KOCAELİ ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ

Bilgisayar Mühendisliğinde Matematik Uygulamaları

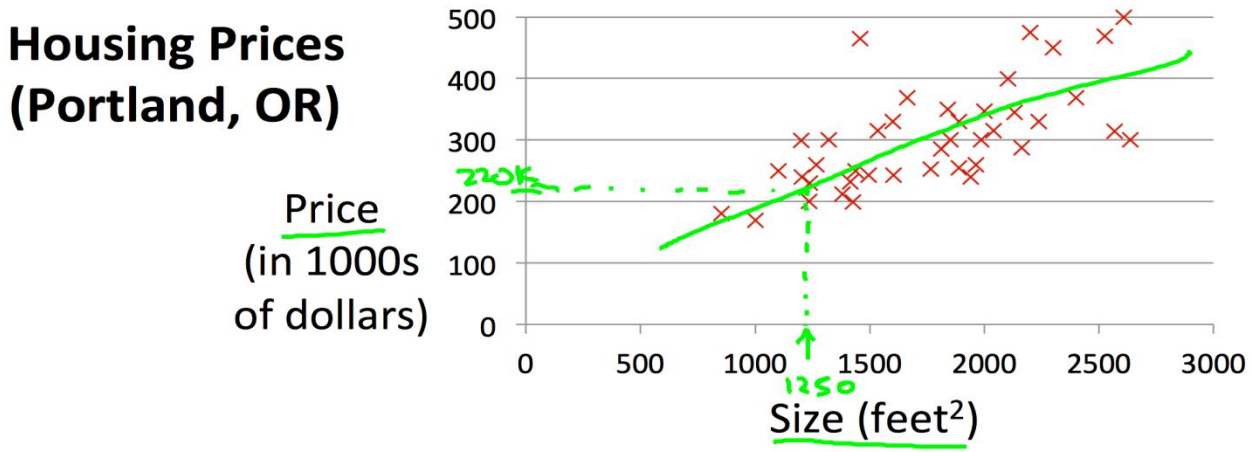
9. Hafta

Yrd. Doç. Dr. A. Burak İNNER

Kocaeli Üniversitesi Bilgisayar Mühendisliği
Yapay Zeka ve Benzetim Sistemleri Ar-Ge Lab.
http://yapbenzet.kocaeli.edu.tr

Regresyon Analizi

Bir veri tablosuna en uygun fonksiyonu bulma sürecine regresyon analizi denir.
Regresyon analizi kavramsal olarak değişkenler arasındaki fonksiyonel ilişkiyi araştırmak için kullanılan istatistiksel bir yöntemdir.
Örnek olarak bir evin satış fiyatını konum,bina yaşı,cephe gibi özellikler ile ilişkilendirebilmek verilebilir.
Regresyon analizi genellikle bir problemin tanımlanması ile başlar.
Bağımlı ve bağımsız değişkenlerin uygunca belirlenmesi gerekir.
Bağımsız (Independent) değişkenler bağımlı değişkenlerin tahmininde kullanılır.



Regresyon Modelleri

Regresyon analizi elimizdeki değişken sayısına göre farklı modellere sahip olabilir.

Tek değişkenli(one variable) lineer model

$h_\theta(x) = \theta_0 + \theta_1 x_1$

Tek değişkenli polinomal model

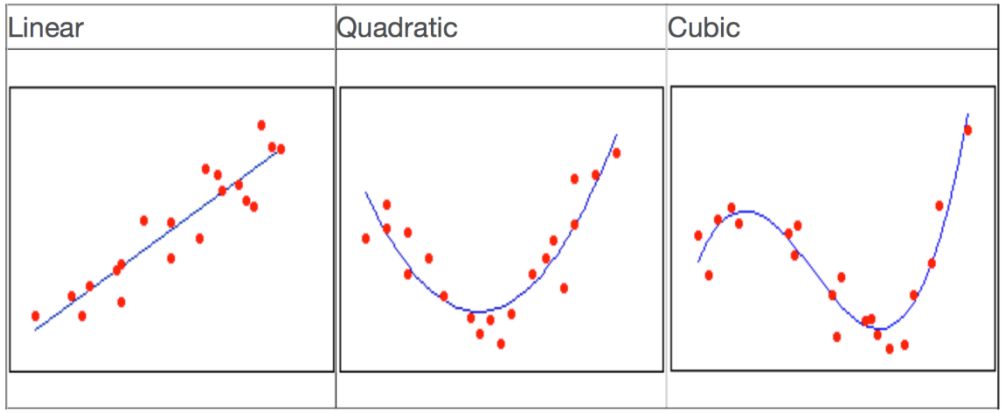
$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 {x_1}^2$

Tek değişkenli kareköksel model

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 \sqrt{x_1}$

Çok değişkenli (multi veriable) lineer model

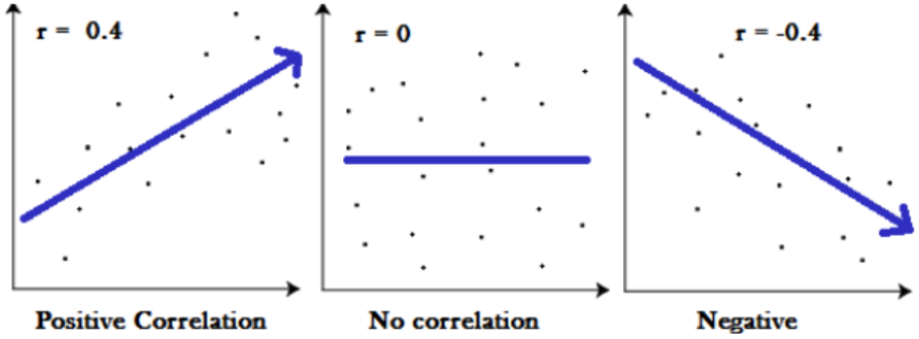
$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n$



Korelasyon

Korelasyon, iki yada daha fazla veri arasındaki anlamlı ilişkiyi gösteren ifadedir. Korelasyon veriler arasındaki ilişkinin gücünü ve bu ilişkinin yönünü gösterir.Korelasyon 1 ve -1 arasında değerler alır. 1 ve -1'e yaklaştıkça ilişkinin gücü artar, 0'a yaklaştıkça ilişkinin gücü azalır. "" ve "-" korelesyonun yönünü belirtir. "+" işaretli ise pozitif yönlü ilişki, "-" ise negatif yönlü bir ilişkiyi anlatır.

İki değişken(biri bağımlı, biri bağımsız) arasındaki korelasyon düşünülürse; pozitif yönlü korelasyon iki veriden bir artarken diğerininde artış gösteriyor olmasını gösterir, negatif yönü korelasyon ise bir değişkeni artış gösterirken diğerinin azaldığını işaret eder.



Graphs showing a correlation of -1 (a negative correlation), 0 and +1 (a positive correlation)

Pearson Korelasyonu

Pearson korelasyon sayısı istatistikte çok geniş kullanım alanın sahiptir. Hesaplaması şu şekilde yapılır.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r = Pearson r korelasyonu

N = Data setindeki değerlerin sayısı

∑xy = x ve y değerlerinin çarpımlarının toplamı

∑x = x'lerin toplamı

∑y = y'lerin toplamı

∑x2= x'lerin kareleri toplamı

∑y2= y'lerin kareleri toplamı

Etki aralığı(Effect size)

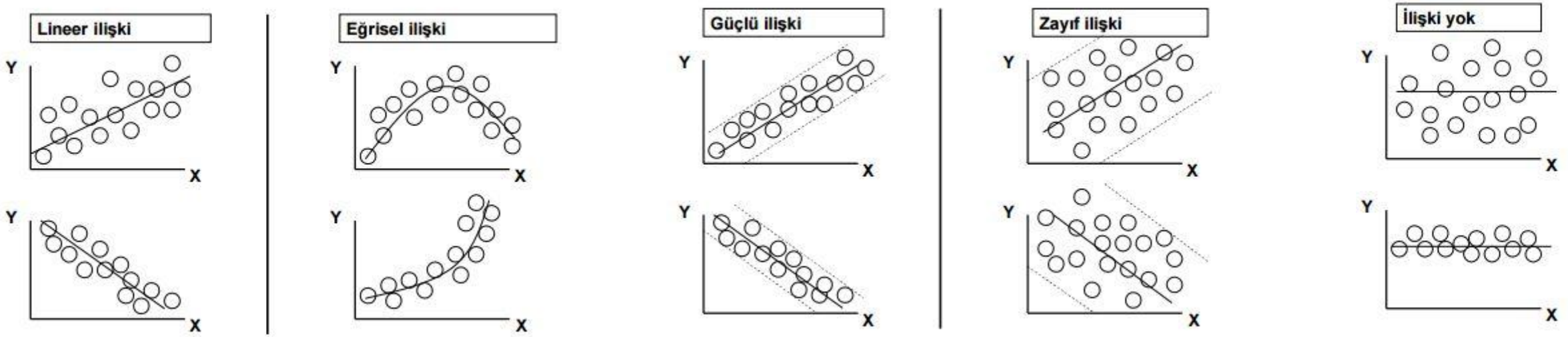
.10 ile .29 aralığı az birliktelik(small association)

.30 ile .50 aralığı orta birliktelik(medium association)

.50 ve üzeri yüksek birliktelik (large association) gösterir

Kaynak:http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

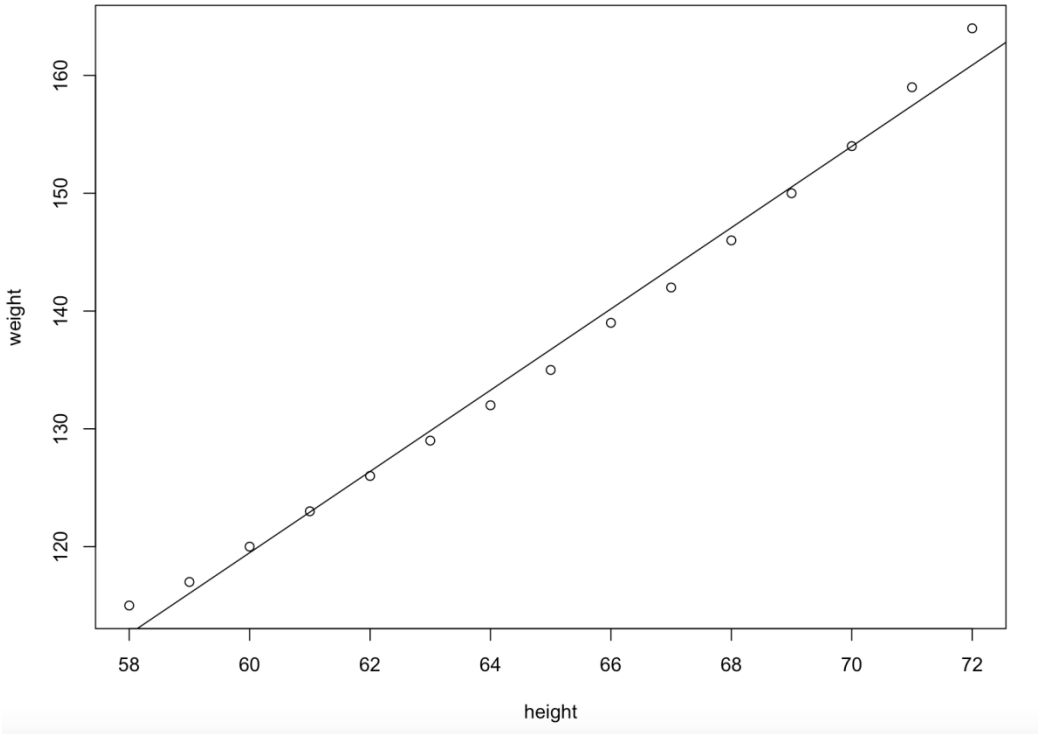
Korelasyon


Lineer ilişki
Y
X
Y
X
Eğrisel ilişki
Y
X
Y
X
Güçlü ilişki
Y
X
Y
X
Zayıf ilişki
Y
X
Y
X
İlişki yok
Y
X
Y
X

Korelasyon örnek R Kodu

```
cor(women$height,women$weight , method = "pearson")
plot(women)
abline(lm(weight ~height , data=women))
```

```
cor(women$height,women$weight,method = ”pearson”)
[1] 0.9954948
```



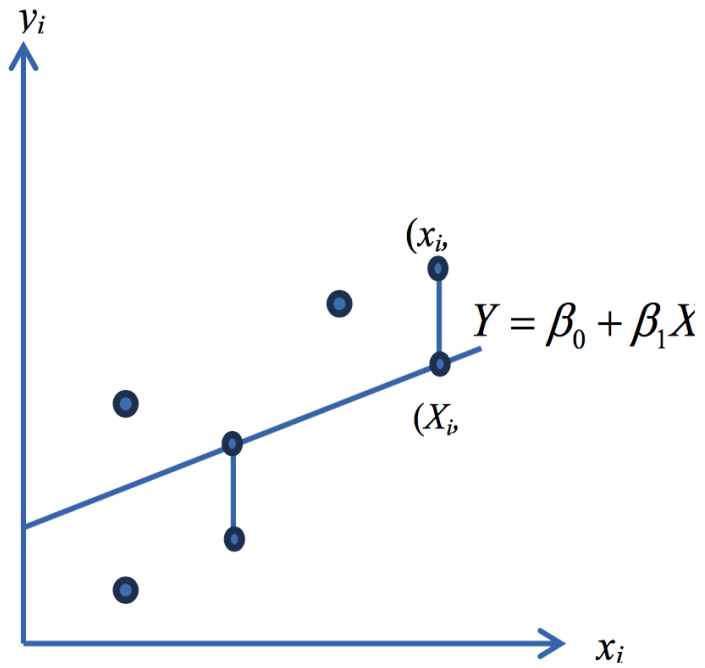
Parametre Kestirimi

Hipotez

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, (i = 1,2,\ldots\ldots,n)$$

Regresyon modeli aşağıdaki hata fonksiyonunu minimize etmeyi amaçlar

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$



Parametre Kestirimi

Parametrelere göre kısmi türev alırsak

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_t - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 -)\beta_1 x_i)x_i.$$

Bu türevler “0” ‘a eşit olmalı

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Parametre Kestirimi

Bu çözümlerin sonucunda

$b_0 = \bar{y} - b_1\bar{x}$ b_0 of β_0 and b_1 of β_1

$b_1 = \frac{s_{xy}}{s_{xx}}$

s_{xy} ve s_{xx}

$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$, $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Elde edilir.

Kaynak:http://home.iitk.ac.in/~shalab/econometrics/Chapter2-Econometrics-SimpleLinearRegressionAnalysis.pdf

Polinom Regresyonu

Bazı durumlarda tüm verileri ideal bir doğrusal işlev ile tanımlamak her zaman mümkün değildir, bu durumda veri bir eğri işlevi ile ifade edilebilir.

$$y= \alpha_0+\alpha_1 x+\alpha_2 x^2+\varepsilon$$

En küçük kareler işlemi,yüksek dereceli polinomlarla verilere eğri uydurma için kullanılabilecek şekilde kolayca genişletilebilir.

Polinom Regresyonu

Artıkların karelerinin toplanması durumunda

$$S_T = \sum \left(y_i - a_0 + a_1 x_i + a_2 x_i^{\ 2} \right)^2$$

olur. Elde ettiğimiz bu eşitliğin polinomun bilinmeyen katsayılarının her birine göre türevini alalım:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum \left(y_i - a_0 + a_1 x_i + a_2 x_i^{\ 2} \right)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i \left(y_i - a_0 + a_1 x_i + a_2 x_i^{\ 2} \right)$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^{\ 2} \left(y_i - a_0 + a_1 x_i + a_2 x_i^{\ 2} \right)$$

Polinom Regresyonu

Bu denklemler sıfıra eşitlenebilir ve yeniden düzenlenerek aşağıdaki normal denklemler elde edilebilir.

$$(n)a_o+(\sum x_i)a_1 + (\sum {x_i}^2)a_2 = \sum y_i$$
$$(\sum x_i)a_o+(\sum {x_i}^2)a_1 + (\sum {x_i}^3)a_2 = \sum x_i y_i$$
$$(\sum {x_i}^2)a_o+(\sum {x_i}^3)a_1 + (\sum {x_i}^4)a_2 = \sum {x_i}^2\, y_i$$

Buradaki tüm toplamlar i=1'den n'ye kadardır. Dikkat ederseniz yukarıdaki üç denklem doğrusaldır ve üç bilinmeyeni vardır:a_0, a_1 ve a_2 . Bilinmeyenlerin katsayıları,gözlenmiş verilerden doğrudan hesaplanabilir.,

Bu durumda, gördüğünüz gibi bir en küçük kareler,ikinci derece polinomu belirleme problemi, eşzamanlı üç doğrusal denklem sisteminin çözümüne eşdeğerdir.

Polinom Regresyonu

İki boyutlu durum kolaylıkla, aşağıdaki gibi m'inci dereceden polinoma genelleştirilebilir.

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e$$

Yukarıdaki analiz buradaki daha genel duruma kolaylıkla genişletilebilir. Böylece, m'inci dereceden bir polinomun katsayılarının belirlenmesi hemen söyleyebileceğimiz gibi, m+1 eşzamanlı doğrusal denklem sisteminin çözümüne eşdeğerdir. Bu durumda, standart hata aşağıdaki şekilde formülleştirilebilir.

$$S_{\frac{y}{x}} = \sqrt{\frac{S_r}{n-(m+1)}}$$

Polinom Regresyonu

Bu değer S_r 'yi hesaplamak için $(m+1)$ adet veri kaynaklı katsayı -$a_0, a_1, \ldots, a_m$ kullanıldığı için $n-(m+1)$'e bölünmüştür; böylece $m+1$ dereceden serbestliği kaybetmiş oluyoruz.

Polinom Regresyonu

Örnek: Tablonun ilk iki sütunundaki verilere ikinci dereceden bir polinom uydurun.

x_i	y_i	$(y_i - \bar{y})^2$	$(y_i - a_0 - a_1 x_i - a_2 {x_i}^2)$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
$\sum$	152.6	2513.39	3.74657

Tablo 1.1 : İkinci dereceden en küçük karelerin hata analizi için hesaplamalar

Polinom Regresyonu

Çözüm: Verilerden:

$$m = 2 \qquad \sum x_i = 15 \qquad \sum x_i^4 = 979$$
$$n = 6 \qquad \sum y_i = 152.6 \qquad \sum x_i y_i = 585.6$$
$$\bar{x} = 2.5 \qquad \sum x_i^2 = 55 \qquad \sum x_i^2 y_i = 2488.8$$
$$\bar{y} = 25.433 \qquad \sum x_i^3 = 225$$

bulunur. Dolayısıyla, eşzamanlı doğrusal denklemler şu şekilde elde edilir.

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{Bmatrix}$$

Polinom Regresyonu

Çözüm devamı:

Gauss eleme tekniği kullanarak bu sistemin çözümünden

$$a_0 = 2.47857$$
$$a_1 = 2.35929$$
$$a_2 = 1.86071$$

Böylece bu problem için ikinci dereceden denklem

$$y = 2.47857 + 2.35929x + 1.86071x^2$$

Olur.

Polinom Regresyonu

Çözüm devamı:

Regresyon polinomuna dayalı olarak tahminin standart hatası şöyledir:

$$S_{\frac{y}{x}} = \sqrt{\frac{3.74657}{6-3}} = 1.12$$

Determinasyon katsayısı ise:

$$r^2 = \frac{2513.39 - 3.74657}{2513.39} = 0.99851$$

Polinom Regresyonu

Çözüm devamı:

Korelasyon katsayısı $r = 0.99925$'dir. Bu sonuçlar orijinal belirsizliğin yüzde 99.925'nin modelle açıklandığını göstermektedir ve ikinci dereceden denklemin mükemmel bir uyum verdiği yorumunu destekler.

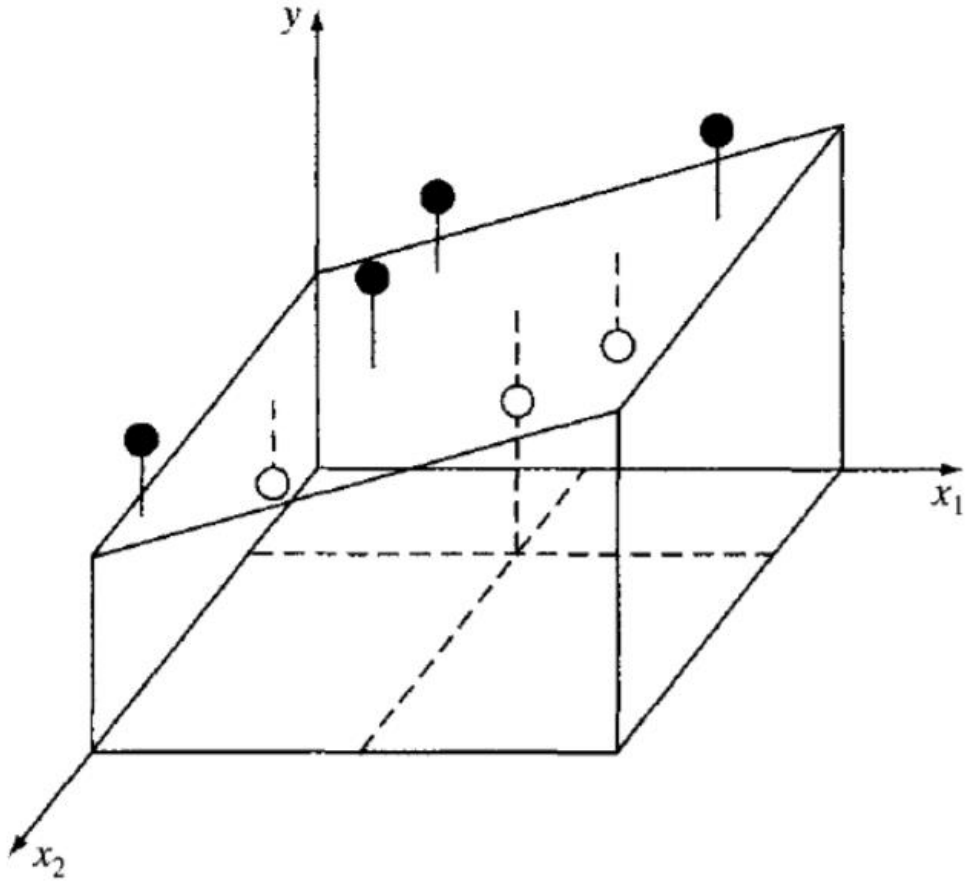
Çoklu Doğrusal Regresyon

Doğrusal regresyonun kullanışlı bir uzantısı, y'nin iki veya daha çok bağımsız değişkenin doğrusal fonksiyonu olduğu durumdur. Örneğin, y aşağıdaki gibi x_1 ve x_2 'nin doğrusal fonksiyonu olabilir.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$$

Bu tür bir denklem, incelenen değişkenin genellikle diğer iki değişkenin bir fonksiyonu olduğu durumda deneysel eğri uydurmak için özellikle kullanışlıdır. Bu iki boyutlu durum için regresyon doğrusu bir düzleme dönüşür.

Çoklu Doğrusal Regresyon



Şekil 1.1: y; x_1 ile x_2 'nin doğrusal fonksiyonu olduğunda, çoklu doğrusal regresyonunun grafik açıklaması

Çoklu Doğrusal Regresyon

Daha önceki durumlarda olduğu gibi, katsayıların en iyi değeri, artıkların karelerinin toplamı yardımıyla belirlenir.

$$S_r = \sum_{i=1}^{n}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})^2$$

Bilinmeyen katsayıların her birine göre türev alınarak

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} x_{1i}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S_r}{\partial a_2} = -2\sum_{i=1}^{n} x_{2i}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

bulunur.

Çoklu Doğrusal Regresyon

- Artıkların karelerinin toplamını minimum yapan katsayılar, kısmi türevlerin sıfıra eşitlenmesi ve sonucunda aşağıdaki gibi matris formunda ifade edilmesiyle elde edilir.

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum {x_{1i}}^2 & \sum x_{1i}\,x_{2i} \\ \sum x_{2i} & \sum x_{1i}\,x_{2i} & \sum {x_{2i}}^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1i}\,y_i \\ \sum x_{2i} y_i \end{Bmatrix}$$

Çoklu Doğrusal Regresyon

Örnek: Aşağıdaki veriler $y = 5 + 4x_1 - 3x_2$ denkleminden hesaplanmıştır:

x_1	x_2	y
0	0	5
2	1	10
2.5	2	9
1	3	0
4	6	3
7	2	27

Bu verilere bağıntı uydurmak için çoklu doğrusal regresyon kullanın.

Çoklu Doğrusal Regresyon

Çözüm: Matris formunda elde etmemiz için ilk önce gerekli hesaplamaları aşağıdaki tabloda yaparız.

toplam

y	x_1	x_2	x_1^2	x_2^2	x_1x_2	x_1y	x_2y
5	0	0	0	0	0	0	0
10	2	1	4	1	2	20	10
9	2.5	2	6.25	4	5	22.5	18
0	1	3	1	9	3	0	0
3	4	6	16	36	24	12	18
27	7	2	49	4	14	189	54
54	16.5	14	76.25	54	48	243.5	100

Çoklu Doğrusal Regresyon

$$\begin{bmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix} = \begin{Bmatrix} 54 \\ 243.5 \\ 100 \end{Bmatrix}$$

Matris formunu bu şekilde elde ederiz. Bu sistem Gauss eleme gibi bir yöntemle çözülürse

$$a_0 = 5$$
$$a_1 = 4$$
$$a_2 = -3$$

bulunur. Bu değerler verilerin türetildiği orijinal denklemle tutarlıdır.

Çoklu Doğrusal Regresyon Algoritması

DO $i = 1$, order+1
  DO $j = 1, i$
    sum = 0
    DO k = 1,n
      sum=sum+$x_{i-1,k}x_{j-1,k}$
    END DO
    $a_{i,j} = sum$
    $a_{j,i} = sum$
  END DO
  sum = 0
  DO k = 1,n
    sum = sum+$y_k x_{j-1,k}$
  END DO
  $a_{i,order+2} = sum$
END DO»

Çok değişkenli Regresyon (Multivariate Linear regression)

Çoklu regresyon modeli için hiptez

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots\ldots + \theta_n x_n$$

Örneğin bir evin temel değeri, evin büyüklüğü(m^2), kat sayısı, oda sayısı, .. vb değişkenler ile evin değerinin tahmin edilmesi düşünülebilir.

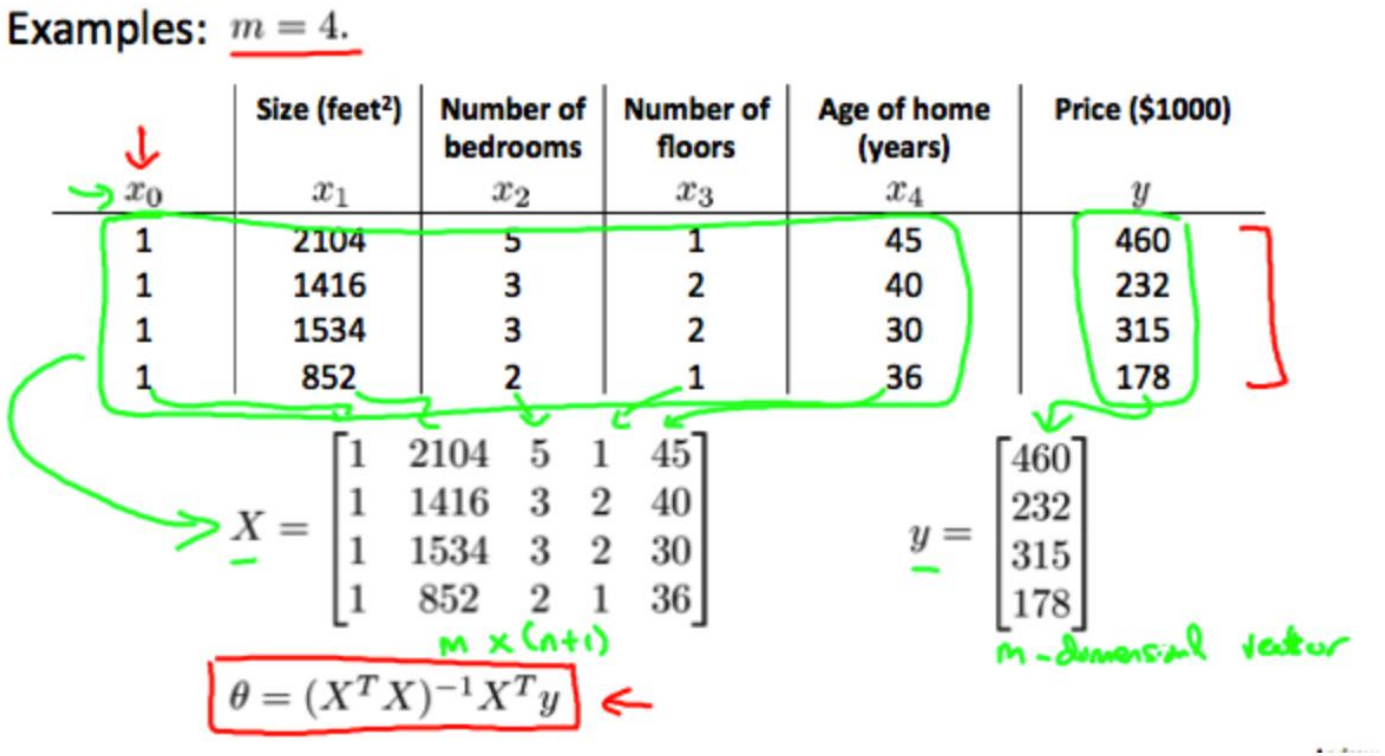
Matris çarpımı şeklinde gösterimi

$$h_\theta(x) = \begin{bmatrix} \theta_0 & \theta_1 & \cdots & \theta_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = \theta^T x$$

Çok değişkenli Regresyon (Multivariate Linear regression)

Çok değişkenli lineer regresyonun parametrelerinin hesaplama yöntemlerinden bir tanesi normal hesaplama yöntemi (Normal Equation) şu şekilde

$\theta = (X^T X)^{-1} X^T y$



Çok değişkenli Regresyon (Multivariate Linear regression)

Burada X^TX ifadesinin tersini almada problem yaşanabilir. Bu matrisin tersinin alınabilmesi için kare matris olması gerek. Fakat kare matris olmayabilir. Bu durumu çözmek için farklı matematiksel yöntemler mevcut. Matlab eğer tersi alınamayan bir fonksiyon ise hata verecektir. Bunun için matlab'daki

pinv()

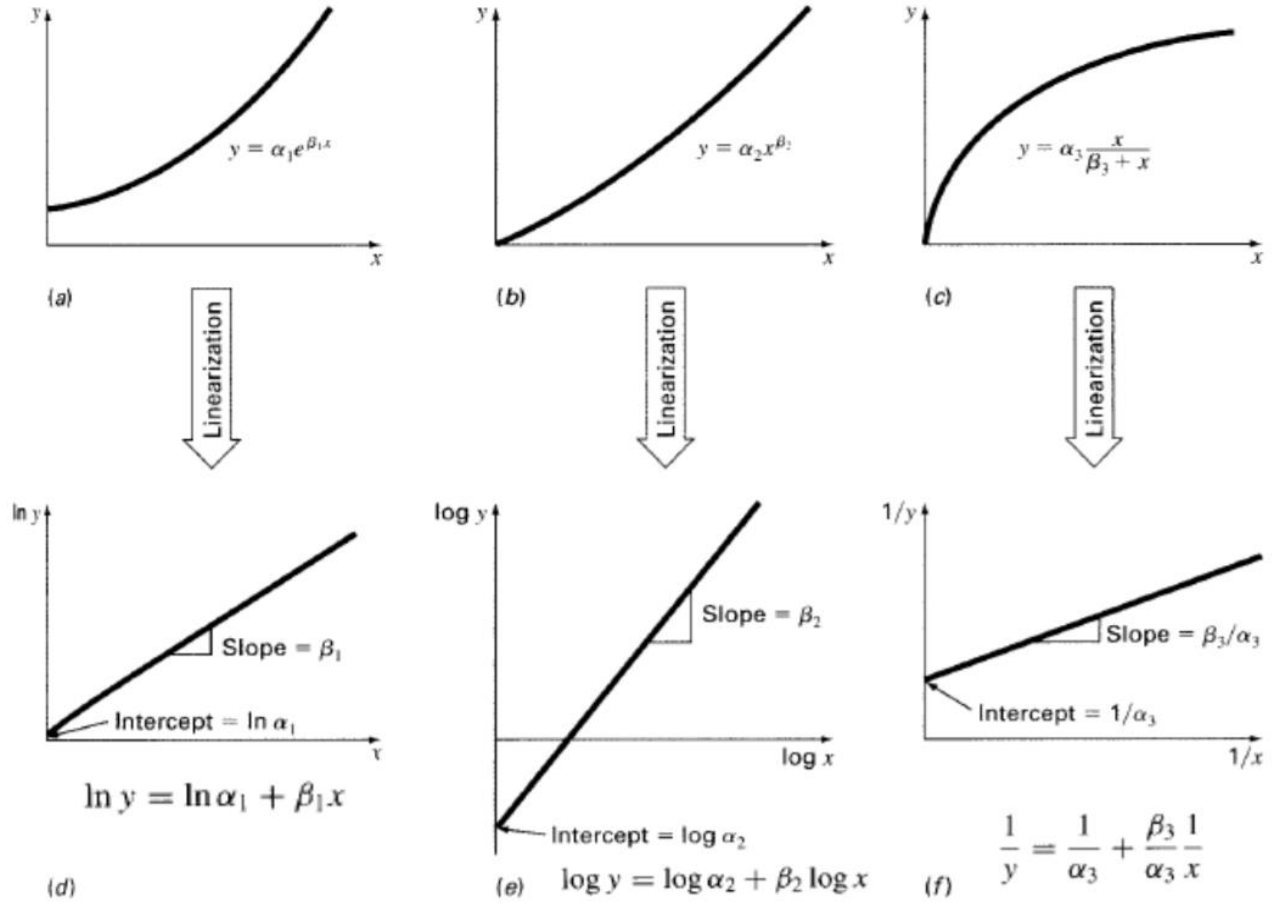
fonksiyonu bizim için bu problemi ortadan kaldıracaktır.

$$\theta = (X^T X)^{-1} X^T y$$

Doğrusal Olmayan Bağıntıların Doğrusallaştırılması

Doğrusal regresyon verilere en iyi doğruyu uydurmak için güçlü bir tekniktir. Ancak, bağımlı ve bağımsız değişken arasındaki bağıntının doğrusal olduğu gerçeğine dayanmaktadır. Bu her zaman geçerli değildir ve herhangi bir regresyon analizinin ilk adımı, verilerin grafiğini çizmek ve doğrusal bir modelin uygulanıp uygulanamıyacağı görsel olarak incelemek olmalıdır.

Doğrusal Olmayan Bağıntıların Doğrusallaştırılması



(a)Üstel denklem,(b) üslü denklem,(c)doymuş büyüme oranı denklemi.(d),(e) ve (f) bu denklemlerin doğrusallaştırılmış hali.

Doğrusal Olmayan Bağıntıların Doğrusallaştırılması

$$y = a_1 e^{b_1 x}$$

Burada a_1 ve b_1 sabitlerdir. Bu model, mühendisliğin birçok alanında kendi büyüklükleri ile orantılı bir hızda artan veya azalan nicelikleri karakterize etmek için kullanılır. Örneğin, nüfus artış oranı veya radyoaktif bozunma bu tür davranışa örnek gösterebilir.Yukarıdaki (a) numaralı şekilde gösterildiği gibi, denklem, y ve x arasında doğrusal olmayan bir bağıntıyı ifade etmektedir.

Doğrusal Olmayan Bağıntıların Doğrusallaştırılması

Örneğin bir önceki sayfada bulunan denklem doğal logaritması alınmak suretiyle doğrusallaştırılabilir.

$$\ln y = \ln a_1 + b_1 x \ln e$$

Ancak $\ln e = 1$ olacağı için

$$\ln y = \ln a_1 + b_1 x$$

olacaktır. Böylece x'e göre $\ln y$'nin çizimi,eğimi b_1 ve kesme noktası $\ln a_1$ olan düz bir doğru verecektir.

Teşekkürler.

Dersin Sonu

Kocaeli Üniversitesi Bilgisayar Mühendisliği
Yapay Zeka ve Benzetim Sistemleri Ar-Ge Lab.
http://yapbenzet.kocaeli.edu.tr/