Tek denklemli Regresyon Modeli

Ekonometri konularıyla ilgili literatürde en çok rastlanan ve ilgi gören model türü tek denklemli doğrusal regresyon modelleridir. 1980 sonrası teorik ve uygulamalı ekonometrideki hızlı gelişme ile birlikte iktisadi gerçekler ve veriler daha komplike modeller ile izah edilebiliyorsa da, tek denklemli modeller hala önemli bir yer tutmaktadır. Bunun sebepleri:

1. Geleneksel iktisat teorisi bir sonucu bir dizi sebebe bağlamakta ve bunu tek denklemli bir model çerçevesinde ele alan bir tutuma sahip bulunmaktadır.

2. Gerçekler çok defa basit kalıplara sokulamayacak kadar karmaşık olmasına rağmen, iktisadi olaylarda sebep-sonuç (kozalite) ilişkilerinin doğrusal bir model içinde sunulması önemli ölçüde kolaylık sağlamaktadır.

Tek denklemli regresyon modellerinde yalnız bir bağımlı değişken vardır ve bu bağımlı değişken bir ya da birden çok bağımsız değişkenin doğrusal fonksiyonu olarak belirtilmiştir.

Tek denklemli regresyon modellerinde bağımsız değişkenlerden bağımlı değişkene doğru tek yönlü bir ilişkinin olduğu varsayılmaktadır. Modelin bağımsız kabul edilen değişkenleri, modelin bağımlı değişkeni tarafından etkilenmeyecektir. Sistemde geri tepme (feedback) ve karşılıklı ilişki olmayacaktır. Regresyon modellerinin nedensellik ilişkisi üzerine kurulur. Nedensellik ilişkisi kuramsal ve önsel (a'priori) olmalıdır. İstatistiksel açıdan anlamlı bir ilişki kendi başına bir nedensellik ilişkisi anlamı taşımaz.

Basit ve Çok Değişkenli Regresyon Modelleri

İktisadi ilişkilerin ölçülmesinde ilk adım, iktisadi ilişkileri yansıtan modellerde yer alacak değişkenlerin tanımlanmasıdır. Bir ana kütlenin veya ondan çekilecek bir örneğin birimleri bir veya birden çok özelliği bakımından gözlemlenebilir. Bu özellikler nicel ve/veya nitel olabilir. Modelde yer alan değişken sayısına göre regresyon modelleri basit regresyon ve çok değişkenli regresyon olmak üzere ikiye ayrılır. Bağımlı değişken sadece bir bağımsız değişken tarafından açıklanıyorsa **basit regresyon modeli**, birden fazla değişken tarafından açıklanıyorsa **çok değişkenli regresyon modeli** olarak adlandırılır.

İlişkileri mümkün olduğu kadar basitleştirebilmek için, tek bir ilişki üzerinde durulan ve bu ilişkinin de sadece iki değişkeni içerdiği şeklindeki varsayımdan hareketle, en basit durum basit regresyon modeli;

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad i = 1, 2......n$$

ile gösterilirken, ikiden fazla değişkenin yer aldığı çok değişkenli regresyon modeli ise;

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...+ \beta_{k-1} X_{(k-1)i} + u_i \qquad i = 1, 2......n$$

gibi gösterilir.

Yukarıdaki modellerde de görüldüğü üzere, basit regresyon modelinde bağımlı değişkeni açıklayan sadece bir tane bağımsız değişken (X) yer alırken, çok değişkenli regresyon modelinde birden fazla bağımsız değişken ($X_1, X_2...........X_{k-1}$) vardır. i ise 1'den n'e kadar olan gözlemleri ifade etmektedir. Her iki modelde de sadece tek bir bağımlı değişken (Y) yer almaktadır.

Örneğin, bir malın talebi (D) en basit ifade ile malın fiyatının (P) fonksiyonudur. Malın fiyatı artarsa malın talebi düşer, malın fiyatı düşerse de talebi artar. Fiyattan talebe doğru bir nedensellik söz konusu olduğuna göre, talep bağımlı değişken, fiyat ise bağımsız değişkendir. Buna göre basit regresyon modeli aşağıdaki gibi kurulacaktır.

$$D_t = \beta_1 + \beta_2 P_t + u_t \qquad t = 1, 2......t$$

Ancak bir malın talebi, malın fiyatının yanı sıra bu malı talep eden kişilerin gelir düzeyine (Y), malın tamamlayıcı malı varsa bu tamamlayıcı malın fiyatına (P_T) ve nihayet rakip mal varsa rakip malın fiyatına (P_R) bağlı olarak değişkenlik gösterecektir. İlgili değişkenler modele dahil edildiğinde, model çok değişkenli regresyon modeli olarak adlandırılacak ve aşağıdaki şekilde gösterilecektir.

$$D_t = \beta_0 + \beta_1 P_t + \beta_2 Y_t + \beta_3 P_{Tt} + \beta_4 P_{Rt} + u_i \qquad t = 1, 2......t$$

Yukarıdaki model çok değişkenli regresyon modeli olup; malın talebi dört tane bağımsız değişken ve hata teriminin (u_i) bir fonksiyonudur.

Regresyon Analizi

Regresyon analizi genel olarak nicel değişkenler arasındaki ilişkilerin incelenmesi olarak tanımlanmaktadır. Regresyon, bağımlı ve bağımsız değişkenler arasındaki bağlantıyı kuran parametrelerin değerinin tahmin imkânını araştırır.

Regresyon analizinde, değişkenler arasındaki ilişki fonksiyonel ya da kesin ilişkiler olmayıp **istatistiksel ilişkilerdir**. Değişkenler arasındaki istatistiksel ilişkilerde, genellikle **stokastik (tesadüfi-rastlantısal) değişkenler** yani olasılık dağılımı olan değişkenler kullanılır.

Fonksiyonel ya da kesin ilişkilerde de değişkenler kullanılır, ancak bunlar tesadüfi ya da stokastik değil, deterministik değişkenlerdir.

Regresyon analizi, korelasyon ile yakından ilişkili, ancak kavramsal olarak çok farklıdır. Regresyonun korelasyona benzer yanı ikisinin de değişkenler arasında birlik ve yakınlığı aramalarıdır. Aralarındaki fark ise, regresyonun bir sebep-sonuç modeli içinde yani nedensellik ilişkisi içinde değişkenler arasındaki bağlantıyı aramasına karşılık, korelasyon analizi bu şekilde bir sebep-sonuç ilişkisi olmadan, değişkenler arasındaki doğrusal ilişkinin yön ve derecesinin saptanmasını sağlar.

Basit korelasyon analizinin amacı, yukarıda da ifade edildiği üzere iki değişken arasındaki doğrusal ilişkinin yönünü ve derecesini ölçmektir (Hatırlanacağı üzere korelasyon katsayısı $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$ formülü ile hesaplanmaktadır.) Formülde yer alan unsurlar ise $x_i = X_i - \bar{X}$ $y_i = Y_i - \bar{Y}$ 'e eşittir. Korelasyon katsayısı (r), $-1 \le r \le 1$ değerleri arasında yer alır). Regresyon analizinde ise bağımlı değişkenin ortalama değeri, bağımsız değişken(ler)in değişmeyen değerlerine dayanılarak tahmin edilmektedir.

Regresyon analizinde bağımlı ve bağımsız değişkenlerin ele alınmasında **asimetri** söz konusudur. Bağımlı değişkenin istatistiksel, tesadüfi olduğu yani olasılık dağılımı bulunduğu varsayılır. Bu varsayım X bağımsız değişkeninin sabit bir değeri için ana kütlede bağımlı değişkenin en az iki farklı değerinin bulunduğu anlamına gelmektedir. Tesadüfi ilişkinin bir gereği olan bu durum bir dağılım oluşturacaktır. Öte yandan bağımsız değişkenlerin yenilenen örneklemlerde değişmeyen değerler aldıkları varsayılmaktadır. Korelasyon analizinde herhangi iki değişkeni **simetrik** olarak ele alabiliriz. Bağımlı ve bağımsız değişken ayrımı yoktur. Her iki değişkenin de tesadüfi olduğu varsayılmıştır.

Bir regresyon modelinin kurulabilmesi için önsel olarak;

1. Sebep-sonuç ilişkisine göre bağımlı bağımsız değişken ayrımının

2. Bağımsız değişkenlerin ve sayısının

3. Modelin fonksiyonel biçiminin

belirlenmesi gerekir. Ancak sebep-sonuç ilişkisi ekonometride regresyonun dışında olup çoğu kez iktisat kuramı tarafından saptanmaktadır.

Anakütle Regresyon Modeli

Regresyon çözümlemesi büyük ölçüde, bağımsız değişkenin değeri bilindiği ya da sabit olduğu zaman bağımlı değişkenin ana kütledeki ortalama değeri ile ilgilenir.

Basit regresyon modeli, bağımlı Y değişkeni ile bağımsız X değişkeni arasında bağlantı sağlamaktadır. X'in her sabit değeri için ana kütlede rastlantısal ilişki gereği bağımlı değişkenin en az iki değerinin bulunması zorunludur. Böylece her X_i değişkeni için Y_{ij} değerleri elde edilecek, bu da bir dağılım oluşturacaktır.

Rastlantısal ilişkiyi açıklayabilmek için belirli bir zaman boyutu içinde, hane halkları yatay kesitinde hipotetik veriler kullanılarak haftalık tüketim harcamaları ile harcanabilir gelir arasında ilişki incelenecektir. Tüketim harcamaları harcanabilir gelirin bir fonksiyonudur. Buna göre, X haftalık harcanabilir gelir, Y ise tüketim harcamalarıdır. Öncelikle 60 aileden oluşan **ana kütle,** gelirleri yaklaşık olarak aynı olan ailelerden oluşan 10 ayrı gruba ayrılır ve **alt ana kütle** olarak adlandırılan her grup aynı gelir düzeyindeki farklı ailelerin tüketim harcamalarını içerir.

n=60 aile

X_i= Haftalık harcanabilir gelir (Dolar)

Y_i= Haftalık tüketim harcamaalarını (Dolar) göstermektedir.

Haftalık harcanabilir gelir

Y_{ij} \ X_i	80	100	120	140	160	180	200	220	240	260
Haftalık tüketim harcaması	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
n_i	5	6	5	7	6	6	5	7	6	7
Toplam	325	462	445	707	678	750	685	1043	966	1211

Haftalık harcanabilir gelir ve tüketim harcaması

Tablo 1'deki her bir sütun yukarıda da ifade edildiği üzere alt ana kütle olarak adlandırılmakta ve sabit bir X_i değerlerine karşılık gelen Y_i bağımlı değişkeninin **koşullu**

dağılımını vermektedir. Buna göre, yukarıdaki her bir sütun belli bir gelir düzeyine karşılık gelen tüketim harcamalarının koşullu dağılımını vermektedir. Tablodan da açıkça görüldüğü üzere, belirli bir gelir düzeyine sahip (X_i) bir gruba ait tüm hane halklarının rassal veya tepkisel satın alma gibi pek çok diğer faktörden dolayı aynı tüketim harcamasında (Y_i) bulunması beklenemez. Diğer bir ifade ile bağımsız değişkenin (gelir) sabit değerine karşın bağımlı değişkenin (tüketim) birbirinden farklı değerleri vardır. Örneğin, 80 $ eşit gelire sahip 5 aile, 55$, 60$, 65$, 70$ ve 75 $ olmak üzere farklı tüketim harcamalarında bulunmaktadır. Farklı tüketim harcamaları ise bir olasılık dağılım (olasılık yoğunluk fonksiyonu) oluşturmaktadır. Bazıları daha fazla, bazıları daha az harcarlarsa da harcama rakamlarının söz konusu gelir düzeyini hedef alan bir değer etrafında toplanacakları beklenebilir. Böylece olasılık dağılımlı her alt ana kütle için **beklenen değer** (koşullu ortalama) $E\left(Y_i \middle| X = X_i\right)$ (veya $E\left(Y_i \middle| X_i\right)$, $E\left(Y\right)$, $\mu_{Y|X}$ olarak gösterilir) hesaplanır.

Beklenen değer (Koşullu ortalama): Bir rassal değişkenin beklenen değeri, onun "ortalama" değeri olarak isimlendirilir ve bu rassal değişkenin olasılık dağılımının merkezi olan anakütle ortalamasının gerçek bir değeridir. Bu, sayısal değerlerin aritmetik ortalaması olan *"örnek ortalaması"* ile aynı değildir. Kısaca Y rastlantısal değişkeninin çok sayıda veya meydana gelişte alacağı değerlerin ortalamasıdır.

Beklenen değerin hesaplanabilmesi için öncelikle Y'nin her alt ana kütle için koşullu olasılıklarının bulunması gerekir. Her alt ana kütle için Y'nin **koşullu olasılığı** $P(Y|X)$ ile gösterilir. Haftalık harcanabilir gelir (Y) 80$ birbirinden farklı 5 tane haftalık tüketim harcaması olduğuna göre, bunların her birinin gerçekleşme olasılığı 1/5'dir. Diğer bir ifade ile 80$ geliri olan 5 aileden oluşan bir alt ana kütleden rastlantısal olarak seçilen bir ailenin 55$ haftalık tüketim harcaması olan aile olma koşullu olasılığı 1/5 'e eşittir ve aşağıdaki gösterilir:

$$P\left(Y = 55 \middle| X = 80\right) = 1/5$$

Diğer bir alt ana kütle için örnek verecek olursak, haftalık harcanabilir geliri 220$ olan 7 aile vardır. Bu ailelerin tüketimleri birbirlerinden farklıdır ve her birinin gerçekleşme olasılığı 1/7'dir. Buna göre 220$ haftalık geliri olan 7 aileden oluşan bir alt ana kütleden rastlantısal olarak seçilen bir ailenin 157$ haftalık tüketim harcaması olan aile olmasının koşullu olasılığı 1/7'ye eşittir.

$$P\left(Y = 157 \middle| X = 220\right) = 1/7$$

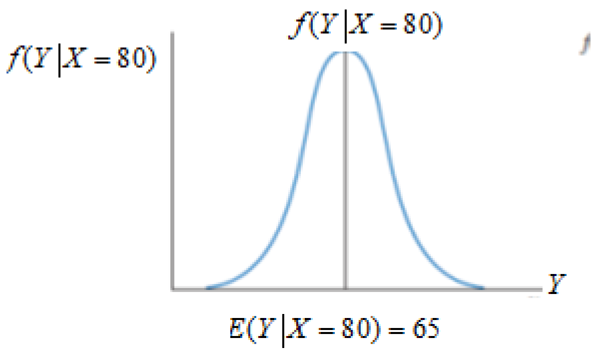
Y’nin her bir koşullu olasılık dağılımı için, **koşullu ortalama (beklenen değer)** $E\left(Y \mid X = X_i\right)$ hesaplanacaktır. Her beklenen değer, X_i’nin doğrusal bir fonksiyonudur.

X=80 $ iken Y’nin koşullu ortalaması veya diğer bir ifade ile beklenen değeri:

$$E\left(Y \mid X = 80\right) = 55\frac{1}{5} + 60\frac{1}{5} + 65\frac{1}{5} + 70\frac{1}{5} + 75\frac{1}{5} = (55 + 60 + 65 + 70 + 75)\frac{1}{5} = 65$$

$$E\left(Y \mid X = 80\right) = 65$$

‘e eşittir.



Şekil: Gelir 80$ iken tüketim harcaması Y’nin olasılık dağılımı $f(Y \mid X = 80)$

Kısaca aynı 80 $ gelir seviyesindeki 5 ailenin birbirinden farklı tüketim harcamalarının koşullu ortalaması 65 $’a eşittir.

Diğer bir alt ana kütle, X=220 $ iken Y’nin koşullu ortalaması veya beklenen değeri:

$$E\left(Y \mid X = 220\right) = (135 + 137 + 140 + 152 + 157 + 160 + 162)\frac{1}{7} = 149$$

$$E\left(Y \mid X = 220\right) = 149$$

‘a eşittir.

Böylece 10 alt ana kütle için 10 tane koşullu ortalama (beklenen değer) hesaplanacaktır. Hesaplanan koşullu ortalamalar tabloda yer almaktadır.

Bu uygulamada alt ana kütlelerin birim mevcutları farklıdır. Alt ana kütle birim mevcutlarının aynı olması halinde, ana kütle içinde bütün alt ana kütlelerin koşullu olasılıkları aynı olacaktır.

$P(Y/X)$ \ X_i	80	100	120	140	160	180	200	220	240	260
Koşullu Olasılıklar	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
Y'nin koşullu ortalamaları	65	77	89	101	113	125	137	149	161	173
Y'nin koşullu varyansları= σ^2	50	66	46.4	133.42	57.3	116.6	82.4	112	310.7	216.5

Tablo: Koşullu Olasılıklar

Her alt ana kütlenin bir dağılımı olduğuna göre, alt ana kütlelerin koşullu varyansları ve dolayısıyla standart sapması hesaplanabilir. Varyans ve standart sapma olasılık dağılımının yayılışını ölçtüğünü bildiğimize göre, Y 'nin koşullu varyansı $Var(Y|X = X_i)$, hane halkı tüketimlerinin (Y), ortalamaları $E(Y|X = X_i)$ etrafındaki yayılımını ölçer. Alt ana kütle için Y 'nin koşullu varyansı ise:

$$\sigma_i^2 = E\left[Y_i - E\left(Y_i | X = X_i\right)\right]^2 = \frac{\sum_{i=1}^{n}\left[Y_i - E\left(Y_i | X = X_i\right)\right]^2}{N}$$

ile ifade edilmektedir. Örneğin

X=80 iken Y'nin koşullu varyansı:

$E\left[Y_i - E\left(Y | X = 80\right)\right]^2 =$ $E\left(Y | X = 80\right) = 65$ olduğu bilindiğine göre

$$E\left[Y_i - 65\right]^2 = \frac{(55-65)^2 + (60-65)^2 + (65-65)^2 + (70-65)^2 + (75-65)^2}{5} = 50$$

'e eşittir

X=100 iken Y'nin koşullu varyansı ise

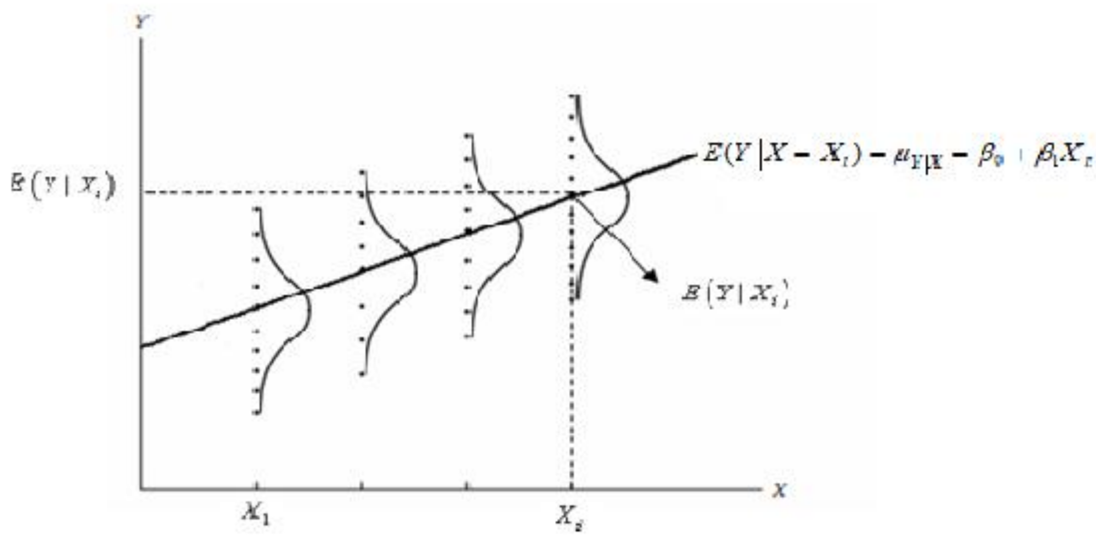
$$E\left[Y_i - E\left(Y | X = 100\right)\right]^2 =$$

$$E\left[Y_i - 77\right]^2 = \frac{(65-77)^2 + (70-77)^2 + (74-77)^2 + (80-77)^2 + (85-77)^2 + (88-77)^2}{6} = 66$$

’ya eşittir.

Hesaplamalara göre alt ana kütle varyanslarının farklı olduğu sonucuna varılmıştır.

Ana kütle regresyon fonksiyonu, X_i veriyken Y’nin ana kütle ortalama dağılımının X_i ile fonksiyonel ilişkili olduğunu gösterir.



Şekil: Ana kütle regresyon doğrusu

Yani X’teki değişmeye karşılık Y’nin ortalama tepkisini gösterir. Böylece, **doğrusal ana kütle regresyon modeli,**

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

şeklinde ifade edilmektedir. Ana kütle regresyon modeli, ana kütle regresyon denklemi olarak da adlandırılmaktadır. Ana kütle regresyon modelinde yer alan β_0 ve β_1 modelin bilinmeyen parametreleridir. β_0 sabit (otonom) parametre, β_1 ise eğim parametresi olarak adlandırılır. β_0 parametresi, ana kütle regresyon doğrusunun koordinat sisteminde Y eksenini kestiği nokta, β_1 ise ana kütle regresyon doğrusunun eğimidir. β_1, X bağımsız değişkenindeki 1 birimlik değişme gerçekleştiğinde $E(Y \mid X_i)$ deki değişme miktarını, diğer bir ifade ile marjinal değişmeyi göstermektedir ve aşağıdaki gibi ifade edilmektedir.

$$\beta_1 = \frac{\Delta E(Y \mid X_i)}{\Delta X} = \frac{dE(Y \mid X_i)}{dX}$$

Doğrusallıktan kasıt, Y nin beklenen değerinin X_i ve parametrelerin doğrusal bir fonksiyonu olmasıdır.

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

Y ve X değişkenleri doğrusal olmayabilir.

Model parametreleri doğrusal mı?	Model değişkenlerde doğrusal mı?	
	Evet	**Hayır**
Evet	DRM	DRM
Hayır	DORM	DORM

Açıklama:DRM= Doğrusal Regresyon modeli
DORM= Doğrusal olmayan Regresyon Modeli

$E(Y \mid X) = \beta_0 + \beta_1 X$ denkleminde β_0 ve β_1 parametrelerinin üstlerinin 1 olması dolayısıyla model doğrusaldır. Hem parametreler hem de değişkenler açısından doğrusal olan regresyon modelleri birinci mertebeden regresyon modelleri olarak adlandırılır.

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i^2$$

İkinci dereceden doğrusal bir regresyon modelidir. Bağımsız değişkenin üstlerinin azamisi modelin mertebesini verir.

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

X=0 olduğunda Y'nin beklenen değeri ($E(Y \mid X_i = 0)$) β_0'a eşittir.

$$E(Y \mid X_i = 0) = \beta_0$$

ancak β_0'ı bu açıdan yorumlanması, X=0 etrafında veri bulunmasına bağlıdır. Aksi halde bu yorum geçerli değildir.

Basit Doğrusal Regresyon Modelinin Temel Varyasımları 1

1. *X*'in her bir değeri için, *Y*'nin ortalama değeri, aşağıdaki doğrusal regresyon fonksiyonu ile elde edilir,

   $$E(Y \mid X = X_i) = \beta_0 + \beta_1 X_i$$

2. *X*'in her bir değeri için, *Y*'nin değerleri, ortalama değerleri etrafında dağılır ve aynı varyansa sahip olasılık dağılımlarını izler,

$$E\left[Y_i - E\left(Y_i \middle| X = X_i\right)\right]^2 = Var(Y|X) = \sigma^2$$

3. *Y*'nin örnek değerlerinin hepsi korelasyonsuz ve sıfır kovaryansa sahiptir –ki bu durum, bunlar arasında doğrusal bir birliktelik olmadığı anlamına gelir.

   $$Cov(Y_i, Y_j | X_i, X_j) = E\left\{\left[Y_i - E\left(Y_i \middle| X = X_i\right)\right]\left[Y_j - E\left(Y_j \middle| X = X_i\right)\right]\right\} = E(Y_i, Y_j) \qquad i \neq j$$

   $$Cov(Y_i, Y_j) = 0$$

   Bu varsayım, *Y*'nin tüm değerlerinin istatiksel olarak bağımsız olduğu varsayımı yapılarak daha güçlü yapılabilir.

4. *X* değişkeni rassal değildir diğer bir ifade ile tekrarlanan örneklerde değeri değişmez ve en azında iki farklı değer almak zorundadır.

   $$Var(X) = \frac{\sum_{i=1}^{n}(X_i - \mu_x)^2}{n} > 0 \qquad \mu_x \text{ veya } \bar{X}, \quad X \text{ değişkeninin ortalaması}$$

   Aksi taktirde ana kütle parametrelerinin tahmini olanaksız hale gelecektir. *X*'ler aynı ise $X_i = \mu_x$ olacak X'deki değişim sıfıra eşit olacaktır. Regresyon analizinde hem *Y* hem de *X*'te değişim zorunludur.

5. *X*'in her bir değeri için, *Y* değerleri, ortalamaları etrafında normal dağılır.

   $$Y \sim N\left[(\beta_0 + \beta_1 X), \sigma^2\right]$$

<u>Ana kütle regresyon denklemi</u>

Klasik doğrusal temel varsayımlarından elde edilen bilgilerle, yukarıdaki hipotetik verilerin kullanıldığı ana kütle için, ana kütle regresyon denklemini oluşturalım.

Öncelikle, hatırlanacağı üzere, ana kütle regresyon modeli,

$$E(Y|X) = \beta_0 + \beta_1 X$$

veya

$$\mu_Y = \mu_{YX_i} = \beta_0 + \beta_1 \mu_X$$

gibi gösterilmektedir. Amacımız β_0 ve β_1 'in sayısal değerlerini bulmaktır. Yukarıda her alt ana kütle için koşullu ortalama (beklenen değeri) hesaplanmıştı.

Öncelikle bu alt ana kütle koşullu ortalamaları hesaplanarak ana kütle regresyonun doğrusunun eğimini veren β_1 hesaplanacaktır:

$$\beta_1 = \frac{E\left(Y \middle| X = 100\right) - E\left(Y \middle| X = 80\right)}{X_2 - X_1} = \frac{77-65}{100-80} = \frac{12}{20} \quad \text{den}$$

$\beta_1 = 0.6$

veya başka iki noktadan;

$$\beta_1 = \frac{E\left(Y \middle| X = 220\right) - E\left(Y \middle| X = 200\right)}{X_2 - X_1} = \frac{149-137}{220-200} = \frac{12}{20} \quad \text{den}$$

$\beta_1 = 0.6$

veya,

$$\beta_1 = \frac{E\left(Y \middle| X = 120\right) - E\left(Y \middle| X = 80\right)}{X_3 - X_1} = \frac{89-65}{120-80} = \frac{24}{40} \quad \text{den}$$

$\beta_1 = 0.6$ ‘ye eşittir.

Bağımsız değişkenin(*X*) koşulsuz ortalaması$\left(\mu_x\right)$:

$$\mu_x = \frac{5}{60}80 + \frac{6}{60}100 + \ldots\ldots\ldots\ldots\ldots\ldots \frac{7}{60}260 = \frac{10420}{60}$$

$\mu_x = 173.67$

olarak hesaplanır.

Bağımlı değişkenin (*Y*) koşulsuz ortalaması$\left(\mu_Y\right)$:

$$\mu_Y = \frac{\sum\sum Y_{ij}}{N} = \frac{7272}{60} = 121.2$$

sonucuna ulaşılır. Bu değer aynı zamanda Y’nin koşullu ortalamasına($E(Y \mid X = X_i) = \mu_{Y|X}$) eşittir.

veya,

$$\mu_{Y|X} = \frac{5}{60}65 + \frac{6}{60}77 + \frac{5}{60}89 \ldots\ldots\ldots\ldots\ldots\ldots \frac{6}{60}161 + \frac{7}{60}173 = \frac{7272}{60} = 121.2$$

hesaplanır. Y ve X’in koşulsuz ortalamaları ve$\beta_1$hesaplandığınaβ_0 göre aşağıdaki ana kütle regresyon modeli ile hesaplanabilir.

$$\mu_{Y|X} = \mu_Y = \beta_0 + \beta_1 \mu_x$$

$$121.1 = \beta_0 + 0.6 \times 173.67$$

$$\beta_0 = 17$$

$$E(Y \mid X_i) = 17 + 0.6 X_i$$

olarak bulunmuş olur. Anakütle regresyon denklemine göre gelir 1 birim değiştiğinde tüketim 0.6 birim artacaktır.Tanımgereği 0.6 marjinal tüketim meyline eşittir. Otonom parametre ise zorunlu tüketim harcamalarının 17 birim olduğunu göstermektedir.

Hata terimi

Regresyon analizinin özü, bağımlı değişken *Y* ile ilgili herhangi bir gözlemin, sistematik bileşen ve rassal bileşen olarak iki kısma ayrılabilmesidir. *Y*'nin sistematik bileşini, kendi ortalamasıdır, $E(Y|X) = \beta_0 + \beta_1 X$ -ki bu ortalama matematiksel bir beklenti olduğu için rassal değildir. *Y*'nin rassal bileşeni, *Y* ve koşullu ortalama değeri $E(Y|X)$ arasındaaki farktır. Bu **rassal hata terimi** olarak isimlendirilir.

Ana kütle regresyon modeli $E(Y \mid X_i) = \beta_0 + \beta_1 X_i$ olduğuna göre, bağımlı değişkenin gözlemlenen değeri Y_i aşağıdaki gibi yazılabilir.

$$Y_i = E(Y \mid X_i) + u_i$$

Yukarıdaki eşitliğine göre hata terimi (u_i), bağımlı değişkenin gözlemlenen değeri ile beklenen değeri arasındaki farka eşittir.

$$u_i = Y_i - E(Y \mid X_i) = Y_i - \beta_0 - \beta_1 X_i$$

Rassal hata terimlerini simgeleyen u_i, bağımlı değişken ile bağımsız değişken arasında ortalama ilişkiyi gösteren ($\beta_0 + \beta_1 X_i$) kısım arasındaki farka eşittir. Alt örnekler itibariyle dikkate alındığında ise her alt ana kütlenin birimlerinin alt ana kütle ortalamasından sapmalarını göstermekte, eğer $Y_i > E(Y \mid X_i)$ ise artı,$Y_i < E(Y \mid X_i)$ ise eksi değer almaktadır.

Böylece birinci alt ana kütle için hata terimleri sırasıyla

55-65= -10

60-65=-5

65-65=0

70-65=5

75-65=10

Böylece ikinci alt ana kütle için hata terimleri sırasıyla

65-77=-12

70-77=-7

74-77=-3

80-77=3

85-77=8

88-77=11

olarak hesaplanır. u_i artı ve eksi değer alabilen gözlenemeyen tesadüfi bir terimdir. Ana kütle hata terimi olarak adlandırılır.

Rassal hata (u_i), bağımlı değişken (Y_i) ile aynı özellikleri göstermesine rağmen gözlemlenemeyen rastlantısal değişkendir.

$$Y_{ij} = \beta_1 + \beta_2 X_i + u_{ij}$$

X=80 iken *u*'nin koşullu ortalaması (beklenen değer):

$$E\left(u \middle| X = 80\right) = 10\frac{1}{5} + 5\frac{1}{5} + 0\frac{1}{5} + (-5)\frac{1}{5} + (-10)\frac{1}{5} = \left[10 + 5 + 0 + (-5) + (-10)\right]\frac{1}{5} = 0$$

$$E\left(u \middle| X = 80\right) = 0$$

X=100 iken *u*'nin koşullu ortalaması (beklenen değer):

$$E\left(u \middle| X = 100\right) = (-12)\frac{1}{6} + (-7)\frac{1}{6} + (-3)\frac{1}{6} + 3\frac{1}{6} + 8\frac{1}{6} + 11\frac{1}{6} = \left[(-12) + (-7) + (-3) + 3 + 8 + 11\right]\frac{1}{6} = 0$$

$$E\left(u \middle| X = 100\right) = 0$$

X=80 iken *u*'nin koşullu varyansı:

$E\left[u_i - E\left(u \middle| X = 80\right)\right]^2 =$ $E\left(u \middle| X = 80\right) = 0$ olduğu bilindiğine göre

$$E\left[u_i - 0\right]^2 = \frac{(-10)^2 + (-5)^2 + (0)^2 + (5)^2 + (10)^2}{5} = 50$$

X=100 iken *u*'nin koşullu varyansı ise

$$E\left[u_i - E\left(u \middle| X = 100\right)\right]^2 = E\left[u_i - 0\right]^2 = \frac{(-12)^2 + (-7)^2 + (-3)^2 + (3)^2 + (8)^2 + (11)^2}{6} = 66$$

Hata terimi, rastlantısal bir değişken olduğuna göre bağımlı değişken (*Y*) gibi hata terimi de olasılık dağılımına sahiptir ve dolayısıyla koşullu ortalaması (beklenen değeri) ve varyansı hesaplanabilir. Hata teriminin (u_i) koşullu ortalaması (beklenen değeri) sıfıra eşittir. Şöyle ki;

$$Y_i = E(Y \mid X_i) + u_i$$

Eşitliğinin her iki yanının beklenen değeri alınır.

$$E(Y_i \mid X_i) = E[E(Y \mid X_i)] + E(u_i \mid X_i)$$
$$= E(Y \mid X_i) + E(u_i \mid X_i)$$

$E[E(Y \mid X_i)] = E(Y \mid X_i)$ olduğundan

$E(u_i \mid X_i) = 0$ olacaktır.

<u>Basit Doğrusal Regresyon Modelinin Temel Varyasımları II</u>

1. *X*'in her bir değeri için, *Y*'nin değeri,

   $Y_i = \beta_0 + \beta_1 X_i + u_i$
2. Rassal hata *u*'nin beklenen değeri,

   $E(u_i) = 0$

Ana kütle regresyon doğrusunun bağımlı değişkenin beklenen değerlerinden geçtiği varsayımı, hata terimlerinin beklenen değerlerinin sıfır olduğu anlamına gelmektedir.

Böylece $E(u_i) = 0$ varsayımı aşağıdaki varsayıma eşdeğerdir:

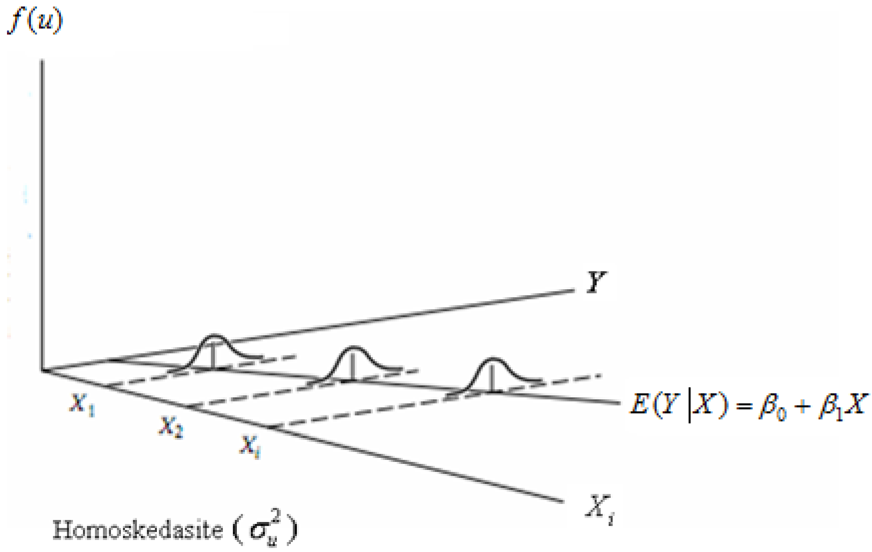
$E(Y_i) = \beta_0 + \beta_1 X_i$

3. Rassal hata *u*'nin varyansı,

   $E\left[u_i - E\left(u_i \middle| X = X_i\right)\right]^2 = Var(u|X) = \sigma^2$

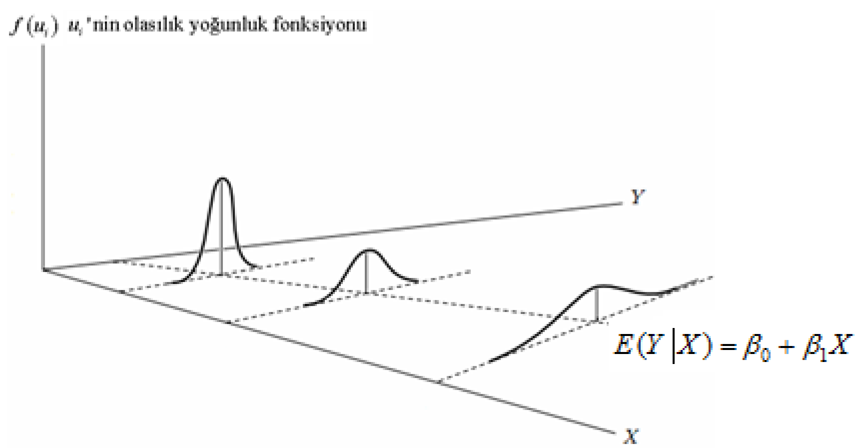
   $Var(u|X) = Var(Y|X) = \sigma^2$

Rassal değişkenler *Y* ve *u* yalnızca bir sabit ile farklı olabildikleri için, aynı varyansa sahiptirler.



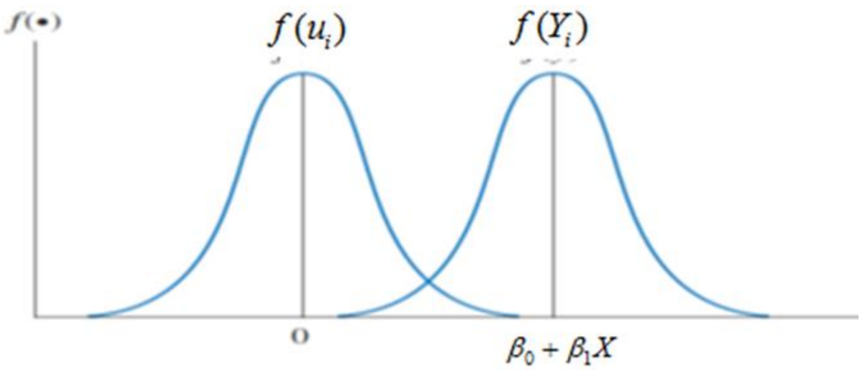
Yukarıdaki şekilde görüldüğü üzere her *X* değerine karşılık gelen hata terimleri alt ana kütlelerin ortalamaları etrafında aynı dağılımı göstermektedir. Alt ana kütlelerin hata terimlerinin aynı değişim aralığında içinde yer almaları yani varyanslarının eşit olması, homoskedasite veya sabit varyans olarak adlandırılmaktadır.

X'in değeri artarken, alt ana kütle hata terimlerinin varyanslarıküçüldüğü görülmektedir.X'in değeri artarken, alt ana kütle hata terimlerinin varyanslarının büyüdüğü durum da söz konusu olması mümkündür.X'in aldığı değere göre hata teriminin varyansının değişmesi ise heteroskedasite veya değişen varyans durumudur. Alt ana kütlelerin birim sayısı eşit olsa bile varyansları eşit olmayabilir.



Yukarıdaki şekilde ise *X*'in değeri artarken, alt ana kütle hata terimlerinin varyanslarının arttığı görülmektedir.*X*'in değeri artarken, alt ana kütle hata terimlerinin varyanslarının küçüldüğü durum da söz mümkündür. X'in aldığı değere göre hata teriminin varyansının değişmesi ise heteroskedasite veya değişen varyans durumudur. Alt ana kütlelerin birim sayısı eşit olsa bile varyansları eşit olmayabilir.

Regresyon doğrusu etrafında dağılmayı gösteren bağımlı değişken Y'nin koşullu varyansı diğer bir ifade ile ana kütle regresyon varyansı ($\sigma^2_{Y|X}$)hata teriminin varyansına($\sigma^2_{u|X}$) eşittir.



4. Rassal hatalar u_i ve u_j nin herhangi bir çifti arasındaki kovaryans,

$$Cov(u_i, u_j) = Cov(Y_i, Y_j) = 0 \qquad i \neq j$$

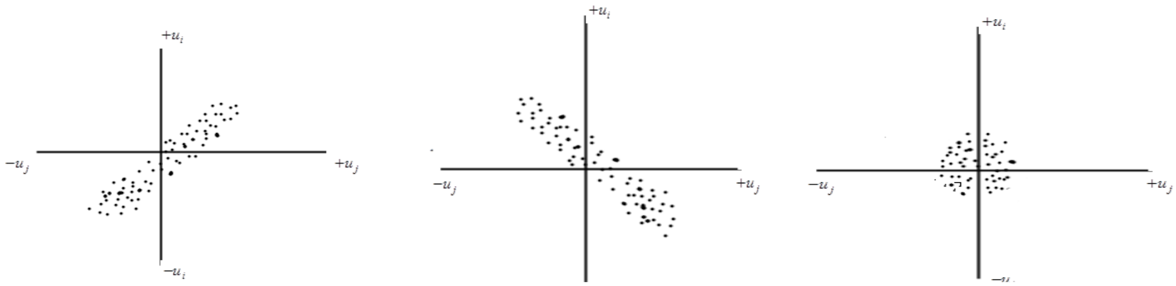
Hata terimleri arasında otokorelasyon(ardışık bağımlılık) yoktur. Bu varsayıma göre X_i ve X_j gibi iki X verilmişken u_i ile u_j arasında ilişki yoktur. Hata terimlerinin birbirleri ile ilişkisiz olması diğer bir ifade ile birbirlerinden bağımsız olmaları ancak $Cov(u_i u_j) = 0$ olması ile mümkündür. Sembolik olarak

$$Cov(u_i u_j | X_i\, X_j) = E\{[u_i - E(u_i | X_i)][u_j - E(u_j | X_j)]\} \qquad i \neq j$$

$$= E(u_i u_j) = 0$$

ile gösterilir.

Bu varsayım gereği, X_i veriyken herhangi iki Y değerinin ortalamalarından sapmaları belli bir kural izlemezler. Varsayımın gerçekleşmemesi durumunda modelin aleyhine negatif veya pozitif otokorelasyon gerçekleşmektedir.



Pozitif Otokorelasyon Negatif Otokroelasyon Otokorelasyon Olmama

Şekil-3.3: Otokorelasyonun çeşitleri

Ana kütle regresyon modelinde($Y_t = \beta_0 + \beta_1 X_t + u_t$), Y_t bağımlı değişkeninin değeri, X_t bağımsız değişkenine ve u_t hata terimine bağlıdır. Ancak u_t ile u_{t-1} arasında otokorelasyon varsa, örneğin u_t ile u_{t-1}pozitif olarak birbirleriyle ilişkiliyse, Y_t sadece X_t ve u_t değil, aynı zamanda u_{t-1} ile de ilişkili olacaktır. Bu durumda bağımsız değişken ve hata teriminin, bağımlı değişken üzerindeki bireysel etkilerini tespit etmek mümkün olmayacaktır.

5. *u* değerleri, ortalamaları etrafında normal dağılır.

$u \sim N\left(0, \sigma^2\right)$

Y değerleri normal dağılırsa veya tersi durum geçerlidir.

6. Hata terimi(u_i) ile *X* bağımsız değişkeni ilgisizdir. Bu varsayım hata terimi ile bağımsız değişkenin otokorelasyonsuz olduğunu ifade etmektedir.

$$Cov(u_i, X_i) = 0$$

$$Cov(u_i X_i) = E\left\{\left[u_i - E(u_i | X_i)\right]\left[X_i - E(X_i)\right]\right\} \qquad E(u_i | X_i) = E(u_i) = 0\text{'dan}$$

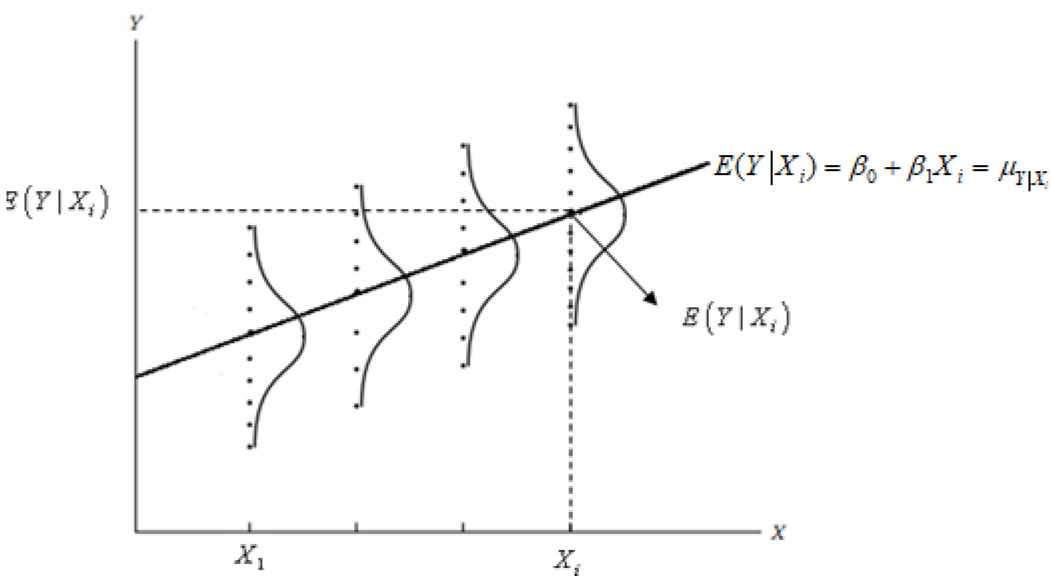
$$= E\left[u_i (X_i - E(X_i)\right] = 0 \qquad E(X_i) \text{ olasılıklı değişken değil } X_i = \bar{X}$$

$$= \mathrm{E}(u_i X_i) - E(X_i) E(u_i)$$

$$= \mathrm{E}(u_i X_i) = 0$$

Bu varsayım bağımsız değişken (sistematik kısım) ve hata teriminin(tesadüfi kısım) bağımlı değişken üzerindeki etkisi ayrı ayrı ölçülebilir ve toplanabilir olduğu anlamına gelmektedir. Eğer hata terimi(u_i) ile *X* bağımsız değişkeni ilişkili ise bağımlı değişken üzerindeki bireysel etkilerini tespit etmek mümkün olamayacaktır. X ile u_t pozitif yönde ilişkiliyse X artarken u_t de artacak, X azalırken u_t de azalacaktır. X ile u_t negatif yönde ilişkiliyse X artarken u_t de azalacak, X azalırken u_t artacaktır.

Ana kütle regresyon doğrusunun bağımlı değişkenin beklenen değerlerinden geçtiği varsayımı, hata terimlerinin beklenen değerlerinin sıfır olduğu anlamına gelmektedir.



Ana kütle Regresyon Fonksiyonu

Yukarıdaki şekilde görüldüğü üzere n sayıdaki (yukarıdaki örnekte n 60'a eşittir.) Y ve X gözlem çiftleri bir dağılm diyagramı üzerinde gösterilebilir. Geometrik olarak, ana kütle regresyon doğrusu, açıklayıcı değişkenlerin sabit değerlerine karşılık gelen bağımlı değişkenin koşullu ortalamalarından (beklenen değerlerinden) geçer.

Olasılık dağılımı rastlantısal değişkenin bütün mümkün değerlerini ve bu değerlerin olasılıklarını ortaya koymaktadır.

Olasılık dağılımının beklenen değeri (koşullu ortalaması, matematik ümidi)

$E(Y) = \mu_Y = \sum Y_i P(Y_i)$ $\qquad$ Y_i : olası değerler $\qquad$ $P(Y_i)$: koşullu olasılık

Olasılık dağılımının koşullu varyansı

$$Var(Y) = \sigma_Y^2 = E\left[Y_i - E(Y)\right]^2 = \sum (Y_i - E(Y))^2 P(Y_i)$$

Örnek

	X_i		
	50	**100**	**200**
Y_{ij}	10	18	23
	14	12	19

İspat

$\beta_0 + \beta_1 X_i$, X_i ile Y_i arasınadi ortalama ilişkiyi yansıtmaktadır.

i. gözlemde *Y* için aşağıdaki denklem yazılır.

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

Heriki tarafın beklenen değeri alınır.

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + u_i)$$

$$E(Y_i) = \beta_0 + \beta_1 X_i + E(u_i) \quad \text{burada} \quad E(u_i) = 0$$

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

ve

$$u_i = Y_{ij} - E(Y_{ij} \mid X_i)$$

Heriki tarafın beklenen değeri alınır.

$$E(u_i) = E\left[Y_{ij} - E(Y_{ij} \mid X_i)\right]$$

$$E(u_i) = E(Y_{ij}) - E\left[E(Y_{ij} \mid X_i)\right]$$

burada $E(u_i) = 0$ ve $E\left[E(Y_{ij} \mid X_i)\right] = E(Y_{ij} \mid X_i)$

$$0 = E(Y_{ij}) - E(Y_{ij} \mid X_i)$$

$$E(Y_{ij}) = E(Y_{ij} \mid X_i)$$

Sonuç olarak

$$\mu_Y = \mu_{Y|X} = \beta_0 + \beta_1 \mu_X$$

yazılır.

a) Hipotetik verilerin yer aldığı yukarıdaki ana kütle için ana kütle regresyon denklemini $E(Y \mid X = X_i) = \beta_0 + \beta_1 X_i$ yazın.

Öncelikle koşullu olasılıkla hesaplanır.

$P(Y=10|X=50)=1/2$ $\quad P(Y=18|X=100)=1/2$ $\quad P(Y=23|X=200)=1/2$

$P(Y=14|X=50)=1/2$ $\quad P(Y=12|X=100)=1/2$ $\quad P(Y=19|X=200)=1/2$

	X_i		
	50	**100**	**200**
$P(Y_i \mid X=X_i)$	1/2	1/2	1/2
	1/2	1/2	1/2

Her alt ana kütle için *Y*'nin beklenen değeri hesaplanır.

$$E(Y|X=50)=(10+14)\times\frac{1}{2}=12$$

$$E(Y|X=100)=(18+12)\times\frac{1}{2}=15$$

$$E(Y|X=200)=(23+19)\times\frac{1}{2}=21$$

	X_i		
	50	**100**	**200**
Y_{ij}	10	18	23
	14	12	19
n	2	2	2
$E(Y \mid X=X_i)$	12	15	21

$$E(Y|X=X_i)=\beta_0+\beta_1 X_i \qquad \text{veya} \qquad \mu_Y=\mu_{Y|X}=\beta_0+\beta_1\mu_X$$

Öncelikle eğim parametresi olan β_1 hesaplanır.

$$\beta_1=\frac{E(Y|X=100)-E(Y|X=50)}{100-50}=\frac{15-12}{100-50}=\frac{3}{50}$$

$$\beta_1=0,06$$

X'in koşulsuz ortalaması:

$$\mu_X=\left[(2\times 50)+(2\times 100)+(2\times 200)\right]/6=116,67$$

Y'in koşulsuz ortalaması:

$$\mu_Y = \frac{\sum Y_{ij}}{n} = \frac{96}{6} = 16$$

Y'in koşullu ortalaması:

$$\mu_{Y|X} = \left(\frac{2}{6}\times 12\right) + \left(\frac{2}{6}\times 15\right) + \left(\frac{2}{6}\times 21\right) = 16$$

$$\mu_{Y|X} = \beta_0 + \beta_1 \mu_X$$

$$16 = \beta_0 + (0,06 \times 116,67)$$

$$\beta_0 = 9$$

$E(Y|X = X_i) = 9 + 0,06X_i$ Ana kütle regresyon denklemi

b) Hata teriminin beklenen değerinin sıfır olduğunu gösteriniz.

$Y_i = \beta_0 + \beta_1 X_i + u_i$ $\quad$ $Y_i = E(Y|X = X_i) + u_i$ $\quad$ $u_i = Y_i - E(Y|X = X_i)$

	X_i		
	50	**100**	**200**
Y_{ij}	10	18	23
	14	12	19
$E(Y\|X = X_i)$	12	15	21
u_{ij}	-2	3	2
	2	-3	-2
	1/2	1/2	1/2
$P(u = u_i\|X = X_i)$	1/2	1/2	1/2

Koşullu olasılıklar

$P(u = -2|X = 50) = 1/2$ $\quad$ $P(u = 3|X = 100) = 1/2$ $\quad$ $P(u = 2|X = 200) = 1/2$

$P(u = 2|X = 50) = 1/2$ $\quad$ $P(u = -3|X = 100) = 1/2$ $\quad$ $P(u = -2|X = 200) = 1/2$

Her alt ana kütle için u'nin beklenen değeri hesaplanır.

$$E(u|X = 50) = [(-2) + (2)] \times \frac{1}{2} = 0$$

$$E(u|X = 100) = [(3) + (-3)] \times \frac{1}{2} = 0$$

$$E(u|X = 200) = [(2) + (-2)] \times \frac{1}{2} = 0$$

	X_i		
	50	**100**	**200**
Y_{ij}	10	18	23
	14	12	19
$E(Y \mid X = X_i)$	12	15	21
u_{ij}	-2	3	2
	2	-3	-2
$E(u \mid X = X_i)$	0	0	0

c) Bağımlı değişken (Y) ile rassal hatanın (u) koşullu varyanslarının eşit olduğunu gösterin

Bağımlı değişkenin koşullu varyansı:

$$E\left[Y_i - E\left(Y \mid X = 50\right)\right]^2 = \left[(10-12)^2 + (14-12)^2\right] \times \frac{1}{2} = 4$$

$$E\left[Y_i - E\left(Y \mid X = 100\right)\right]^2 = \left[(18-15)^2 + (12-15)^2\right] \times \frac{1}{2} = 9$$

$$E\left[Y_i - E\left(Y \mid X = 200\right)\right]^2 = \left[(23-21)^2 + (19-21)^2\right] \times \frac{1}{2} = 4$$

Hata teriminin koşullu varyansı

$$E\left[u_i - E\left(u \mid X = 50\right)\right]^2 = \left[(-2-0)^2 + (2-0)^2\right] \times \frac{1}{2} = 4$$

$$E\left[u_i - E\left(u \mid X = 100\right)\right]^2 = \left[(3-0)^2 + (-3-0)^2\right] \times \frac{1}{2} = 9$$

$$E\left[u_i - E\left(u \mid X = 200\right)\right]^2 = \left[(2-0)^2 + (-2-0)^2\right] \times \frac{1}{2} = 4$$

	X_i		
	50	**100**	**200**
Y_{ij}	10	18	23
	14	12	19
$E(Y \mid X = X_i)$	12	15	21
$E\left[Y_i - E\left(Y \mid X = X_i\right)\right]^2$	4	9	4
u_{ij}	-2	3	2
	2	-3	-2
$E(u \mid X = X_i)$	0	0	0

$E\left[u_i - E\left(u \middle| X = X_i\right)\right]^2$	4	9	4

İspat

$$E\left[Y_i - E\left(Y \middle| X = X_i\right)\right]^2 = E\left[u_i - E\left(u \middle| X = X_i\right)\right]^2 = \sigma^2$$

$$\sigma^2_{Y|X} = \sigma^2_{u|X}$$

$$\sigma^2_{Y|X} = E\left[Y_i - E(Y)\right]^2$$

burada

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \text{ve} \quad E(Y) = \beta_0 + \beta_1 X_i$$

$$\sigma^2_{Y|X} = E\left[(Y_i = \beta_0 + \beta_1 X_i + u_i) - (\beta_0 + \beta_1 X_i)\right]^2$$

$$\sigma^2_{Y|X} = E(u_i)^2 = \sigma^2_u$$

veya

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

burada $\beta_0 + \beta_1 X_i$ sabittir. Birimlere $\beta_0 + \beta_1 X_i$ gibi sabit bir sayı ilave edildiğinde varyans değişmeyeceğinden dolayı

$$Var(Y|X_i) = Var(u_i|X_i) = \sigma^2$$

Bağımlı değişken ile rassal hata aynı dağılıma sahiptir. Ortalamaları farklıdır.

$$Y_i \sim N((\beta_0 + \beta_1 X_i), \sigma^2) \qquad u_i \sim N(0, \sigma^2)$$

Örnek Kütle Regresyon Modeli

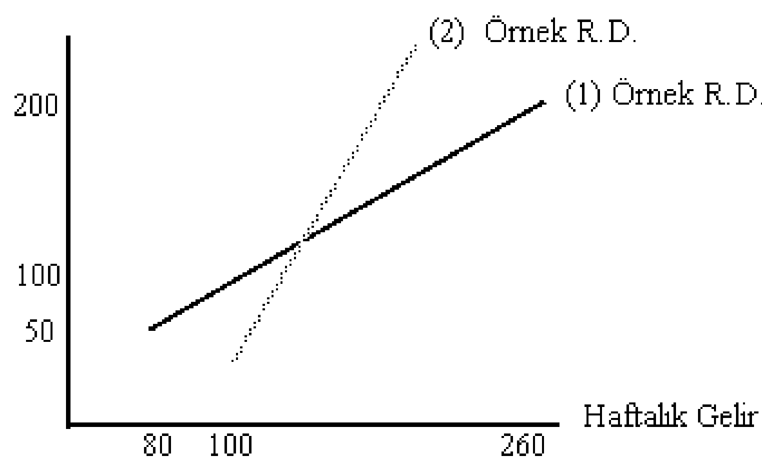
Uygulamada ana kütlenin gözlemlenmesi ve $\beta_0 + \beta_1 X_i$ doğrusunun bilinmemesi nedeniyle, anakütle yerine çoğu kez örnekten hareket edilir. β_0 , β_1 ve hata teriminin varyansı (σ^2) bilinmeyen parametrelerdir. Bu parametreler ana kütle yerine örnek gözlemlerine dayanılarak istatistiksel olarak tahmin edilir. Genellikle X'in sabit değerleri için, *Y* örnekleme değerleri bulunmaktadır.

Ana kütleden çekilen rastlantısal çekilen

Örnek 1 **Örnek 2**

Y	X
65	80
80	100
79	120
113	140
125	160
115	180
144	200
157	220
155	240
178	260

Y	X
55	80
74	100
90	120
103	140
107	160
135	180
144	200
160	220
189	240
150	260



Şekil: Örnek regresyon doğruları

Yukarıdaki iki farklı örneklemde de görüldüğü üzere X değeri, tekrarlanan örneklemlerde değişmemekte, sabit kalmaktadır. Ancak Y stokastik (rastlantısal) bir değişken olduğu için X in her bir değerine birden fazla Y değeri karşılık gelmekte ve dolayısıyla yukarıda da görüldüğü üzere yinelenen örneklemlerde Y'nin değeri değişmektedir.

Regresyon analizdeki birinci amaç, örneklem verilerinin kullanılarak tahmin edilen örnek regresyon fonksiyonuna dayanarak (diğer bir ifade ile örnekten tahmin edilen bilgilerle) ana kütle regresyon fonksiyonunu tahmin etmektir.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$\hat{Y}_i$ ana kütle regresyon modelinde $E(Y \mid X)$ nın tahminidir. $\hat{\beta}_0$, $\hat{\beta}_1$ sırasıyla β_0 ve β_1 nin tahmini verir.

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

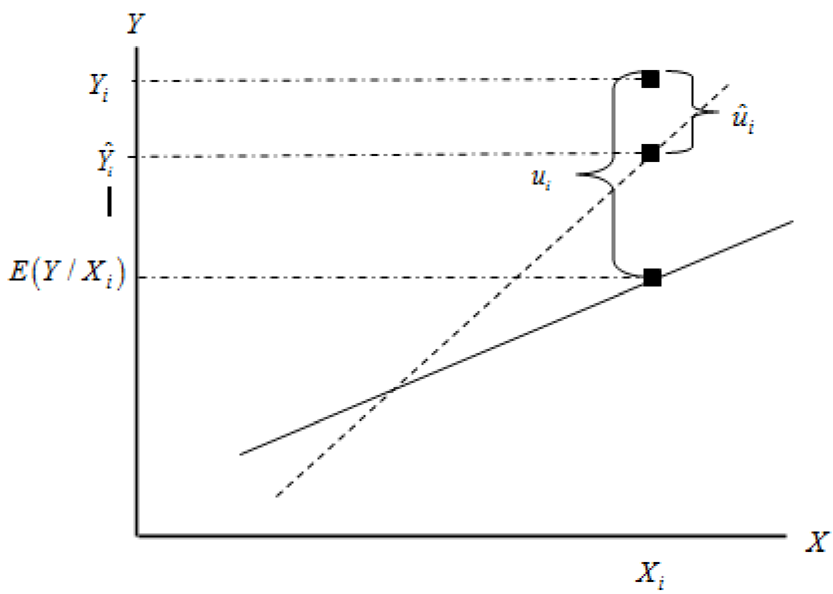
$\hat{u}_i$ kalıntı olarak adlandırılmakta olup, ana kütledeki hata terimini u_i 'nin tahminidir. Böylece ana kütle regresyon fonksiyonu

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

yerini örnek kütle regresyon fonksiyonu

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

alacaktır. Burada önemli bir husus, ana kütle yerine örnekten hareket edildiği durumda ana kütle regresyon denklemi ve dolayısıyla ana kütle hata terimi u_i bilinmediği için u_i 'nin varyansı hesaplanamaz. Örnek regresyon fonksiyonu mümkün olduğunca ana kütle regresyon fonksiyonuna yakın tahmin edilmelidir.



Şekil-Anakütle ve örnek regresyon doğruları

Yukarıdaki şekilde görüldüğü üzere ana kütle regresyon doğrusu her alt ana kütle için hesaplanan bağımlı değişkenin beklenen değerlerinden geçerken, örnek regresyon doğrusu bağımlı değişkenin tahmini değerlerinden geçmektedir.

Hata Teriminin Kaynakları

1. Spesifikasyon (belirlenme) hataları

- Dışlanmış değişken

- Gereksiz değişken

- Matematiksel biçimleme hatası: değişkenler arası fonksiyonel ilişki doğrusal olmadığı halde doğrusal olarak yapılan biçimleme hatası

-Eşanlı modelin tek denklem ile ifade edilmesi

2. Ölçme ve birleştirme hataları

3. Aynı iktisadi olay için bireylerin davranışındaki farklılıklar.