10.1. Basit Korelasyon Analizi

Basit korelasyon analizi, aralarında ilişki olduğu düşünülen en az aralık ya da oran ölçeğinde toplanmış iki değişkenin birlikte değişme derecesini gösteren bir analizdir. Bu tanım içinde yer alan, 'aralarında ilişki olduğu düşünülen' ifadesine özellikle dikkat çekmek istiyoruz.

Korelasyon analizi, aralarında ilişki olduğu bir teori tarafından kabul edilen ya da en azından akla ve mantığa uygun olarak aralarında ilişki olduğu düşünülen iki değişkenin birlikte değişme derecesidir.

Korelasyon analizi iki değişkenin gözlem değerlerinin birlikte nasıl bir değişim içinde olduğunu araştırır. Dolayısıyla, değişkenlerin aldıkları değerlerin seyrine bakar. Bu sebeple, eğer serilerin gözlem değerlerinin seyri benzerlik gösteriyorsa, gerçekte ilişkisiz iki seri arasında da güçlü korelasyon çıkabilir. Bu durumu bir örnek ile açıklamaya çalışalım.

İstanbul'da son 10 yıl içinde şehir hatları vapurlarının yolcu sayıları ile İstanbul'daki evlenme sayısı arasında bir ilişki olup olmadığını korelasyon analiziyle araştırdığımızı ve yüksek pozitif ilişki bulduğumuzu varsayalım. İki değişken arasında güçlü pozitif ilişki olduğunu gösteren korelasyon katsayısına bakarak, vapura binen insan sayısı arttıkça evlenme sayısı artıyor, gibi bir sonuca varırız ki bu tebessüme sebep olacak türden bir yorum anlamı taşır. Vapura binen insan sayısı ile yapılan evlilik sayısı arasında bir ilişki olduğunu düşünmemizi sağlayacak ne bir teori ne de bir mantıklı kurgu bulunmamaktadır. Nitekim, bu iki değişken arasında güçlü bir ilişki olduğu sonucunu doğuran asıl sebep, İstanbul'un nüfus artışıdır. Nüfus arttıkça daha çok insan vapura binmekte, nüfus arttıkça daha çok insan evlenmektedir. Dolayısıyla burada vapura binen insan sayısı ile evlenen insan sayısı üzerinde ortak bir nüfus etkisi bulunmaktadır ve bu etki göz ardı edildiği taktirde sanki bu iki değişken birbirini etkiliyormuş gibi bir sonuç ortaya çıkmaktadır. Şüphesiz bu tamamen yanlış bir değerlendirme ve yorum olacaktır. Dolayısıyla, iki değişken arasında mantıksal bir ilişki bulunması korelasyon analizi için vazgeçilmez bir gerekliliktir.

İki değişken arasındaki ilişki güçlü bir ilişki olabileceği gibi orta düzeyde ya da zayıf bir ilişki de olabilir. Örneğin daha önce de değindiğimiz gelir ile tüketim ilişkisi, tasarruf ile tüketim ilişkisinden daha güçlü bir ilişkidir.

İki değişken arasındaki doğrusal ilişkinin derecesi basit korelasyon katsayısı ile ölçülür. İki değişkenin birlikte değişme derecesini ölçmek amacıyla hesaplanan korelasyon katsayısına **basit korelasyon katsayısı** denir (r) ile gösterilir, çok sayıda değişkenin birlikte değişme derecesini ölçen katsayıya ise **çoklu korelasyon katsayısı** denmektedir ve bu katsayı (R) ile gösterilmektedir. Bu kitap kapsamında sadece basit korelasyon katsayısı üzerinde durulacaktır.

Korelasyon katsayısı -1 ile +1 arasında değer alabilen bir ölçüdür. Başka bir deyişle, korelasyon katsayısının alabileceği en küçük değer -1 ve en büyük değer de +1 olmaktadır.

Korelasyon katsayısının -1 ve +1'e eşit olduğu durumlarda deterministik ilişki söz konusudur ve iki değişken arasında mutlak kesinlikte bir ilişkiden söz edilmektedir. Yukarıda verdiğimiz örnekte, iki hidrojen ve bir oksijen atomunun birleşerek bir su molekülü oluşturduğunu hatırlayalım. 4 hidrojen atomu ile 2 oksijen atomu tepkimeye girecek ve iki molekül su oluşturacaktır, benzer şekilde 20 hidrojen atomu 10 oksijen atomuyla, 100 hidrojen

atomu da 50 oksijen atomuyla tepkimeye girecektir. Görüldüğü gibi, tepkimeye girecek oksijen atomu sayısını biliyorsak, hidrojen atomu sayısını ve oluşacak su molekülü sayısını hatasız olarak öngörebiliriz. Ya da kaç molekül su oluşmasını istiyorsak bunun için gerekli hidrojen ve oksijen atomu sayılarını yine mutlak kesinlikte öngörebiliriz. İşte bu tür matematik kesinlik içinde gerçekleşen ilişkilere **deterministik ilişki** denir ve burada bir tahmin söz konusu değildir.

İki değişken arasında deterministik ilişki söz konusu olduğunda bir değişkenin değerini bilmek diğerinin de değerini mutlak kesinlikte bilmek anlamına gelir ve bu durumda korelasyon katsayısının değeri + 1 veya – 1 olup iki değişken arasında % 100 ilişki durumunu ifade etmektedir.

Deterministik ilişkiler sadece pozitif bilimlerde söz konusudur ve konusu kollektif olaylar olan istatistiğin uygulama alanlarında deterministik ilişkilere rastlanmaz. Zira, kollektif olaylar çok sayıda faktörün etkisi altındadır ve bu sebeple kollektif olaylar söz konusu olduğunda matematik kesinlik içinde oluşan ilişkilere rastlamak mümkün değildir.

Aynı anne babadan dünyaya gelen ikiz çocukları düşünelim. Çocuklar aynı zamanda ve aynı koşullarda dünyaya gelmiş, aynı ailenin kültür ve geleneği ile yetişmiş olsalar bile davranışlarının bire bir aynı olmasını bekleyemeyiz. Bu iki çocuk aynı eğitim sürecinden geçseler, aynı öğretmenlerden dersler alsalar ve aynı şekilde çalışsalar bile bütünüyle aynı başarıyı göstermeleri ve aynı hayatı yaşamaları söz konusu değildir. Dolayısıyla, iki kollektif olay yani iki değişken arasında bir ilişki ve ortak bir yönelim olmakla beraber farklılıklar da bulunmaktadır. İşte bu farklılıklar sonucu korelasyon katsayısının –1 veya + 1 çıkması mümkün değildir.

Korelasyon katsayısının $-1 < r < +1$ olması durumu iki değişken arasında olasılık içeren bir ilişki olduğunu ifade eder ki bu tür ilişkilere **stokastik ilişki** denmektedir.

Başka bir deyişle, iki değişken arasındaki ilişki olduğunda bu iki değişken özünde ortak bir yönelim içinde bulunmaktadır. Ancak bununla beraber, iki değişken üzerinde, içinde bulundukları bu ortak davranış dışında etkiler de söz konusu olmaktadır. İşte iki değişken arasındaki bu genel yönelim dışında pek çok faktörün de etkisinin bulunması ilişkinin stokastik ilişki olarak tanımlanmasına sebep olmaktadır.

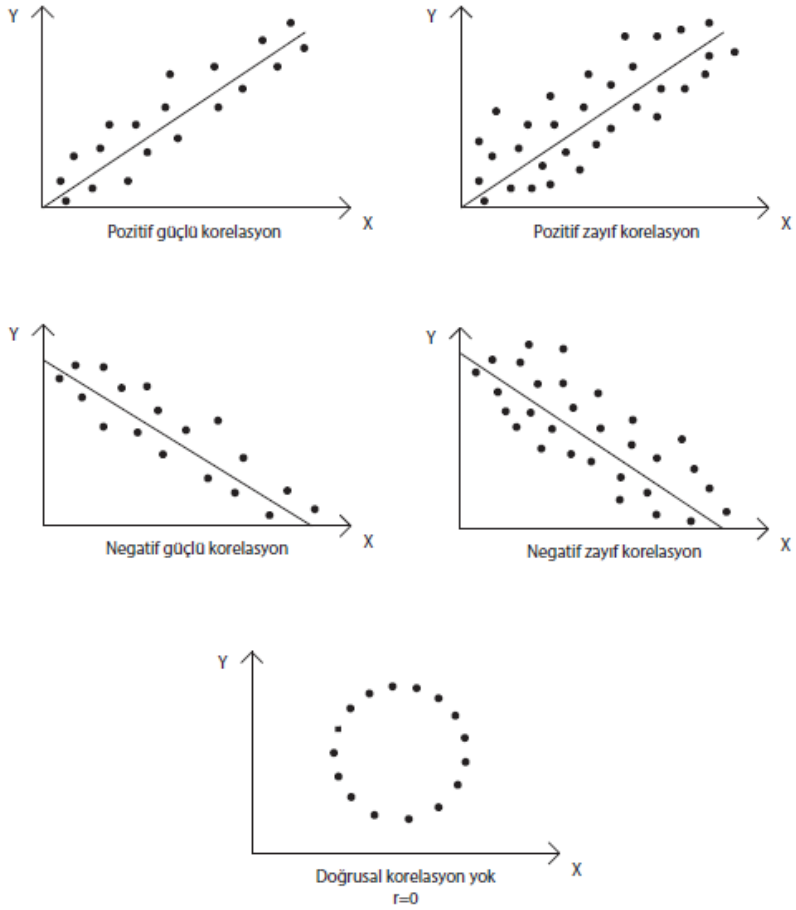
Korelasyon katsayısının pozitif değer alması, iki değişken arasında pozitif bir ilişki bulunduğunu ve iki değişkenin birlikte artıp birlikte azaldığını ifade etmektedir. Daha önce verdiğimiz gelir-tüketim ilişkisi örneği pozitif ilişkiye örnektir ve bu durumda korelasyon katsayısının işareti + olmaktadır.

Korelasyon katsayısının negatif değer alması, iki değişken arasında negatif yönlü bir ilişki bulunduğunu ifade etmekte ve bu durumda değişkenlerden biri artarken diğerinin azalmakta olduğu anlaşılmaktadır. Fiyat ile talep arasındaki ilişki negatif ilişkiye bir örnektir ve bu durumda korelasyon katsayısı – çıkmaktadır.

Korelasyon katsayısının sıfır çıkması ise, iki değişken arasında ilişki olmadığını değil, iki değişken arasında doğrusal bir ilişki bulunmadığını ifade etmektedir. Dolayısıyla, korelasyon katsayısı sıfır ise iki değişken arasında doğrusal formda ilişki yoktur ancak örneğin, ikinci dereceden ya da üstel bir ilişki olabilir.

Korelasyon katsayısının bire yakın değerleri iki değişken arasında güçlü bir ilişki bulunduğunu ifade eder. Benzer şekilde korelasyon katsayısı bir'den uzaklaştığı ölçüde de iki değişken arasındaki ilişkinin zayıf ilişki olduğu anlaşılır. Korelasyon katsayısının sıfıra yaklaşması ise iki değişken arasında zayıf düzeyde doğrusal ilişki bulunabileceğinin göstergesidir.

Korelasyon katsayısının yukarıda açıklanan güçlü ya da zayıf, pozitif veya negatif yönlü bir ilişkiyi göstermesi durumlarını aşağıdaki serpilme diyagramları ile gösterelim:



Şekil 3.1: Serpilme diyagramının iki değişken arasında bulunan korelasyona göre görüntüsü

Korelasyon katsayısı, sebep-sonuç ilişkisine dayanmaz. Burada kastedilen iki değişkenin birlikte değişme derecesidir ve hangi değişkenin sebep hangisinin sonuç olduğu önemli değildir. Dolayısıyla, X ve Y arasında hesaplanacak korelasyon katsayısı Y ile X arasında hesaplanan korelasyon katsayısına eşit olmaktadır.

İki değişken arasındaki ilişkiyi olasılık teorisinde öğrendiğimiz şekilde koşullu bir ilişki gibi düşünür ve ifade edersek, $r_{Y/X} = r_{X/Y}$'dir. Yani X değişkeni sebep Y değişkeni sonuç kabul edildiğinde hesaplanan korelasyon kaysayısı ($r_{Y/X}$), tam tersi durumda yani, Y değişkeni sebep X değişkeni ise sonuç alındığında hesaplanan korelasyon katsayısına ($r_{X/Y}$) eşit olmaktadır. Bu durum iki değişken arasında hesaplanan korelasyon katsayısının simetrik olduğunu ifade eder.

Korelasyon katsayısı aralarında ilişki olduğu düşünülen iki değişkenin tüm gözlem değerleri kullanılarak doğrudan anakütle için hesaplanabileceği gibi, anakütle içinden seçilecek örnek kütlede yer alan gözlem değerlerinden hareketle de hesaplanabilir. Korelasyon katsayısının örnek kütleden hesaplanması durumu uygulamada daha çok tercih edilen bir yöntemdir ve durumda örnek kütleden hareketle hesaplanan korelasyon katsayısı anakütle korelasyon katsayısının tahmini değerini oluşturmaktadır.

Bu bağlamda, ilerleyen kısımlarda görülecek korelasyon katsayısına yönelik örnek çözümlerin örnek kütle korelasyon katsayısı hesabına yönelik olduğunu da ifade etmek isteriz.

Korelasyon katsayısı hesabında kullanılmak üzere geliştirilmiş ve birbirinden türetilmiş çok sayıda formül bulunmaktadır. Bu formüller içinde en çok kullanılan iki tanesi üzerinde duracağız.

Öncelikle, korelasyon katsayısının sebep-sonuç ilişkisine dayanmadığını ve hangi değişkene sebep hangi değişkene sonuç olarak yaklaştığımızın bir önemi olmadığını bir kez daha ifade etmek istiyoruz. Ancak, bir sonraki bölümde inceleyeceğimiz regresyon analizi sebep –sonuç ilişkisine dayanan bir analiz yöntemi olduğundan, sebep değişken X, sonuç değişken ise Y sembolü ile ifade edilmektedir ve bu durum evrensel literatürde de kabul görüp kullanılmakta olan bir yaklaşımdır.

Korelasyon analizinde sebep – sonuç değişken ayrımı gerekmese de, bir sonraki aşamada regresyon analizini öğrenip uygulayacağımız için, şimdiden sebep değişkeni X, sonuç değişkeni Y ile göstermeyi tercih edeceğiz.

10.1.1. Korelasyon Katsayısının Gözlem Değerleri Üzerinden Hesaplanması

X ve Y ile sembolize edilen iki değişken arasındaki korelasyon katsayısı doğrudan gözlem değerleri kullanılarak aşağıdaki formül yardımıyla hesaplanabilir:

$$r = \frac{\sum X_i Y_i \; - \; n\bar{X}\bar{Y}}{\sqrt{\left(\sum {X_i}^2 - \; n\bar{X}^2\right)\left(\sum {Y_i}^2 - \; n\bar{Y}^2\right)}}$$

Formülde yer alan;

$\bar{X}$: X serisinin ortalamasını,

$\bar{Y}$: Y serisinin ortalamasını ve

n : gözlem sayısını ifade etmektedir.

10.1.2. Korelasyon Katsayısının Ortalamadan Sapmalar Üzerinden Hesaplanması

Korelasyon katsayısı, X ve Y değişkenlerinin aritmetik ortalamaları hesaplanıp, her bir gözlem değerinin aritmetik ortalamadan farklarının hesaplanması ile elde edilen aritmetik ortalamadan sapma serileri kullanılarak da hesaplanabilir.

Aritmetik ortalamadan sapmalar üzerinden korelasyon katsayısı aşağıdaki formül yardımıyla hesaplanır:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Formülde yer alan;

$(X_i - \bar{X})$, X değişkeninin aritmetik ortalamadan sapmalarını,

$(Y_i - \bar{Y})$, Y değişkeninin aritmetik ortalamadan sapmalarını ifade etmektedir.

Korelasyon katsayısının aritmetik ortalamadan sapmalar yöntemine göre hesaplanması sırasında sıkça yapılan bir hatayı vurgulamak ve formülün pay kısmında yer alan $(X_i - \bar{X})(Y_i - \bar{Y})$ ifadesine özellikle dikkat çekmek istiyoruz.

Bilindiği gibi gözlem değerlerinin aritmetik ortalamadan sapmaları toplamı sıfırdır. Yani, X değişkeninin aritmetik ortalamadan sapmaları toplamı ile,

$$\sum(X_i - \bar{X}) = 0$$

Y değişkeninin aritmetik ortalamadan sapmaları toplamı,

$$\sum(Y_i - \bar{Y}) = 0$$

çarpıldığında,

$$\sum(X_i - \bar{X}) \,.\, \sum(Y_i - \bar{Y}) = 0$$

sonucuna ulaşılmaktadır. Ancak, yukarıdaki formülün pay kısmına dikkat edilecek olursa, X ve Y değişkenlerinin ortalamadan sapmaları hesaplandıktan sonra, karşılıklı olarak sapma değerleri çarpılmaktadır. Başka bir deyişle X ve Y serilerinin ortalamadan sapmaları hesaplanıp, sapma değerleri toplanarak (ki toplam değerleri sıfırdır), toplam değerleri çarpılmamaktadır. Formülün pay kısmı,

$$\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

şeklinde olup her bir değişken için hesaplanan sapma değerlerinin (X,Y) ikilileri için karşılıklı olarak çarpılmasını ve daha sonra toplanmasını ifade etmektedir. Bu konuya özellikle dikkat çekmek istiyoruz.

10.2. Basit Korelasyon Katsayısı Hesabına Yönelik Uygulamalar

Şimdi yukarıda verdiğimiz iki korelasyon katsayısı formülünü kullanarak örnek uygulamalar yapalım:

Örnek:

Aşağıda bir firmanın 2013-2016 yılları arasında yaptığı reklâm harcamaları ve satış rakamları veriliyor. Buna göre, reklâm harcamaları ile satışlar arasındaki ilişkinin derecesini yani korelasyon katsayısını hesaplayarak yorumlayalım.

Örneğimizde reklâm harcamaları ve satış değişkenleri verilmiş ve korelasyon katsayısının hesaplanması isteniyor. Koralasyon katsayısı sebep –sonuç ilişkisine dayanmayıp sadece iki değişkenin birlikte değişme ölçüsünü verir. Dolayısıyla, hangi değişkene X hangi değişkene Y dediğimizin bir önemi yoktur ve her iki koşulda da korelasyon katsayısı aynı değeri alır. Ancak bir sonraki analizimiz regresyon çözümlemesi olacağından değişken tanımlamasını burada yaparak sebep değişkeni X, sonuç değişkeni Y ile sembolize ediyoruz.

Reklâm harcamaları satışları etkilediği için sebep değişken olarak düşünülerek X değişkeni, satışlar ise reklam harcamaları sonucunda oluştuğu için Y değişkeni olarak tanımlanmıştır.

Yıllar	Reklam Harcamaları *(X_i)*	Satışlar *(Y_i)*	X_i^2	Y_i^2	X_iY_i
2013	8	60	64	3600	480
2014	12	80	144	6400	960
2015	15	110	225	12100	1650
2016	25	150	625	22500	3750
	60	400	1058	44600	6840

Gözlem değerlerinin kendisi kullanılarak korelasyon hesabı aşağıdaki formül ile yapılmaktadır:

$$r = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}}$$

Formülden de görüleceği gibi, X ve Y değişkenlerinin ortalamalarını, kareleri toplamlarını ve X ile Y çarpım toplamını hesaplayarak formülde yerine koymamız gerekiyor.

X ve Y değişkenlerinin ortalamalarını:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{60}{4} = 15$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{400}{4} = 100$$

olarak hesaplıyoruz. Ayrıca yukarıdaki tablodan da görüleceği gibi, korelasyon formülünde yer alan diğer değerler de aşağıdaki gibi bulunuyor.

$$\sum X_i Y_i = 6840$$

$$\sum {X_i}^2 = 1058$$

$$\sum {Y_i}^2 = 44600$$

Şimdi hesapladığımız tüm değerleri korelasyon katsayısı formülünde yerine koyalım:

Korelasyon katsayısı,

$$r = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum {X_i}^2 - n\bar{X}^2\right)\left(\sum {Y_i}^2 - n\bar{Y}^2\right)}}$$

$$r = \frac{6840 - 4 \,.\, 15 \,.\, 100}{\sqrt{(1058 - 4 \,.\, 15^2)(44600 - 4 \,.\, 100^2)}}$$

$$r = \frac{6840 - 6000}{\sqrt{(1058 - 900)(44600 - 40000)}}$$

$$r = \frac{840}{\sqrt{158 \;.\; 4600}} = +\frac{840}{852} = +0{,}9859 \quad yani \quad \%\,98{,}59$$

olarak hesaplanmaktadır.

Korelasyon katsayısının değeri 0,9859 bulunmuştur ve bu değer reklâm harcamaları ile satış değişkenlerinin % 98,59 oranında birlikte değiştiklerini gösterir. Korelasyon katsayısı bir'e çok yakın ve pozitif bir değerdir. Dolayısıyla, iki değişken arasında oldukça güçlü ve pozitif yönlü bir ilişki vardır ve iki değişken birlikte artmakta ve birlikte azalmaktadır.

Aynı örneği, korelasyon katsayısı hesabında kullanılabilecek ikinci formülümüz olan ortalamadan sapmalar formülüne göre de çözelim.

Bunun için öncelikle X ve Y serilerinin ortalamadan sapmalarını hesaplamamız gerekiyor. Bir önceki çözümde X ve Y değişkenlerinin ortalamalarını,

$$\bar{X} = \frac{\sum X_i}{n} = \frac{60}{4} = 15$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{400}{4} = 100$$

şeklinde hesapladığımızı hatırlayalım.

X ve Y değişkenlerinin gözlem değerlerinin ortalamadan farklarını alarak sapma serilerini hesaplayalım ve daha sonra aşağıdaki tabloda detaylarını görebileceğimiz formül için gerekli diğer işlemleri gerçekleştirelim.

Yıllar	Rklm Harc. (X_i)	Satışlar (Y_i)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
2013	8	60	-7	49	-40	1600	280
2014	12	80	-3	9	-20	400	60
2015	15	110	0	0	10	100	0
2016	25	150	10	100	50	2500	500
	60	400	0	158	0	4600	840

Ortalamadan sapmalar yöntemine göre korelasyon katsayısının formülü aşağıdaki gibidir:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Formül için gerekli olan ve yukarıdaki tabloda ayrıntılı hesaplamalarını görebileceğiniz değerleri yerine koyarak korelasyon katsayısı,

$$r = \frac{840}{\sqrt{158 \,.\, 4600}} = + \frac{840}{852} = +0{,}9859$$

olarak hesaplanmaktadır. Reklâm harcamaları ile satışlar arasındaki ilişkinin derecesi % 98,59 düzeyindedir ve bu, iki değişkenin % 98,59 oranında birlikte değiştiğini göstermektedir. Korelasyon katsayısı pozitif ve güçlü ilişkiyi işaret etmektedir. Dolayısıyla, reklâm harcamaları ile satışlar arasında oldukça yüksek ve pozitif yönlü ilişki vardır ve iki değişken birlikte artmakta ve azalmaktadır.

Örnek:

Aşağıda aynı büyüklükte 5 ekim alanında kullanılan gübre miktarı ve gerçekleştirilen buğday üretimi değerleri veriliyor. Buna göre, gübre kullanımı ile buğday üretimi arasındaki ilişkinin derecesini (korelasyon katsayısını) hesaplayarak yorumlayalım.

Korelasyon analizi açısından hangi değişkenin sebep hangi değişkenin sonuç olduğunun önemli olmadığını bir kez daha yineleyelim. Ancak, bir sonraki aşamada aynı örneği regresyon

çözümlemesinde de kullancağımız için değişkenleri isimlendirirken sebep – sonuç ilişkisini dikkate alıyor ve gübre kullanımı değişkenini buğday üretimi üzerinde etkili bir değişken olması sebebiyle X ile ve buğday üretimini de sonuç değişken olması nedeniyle Y ile ifade ediyoruz.

Gübre Kullanımı (X_i)	Buğday Üretimi (Y_i)	X_i^2	Y_i^2	X_iY_i
12	40	144	1600	480
15	80	225	6400	1200
18	100	324	10000	1800
20	120	400	14400	2400
35	160	1225	25600	5600
100	500	2318	58000	11480

Öncelikle gözlem değerlerinin kendisinin kullanıldığı korelasyon katsayısı formülüne göre çözüm yapalım. Kullanacağımız formül için gerekli alt işlemler yukarıdaki tablodan takip edilebilir.

X ve Y değişkenlerinin ortalamaları:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{100}{5} = 20$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{500}{5} = 100$$

olarak hesaplanıyor. Ayrıca yukarıdaki tablodan da görüleceği gibi, korelasyon formülünde yer alan diğer değerler de aşağıdaki gibi bulunuyor:

$$\sum X_iY_i = 11480$$

$$\sum X_i^2 = 2318$$

$$\sum Y_i^2 = 58000$$

Şimdi hesapladığımız tüm değerleri korelasyon katsayısı formülünde yerine koyalım:

$$r = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}}$$

$$r = \frac{11480 - 5.20.100}{\sqrt{(2318 - 5.20^2)(58000 - 5.100^2)}}$$

$$r = \frac{11480 \; - \; 10000}{\sqrt{(2318 - 2000)(58000 - 50000)}}$$

$$r = + \frac{1480}{\sqrt{(318)(8000)}} = + \frac{1480}{\sqrt{2544000}} = + \frac{1480}{1594{,}99} = +0{,}9279$$

Çözümden de görüleceği gibi gübre kullanımı ile buğday üretimi arasındaki korelasyon katsayısı + 0,9279 olarak hesaplanmıştır. Dolayısıyla, gübre kullanımı ile buğday üretimi arasında pozitif yönlü ve güçlü bir ilişki vardır ve bu iki değişken % 92,79 oranında aynı yönde ve birlikte değişmektedir. Yani, iki değişken % 92,79 oranında birlikte artmakta ve birlikte azalmaktadırlar.

Aynı örneği ortalamadan sapmalar yöntemine göre de çözelim:

Ortalamadan sapmalar yöntemine göre korelasyon katsayısı hesaplayabilmek için öncelikle değişkenlerin ortalamaları ve daha sonra da ortalamadan sapmalarının hesaplanması gerekir. Değişkenlerin ortalamalarını bir önceki çözüm için hesaplamıştık.

Ortalamadan sapmalarla korelasyon katsayısı hesaplayabilmek için gerekli işlemler aşağıdaki tablo ile gösterilmektedir:

Gübre Kullanımı (X_i)	Buğday Üretimi (Y_i)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
12	40	-8	64	-60	3600	480
15	80	-5	25	-20	400	100
18	100	-2	4	0	0	0
20	120	0	0	20	400	0
35	160	15	225	60	3600	900
100	500	0	318	0	8000	1480

Ortalamadan sapmalar yöntemine göre korelasyon katsayısının formülü aşağıdaki gibidir:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Formül için gerekli olan ve yukarıdaki tabloda ayrıntılı hesaplamalarını görebileceğiniz değerleri yerine koyarak korelasyon katsayısını,

$$r = + \frac{1480}{\sqrt{(318)(8000)}} = + \frac{1480}{\sqrt{2544000}} = + \frac{1480}{1594{,}99} = +0{,}9279$$

olarak hesaplıyoruz.

Yukarıdaki çözümden de görüleceği gibi gübre kullanımı ile buğday üretimi arasındaki korelasyon katsayısı + 0,9279 olarak hesaplanmaktadır. Dolayısıyla, gübre kullanımı ile buğday üretimi arasında pozitif yönlü ve güçlü bir ilişki vardır ve bu iki değişken % 92,79 oranında aynı yönde ve birlikte değişmektedir. Yani, gübre kullanımı ve buğday üretimi değişkenleri % 92,79 oranında birlikte artmakta ve birlikte azalmaktadırlar.

Aynı örnek verilerine, gerek gözlem değerlerinin kendisinin kullanıldığı korelasyon formülüne göre gerekse ortalamadan sapmaların kullanıldığı korelasyon formülüne göre çözüm yapıldığında tamamen aynı sonuçların elde edildiği görülmektedir.

Ortalamadan sapmalara göre çözüm yönteminin, değişkenlere ilişkin ortalamaların tam sayı olması halinde kullanılması önerilmektedir. Aksi taktirde ortalamanın ondalık değerli olması, ortalamadan farkların da ondalık değerli olmasına sebep olacak ve çözüm için daha fazla zaman harcamak gerekecektir.

8.1. Regresyon Analizi

Regresyon analizi değişkenler arasındaki ilişkinin araştırılmasında kullanılan istatistiksel bir araçtır. Bu araçla, bir değişkenin diğer değişken üzerindeki nedensel ilişkisi araştırılır. İncelenen ilişkideki değişkenler aralarındaki ilişki göz önüne alınarak değişkenler bağımlı ve bağımsız olarak isimlendirilir. Önceki derslerde de değinildiği gibi istatistiğin öncelikli ilgi alanını rastlantı değişkeninin davranışını bir modelle tahmin etmek oluşturur. Davranışı tahmin edilecek olan rastlantı değişkeni bir diğer değişken(ler)in fonksiyonu olarak gösterilebilir ve bu değişken bağımlı olarak isimlendirilir ve Y ile gösterilir. Bağımlı değişkeni etkileyen değişken ise X ile gösterilir ve bağımsız değişken olarak isimlendirilir. Yani, bağımlı değişken, bağımsız değişken(ler) tarafından açıklanmaya çalışılır ve açıklayıcı değişkenlerin modelde bilinen sabitler olduğu varsayılır.

Model şu şekilde gösterilir; $Y_i = b_0 + b_1 X_i + \varepsilon_i$

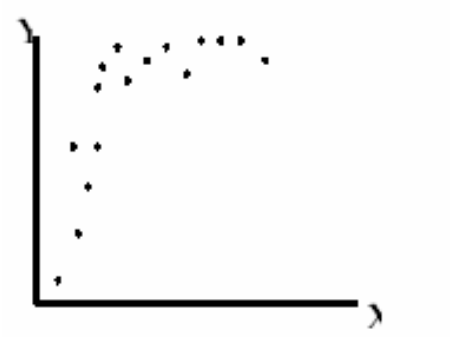
burada b_0 sabit katsayıyı gösterir buna başlangıç parametresi de denir, b_1 ise eğim parametresidir. X'deki 1 birimlik değişmenin Y üzerinde nasıl bir değişim yaptığını gösterir. Denklemdeki ε_i ise daha sonra da açıklanacağı gibi hata terimine karşılık gelir.

Örneğin, tüketim ve gelir üzerine yapılan bir çalışmada bağımsız değişken gelir, bağımlı değişkense tüketimdir ya da bir hastaya uygulanan ilacın dozu ve hastanın iyileşme süreci çalışmasında bağımsız değişken ilacın dozu ve bağımlı değişkense hastanın iyileşme süreci olur.

Regresyon analizi, bilinen gerçekleşen olaylar sonucunda elde edilen bulgulardan yola çıkarak gelecekteki olaylarla ilgili tahmin yani öngörü yapılmasını sağlar. Regresyon modelinde amaç, koşullar değiştiğinde bağımlı değişkenin ortalamasının $E(Y_i)$ 'nin nasıl değiştiğini tanımlamaktır.

Değişkenler arasındaki ilişki deterministik (kesin) ya da olasılıksal (istatistiksel) olarak isimlendirilir. Deterministik ilişkileri açıklamada kullanılan matematiksel fonksiyondan farklı olarak, regresyon analizi olasılıksal ilişkileri açıklar. Arz-talep, gelir-tüketim gibi ilişkilerin modellenmesinde deterministik ilişkiye sahip değişkenler yerine istatistiksel ilişkiye sahip değişkenler kullanılır.

Söz konusu ilişkide bağımsız değişken sayısının bir tane olması basit regresyonla, birden fazla olması ise çoklu regresyonla açıklanır. Regresyon modeli, doğrusal yapıda olabileceği gibi parabolik, logaritmik, üstel biçimli de olabilir. Modelde bir bağımlı ve bir bağımsız değişken söz konusu olduğunda, yani basit regresyon söz konusu iken serpilme diyagramı kullanılarak uygun model seçimi yapılabilir. Serpilme diyagramı, i. gözlemin bağımlı değeri y_i ve bağımsız değeri x_i olmak üzere tüm gözlem çiftleri üzerinden, her ikili yani Y ve X değişkenlerinin aldığı tüm değerler birer nokta ile temsil edilecek şekilde çizilir. Diyagramdaki dağılıma bakılarak uygun model belirlenir.



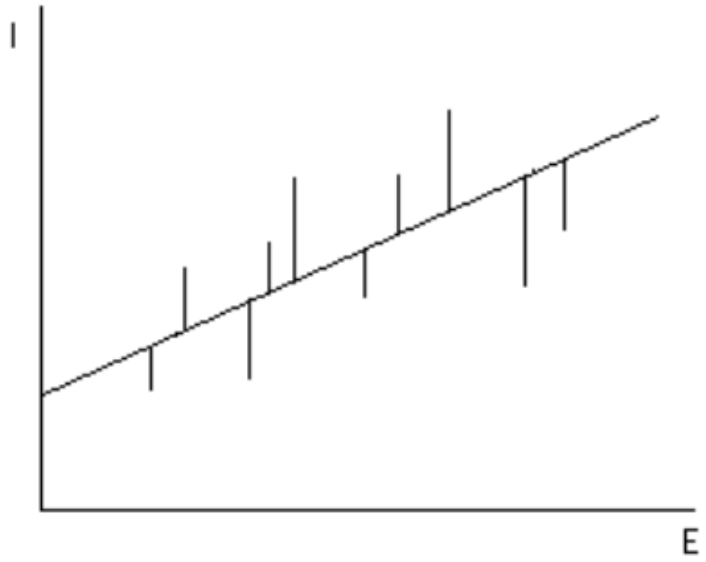
Yukarıdaki serpilme diyagramlarında noktaların ortasından geçecek olan eğri dikkate alınır ve bu eğri incelenen ilişki biçimi hakkında bilgi verir. Buna göre, ilk çizimde noktaların bir doğru etrafında toplandığı söylenebilir ve değişkenler arasında aynı yönlü doğrusal bir ilişkinin varlığından söz edilebilir. İkinci çizimse ters yönlü doğrusal bir ilişkinin varlığını gösterir. Üçüncü çizimde doğrusal olmayan bir ilişkinin varlığı söz konusudur. Son çizim

dikkate alındığında ise bir eğri oluşturmak mümkün görünmemektedir, değişkenler arasında bir ilişkinin olmadığı sonucuna varılır.

Serpilme diyagramı çizimi sonrasında uygun modele karar verilir ve modeldeki parametreler tahmin edilir. Tahmin sürecinde çeşitli kriterler doğrultusunda kullanılan yöntemler olmakla beraber, burada bu yöntemlerden sadece En Küçük Kareler Yöntemine (EKK) değinilecektir.

8.2. Regresyon Katsayılarının Tahmini

EKK ile bulunacak eğrinin her (xi,yi) gözlem çiftine karşılık gelen nokta ile bu noktanın EKK ile elde edilecek eğri üzerindeki dik izdüşümü arasındaki farklar toplamı sıfır olmalıdır. Bu farklar, yani Y_i değerlerinin regresyon doğrusuna olan uzaklığı, daha sonrada bahsedileceği gibi 'hata' olarak isimlendirilir. Y_i değerlerinin regresyon doğrusu üzerindeki görüntüsü $\hat{Y}_i$ (tahmini Y_i) ile arasındaki fark hataya karşılık gelir. İdeal regresyon doğrusu, bu farkların karelerinin toplamını $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, minimum verenle elde edilir.



Hata kareler toplamının minimum olabilmesi için, sabit ve eğim parametrelerine göre türevleri alınarak sıfıra eşitlenir:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

$$\frac{\partial(\sum e_i^2)}{\partial b_0} = 2\sum (Y - b_0 - b_1 X)(-1) = 0$$

$$\frac{\partial(\sum e_i^2)}{\partial b_1} = 2\sum (Y - b_0 - b_1 X)(-x) = 0$$

Bu işlem sonrasında elde edilen denklemler Normal Denklemler olarak isimlendirilir:

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

Cramer yöntemine göre bu denklemler çözüldüğünde;

$$b_0 = \frac{\begin{vmatrix} \sum Y & \sum X \\ \sum YX & \sum X^2 \end{vmatrix}}{\begin{vmatrix} n & \sum X \\ \sum X & \sum X^2 \end{vmatrix}} = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

$$b_1 = \frac{\begin{vmatrix} n & \sum Y \\ \sum X & \sum XY \end{vmatrix}}{\begin{vmatrix} n & \sum X \\ \sum X & \sum X^2 \end{vmatrix}} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

elde edilir. Ancak eğim ve sabit parametre tahminlerini ortalamadan sapmalar üzerinden giderekte belirlemek mümkündür:

$$x = X - \bar{X}$$

$$y = Y - \bar{Y}$$

Eğim parametresi bu x ve y üzerinden gidilerek belirlenir.

$$\frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{n\sum (x+\bar{X})(y+\bar{Y}) - [\sum (x+\bar{X})(y+\bar{Y})]}{n\sum (x+\bar{X})^2 - (\sum (x+\bar{X}))}$$

$$= \frac{n\sum (xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + n\bar{X})(\sum y + n\bar{Y})}{n\sum (x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + n\bar{X})^2}$$

$$= \frac{n\sum xy + n\bar{Y}\sum x + n\bar{X}\sum y + n^2\bar{X}\bar{Y} - (\sum x + n\bar{X})(\sum y + n\bar{Y})}{n\sum x^2 + 2n\bar{X}\sum x + n^2\bar{X}^2 - (\sum x + n\bar{X})^2}$$

$$\sum x = \sum (X - \bar{X}) = 0$$

$$\sum y = \sum (Y - \bar{Y}) = 0$$

olduğundan eşitlikler düzenlenirse;

$$b_1 = \frac{n\sum xy + n^2\bar{X}\bar{Y} - n^2\bar{X}\bar{Y}}{n\sum x^2 + n^2\bar{X}^2 - n^2\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Sabit parametre için,

$\sum Y_i = nb_0 + b_1 \sum X_i$ olduğundan, eşitliğin her iki tarafı n'e bölünürse,

$$\bar{Y} = b_0 + b_1\bar{X}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

elde edilir.

Örnek: Aşağıda bir sınıftaki öğrencilerin muhasebe ve matematik derslerine ait veri bulunmaktadır. Muhasebe dersinden başarının matematik dersinden başarıya bağımlı olup olmadığını sınamak için regresyon denklemini oluşturunuz.

Muhasebe	**Matematik**
1	2
2	3
3	5
5	6
6	7
7	10
8	7
8	8

Y	**X**	**Y^2**	**X^2**	**YX**	**X-Xort=x**	**Y-Yort=y**	**xy**	**x^2**
1	2	1	4	2	-4	-4	16	16
2	3	4	9	6	-3	-3	9	9
3	5	9	25	15	-1	-2	2	1
5	6	25	36	30	0	0	0	0
6	7	36	49	42	1	1	1	1
7	10	49	100	70	4	2	8	16
8	7	64	49	56	1	3	3	1
8	8	64	64	64	2	3	6	4
40	**48**	**252**	**336**	**285**	**0**	**0**	**45**	**48**

Yort=5

Xort=6

Normal denklemlerden elde edilen eşitlikler kullanılarak katsayılar şöyle tahmin edilmiştir:

$$b_0 = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n\sum X^2 - (\sum X)^2}$$

$$b_0 = \frac{40(336) - 48(285)}{8(336) - 48^2} = -0.625$$

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b_1 = \frac{8(285) - 48(40)}{8(336) - 48^2} = 0.9375$$

Ortalamadan sapmalarla,

$$b_1 = \frac{\sum xy}{\sum x^2} = \frac{45}{48} = 0.9375$$

$$b_0 = \overline{Y} - b_1\overline{X} = 5 - 0.9375(6) = -0.625$$

yine aynı donuca ulaşılmıştır. Tahmin edilen regresyon denklemi şöyledir:

$$\hat{y} = -0.625 + 0.9375$$

Örnek:

Aşağıda bir eyaletteki suç ve işsizlik oranlarına ilişkin veri mevcuttur.

işşizlik oranı	**suç oranı**
0,8	3
1,4	6
2,3	7
3,5	15
4,5	19

İşsizlik ve suç işleme oranları arasındaki ilişkiyi gösteren regresyon denklemini oluşturunuz.

Sorunun çözümünün ilk aşamasında bağımlı ve bağımsız değişkenleri belirleyelim. İşssizlik oranı bağımsız değişkendir (x), suç oranını etkiler ki bu da bağımlı değişken (Y) olarak isimlendirilir. İlgili kolonların toplam ve çarpımlarının toplamına ait bilgiler aşağıdaki gibi özetlenmiştir.

$$\sum Y = 50 \quad \sum X = 12.5 \quad \sum XY = 164.9$$
$$\sum X^2 = 40.39 \quad \sum Y^2 = 680$$
$$\sum (X - \bar{X})(Y - \bar{Y}) = 39.9$$
$$\sum (X - \bar{X})^2 = 9.14 \quad \sum (Y - \bar{Y})^2 = 180$$

$$b1 \equiv \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{39.9}{9.14} = 4.365$$

$$b_0 = \frac{50}{5} - 4.365(\frac{12.5}{5}) = -0.9$$

$$Y = -0.9 + 4.365X$$

İşsizlik oranı 1 birim arttığında suç oranı 4.365 birim artar.

Ortalamadan sapmalar serisi yerine orijinal seriden hareket edilirse eğim katsayısı şöyle bulunacaktır:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5(164.9) - (12.5)(50)}{5(40.39) - 12.5^2} = 4.365$$

Örnek:

Bir firmanın reklam harcamaları ve satış rakamlarına ilişkin veri mevcuttur. Regresyon denklemini oluşturunuz

Reklam Harca.	**Satışlar**
1.6	6
2.8	12
4.6	14
7	30
9	38

Bağımlı değişken satışlardır., Reklam harcamaları ise bağımsız değişkendir.

$$\sum Y = 100 \quad \sum X = 25 \quad \sum XY = 659.6$$
$$\sum X^2 = 161.56 \quad \sum Y^2 = 2720$$
$$\sum (X - \bar{X})(Y - \bar{Y}) = 159.6$$
$$\sum (X - \bar{X})^2 = 36.56 \quad \sum (Y - \bar{Y})^2 = 720$$

$$b_1 = \frac{159.9}{36.56} = 4.37$$
$$b_0 = \frac{100}{5} - 4.37(\frac{25}{5}) = -1.8$$
$$Y = -1.8 + 4.37X$$

Reklam harcamaları 1 birim artarsa satış 4.37 birim artar.

Diğer yaklaşımla:

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5*659.6 - 25*100}{5*161.56 - 25^2} = 4.37$$

$$b_0 = \frac{\sum Y \sum X^2 - \sum X \sum XY}{n\sum X^2 - (\sum X)^2} = -1.8$$

Örnek: Aşağıdaki x,y serilerinden hareketle regresyon denklemini bularak yorumlayınız.

x	y
10.2	7
8.4	5
6.2	4
4.2	1
11	8

$$n = 5, \quad \sum X = 85, \quad \sum Y = 45$$
$$\sum XY = 789, \quad \sum X^2 = 1475, \quad \sum Y^2 = 425$$

$$b_1 = \frac{789 - 5(85/5)(45/5)}{1475 - 5(85/5)^2} = 0.8$$

$$b_0 = 45/5 - (0.8)(85/5) = -4.6$$

Örnek: Aşağıda verilen x,y ikilisinden hareketle regresyon denklemini bulunuz.

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

$$n = 5, \quad \sum X = 40, \quad \sum Y = 25$$

$$\sum XY = 230.4, \quad \sum X^2 = 351.68, \quad \sum Y^2 = 155$$

$$b_1 = \frac{230.4 - 5(40/5)(25/5)}{351.68 - 5(40/5)^2} = 0.96$$

$$b_0 = 25/5 - (0.96)(40/5) = -2.677$$

$$y = -2.677 + 0.96x$$