Regresyon ve Korelasyon Analizi

Yrd. Doç. Dr. Emre ATILGAN

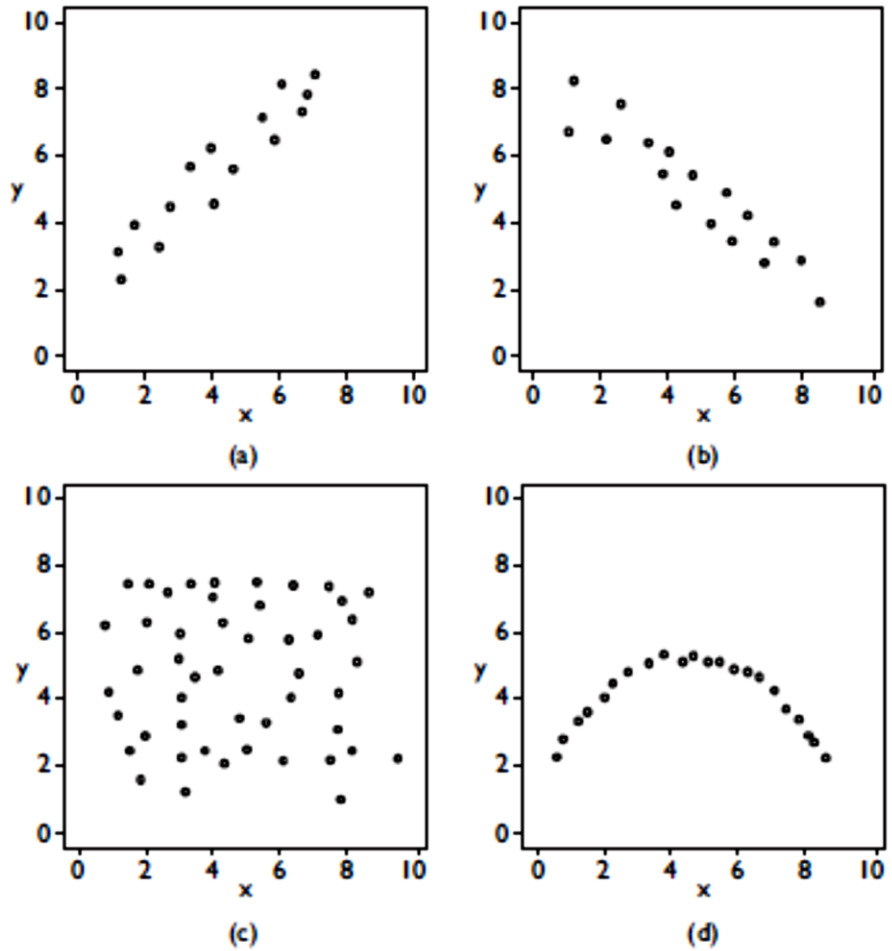
KORELASYON VE REGRESYON

Regresyon Analizi: iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan analiz metodudur. Eğer tek bir değişken kullanılarak analiz yapılıyorsa buna tek değişkenli regresyon, birden çok değişken kullanılıyorsa çok değişkenli regresyon analizi olarak isimlendirilir. Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı, eğer ilişki var ise bunun gücü hakkında bilgi verir.

Korelasyon: iki rassal değişken arasındaki doğrusal ilişkinin yönünü ve gücünü belirtir. Genel istatistiksel kullanımda korelasyon, bağımsızlık durumundan ne kadar uzaklaşıldığını gösterir.

KORELASYON ANALİZİ

Korelasyon analizi genel olarak değişkenler arasındaki ilişkinin incelenmesidir.



a) Pozitif ilişki
b) Negatif ilişki
c) Rassal İlişki (İlişkisiz)
d) Rassal (eğrisel) İlişki

Pearson Korelasyon Katsayısı (r)

Ölçümle belirtilen iki değişken arasındaki <u>doğrusal</u> ilişkinin kuvveti (derecesi) ve yönü hakkında bilgi verir.

$$-1 \leq r \leq 1$$

arasında değişir.

İlişkiler aşağıdaki gibi nitelendirilebilir.

r	İlişkinin derecesi
0.90 to 1.00	Çok kuvvetli
0.70 to 0.89	Kuvvetli
0.50 to 0.69	Orta
0.30 to 0.49	Düşük
0.00 to 0.29	Zayıf

r’nin Hesaplanması

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right) \left(\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \right)}}$$

r’nin Anlamlılığı

Korelasyon katsayısının (r) anlamlı olup olmadığı (sıfırdan farklı olup olmadığı) t dağılımı yardımı ile test edilebilir.

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Karşılaştırma

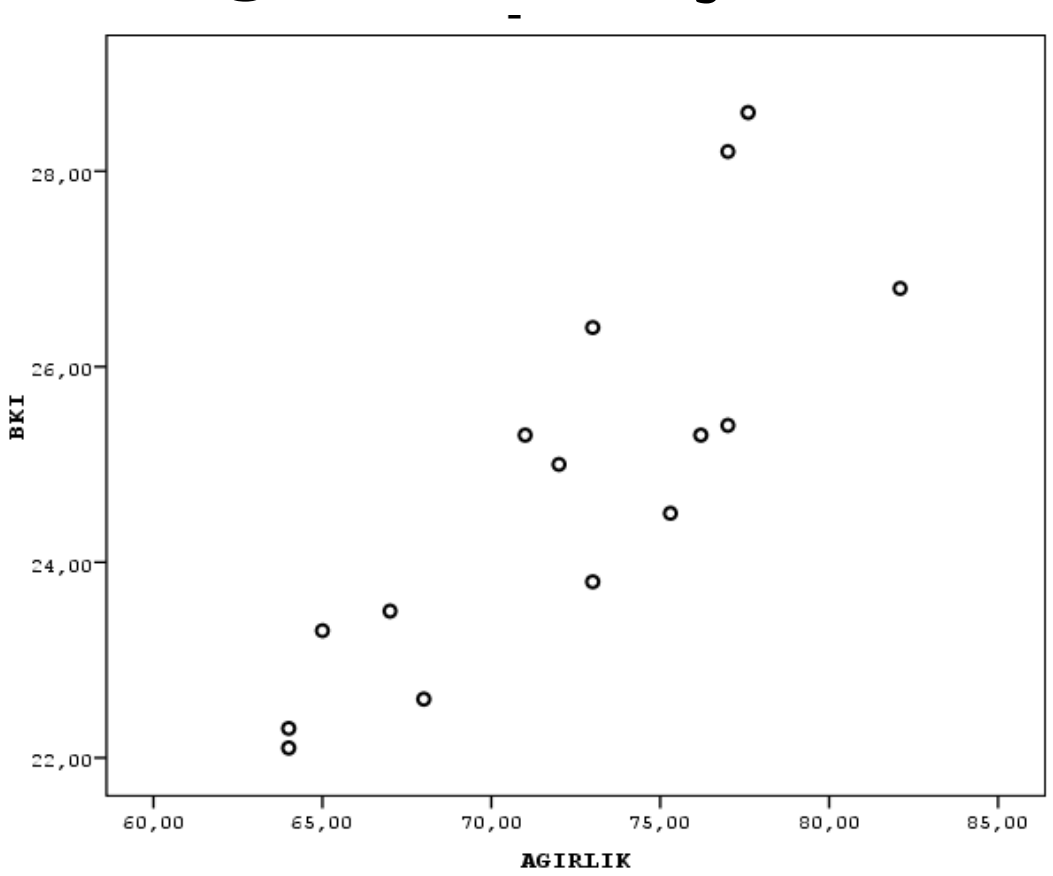
- Hesapla bulunan t istatistiği, belirlenen yanılma düzeyinde n-2 serbestlik dereceli t tablo istatistiği ile karşılaştırılır.
- $t_{Hesap} > t_{Tablo}$ ise iki değişken arasındaki ilişkinin sıfırdan farklı olduğu söylenir.

Örnek:

Ağırlık-BKI İlişkisi

Ağırlık (x)	BKI (y)
64.0	22.1
75.3	24.5
73.0	23.8
82.1	26.8
76.2	25.3
.	.
.	.
77.6	28.6

Ağırlık-BKI İlişkisi



$$r = \frac{27047.85 - \frac{(1082.20) \times (373.10)}{15}}{\sqrt{\left(78509.70 - \frac{(1082.20)^2}{15}\right) \times \left(9337.03 - \frac{(373.10)^2}{15}\right)}} = 0.83$$

İLİŞKİ(KORELASYON) MATRİSİ

	AĞIRLIK	BKI
AĞIRLIK	1,0000	0,83
BKI	0,83	1,0000

Anlamlılığın Test Edilmesi

1. ***Hipotezlerin Kurulması.***

H_0: *İki değişken arasında ilişki yoktur (ρ=0).*

H_1: *İki değişken arasında pozitif ilişki vardır (ρ>0)*

2. ***Test istatistiğinin elde edilmesi.***

$$t = \frac{0.83}{\sqrt{\frac{1-0.83^2}{15-2}}} = 5.39$$

3. Yanılma düzeyi alfa=0.05 alınmıştır.

4. Karar için:

Sd=n-2=15-2=13'tür.

13 serbestlik dereceli tek yönlü t tablo istatistiği 1.77 olarak bulunur.

Karar:

$t_{Hesap}=5.39>t_{Tablo}=1.77$ olduğu için H_0 hipotezi reddedilir ve r'nin sıfırdan büyük bir değer olduğu söylenir ($p<0.05$).

Açıklayıcılık (Belirlilik) Katsayısı (R^2)

- Açıklayıcılık (belirtme) katsayısı (R^2), değişkenleri bağımlı-bağımsız değişken olarak düşündüğümüzde bağımlı değişkendeki toplam değişimin yüzde kaçının bağımsız değişken tarafından açıklanabildiğini belirtir.
- İki değişken arasında doğrusal ilişki olması durumunda, korelasyon katsayısının karesi açıklayıcılık katsayısına eşittir.

$$\mathbf{R^2=r^2}$$

- R^2 değeri 0 ile +1 arasında değişir.
- R^2 değerinin 1'e yaklaşması, bağımlı değişkendeki değişimin büyük bir bölümünün bağımsız değişken tarafından açıklandığını gösterir.

Spearman Sıra Korelasyon Katsayısı (r_s)

- Pearson korelasyon katsayısının parametrik olmayan karşılığıdır.
- Değişkenlerin biri ya da her ikisinin normal dağılmadığı durumlarda kullanılabileceği gibi doğrudan sıralı (ordinal) olarak elde edilen ya da belli bir kritere göre sıralanmış olan iki değişkenin ilişki miktarını belirlemek amacı ile de kullanılır.

REGRESYON ANALİZİ

- İki değişken arasındaki korelasyon katsayısı yeterince büyükse, kolay elde edilen bir x değişkeni değeri yardımıyla elde edilmesi zor olan bir y değişkeni değeri kestirilebilir. Bu kestirim regresyon çözümlemesi yardımıyla yapılır.

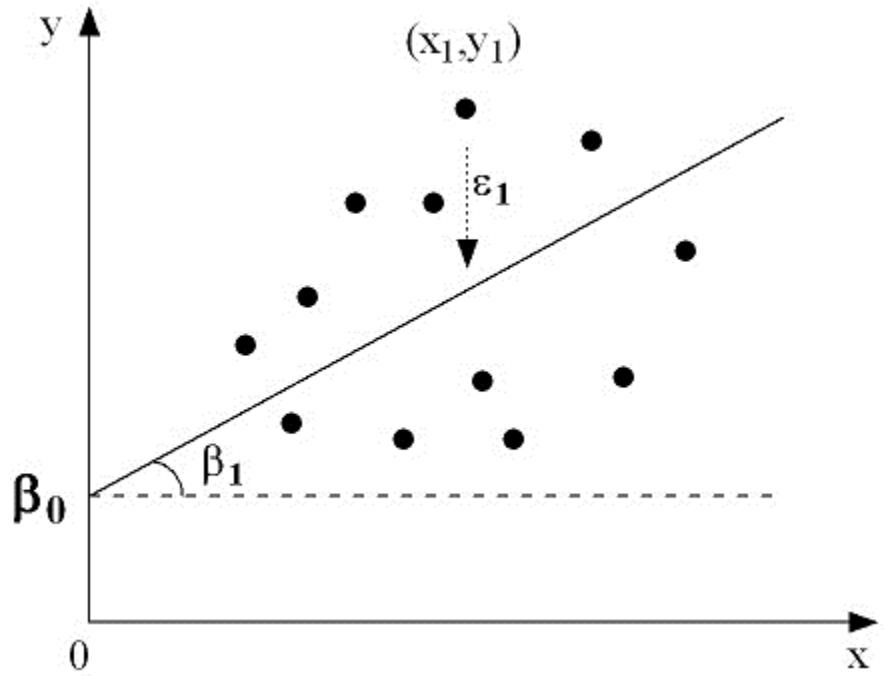
Ör:

✓İnsanların boyları ile kiloları
✓Futbol takımlarının çalışma süreleri ve maç skorları toplamları
✓Öğrencilerin çalışma miktarları ve sınav notları
✓Bir malın fiyatı ve talep miktarı
✓Bir ürünün verimi ve verilen gübre miktarı, vb.

REGRESYON ANALİZİ

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

Burada;
X: Bağımsız (Açıklayıcı) Değişken
Y: Bağımlı (Açıklanan;Etkilenen;Cevap) Değişken
β_0: X=0 olduğunda bağımlı değişkenin alacağı değer (kesim noktası)
β_1: Regresyon Katsayısı [tan(β_1) regresyon doğrusunun eğimini verir]
ε : Hata terimi (Ortalaması=0 ve Varyansı=σ^2'dir)



Regresyon Katsayısı (β_1) :

Bağımsız değişkendeki bir birimlik değişimin, bağımlı değişkendeki yaratacağı ortalama değişimi göstermektedir.

ϵ: (Hata terimi):

Her bir gözlem çiftindeki bağımlı değişkene ilişkin gerçek değer ile modelden tahmin edilen değer arasındaki farktır.

$$\varepsilon_i = (\beta_0 + \beta_1 X) - Y_i$$

$$\hat{Y}_i$$

Tanımlanan Regresyon Modeli Kitleden seçilen n gözlemli örneklem için;

$$\hat{Y} = b_o + b_1 X$$ **biçimindedir**

Yukarıdaki Doğrusal Regresyon Modeli Gözlemler için ;

$$\hat{y}_i = b_0 + b_1 x_i + e_i \qquad \text{i = 1 ,..., n}$$

Kesim Noktası ve Regresyon Katsayısının Tahmin Yöntemi

- Doğru ve güvenilir bir regresyon modelinde amaç, gerçek gözlem değeri ile tahmin değeri arasında fark olmaması yada farkın minimum olmasıdır. Bunun için çeşitli tahmin yöntemleri geliştirilmiştir. Bu yöntemlerden biri **"En Küçük Kareler"** kriteridir.

$$\sum_{i=1}^{n} e_i^2 \quad _i = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Bu farkın en küçük olması amaçlanır

En Küçük Kareler Yöntemi ile Bulunan Tahminler

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(n\,\bar{x}\,\bar{y}\right)}{\sum_{i=1}^{n} x_i^2 - \left(n\,\bar{x}^2\right)}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Değişkenler birlikte artıyor artıyor yada birlikte azalıyor ise **“b_1 pozitif değerli”**dir.
- Değişkenlerden biri artarken diğeri azalıyor ise **“b_1 negatif değerli”**dir.

Basit Doğrusal Regresyon Analizinde Hipotez Testleri

F – Testi: Doğrusallıktan Ayrılışın Önem Kontrolü

H_0: Gözlenen Noktaların Regresyon Doğrusuna Uyumu Önemsizdir (Model geçersizdir)

H_1 : Gözlenen Noktalar Regresyon Doğrusu ile tanımlanabilir (Model Geçerlidir)

Eğer F_H=(RKO / RAKO) > $F_{(1;n-2;\ \alpha)}$ ise

Ho Hpotezi RED Edilir.

Basit Doğrusal Regresyon Analizinde Hipotez Testleri

t – Testi: Regresyon Katsayısının Önem Kontrolü

H_0: Regresyon Katsayısı Önemsizdir ($\beta_1=0$)

H_1 : Regresyon Katsayısı Önemlidir ($\beta_1\neq0$)

«Burada, regresyon katsayısının önemsiz olması demek; örneklemin çekildiği kitlede, bağımsız değişkende bir birimlik değişimin, bağımlı değişkende değişiklik yaratamayacağı anlamına gelir.»

Eğer $\mathbf{t_h > t_{(n-2;\,\alpha)}}$ **ise Ho Hipotezi RED** edilir.

Basit Doğrusal Regresyon Analizinde Özel Durum

- Basit Doğrusal regresyonda tek bir bağımsız değişken olması nedeniyle t dağılımı ve F dağılımı arasında aşağıdaki matematiksel eşitlik söz konusudur :

$$t_h^2 = F_h$$

Açıklama (Belirtme) Katsayısı R^2

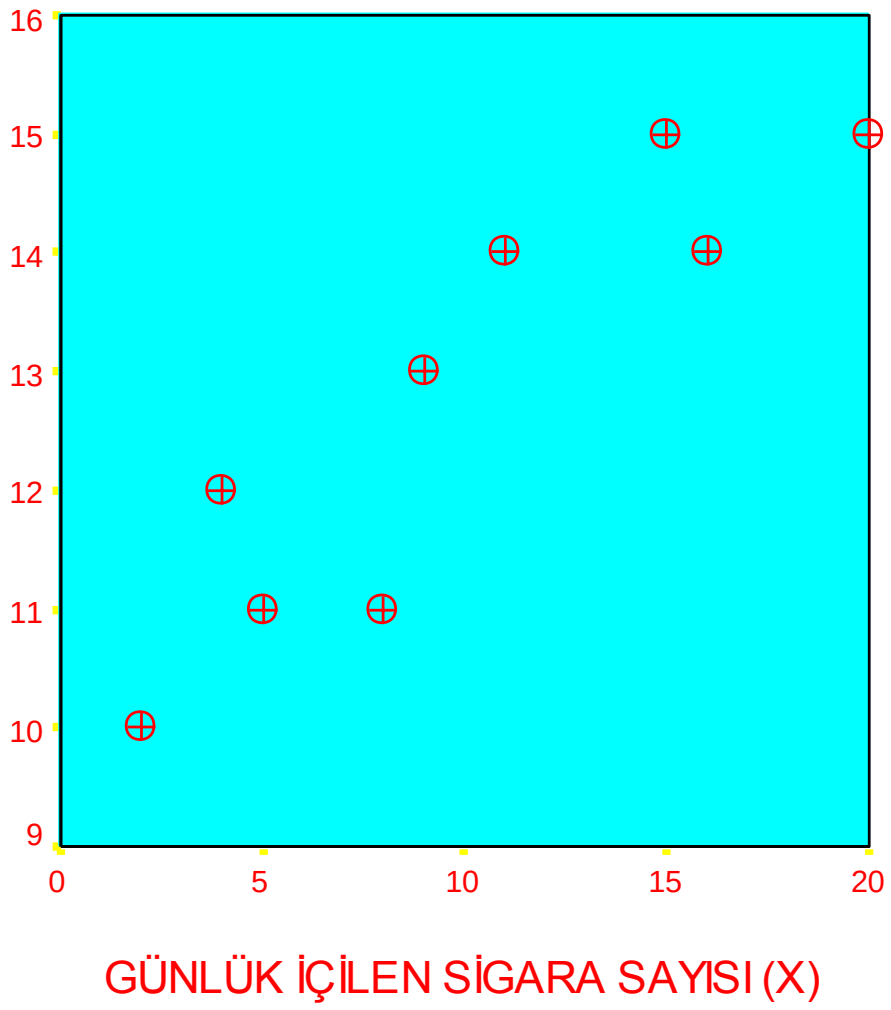
- Yüzde cinsinden ifade edilen açıklama katsayısı, regresyon analizinde önemlidir.

$$o \leq R^2 \leq 1$$

- Açıklama Katsayısı bire yakın bulunur ise, bağımlı değişkendeki değişimin büyük bir kısmı bağımsız değişken tarafından açıklanabilir yorumu yapılabilmektedir.

Örnek:

Günlük içilen sigara sayısı(x)	Sistolik kan basıncı(y)
4	12
11	14
8	11
15	15
5	11
16	14
20	15
9	13
2	10



b_0= 10.004

b_1= 0.277

$$\hat{y}_i = 10.004 + 0.277 x_i$$

r=0.903

$R^2=0.903^2=0.815$

KATSAYILARA İLİŞKİN İSTATİSTİKLER

Değişken	Katsayı	Standart hata	t	p
Sabit	10.004	0.574	17.425	0.000
x	0.277	0.050	5.561	0.001

VARYANS ANALİZİ TABLOSU

DK	KT	Sd	KO	F	P
Toplam	27.556	8			
Regresyon	22.469	1	22.469	30.923	0.001
Artık	5.086	7	5.086		

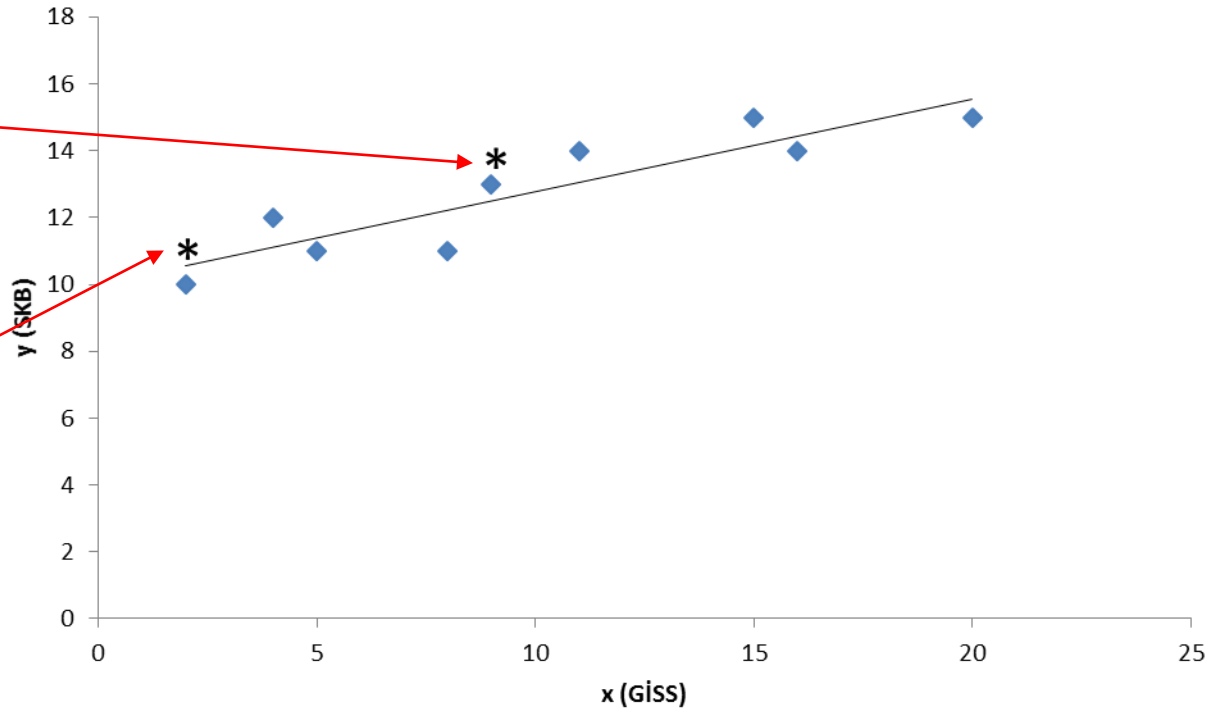
Regresyon Doğrusunun Çizimi

$x_i = 8$ için

$\hat{y}_i = 10.004 + 0.277 \times (8) = 12.22$ cmHg

$x_i = 2$ için

$\hat{y}_i = 10.004 + 0.277 \times (2) = 10.56$ cmHg



ÇOKLU DOĞRUSAL REGRESYON ANALİZİ

- ***Amaç:***
- Kolay elde edilebilir bağımsız değişkenler yardımıyla zor elde edilen bağımlı değişken değerini kestirmek
- Bağımsız değişkenlerden hangisi ya da hangilerinin bağımlı değişkeni daha çok etkilediğini belirlemek

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots\ldots + b_p x_p$$

ÇOKLU DOĞRUSAL REGRESYON ANALİZİ

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_p x_p$$

- ***Bağımlı değişken*** sürekli ya da kesikli sayısal veri tipinde olmalıdır.
- ***Bağımsız değişkenler*** sürekli kesikli ya da nitelik veri tipinde olabilir.
- ***Nitelik bağımsız değişkenler*** olduğunda göstermelik (dummy) değişkenler oluşturulur.