**

Örneğin haftalık geliri 80 dolar olan bir ailenin tüketimi 55 dolar, bir diğerininki 60 dolar.

Tablo, Y'nin koşullu dağılımını gösterir. Yani X değerleri veri iken Y değerlerini verir. Buna dayanarak Y'nin koşullu olasılıklarını bulabiliriz. Ör. X=80 iken Y'nin alabileceği 5 değer vardır. Yani X=80 iken Y'nin bu değerlerden herhangi birisini alma olasılığı 1/5'tir. Örneğin $P(Y=55|X=80)=1/5$, $P(Y=60|X=80)=1/5$ gibi. Bu koşullu olasılıklar Tablo 2.2'de verilmiştir.

Tablo 2.2: Her bir anakütle X değeri için Y'nin anakütle koşullu olasılıkları

X_i →	80	100	120	140	160	180	200	220	240	260
Y_i ↓	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	--	1/6	--	1/7	1/6	1/6	--	1/7	1/6	1/7
	--	--	--	1/7	--	--	--	1/7	--	1/7
Toplam Y_i	**325**	**462**	**445**	**707**	**678**	**750**	**685**	**1043**	**966**	**1211**
Y_i sayısı	**5**	**6**	**5**	**7**	**6**	**6**	**5**	**7**	**6**	**7**
$E(Y\|X_i)$	**65**	**77**	**89**	**101**	**113**	**125**	**137**	**149**	**161**	**173**

Bu koşullu olasılıklara bakarak Y'nin koşullu beklenen değerini bulabiliriz. Yani X herhangi bir X_i değerini alması koşulu altında Y'nin beklenen değeri $E(Y|X=X_i)$ bulunabilir. Örneğin X=80 iken Y'nin koşullu beklenen değeri

$E(Y|X=80)= 55*(1/5)+ 60*(1/5)+65*(1/5)+70*(1/5)+75*(1/5) = 65$ dir.

Bunu tüm X değerleri için yaparsak Y'nin koşullu beklenen değerlerini $E(Y|X_i)$ bulabiliriz. Tüm X değerleri için hesaplanan koşullu beklenen değerler Tablo 2.3'de verilmiştir.

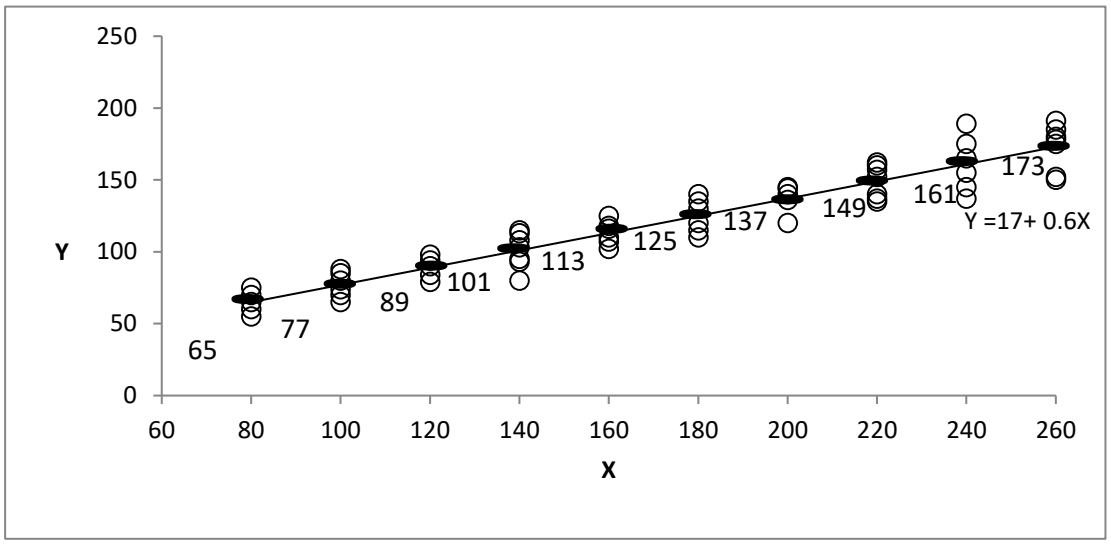
Tablo 2.3: Y'nin anakütle koşullu beklenen değerleri

X_i	$E(Y\|X_i)$
80	65
100	77
120	89
140	101
160	113
180	125
200	137
220	149
240	161
260	173

Tablo 3, Örnek 1'de yer alan 60 aileden oluşan anakütlede X_i ile $E(Y|X_i)$ ilişkisini gösterir. X_i ile $E(Y|X_i)$ arasındaki anakütle ilişkisi *anakütle regresyon fonksiyonu* (ARF) olarak adlandırılır. ARF doğrusal olarak belirlendiğinde $E(Y|X_i) = \beta_0 + \beta_1 X_i$ şeklini almaktadır.

Şimdi bu 60 aile için X_i ile $E(Y|X_i)$ arasındaki ilişkiyi bulmaya çalışalım. Tablodan görülebileceği gibi X 20şer dolar artarken $E(Y|X_i)$ 12 şer dolar artmaktadır. Demek ki bu toplulukta marjinal tüketim eğilimi (β_1=) 12/20=0.60'dır. Bu durumda sabit terim de (β_0=) 17 bulunur. Demek ki bu 60 aile için ARF, $E(Y|X_i) = \beta_0 + \beta_1 X_i = 17 + 0.60X_i$'dir.

Grafik 2.1: Anakütle Verileri, Koşullu Beklenen Değerler E(Y\|X) ve Anakütle Regresyon Fonksiyonu



Grafik 2.1'de yer alan noktalar Örnek 1'de yer alan 60 aile için X ve Y'nin anakütle değerlerini gösterir. Doğru ise 60 aile için anakütle regresyon doğrusudur. Her bir ($E(Y|X_i)$, X_i) ikilisi siyah oval noktalarla gösterilmiştir. Bunların $E(Y|X_i)$ değerleri yanlarında yer almaktadır. Anakütle regresyon doğrusu Tablo 2.3'de yer alan 10 noktanın ($E(Y|X_i)$, X_i ikililerinin) birleşimidir.

Geometrik olarak anakütle regresyon doğrusu, açıklayıcı değişkenlerin veri değerleri için bağımlı değişkenin koşullu olasılıklarını veya beklenen değerlerini veren eğridir. Grafik 2.1'den de görülebileceği gibi her bir X_i değeri için çeşitli anakütle Y değerleri ve tek bir koşullu beklenen değer vardır ve regresyon doğrusu bu beklenen değerlerden geçer[1].

2.3 Hata Terimi ve Anakütle Regresyon Denklemi (ARD)

i. birim (Örnek 1'deki 60 aileden her biri) için rassal ve görünmeyen hata terimi aşağıdaki gibi tanımlanır:

$$u_i = Y_i - E(Y|X_i) \qquad \forall\ i.$$

Hata terimi her bir ailenin Y_i değeri ile yine o ailenin X_i değerine karşılık gelen Y'nin koşullu ortalaması ($E(Y|X_i)$) arasındaki farktır.

[1] Koşullu ortalamalar her zaman düz bir doğru üzerinde olmak zorunda değildir. Buradaki örnek düz çizgi verecek şekilde oluşturulmuştur.

Bu tanımdan yola çıkarak her bir birimin Y_i değeri aşağıdaki gibi yazılabilir.

$Y_i = E(Y|X_i) + u_i$

$= \beta_0 + \beta_1 X_i + u_i$ ($E(Y|X_i) = \beta_0 + \beta_1 X_i$ olduğundan)

Bu denklem *anakütle regresyon denklemi* (ARD) olarak adlandırılır. ARD denklemine göre her bir Y_i iki bileşenden oluşur:

1) $E(Y|X_i) = \beta_0 + \beta_1 X_i$: $X=X_i$ iken Y'nin koşullu ortalaması (geliri X_i olan ailelerin ortalama gelirleri)

2) $u_i = Y_i - E(Y|X_i)$: i. birim için rassal hata terimi (her bir ailenin Y_i değeri ile kendisi ile aynı gelire sahip -geliri X_i olan- ailelerin ortalama gelirleri arasındaki fark).

Yine ARD denkleminden yola çıkarak her bir X_i'ye karşılık gelen hata terimlerinin koşullu beklenen değerinin 0 olduğu sonucuna ulaşırız:

$E(u_i|X_i) = 0 \quad \forall$ i.

Kanıt: ARD denkleminin iki tarafının koşullu beklenen değerini alalım:

$E(Y_i|X_i) = E(E(Y|X_i)|X_i) + E(u_i|X_i)$

$= E(Y|X_i) + E(u_i|X_i)$ ($E(Y|X_i)$ sabit olduğundan)

Buradan $E(u_i|X_i) = 0$ bulunur.

Hata terimleri, anakütle Y_i değerlerini belirleyen, X dışındaki tüm diğer bilinmeyen ve gözlemlenemeyen etkileri temsil eder.

Tablo 2.4: Örnek 1'de Yer Alan 60 Aile için Rassal Hata Terimleri

X_i = 80 için	Y_i	$E(Y\|X_i)$	$u_i = Y_i - E(Y\|X_i=80)$
	55	65	-10
	60	65	-5
	65	65	0
	70	65	5
	75	65	10
Toplam	**325**		**0**
Ortalama	**65**		**0**

X_i = 100 için	Y_i	$E(Y\|X_i)$	$u_i = Y_i - E(Y\|X_i=100)$
	65	77	-12
	70	77	-7
	74	77	-3
	80	77	3
	85	77	8
	88	77	11
Toplam	**462**		**0**
Ortalama	**77**		**0**

X_i = 120 için	Y_i	$E(Y\|X_i)$	$u_i = Y_i - E(Y\|X_i=120)$
	79	89	-10
	84	89	-5
	90	89	1
	94	89	5
	98	89	9
Toplam	**445**		**0**
Ortalama	**89**		**0**

X_i = 140 için	Y_i	$E(Y\|X_i)$	$u_i = Y_i - E(Y\|X_i=140)$
	80	101	-21
	93	101	-8
	95	101	-6
	103	101	2
	108	101	7
	113	101	12
	115	101	14
Toplam	**707**		**0**
Ortalama	**101**		**0**

X_i = 200, 220, 240, 260 için hata terimleri ile koşullu beklenen değerlerinin hesaplanması okuyucuya bırakılmıştır.

2.4 Örneklem Regresyon Fonksiyonu (ÖRF)

Gerçekte tüm anakütleyi gözlemlemek mümkün olmadığından ARF'yi bilmemiz mümkün değildir. Bu nedenle ARF'yi ancak örneklem verileri kullanarak tahmin edebiliriz. Bu amaçla, anakütleden rassal olarak seçilmiş N gözlemli (X_i,Y_i) (i=1, …, N) örneklem oluşturulacaktır.

ARF $E(Y|X_i) = \beta_0 + \beta_1 X_i$ iken örneklem regresyon fonksiyonu (ÖRF) aşağıdaki gibidir.

$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad (i=1, \ldots, N)$$

Burada $\widehat{Y}_i$ ARF'nin $(E(Y|X_i) = \beta_0 + \beta_1 X_i)$ bir tahmini, $\hat{\beta}_0$ sabit terim β_0'ın bir tahmini ve $\hat{\beta}_1$eğim katsayısı β_1'in bir tahminidir.

Örneklem verisinin özellikleri:

- Örneklemi oluşturan gözlemler, anakütle gözlemlerinin bir alt kümesidir.
- Anakütleden farklı rassal örneklemler elde edilebilir. Her bir örneklem β_0 ve β_1 için farklı tahmin değerleri verir. Diğer bir deyişle her bir rassal örneklem farklı bir ÖRF (farklı $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri) verecektir.

Örnek: Örnek 1'de yer alan 60 aileden oluşan anakütleden 10 gözlemli iki faklı rassal örneklem alalım. Örneklem seçilirken 10 adet X değerinin tümü alınmış, bunlara karşılık gelen Y değerleri rassal olarak seçilmiştir.

Tablo 2.5: Örneklem 1 ve Örneklem 2

Örneklem 1		**Örneklem 2**	
X_i	Y_i	X_i	Y_i
80	70	80	55
100	65	100	88
120	90	120	90
140	95	140	80
160	110	160	118
180	115	180	120
200	120	200	145
220	140	220	135
240	155	240	145
260	150	260	175

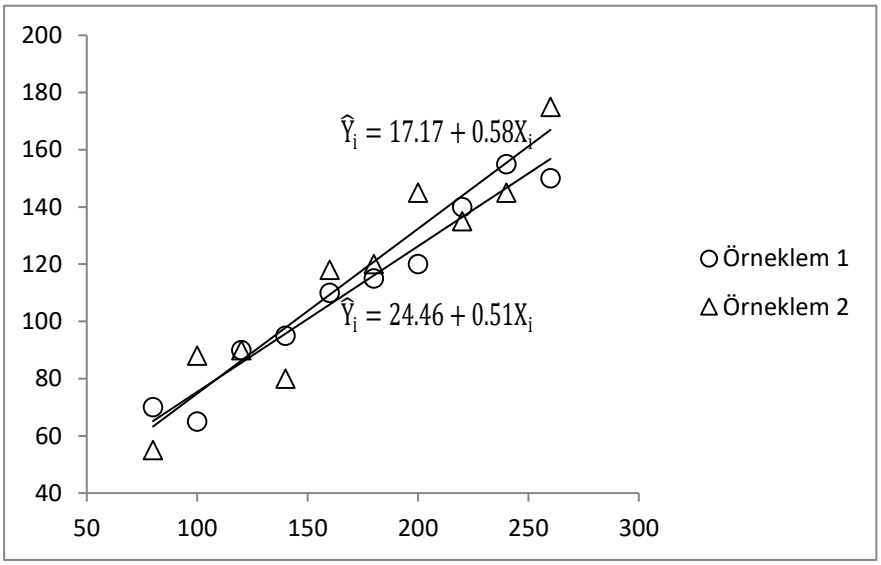
İki örneklemde aynı olan X değerlerine karşılık farklı Y değerleri vardır. Bu nedenle farklı $\hat{\beta}_0$ ve $\hat{\beta}_1$ değerleri elde edilecektir. ÖRF elde edilirken verileri en iyi yansıtacak doğrunun bulunması amaçlanmış ve aşağıdaki ÖRF'ler bulunmuştur.

Örneklem 1: $\hat{Y}_i = 24.46 + 0.51X_i$

Örneklem 2: $\hat{Y}_i = 17.17 + 0.58X_i$

Diğer bir deyişle Örneklem 1 verileri kullanıldığında $\hat{\beta}_0 = 24.46,\ \hat{\beta}_1 = 0.5091$, Örneklem 2 verileri kullanıldığında $\hat{\beta}_0 = 17.17,\ \hat{\beta}_1 = 0.5761$ bulunmaktadır. Grafik 2.2 Örneklem 1 ve Örneklem 2 verileri kullanılarak elde edilen ÖRF'leri vermektedir. Örneklem 1 ile bulunan eğri daha yatık, Örneklem 2 ile bulunan eğri daha diktir.

Grafik 2.2: Örneklem Verileri ve Örneklem Regresyon Fonksiyonları



Hangi ÖRF gerçek ARF'unu yansıtmaktadır? Gerçek anakütle eğrisine bakmadan bunu bilmek mümkün değildir. Bunların gerçek eğriyi yansıttığı düşünülür. Ancak örnekleme yapıldığından bunlar gerçek ARF ile aynı değildir, ARF'ye yalnızca bir yaklaştırmadır (tahmindir).

ÖRF'nun ARF'a ne kadar yakın olduğu, ÖRF'nun örneklem verisi kullanılarak nasıl oluşturulduğuna, diğer bir deyişle $\hat{\beta}_0$ ve $\hat{\beta}_1$ tahmin edicilerinin özelliklerine bağlıdır[2].

[2] Tahmin edici, örneklem verileri kullanılarak anakütle parametrelerinin nasıl tahmin edileceğini gösteren kural, formül veya metoddur.

2.5 Örneklem Regresyon Denklemi (ÖRD)

Örneklem regresyon denklemi (ÖRD) anakütle regresyon denkleminin (ARD) örneklemdeki karşılığıdır.

ARD: $Y_i = E(Y|X_i) + u_i = \beta_0 + \beta_1 X_i + u_i$

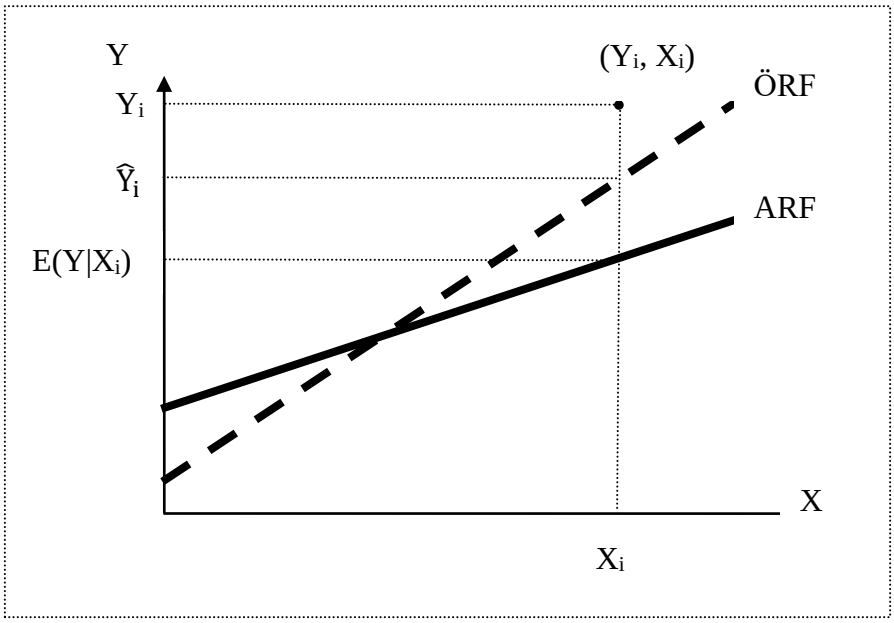
ÖRD: $Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$

Burada $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ARF'nin ($E(Y|X_i) = \beta_0 + \beta_1 X_i$), $\hat{\beta}_0$ sabit terim β_0'ın, $\hat{\beta}_1$eğim katsayısı β_1'in ve $\hat{u}_i$ i örneklem gözlemi için hata terimidir.

ÖRD denklemi, her bir örneklem gözlemini (Y_i) iki bileşenin toplamı olarak belirler:

1) Her bir örneklem X değeri (X_i) için Y'nin tahmin değeri: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
2) i. örneklem gözlemine karşılık gelen hata terimi: $\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ (i=1,…,N)

Grafik 2.3: Anakütle ve Örneklem Regresyon Eğrilerinin Karşılaştırması



Anakütle regresyon eğrisi ARF'nin ($E(Y|X_i) = \beta_0 + \beta_1 X_i$), örneklem regresyon eğrisi ÖRF'nun ($\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$) grafiğidir.

Bundan sonraki bölümlerde amacımız anakütle regresyon eğrisine en yakın sonucu verecek örneklem regresyon eğrisi tahmin yöntemini bulmak olacaktır.