Regresyon ve Korelasyon Analizi

Regresyon Analizi

Regresyon analizi, aralarında sebep-sonuç ilişkisi bulunan iki veya daha fazla değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi kullanark o konu ile ilgili tahminler (estimation) ya da kestirimler (prediction) yapabilmek amacıyla yapılır. Doğada birçok olayda sebep-sonuç ilişkisine rastlamak mümkündür.

Örnek:	Sebep	Sonuç
	Gelir	Harcama
	Yaş	Boy
	Gübre	Verim
	Yem miktarı	Süt miktarı
	Çalışma süresi	Alınan not

Bu analiz tekniğinde iki (basit regresyon) veya daha fazla değişken (çoklu regresyon) arasındaki ilişki açıklamak için matematiksel bir model kullanılır ve bu model regresyon modeli olarak adlandırılır.

Bu kısımda anlaşılması daha kolay olduğu için basit regresyon analizi anlatılmıştır.

Basit regresyon modeli

$$Y=\alpha+\beta X+\varepsilon$$

şeklinde bir bağımlı ve bir de bağımsız değişken içeren bir modeldir. Burada

Y; bağımlı (sonuç) değişken olup belli bir hataya sahip olduğu varsayılır.

X; bağımsız (sebep) değişkeni olup hatasız ölçüldüğü varsayılır.

α; sabit olup X=0 olduğunda Y'nin aldığı değerdir.

β ise regresyon katsayısı olup, X'in kendi birimi cinsinden 1 birim değişmesine karşılık Y'de kendi birimi cinsinden meydana gelecek değişme miktarını ifade eder.

ε; tesadüfi hata terimi olup ortalaması sıfır varyansı σ^2 olan normal dağılış gösterdiği varsayılır. Bu varsayım parametre tahminleri için değil katsayıların önem kontrolleri için gereklidir.

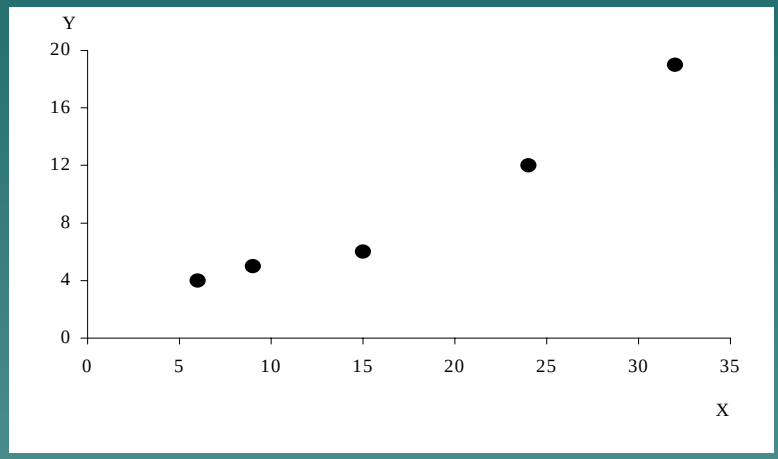
Parametrelerin (Katsayıların) Tahmini

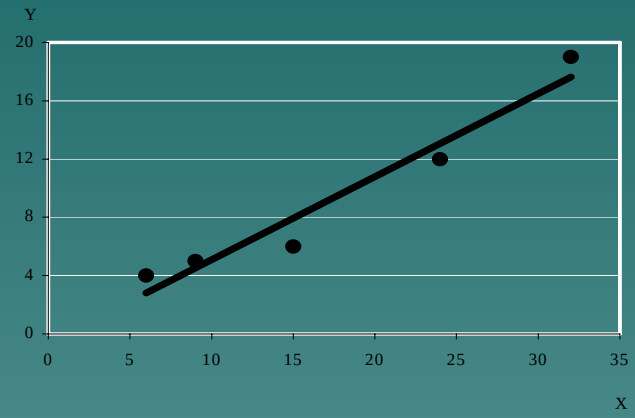
Bir regresyon modeli oluşturulurken genelde en-küçük kareler ve en büyük olabilirlik (maximum likelihood) teknikleri olarak bilinen iki yaklaşımdan birisi kullanılır. Eğer hata teriminin normal dağılım göstermesi şeklinde bir varsayım varsa en büyük olabilirlik, hata teriminin dağılışı ile ilgili herhangi bir varsayım söz konusu değilse en-küçük kareler tekniği kullanılarak parametreler tahmin edilir.

En-küçük kareler tekniği kullanılarak parametrelerin nasıl tahmin edildiğini örnek bir veri grubu üzerinde kısaca özetleyelim.

Boy(cm) (X)	Çevre(cm) (Y)
9	5
15	6
6	4
24	12
32	19

Tabloda verilen X ve Y değişkenlerine ait beş gözlem çifti, koordinat eksenlerine yerleştirildiğinde elde edilen serpme diyagramının Şekil (a)'daki grafik elde edilir.





(a) (b)

Şekil (a)'da verilen noktaları temsil eden regresyon doğrusu oluşturulursa Şekil (b) elde edilir. Uydurulan regresyon doğrusu ile gözlem noktaları arasındaki fark hata (ε) olarak isimlendirilir. Regresyon doğrusuna ait parametreler öyle tahmin edilmelidir ki; doğru ile gözlem noktaları arasındaki fark (hata) en az olsun. Bunu sağlayacak teknik ise en-küçük kareler tekniğidir

Yukarıda verilen basit regresyon modelinden e çekilirse yani;

$$Y = \alpha + \beta X_i + \varepsilon \Rightarrow \varepsilon = Y - \alpha - \beta X_i$$

olur. Burada hata teriminin tüm gözlemler için kareleri alınır ve toplanırsa (Hata Kareler Toplamı (HKT);

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)^2$$

olur.
En-küçük kareler tekniğinde, HKT'nı en küçük yapabilmek için yukarıdaki ifadenin önce α'ya göre türevini alıp sıfıra eşitleyerek

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \alpha} = 0 \Rightarrow \frac{\partial \sum_{i=1}^{n} (Y - \alpha - \beta X_i)^2}{\partial \alpha} = 0$$

$$\Rightarrow \; n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \quad \text{(Normal denklem 1)}$$

daha sonra da β'ya göre türevini alıp sıfıra eşitleyerek

$$\frac{\partial \sum_{i=1}^{n} \varepsilon_i^2}{\partial \beta} = 0 \Rightarrow \frac{\partial \sum_{i=1}^{n} (Y - \alpha - \beta X_i)^2}{\partial \beta} = 0$$

$$\Rightarrow \hat{\alpha} \sum_{i=1}^{n} X_i + \hat{\beta} \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \quad \text{(Normal denklem 2)}$$

şeklinde normal denklemler olarak isimlendirilen iki bilinmeyenli iki denklem elde edilir. Normal denklemlerin iki bilinmeyenli denklerin çözümünde kullanılan değişik yöntemlerden birisi kullanılarak çözümü yapıldığında;

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n} = \frac{Sxy}{Sxx}$$

ve

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}\frac{\sum_{i=1}^{n} X_i}{n} = \overline{Y} - \hat{\beta}\overline{X}$$

eşitlikleri elde edilir. Bu eşitlikler ile regresyon doğrusu denkleminde yer alan ve HKT'nı en az yapacak parametreler tahmin edilir. Böylece, en-küçük kareler regresyon doğrusu denklemi

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X_i$$

Regresyon Katsayısının Önem Testi:

Hipotezler

Ho: $\beta=0$

Hı : $\beta\neq0$

Kullanılacak test istatistiği

$$t=\frac{\hat{\beta}-\beta}{S_{\hat{\beta}}}\sim t_{n-2,\alpha/2}$$ olup burada;

$$S_{\hat{\beta}}=\sqrt{\frac{S_{y.x}^2}{S_{xx}}}=\sqrt{\frac{[S_{yy}-(S_{xy})^2/S_{xx}]/(n-2)}{S_{xx}}}$$

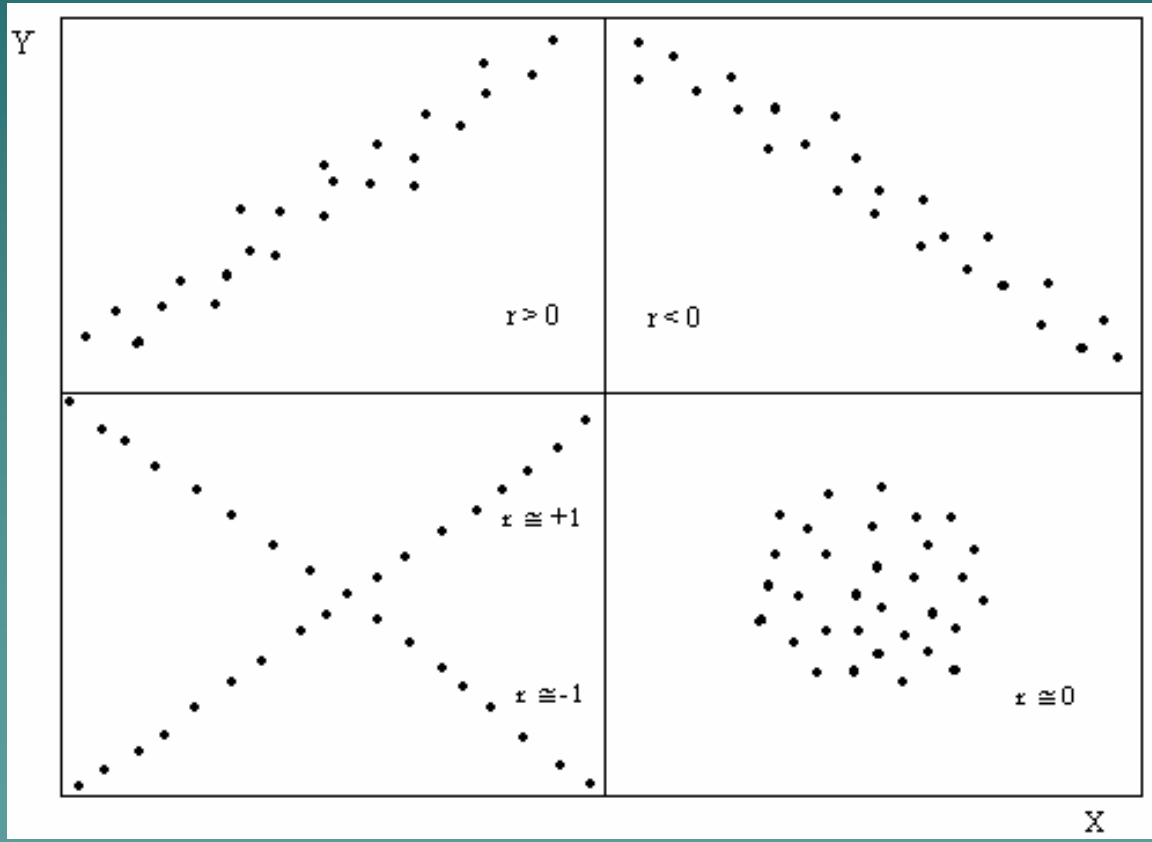
Korelasyon Analizi

İki değişken arasındaki ilişkinin derecesini ve yönünü belirlemek amacıyla kullanılan istatistik yöntemlerden birisidir. Değişkenlerin bağımlı veya bağımsız olması dikkate alınmaz. Değişik şekillerde hesaplanan ve değişik amaçlar için kullanılan Pearson korelasyon katsayısı, Canonical korelasyon katsayısı, kısmi korelasyon katsayısı gibi farklı isimler alan korelasyon katsayıları vardır. Bunlardan Pearson korelasyon katsayısı, r ile gösterilir ve

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - (\sum_{i=!}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 / n\right)\left(\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2 / n\right)}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

formülü ile hesaplanır. Korelasyon katsayısı -1 ile +1 arasında değişen değerler alır (-1≤ r ≤+1). . Katsayı, ilişkinin olmadığı durumda 0, tam ve kuvvetli bir ilişki varsa 1, ters yönlü ve tam bir ilişki varsa -1 değerini alır.

Aralarındaki ilişkinin derecesi araştırılan değişkenlere ait gözlemler serpme diyagramında incelendiğinde, noktaların dağılımına göre korelasyon katsayısının alabileceği değerler Şekilde gösterilmiştir.



Korelasyon katsayısının yorumunu, tam değerler dışında ara değerler için yapmak oldukça güçtür. Ara değerler için katsayı değerlendirirken, örnek gözlem sayısı (n) oldukça önemlidir. Çok fazla gözleme dayanan değerlendirmelerde 0.25'e kadar düşmüş bir korelasyon katsayısı bile anlamlı sayılabilmektedir. Fakat az sayıda, 10-15 gözleme dayanan değerlendirmelerde korelasyon katsayısının 0.71 üstünde olması beklenir. Populasyona göre normal sayılacak kadar bir gözlem sayısı alınarak bakılmış gözlem grupları için genellikle, 0-0.49 arasında ise korelasyon zayıf, 0.5-0.74 arasında ise orta derecede, 0.75-1 arasında ise kuvvetli ilişki vardır denilmektedir. Basit korelasyon analizinden söz edilebileceği gibi, çoklu korelasyon analizi yapmak da mümkündür.

Korelasyon Katsayısının Önem Testi:

Hipotezler

Ho: $\rho=0$

Hı : $\rho\neq0$

Kullanılacak test istatistiği

$$t=\frac{r-\rho}{S_{\bar{r}}}\sim t_{n-2,\alpha/2}$$ olup burada; $$S_{\bar{r}}=\sqrt{\frac{1-r^2}{n-2}}$$

Örnek: Bir balık türü için balığın boyu(cm) ve vücut çevresine (cm) ait değerler aşağıdaki gibidir. Balığın boyu ile vücut çevresi arasındaki ilişkinin doğrusal bir ilişki olduğunu varsayarak

a) Bu ilişkiyi açıklayan regresyon modelini oluşturunuz ve regresyonun önem testini yapınız

b) Korelasyon katsayısını hesaplayarak önem testini yapınız.

Boy(cm) (X)	Çevre(cm) (Y)	XY	X^2	Y^2
9	5	45	81	25
15	6	90	225	36
6	4	24	36	16
24	12	288	576	144
32	19	608	1024	361

Çözüm: (a)

Boy(cm) (X)	Çevre(cm) (Y)	XY	X^2	Y^2
9	5	45	81	25
15	6	90	225	36
6	4	24	36	16
24	12	288	576	144
32	19	608	1024	361
86	46	1055	1942	582

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n}{\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2/n} = \frac{1055 - 86*46/5}{1942 - 86^2/5} = \frac{263.8}{462.8} = 0.57$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = \frac{46}{5} - 0.57 * \frac{86}{5} = -0.604$$

Böylece, en-küçük kareler regresyon doğrusu;

$$\hat{Y} = -0.604 + 0.57 X_i$$

Regresyon Katsayısının Önem Testi:

Hipotezler

Ho: $\beta=0$

Hı : $\beta\neq0$

$\alpha=0.01$

$$t=\frac{\hat{\beta}-\beta}{S_{\hat{\beta}}}=\frac{0.57-0}{0.0779}=7.31$$ > $t_{5-2,0.01/2}=t_{3,0.005}=5.841$ olduğundan Ho RED edilir.

$$S_{\hat{\beta}}=\sqrt{\frac{[S_{yy}-(S_{xy})^2/S_{xx}]/(n-2)}{S_{xx}}}=\sqrt{\frac{[158.8-(263.8)^2/462.8]/(5-2)}{462.8}}=0.0779$$

$$S_{yy}=\sum_{i=1}^{n}Y_i^2-\left(\sum_{i=1}^{n}Y_i\right)^2/n=582-46^2/5=158.8$$

Çözüm (b):

Korelasyon katsayısı;

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{263.8}{\sqrt{462.8 * 158.8}} = 0.973$$

Balığın boyu ile çevresi arasında % 97.3'lük pozitif bir ilişki vardır. Bir diğer ifade ile balığın boyu arttıkça, çevresi de artmaktadır.

Korelasyon Katsayısının Önem Testi:

Hipotezler

Ho: $\rho=0$

Hı : $\rho\neq0$

$\alpha=0.01$

$$t = \frac{r-\rho}{S_{\bar{r}}} = \frac{0.973-0}{\sqrt{\frac{1-0.973^2}{5-2}}} = \frac{0.973}{0.1333} = 7.299$$ > $t_{5-2,0.01/2}=t_{3,0.005}=5.841$ olduğundan Ho RED edilir.

Örnek: İneğin günlük yediği yem(kg) ile verdiği süt(kg) arasında bir ilişki olup olmadığını araştırmak amacıyla yapılan bir denemeden elde edilen veriler aşağıdaki gibir.

a) Bu ilişkiyi açıklayan regresyon modelini oluşturunuz ve regresyonun önem testini yapınız

b) Korelasyon katsayısını hesaplayarak önem testini yapınız.

Yem Miktarı.(kg)	Verdiği Süt Miktarı (kg)
5	12
7	18
9	19
10	22
8	20
6	13
11	25