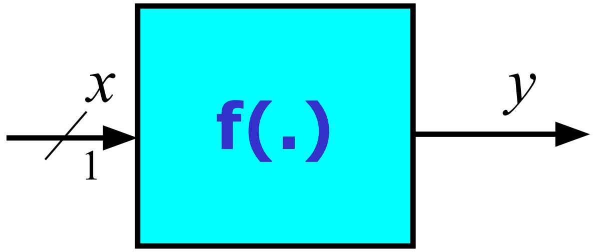
Supervised Machine Learning: Regression

Linear Regression

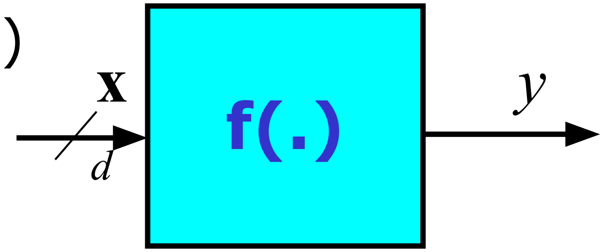
Linear Regression

- Linear approach to model the relationship between a scalar response, (y) (or dependent variable) and one or more predictor variables, (x *or* $\mathbf{x}$) (or independent variables)
- The output is going to be the linear function of input (one or more independent variables)

- Simple linear regression (straight-line regression):
  - Single independent variable (x)
  - Single dependent variable (y)
  - *Fitting a straight-line*



- Multiple linear regression:
  - two or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)
  - *Fitting a hyperplane (linear surface)*



Straight-Line (Simple Linear) Regression

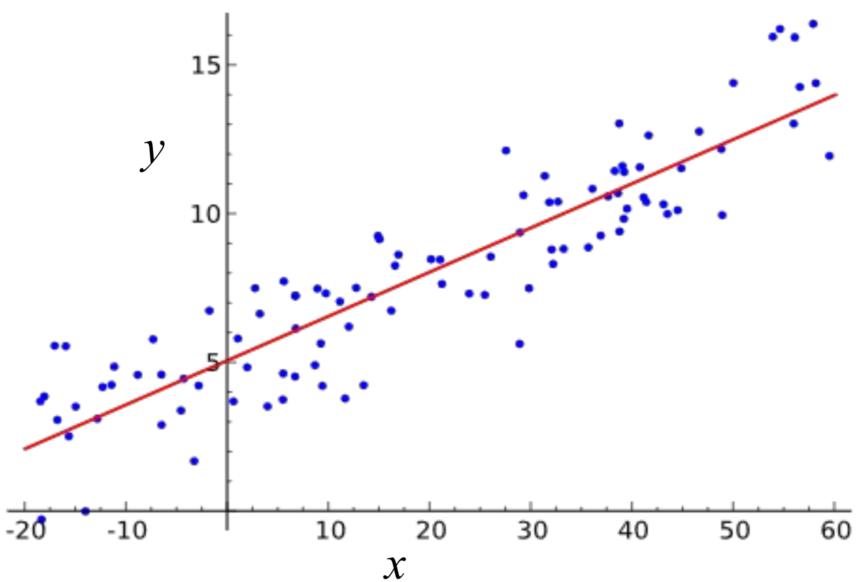
- Given:- Training data: $\mathrm{D} = \{x_n, y_n\}_{n=1}^{N}$, $x_n \in \mathrm{R}^1$ and $y_n \in \mathrm{R}^1$
  - x_n: n^{th} input example (independent variable)
  - y_n: Dependent variable (output) corresponding to n^{th} independent variable
- Example: Predicting the salary given the year of experience

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

- Independent variable:
  - Years of experience
- Dependent variable:
  - Salary

Straight-Line (Simple Linear) Regression

- Given:- Training data: $D = \{x_n, y_n\}_{n=1}^{N}, \; x_n \in R^1$ and $y_n \in R^1$
  - x_n: n^{th} input example (independent variable)
  - y_n: Dependent variable (output) corresponding to n^{th} independent variable
- Function governing the relationship between input and output: $y_n = f(x_n, w, w_0) = w x_n + w_0$
  - The coefficients w_0 and w are parameters of straight-line (regression coefficients) ***- Unknown***



- Function $f(x_n, w, w_0)$ is a linear function of x_n and it is a linear function of coefficients w and w_0
  - ***Linear model for regression***
- The values for the coefficients will be determined by fitting the linear function (straight-line) to the training data

Straight-Line (Simple Linear) Regression: Training Phase

- Given:- Training data: $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}, \; x_n \in \mathbb{R}^1$ and $y_n \in \mathbb{R}^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, w, w_0)$, in the training set for any given value of w and w_0

$$\hat{y}_n = f(x_n, w, w_0) = w\,x_n + w_0$$

$$(\hat{y}_n - y_n)^2 \qquad \forall n = 1, 2, \ldots, N$$

Straight-Line (Simple Linear) Regression: Training Phase

- Given:- Training data: $\mathrm{D} = \{x_n, y_n\}_{n=1}^{N},\ x_n \in \mathrm{R}^1$ and $y_n \in \mathrm{R}^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, w, w_0)$, in the training set for any given value of w and w_0

$$\hat{y}_n = f(x_n, w, w_0) = w\, x_n + w_0$$

$$E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{y}_n - y_n\right)^2$$

Straight-Line (Simple Linear) Regression: Training Phase

- Given:- Training data: $\mathrm{D} = \{x_n, y_n\}_{n=1}^{N},\ x_n \in \mathrm{R}^1$ and $y_n \in \mathrm{R}^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, w, w_0)$, in the training set for any given value of w and w_0

$$\hat{y}_n = f(x_n, w, w_0) = w\, x_n + w_0$$

$$E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(f(x_n, w, w_0) - y_n\right)^2$$

Straight-Line (Simple Linear) Regression: Training Phase

- Given:- Training data: $\mathrm{D} = \{x_n, y_n\}_{n=1}^{N}, \; x_n \in \mathrm{R}^1$ and $y_n \in \mathrm{R}^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, w, w_0)$, in the training set for any given value of w and w_0

$$\hat{y}_n = f(x_n, w, w_0) = w\, x_n + w_0$$

$$\underset{w, w_0}{\text{minimize}}\, E(w, w_0) = \frac{1}{2} \sum_{n=1}^{N} \left(f(x_n, w, w_0) - y_n \right)^2$$

- Minimize the error such that the coefficients w_0 and w represent the parameter of line that best fit the training data

Straight-Line (Simple Linear) Regression: Training Phase

- Given:- Training data: $D = \{x_n, y_n\}_{n=1}^{N}, \; x_n \in R^1$ and $y_n \in R^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, w, w_0)$, in the training set for any given value of w and w_0

$$\hat{y}_n = f(x_n, w, w_0) = w\,x_n + w_0$$

$$\underset{w, w_0}{\text{minimize}}\, E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{y}_n - y_n\right)^2$$

- The derivatives of error function with respect to the coefficients will be linear in the elements of w and w_0
- Hence the minimization of the error function has unique solution and found in closed form

Straight-Line (Simple Linear) Regression: Training Phase

- Cost function for optimization:

$$E(w, w_0) = \frac{1}{2}\sum_{n=1}^{N}\left(f(x_n, w, w_0) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial E(w, w_0)}{\partial w} = 0$ $\frac{\partial E(w, w_0)}{\partial w_0} = 0$

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(w x_n + w_0 - y_n\right)^2}{\partial w} = 0 \qquad \frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(w x_n + w_0 - y_n\right)^2}{\partial w_0} = 0$$

- Solving this give optimal $\hat{w}$ and $\hat{w}_0$ as

$$\boxed{\hat{w} = \frac{\sum_{n=1}^{N}(x_n - \mu_x)(y_n - \mu_y)}{\sum_{n=1}^{N}(x_n - \mu_x)^2}} \qquad \boxed{\hat{w}_0 = \mu_y - \hat{w}\mu_x}$$

- μ_x: sample mean of independent variable x
- μ_y: sample mean of dependent variable y

Straight-Line (Simple Linear) Regression: Testing Phase

- For any test example x, the predicted value is given by:

$$\hat{y} = f(x, \hat{w}, \hat{w}_0) = \hat{w}\, x + \hat{w}_0$$

  - For any $\hat{w}$ and $\hat{w}_0$ are the optimal parameters of the line learnt during training

Evaluation Metrics for Regression: Squared Error and Mean Squared Error

- The prediction accuracy is measured in terms of squared error: $\boxed{E = (\hat{y} - y)^2}$
  - y : actual value
  - $\hat{y}$: predicted value
- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error (RMSE):

$$E_{\text{RMSE}} = \sqrt{\frac{1}{N_t} \sum_{n=1}^{N_t} (\hat{y}_n - y_n)^2}$$

- RMSE expressed in % as:

$$E_{\text{RMSE}}(\%) = \frac{\sqrt{\frac{1}{N_t} \sum_{n=1}^{N_t} (\hat{y}_n - y_n)^2}}{\frac{1}{N_t} \sum_{n=1}^{N_t} y_n} * 100$$

Illustration of Simple Linear Regression: Salary Prediction - Training

Years of experience (x)	Salary (in Rs 1000) (y)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

$$\hat{w} = \frac{\sum_{n=1}^{N} (x_n - \mu_x)(y_n - \mu_y)}{\sum_{n=1}^{N} (x_n - \mu_x)^2} \qquad \hat{w}_0 = \mu_y - \hat{w}\mu_x$$

- μ_x: 9.1
- μ_y: 55.4
- $\hat{w}$: 3.54
- $\hat{w}_0$: 23.21

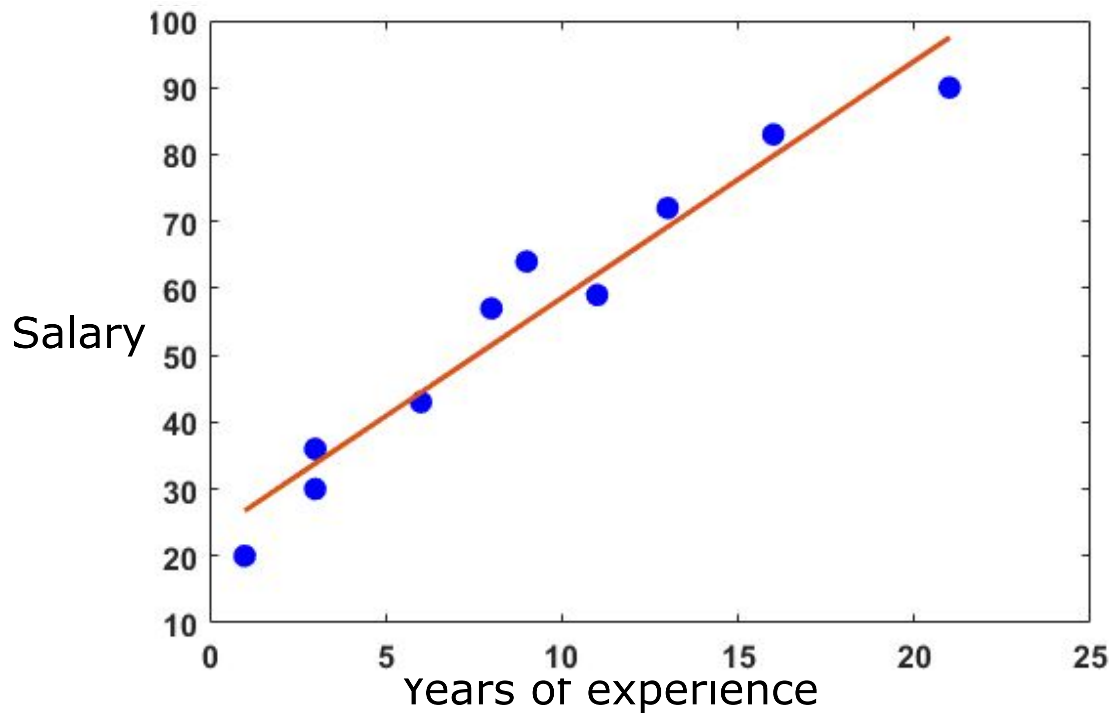
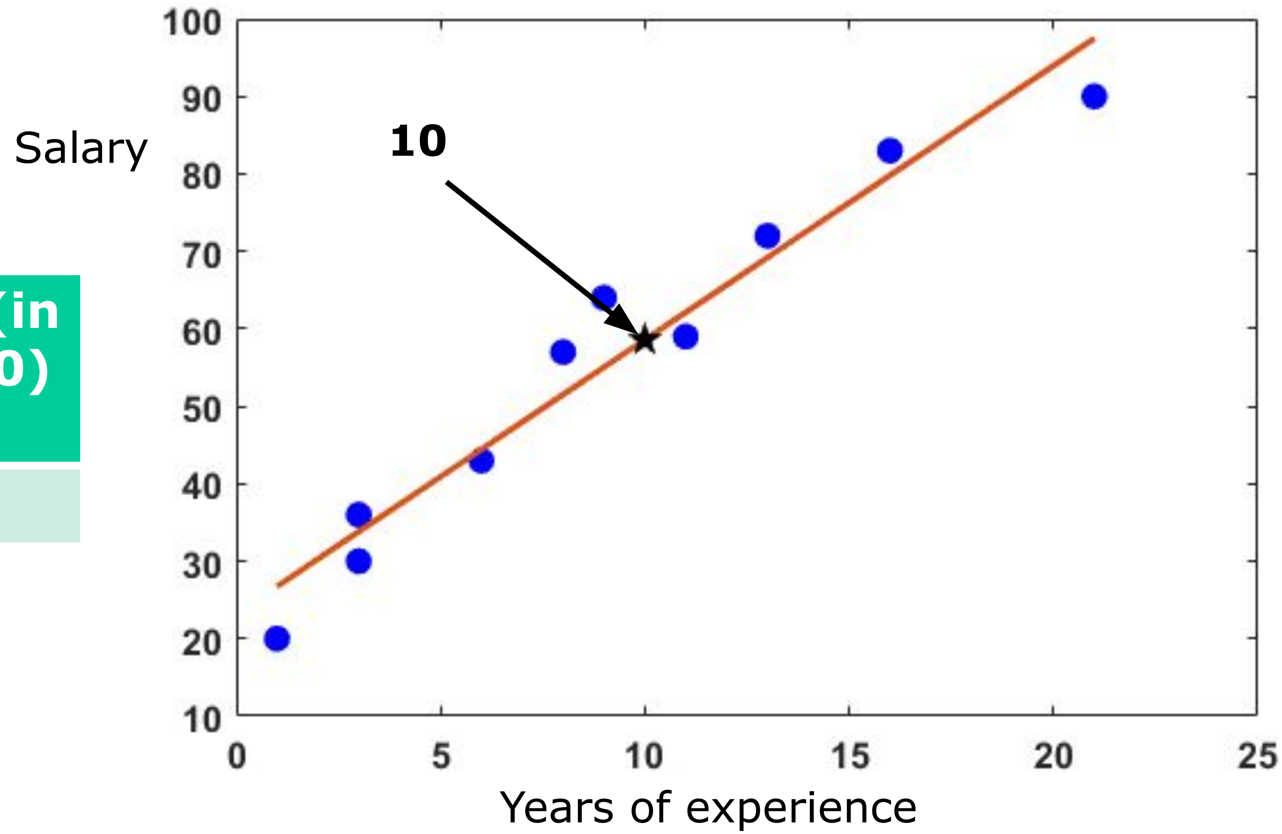


Illustration of Simple Linear Regression: Salary Prediction - Test

- $\hat{w}$: 3.54
- $\hat{w}_0$: 23.21

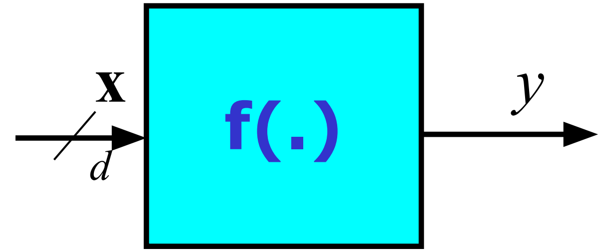
Years of experience (x)	Salary (in Rs 1000) (y)
10	-



- Predicted salary: 58.584
- Actual salary: 58.000
- Squared error: **0.34**

Multiple Linear Regression

- Multiple linear regression:
  - Two or more independent variable ($\mathbf{x}$)
  - Single dependent variable (y)



- Given:- Training data: $D = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, $\mathbf{x}_n \in R^d$ and $y_n \in R^1$
  - d: dimension of input example (number of independent variables)
  - $\mathbf{x}_n$: n^{th} input example (d independent variables)
  - y_n: Dependent variable (output) corresponding to n^{th} input example
- Function governing the relationship between input and output:

$$y_n = f(\mathbf{x}_n, \mathbf{w}) = w_d x_{nd} + \ldots. + w_2 x_{n2} + w_1 x_{n1} + w_0 = \sum_{i=0}^{d} w_i x_{ni} = \mathbf{w}^\mathsf{T} \mathbf{x}_n$$

  - The coefficients w_0, w_1, … , w_d are collectively denoted by the vector $\mathbf{w}$ *- Unknown*
- Function $f(\mathbf{x}_n, \mathbf{w})$ is a linear function of $\mathbf{x}_n$ and it is a linear function of coefficients $\mathbf{w}$
  - ***Linear model for regression***

Linear Regression: Linear Function Approximation

- Linear function:
  - 2 input variable case (3-dimensional space): The mapping function is a **plane** specified by

$$y = f(\mathbf{x}, \mathbf{w}) = w_2 x_2 + w_1 x_1 + w_0 = 0$$

$$\text{where } \mathbf{w} = [w_0, w_1, w_2]^{\mathsf{T}} \text{ and } \mathbf{x} = [1, x_1, x_2]^{\mathsf{T}}$$

  - d input variable case (d+1–dimensional space): The mapping function is a **hyperplane** specified by

$$y = f(\mathbf{x}, \mathbf{w}) = w_d x_d + \ldots. + w_2 x_2 + w_1 x_1 + w_0 = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^{\mathsf{T}} \mathbf{x} = 0$$

$$\text{where } \mathbf{w} = [w_0, w_1, \ldots, w_d]^{\mathsf{T}} \text{ and } \mathbf{x} = [1, x_1, \ldots, x_d]^{\mathsf{T}}$$

Multiple Linear Regression: Training Phase

- The values for the coefficients will be determined by fitting the linear function to the training data
- Given:- Training data: D $=\{\mathbf{x}_n, y_n\}_{n=1}^{N}$, $\mathbf{x}_n \in \mathrm{R}^d$ and $y_n \in \mathrm{R}^1$
- **Method of least squares**: Minimizes the sum of the squared error between
  - all the actual data (y_n) i.e. actual dependent variable and
  - the estimate of line (predicted dependent variable ($\hat{y}_n$)) i.e. the function $f(x_n, \mathbf{w})$, in the training set for any given value of $\mathbf{w}$

$$\hat{y}_n = f(\mathbf{x}_n, \mathbf{w}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}_n + w_0 = \sum_{i=0}^{d} w_i x_i$$

$$\underset{\mathbf{w}}{\text{minimize}}\, E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(\hat{y}_n - y_n\right)^2$$

- The error function is a
  - quadratic function of the coefficients $\mathbf{w}$ and
  - The derivatives of error function with respect to the coefficients will be linear in the elements of $\mathbf{w}$
- Hence the minimization of the error function has unique solution and found in closed form

Multiple Linear Regression: Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{x}_n, \mathbf{w}) - y_n\right)^2$$

- Conditions for optimality: $\dfrac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$

- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\sum_{i=0}^{d} w_i x_{ni} - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

Multiple Linear Regression: Training Phase

- Cost function for optimization:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(f(\mathbf{x}_n, \mathbf{w}) - y_n\right)^2$$

- Conditions for optimality: $\frac{\partial E(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}$

- Application of optimality conditions gives optimal $\hat{\mathbf{w}}$:

$$\frac{\partial \frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{w}^\top \mathbf{x}_n - y_n\right)^2}{\partial \mathbf{w}} = \mathbf{0}$$

$$\boxed{\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}}$$

  – Assumption: $d < N$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1d} \\ 1 & x_{21} & x_{22} & \ldots & x_{2d} \\ - & - & - & - & - \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nd} \\ - & - & - & - & - \\ 1 & x_{N1} & x_{N2} & \ldots & x_{Nd} \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ - \\ y_n \\ - \\ y_N \end{bmatrix}$$

$\mathbf{X}$ is data matrix

Multiple Linear Regression: Testing Phase

- Optimal coefficient vector $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\hat{\mathbf{w}} = \mathbf{X}^{+} \mathbf{y}$$

where $\mathbf{X}^{+} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ is the pseudo inverse of matrix $\mathbf{X}$

- For any test example $\mathbf{x}$, the predicted value is given by:

$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^\top \mathbf{x} = \sum_{i=0}^{d} \hat{w}_i x_i$$

- The prediction accuracy is measured in terms of squared error: $E = \left(\hat{y} - y\right)^2$
- Let N_t be the total number of test samples
- The prediction accuracy of regression model is measured in terms of root mean squared error:

$$E_{\text{RMS}} = \sqrt{\frac{1}{N_t} \sum_{n=1}^{N_t} \left(\hat{y}_n - y_n\right)^2}$$

Illustration of Multiple Linear Regression: Temperature Prediction

Humidity (x_1)	Pressure (x_2)	Temp (y)
82.19	1036.35	25.47
83.15	1037.60	26.19
85.34	1037.89	25.17
87.69	1036.86	24.30
87.65	1027.83	24.07
95.95	1006.92	21.21
96.17	1006.57	23.49
98.59	1009.42	21.79
88.33	991.65	25.09
90.43	1009.66	25.39
94.54	1009.27	23.89
99.00	1009.80	22.51
98.00	1009.90	22.90
99.00	996.29	21.72
98.97	800.00	23.18

- Training:

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y}$$

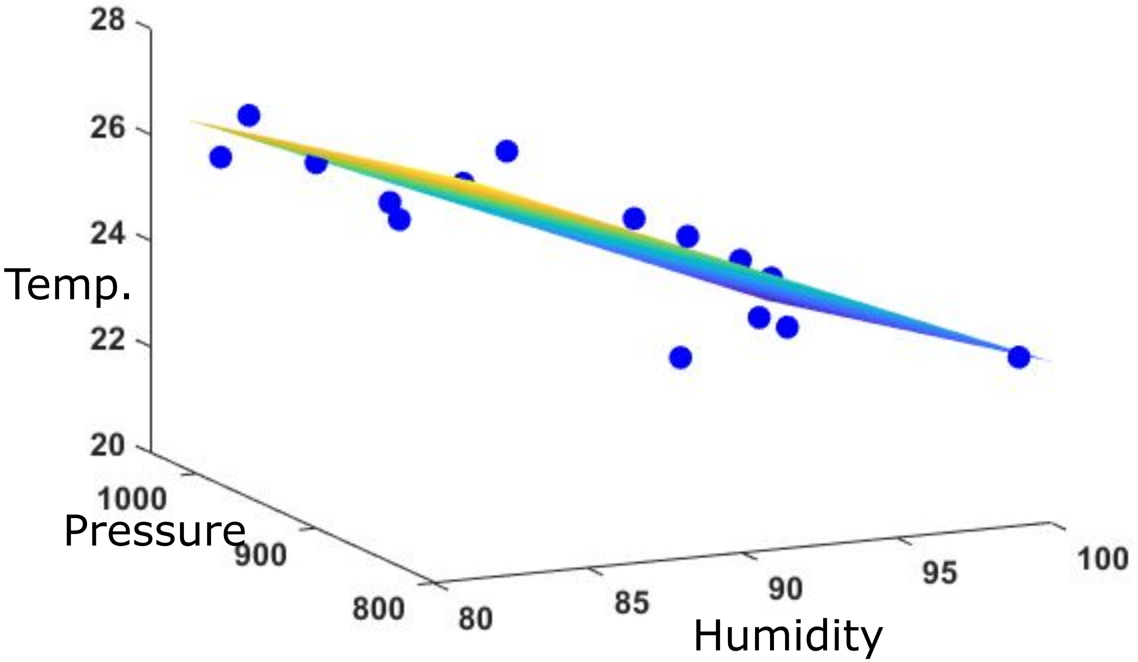
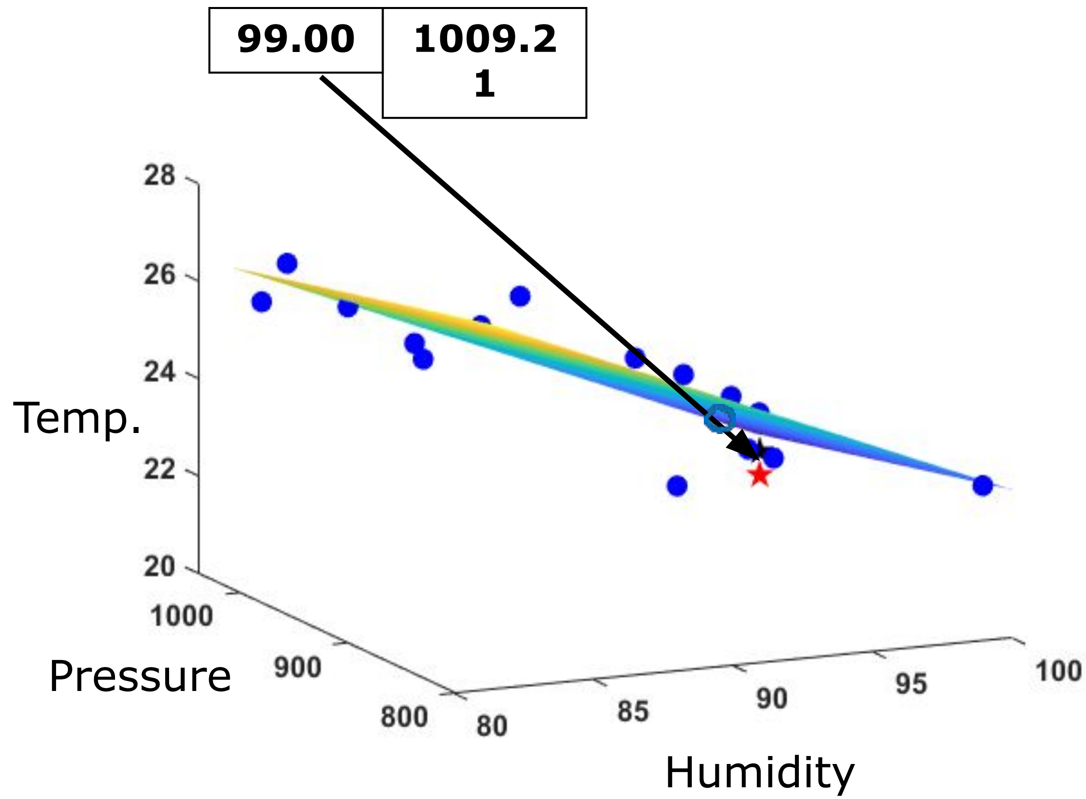


Illustration of Multiple Linear Regression: Temperature Prediction - Test

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{y}$$

Humidity (x_1)	Pressure (x_2)	Temp (y)
99.00	1009.21	-

$$y = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\top}\mathbf{x}$$

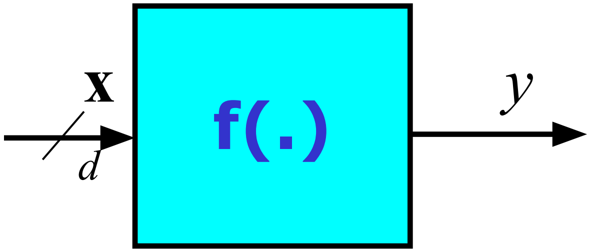


- Predicted temperature: 21.72
- Actual temperature: 21.24
- Squared error: **0.2347**

Application of Regression: A Method to Handle Missing Values

- Use most probable value to fill the missing value:
  - Use regression techniques to predict the missing value (regression imputation)
    - Let x_1, x_2, …, x_d be a set of d attributes
    - Regression (multivariate): The n^{th} value is predicted as

$$y_n = f(x_{n1}, x_{n2}, \ldots, x_{nd})$$



    - Simple or Multiple Linear regression:

$$y_n = w_1 x_{n1} + w_2 x_{n2} + \ldots + w_d x_{nd}$$

    - Popular strategy
    - It uses the most information from the present data to predict the missing values
    - It preserves the relationship with other variables

Application of Regression: A Method to Handle Missing Values

- Training process:
  - Let y be the attribute, whose missing values to be predicted
  - Training examples: All $\mathbf{x}=[x_1, x_2, \ldots, x_d]^T$, a set of d dependent attributes for which the independent variable y is available
  - The values for the coefficients will be determined by fitting the linear function to the training data

1	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	NaN	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	NaN	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	NaN	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	NaN	99	513
16	25-07-2018	23.18182	98.9697	195.75
17	[illegible]	[illegible]	[illegible]	[illegible]

- Dependent variable: Temperature
- Independent variables: Humidity and Rainfall

Application of Regression: A Method to Handle Missing Values

- Testing process (Prediction):
  - Optimal coefficient vector $\mathbf{w}$ is given by

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$$

  - For any test example $\mathbf{x}$, the predicted value is given by:

$$\hat{y} = f(\mathbf{x}, \hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathsf{T}}\mathbf{x} = \sum_{i=0}^{d} \hat{w}_i x_i$$

1	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	NaN	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	NaN	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	NaN	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	NaN	99	513
16	25-07-2018	23.18182	98.9697	195.75

1	Dates	Temperature	Humidity	Rain
2	08-07-2018	25.46875	82.1875	6.75
3	09-07-2018	26.19298	83.1491	1761.75
4	10-07-2018	25.17021	85.3404	652.5
5	11-07-2018	**24.2**	87.6866	963
6	12-07-2018	24.06923	87.6462	254.25
7	13-07-2018	21.20779	95.9481	339.75
8	15-07-2018	23.48571	96.1714	38.25
9	18-07-2018	**21.5**	98.5897	29.25
10	19-07-2018	25.09346	88.3271	4.5
11	20-07-2018	25.39423	90.4327	112.5
12	21-07-2018	**23.7**	94.5378	735.75
13	22-07-2018	22.5098	99	607.5
14	23-07-2018	22.904	98	717.75
15	24-07-2018	**21.6**	99	513
16	25-07-2018	23.18182	98.9697	195.75

Summary: Regression

- Regression analysis is used to model the relationship between one or more independent (predictor) variable and a dependent (response) variable
- Response is some function of one or more input variables
- Linear regression: Response is linear function of one or more input variables
  - If the response is linear function of one input variable, then it is simple linear regression (straight-line fitting)
  - If the response is linear function of two or more input variable, then it is multiple linear regression (linear surface fitting or hyperplane fitting)

Text Books

1. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Third Edition, Morgan Kaufmann Publishers, 2011.

2. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.