Examples: $m = 4$.

x_0	Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix} \qquad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

m × (n+1)   m-dimensional vector

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

Matrix form to be fitted

- For convenience of writing the expressions introduce zeroth variable x_0; which always takes the value 1
- Hypothesis function is $y = y(x_1, x_2, \dots, x_n) = \; b_0x_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n;$
- $\boldsymbol{b} = \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_n \end{pmatrix} \epsilon\, \mathcal{R}^{n+1}$, $\boldsymbol{x} = \begin{pmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{pmatrix} \epsilon\, \mathcal{R}^{n+1}$; $\boldsymbol{b^T} = (b_0, b_1, \ldots, b_n)$ is 1 x (n+1) vector and $\boldsymbol{x}$ is (n+1)x1 vector
- $\boldsymbol{b^T x}$ is 1x1 vector : $= (b_0, b_1, \ldots, b_n) \cdot \begin{pmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{pmatrix}$
- $= b_0x_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$
- $y = \boldsymbol{b^T}\ \boldsymbol{x}$ is the hypothesis function

$\boldsymbol{x_0}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	. . . .	$\boldsymbol{x_n}$	$\boldsymbol{y}$
1	x_{11}	x_{21}	x_{31}	. . .	x_{n1}	y_1
1	x_{12}	x_{22}	x_{32}	. . .	x_{n2}	y_2
1	. . .	. . .	. . .	. . .	. . .	. . .
1	x_{1i}	x_{2i}	x_{3i}	. . .	x_{ni}	y_i
1	. . .	. . .	. . .	. . .	. . .	. . .
1	x_{1m}	x_{2m}	x_{3m}	. . .	x_{nm}	y_m

Least Square method

- $e_i = y_i - \widehat{y_i} \ = y_i - \sum_j b_j \, x_{ji} \ \ for \ i = 1,2, \dots , m$
- ${e_i}^2 = (y_i - \widehat{y_i})^2 = (y_i - \sum_j b_j \, x_{ji})^2 \ for \ i = 1,2, \dots , m$
- SSE $= \sum_i {e_i}^2 = \sum_i (y_i - \sum_j b_j \, x_{ji})^2$
- Observations in n+1 dimensional space
- For least square fit, need to find b_j for $j = 0,1, \dots , n$ so that SSE is minimum. Denote SSE as S.
- Partial derivatives of S with respect to each of the parameters must be zero.

$\frac{\partial S}{\partial b_j} = 0$; for $j = 0,1, \dots , n$

Normal Equations

- $\frac{\partial S}{\partial b_j} = 0$; for $j = 0,1, \dots, n$
- $S = \sum_i (y_i - \sum_j b_j\, x_{ji})^2$;
- $\frac{\partial S}{\partial b_j} = \frac{\partial}{\partial b_j} \sum_i (y_i - \sum_j b_j\, x_{ji})^2$
- $\frac{\partial S}{\partial b_j} = \sum_i \frac{\partial}{\partial b_j} (y_i - \sum_j b_j\, x_{ji})^2$
- $\frac{\partial S}{\partial b_j} = 2 \sum_i \left(y_i - \sum_j b_j\, x_{ji}\right)\left(-x_{ji}\right) = 0$ for $j = 0,1, \dots, n$
- **$\sum_i \left(y_i - \sum_j b_j\, x_{ji}\right)\left(x_{ji}\right) = 0$ for $j = 0, 1, \dots, n$ are n+1 simultaneous linear equations in b_j's for $j = 0, 1, \dots, n$ called normal equations**
- For example for $j = 0$:
- $\sum_i y_i = b_0 \sum_i x_{0i} + b_1 \sum_i x_{1i} + b_2 \sum_i x_{2i} + \dots + b_n \sum_i x_{ni}$
-

Normal Equations

- 3.8 = 5 $b_0 + 40\ b_1 + 45\ b_2$
- 33 = 40 $b_0 + 338\ b_1 + 342\ b_2$
- 27.9 = 45 $b_0 + 342\ b_1 + 549\ b_2$
- Solve by elimination manually (H.W.)
- $\begin{pmatrix} 5 & 40 & 45 \\ 40 & 338 & 342 \\ 45 & 342 & 549 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 3.8 \\ 33 \\ 27.9 \end{pmatrix}$
- $\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 5 & 40 & 45 \\ 40 & 338 & 342 \\ 45 & 342 & 549 \end{pmatrix}^{-1} \begin{pmatrix} 3.8 \\ 33 \\ 27.9 \end{pmatrix}$

Matrix Method to derive normal Equations

- **X** = $\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}$ is m X (n+1) matrix; $\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}$ is m X 1 vector
- $\boldsymbol{b}$ = $\begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{pmatrix} \epsilon\, \mathcal{R}^{n+1}$ is (n+1) X 1 vector; $\hat{\boldsymbol{y}} = \begin{pmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \dots \\ \widehat{y_m} \end{pmatrix}$ is m X 1 vector ; $X^T\ is\ (n+1)x\ m$
- $X^T\ \boldsymbol{y}\ \boldsymbol{is}\ ((n+1)xm\ )$X (m X 1) giving (n+1) X 1 = $\begin{pmatrix} k_0 \\ k_1 \\ \dots \\ k_n \end{pmatrix}$ (say)
- $\boldsymbol{b^T} = (b_0, b_1, \dots, b_n)$ is 1 x (n+1) vector; $\boldsymbol{b^T} X^T\ \boldsymbol{y}$ = $b_0 k_0 + b_1 k_1 + b_2 k_2 + \dots + b_n k_n$

Predicted Values

- $b_0 x_{01} + b_1 x_{11} + b_2 x_{21} + \ldots + b_n x_{n1} = \widehat{y_1}$
- $b_0 x_{02} + b_1 x_{12} + b_2 x_{22} + \ldots + b_n x_{n2} = \widehat{y_2}$
- . . . .
- $b_0 x_{0i} + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_n x_{ni} = \widehat{y_i}$
- ………..
- $b_0 x_{0m} + b_1 x_{1m} + b_2 x_{2m} + \ldots + b_n x_{nm} = \widehat{y_m}$

- $b_0 x_{0i} + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_n x_{ni} = \widehat{y_i} \; fori = 1,2, \ldots, m$
- $\sum_j b_j \, x_{ji} = \widehat{y_i} \; fori = 1,2, \ldots, m; j = 0,1, \ldots, n$

- SSE $=\sum_i {e_i}^2$; $\boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_m \end{pmatrix}$; $\boldsymbol{e^T} = (e_1, e_2, \dots, e_m)$; $\boldsymbol{e^T}\ \boldsymbol{e} = (e_1, e_2, \dots, e_m) \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_m \end{pmatrix}$
- $\boldsymbol{e^T}\ \boldsymbol{e} = {e_1}^2 + {e_2}^2 + \cdots + {e_m}^2 = \sum_i {e_i}^2 = SSE$; $SSE = \boldsymbol{e^T}\ \boldsymbol{e}$
- $e_i = y_i - \widehat{y_i} = y_i - \sum_j b_j x_{ji}\ \ for\ i = 1,2, \dots, m$
- X = $\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}$; $\boldsymbol{b}$ = $\begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{pmatrix}$
- $X\boldsymbol{b} = \widehat{y_i}\ \ e_i = y_i - \widehat{y_i}$ *for i=1,2,…,m* ; $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$

- $S = \ (\boldsymbol{y} - X\boldsymbol{b})^T\ (\boldsymbol{y} - X\boldsymbol{b})$= $(\boldsymbol{y}^T\ - (X\boldsymbol{b})^T)\ (\boldsymbol{y} - X\boldsymbol{b})$
- S= $(\boldsymbol{y}^T\ - \boldsymbol{b}^T X^T)\ (\boldsymbol{y} - X\boldsymbol{b})\ = (\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{b} - \boldsymbol{b}^T X^T\ \boldsymbol{y} + \boldsymbol{b}^T X^T\ X\boldsymbol{b})$
- $\boldsymbol{y}^T X\boldsymbol{b}$ is a scalar $\big((1xm)X(mx(n+1)X((n+1)x1\big)$: 1x1
- $(\boldsymbol{y}^T X\boldsymbol{b})^T\ =\ \boldsymbol{b}^T X^T\ \boldsymbol{y}$ = $\boldsymbol{y}^T X\boldsymbol{b}$ being scalar
- $S = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{b}^T X^T\ \boldsymbol{y} + \boldsymbol{b}^T X^T\ X\boldsymbol{b}$
- $\boldsymbol{b}^T X^T\ \boldsymbol{y}$ = $b_0 k_0\ + b_1 k_1\ +\ b_2\, k_2 + \ldots + b_n k_n$; k’s are constant with respect to b’s
- $\dfrac{\partial \boldsymbol{b}^T X^T\ \boldsymbol{y}}{\partial b_j} =\ k_j\ ;$

- $Contribution\ of\ \boldsymbol{b}^T X^T\ X\boldsymbol{b}$ is $b_0^2 d_0 + b_1^2 d_1 + b_2^2 d_2 + \ldots + b_n^2 d_n$
- $\frac{\partial \boldsymbol{b}^T X^T\ X\boldsymbol{b}}{\partial b_j}$ = $2b_j d_j$
- $\frac{\partial S}{\partial b_j}$ = $-2\ k_j$ +$2b_j d_j$
- S is scalar and a function of $b_0, b_1, \ldots, b_n$
- $\nabla f(x, y, z) = (f_x, f_y, f_z)$ called Grad f , gradient of a function
- In order that S is minimum , necessary condition is ∇S = **0** (zero vetor)

- ∇S= $\frac{\partial S}{\partial \boldsymbol{b}} = \begin{pmatrix} \frac{\partial S}{\partial b_0} \\ \frac{\partial S}{\partial b_1} \\ \dots \\ \frac{\partial S}{\partial b_n} \end{pmatrix}$; $\frac{\partial S}{\partial \boldsymbol{b}} = \begin{pmatrix} -2\, k_0 + 2b_0\, d_0 \\ -2\, k_1 + 2b_1\, d_1 \\ \dots \\ -2\, k_n + 2b_n\, d_n \end{pmatrix}$ = **0**
- $\frac{\partial S}{\partial \boldsymbol{b}} = 0 - 2X^T\, \boldsymbol{y} + 2\, X^T\, X\boldsymbol{b} = 0 \Rightarrow\; X^T\, X\boldsymbol{b} - X^T\, \boldsymbol{y} = 0$
- $X^T\, X\boldsymbol{b} = X^T\, \boldsymbol{y}$ represent normal e
- $\boldsymbol{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^T\, \boldsymbol{y}$

- X = $\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}$; $X^T = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nm} \end{bmatrix}$

$$X^T\,\mathrm{X} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & x_{13} & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nm} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{n1} \\ 1 & x_{12} & x_{22} & \dots & x_{n2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1m} & x_{2m} & \dots & x_{nm} \end{bmatrix}$$

ŋ

The Matrix form

y_1
y_2
…
y_n

1	$x_{1,1}$	$x_{2,1}$	…	$x_{k,1}$
1	$x_{1,2}$	$x_{2,2}$	…	$x_{k,2}$
1	…	…	…	…
1	$x_{1,n}$	$x_{2,n}$	…	$x_{k,n}$

$$\beta = (X'X)^{-1} X'y$$

Comparison between analytical and numerical method

Analytical	Numerical
Normal Equations	Iterative: many iterations
Direct formula/Matrix Formula	Gradient Descent Method
No requirement of choosing learning rate	Need to choose learning rate appropriately
Does not require feature scaling	Requires feature scaling
Becomes slow , when n is large ; involves calculation of inverse of $(X^T X)^{-1}$;nxn matrix, O(n^3),upto n=10,000 fine, n$\geq$ $\mathbf{10000},$ gradient descent advisable	Works well when n is large O(n^2)