Machine Learning
Regression

DR. BHARGAVI R

SCOPE

VIT CHENNAI

Regression - Fundamentals

Regression analysis enables us to -

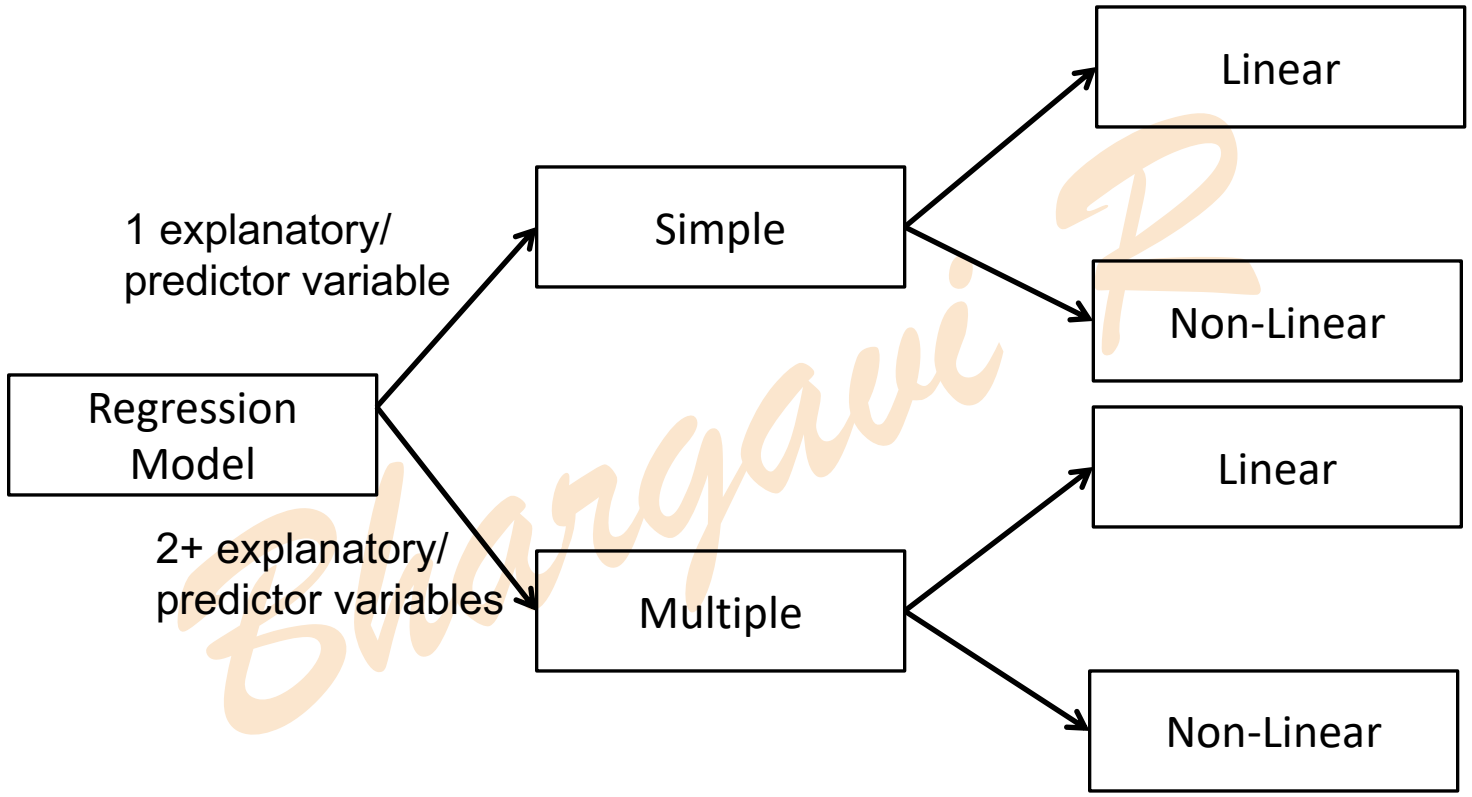
- Develop a mathematical function to predict the values of **Output/Dependent/Target/Response** variable, based on the value of **Input/Independent/ Predictor** variable(s).
- Quantify the effect of changes in the independent variable have on the dependent variable.
- Identify unusual observations.

Fundamentals (cont…)

Applications

- Sales prediction based on size of the shop
- Financial Forecasting: Predicting stock prices, interest rates, or economic indicators.
- Disease Progression: Predicting the progression of diseases based on patient data and medical history.
- Drug Response: Modeling the effectiveness of different drug dosages on patient outcomes.
- Quality Control: Analyzing factors that affect product quality in manufacturing processes.
- Signal Processing: Modeling and predicting signals in communication systems.

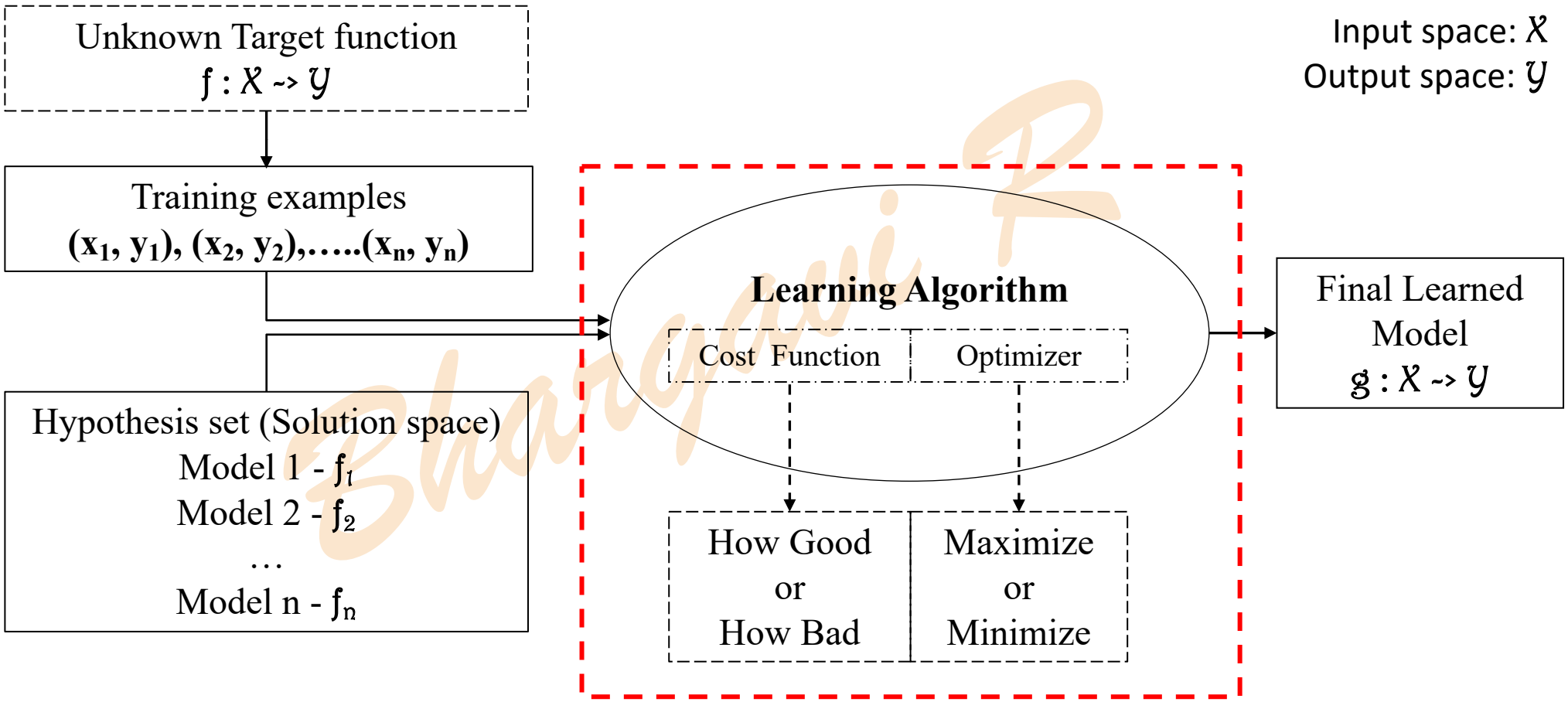
Regression Models - Types



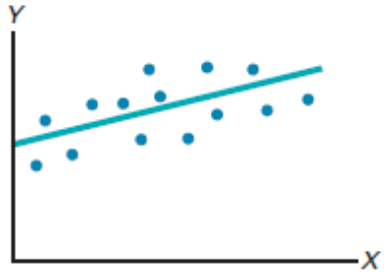
Simple Linear Regression

- Simple and straightforward.
- Predicts a ***quantitative response*** y based on ***single predictor*** variable x
- Assumes a linear relationship between x and y
- So **Hypothesis function** h(x) is $y = h(x) = w_0 + w_1 x$
- w_0 is the **intercept** and w_1 is the **slope**
- w_0 and w_1 are called the **linear regression coefficients** or **model coefficients** or **parameters** of the model
- Different values of w's result in different hypotheses.

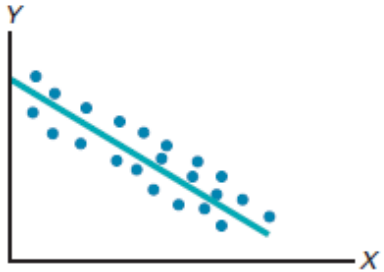
Machine Learning - Process



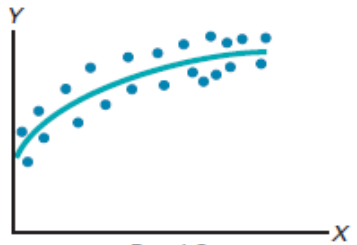
Relationship Between Variables



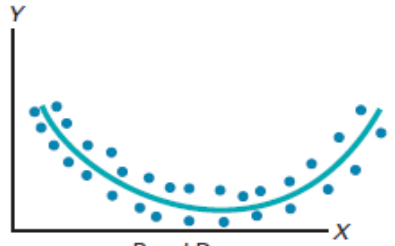
Panel A
Positive linear relationship



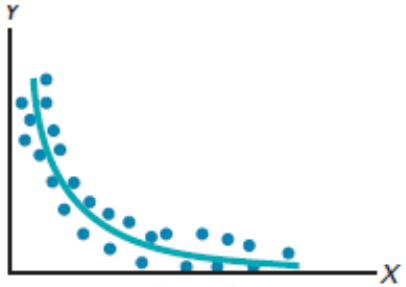
Panel B
Negative linear relationship



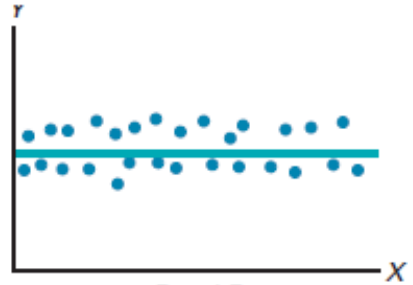
Panel C
Positive curvilinear relationship



Panel D
U-shaped curvilinear relationship



Panel E
Negative curvilinear relationship



Panel F
No relationship between *X* and *Y*

A Simple Use Case

"MyStyle" is a chain of apparel show rooms. Director of planning wants to forecast annual sales for all new stores for strategic planning, based on store size.

To examine the relationship between the store size in square feet and its annual sales, data is collected from already established stores.

Size of show room (Thousands of sq.ft)	Sales (millions of $)	Size of show room (Thousands of sq.ft)	Sales (millions of $)
1.7	3.7	1.1	2.7
1.6	3.9	3.2	5.5
2.8	7.7	1.5	2.9
5.6	9.5	5.2	10.7
1.3	3.4	4.6	7.6
2.2	5.6	5.8	11.8
1.3	3.7	3	4.1

Data

Showroom 1 (x_1 sq.ft, y_1 \$) (1.7 sq.ft , 3.7\$)

Showroom 2 (x_2 sq.ft, y_2 \$) (1.6 sq.ft , 3.9\$)

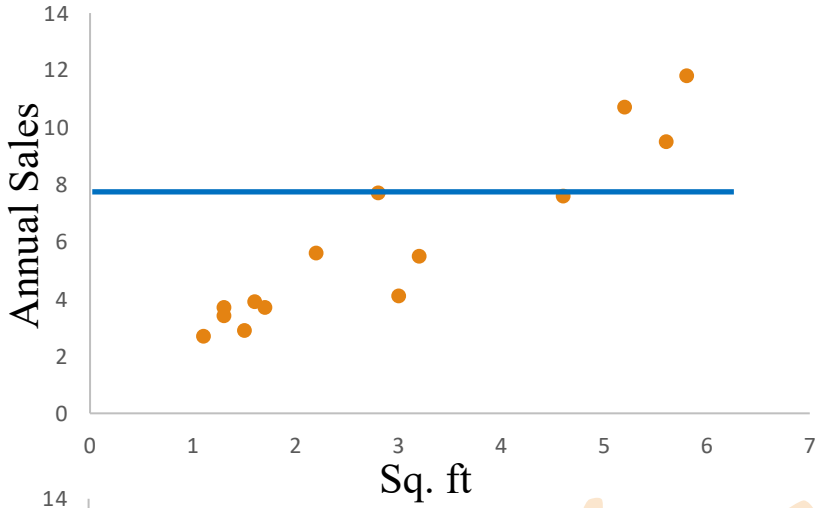
Showroom 3 (x_3 sq.ft, y_3 \$) (2.8 sq.ft , 7.7\$)

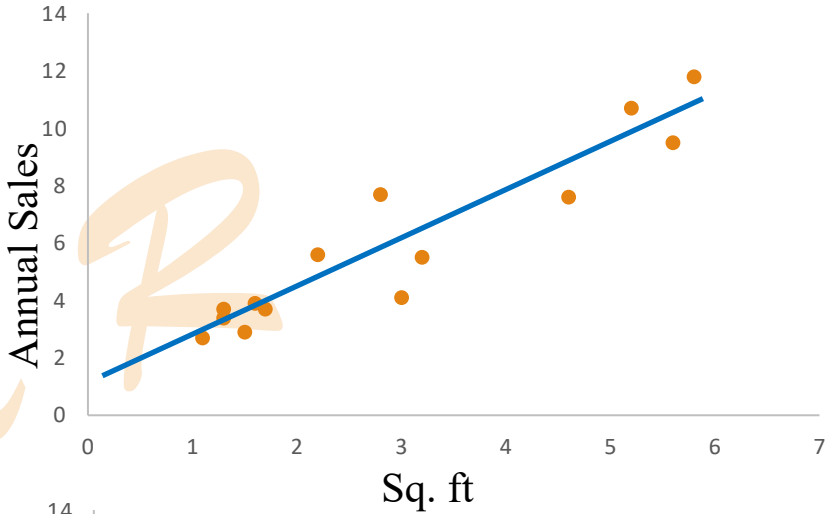
--

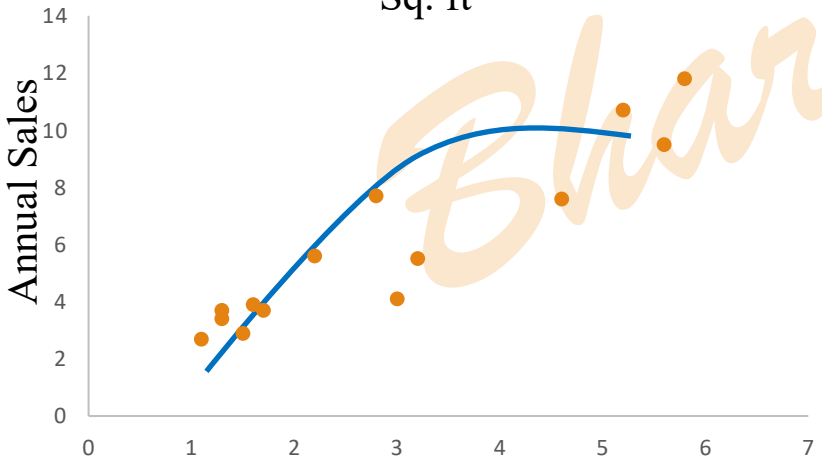
--

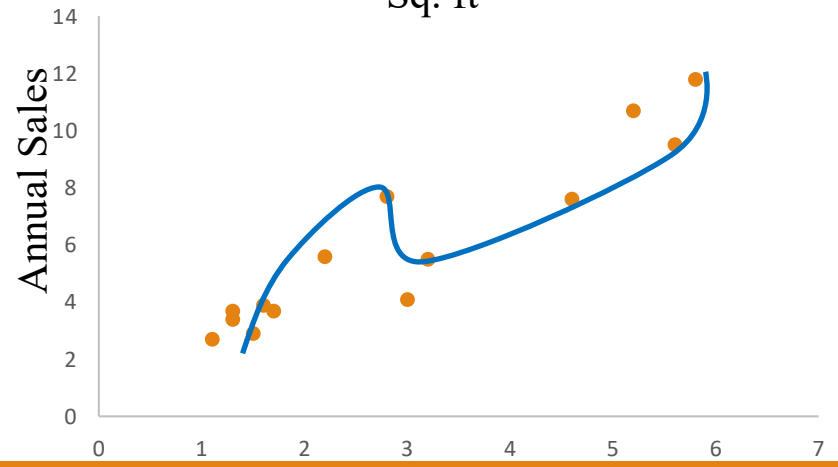
Showroom n (x_n sq.ft, y_n \$) (3 sq.ft , 4.1\$)

Which Model

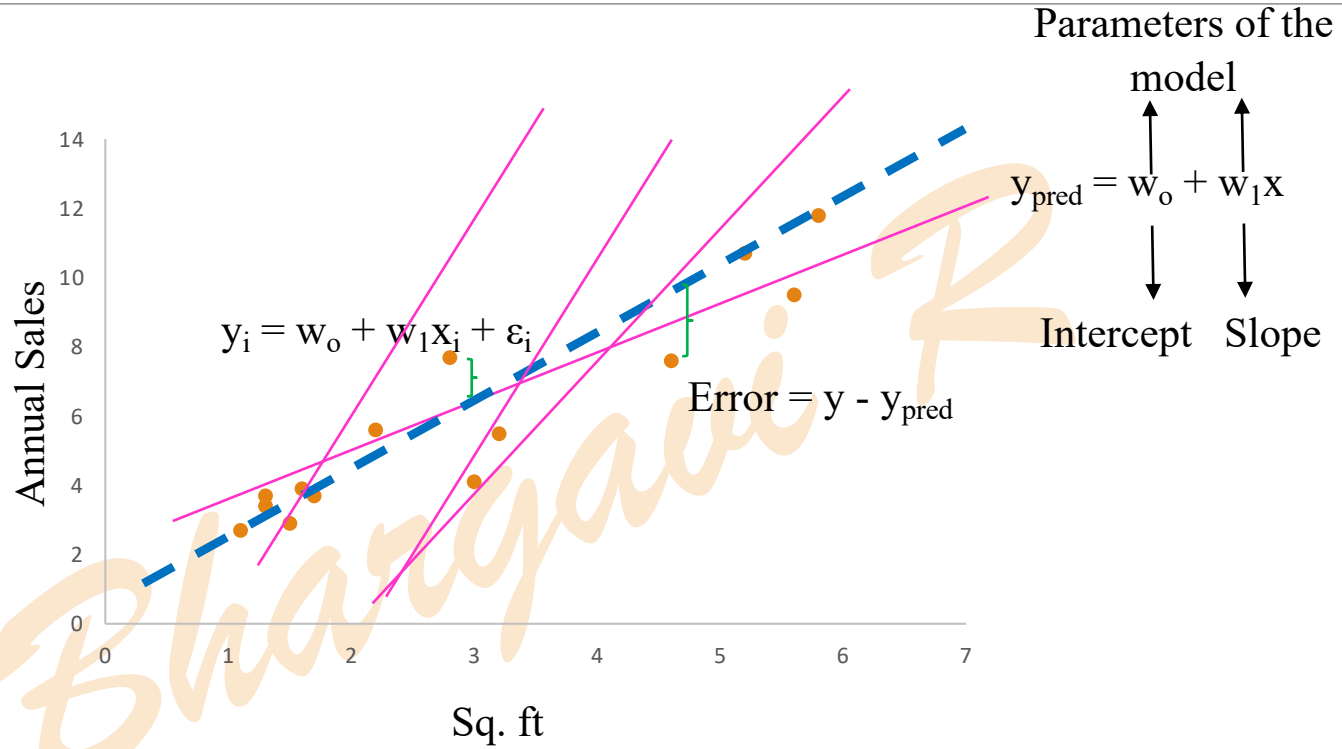








Estimate Best Function



Cost function

- RSS – Residual Sum of Squares
- RSS = (annual sales$_1$ - (w_0+ $w_1$sq. ft$_1$))^2

  + (annual sales$_2$ - (w_0+ $w_1$sq. ft$_2$))^2

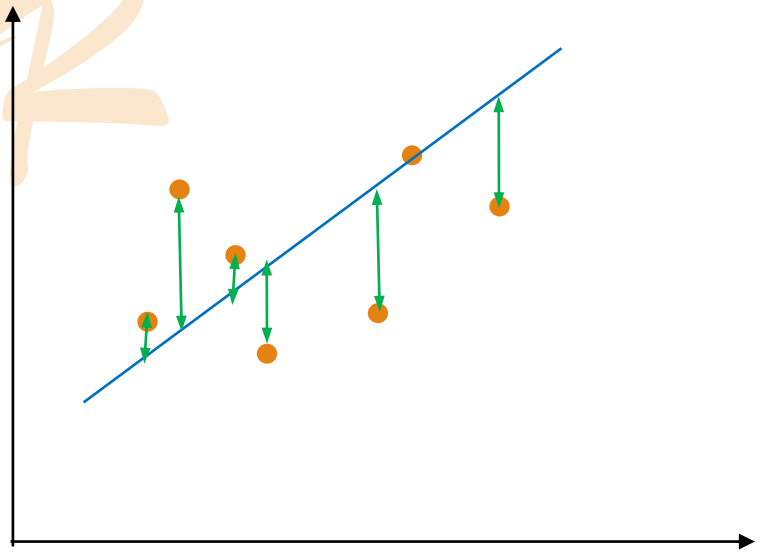
  + (annual sales$_3$ - (w_0+ $w_1$sq. ft$_3$))^2

  + --- + (annual sales$_n$ - (w_0+ $w_1$sq. ftn))^2

RSS = $\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$

RSS = $\sum_{i=1}^{n}(y_i - h(x_i))^2$

Cost function is $\sum_{i=1}^{n}(y_i - h(x_i))^2$



Objective Function

- Choose w_0 and w_1 such that y_{pred} is close to y for the training data.
- Minimize the Residual sum square (RSS)
- Minimize cost
- Objective function is $\min_{w_0\, w_1} \sum_{i=1}^{n} (y_i - h(x_i))^2$

Or $\min_{w_0\, w_1} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$

Prediction

- Model parameters are obtained by minimizing the cost function.
- For any given test input x_t, value of y can be computed by substituting the values of x_t , $\hat{w}_0$ and $\hat{w}_1$ in the fitted regression line.
- let the estimated parameters be $\hat{w}_0 = 0.9645$, $\hat{w}_1 = 1.6699$. Then the annual sales of a show room of size 4000 sq,ft can be predicted as

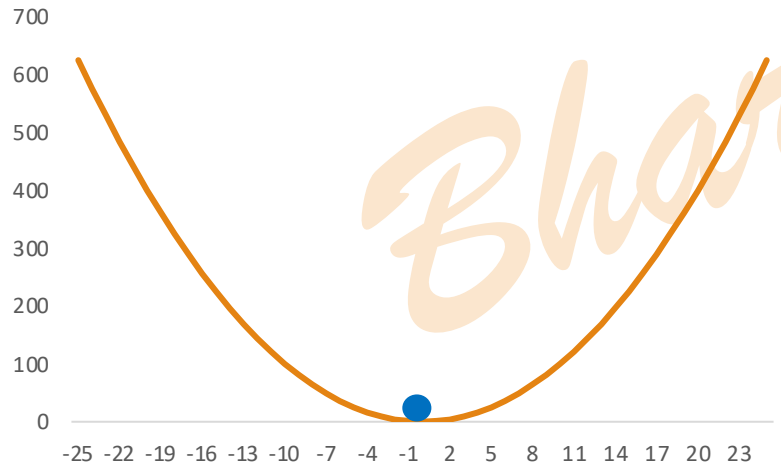
$$y_{pred} \text{ or } \hat{y} = 0.9645 + 1.6699 * 4 = 7.644 \text{ million \$}$$

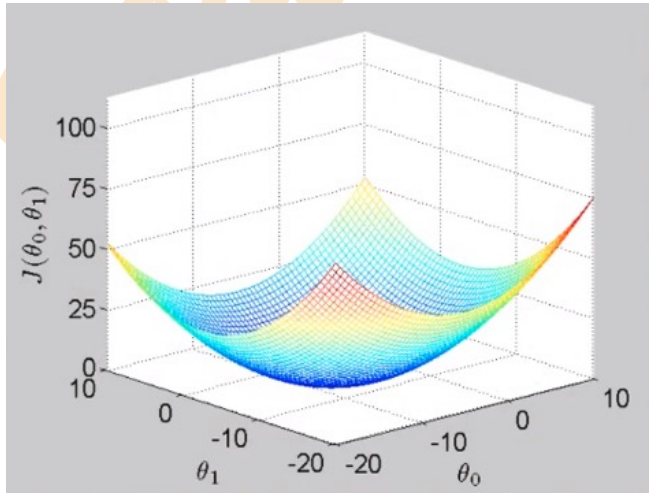
Interpretation of Predicted coefficients

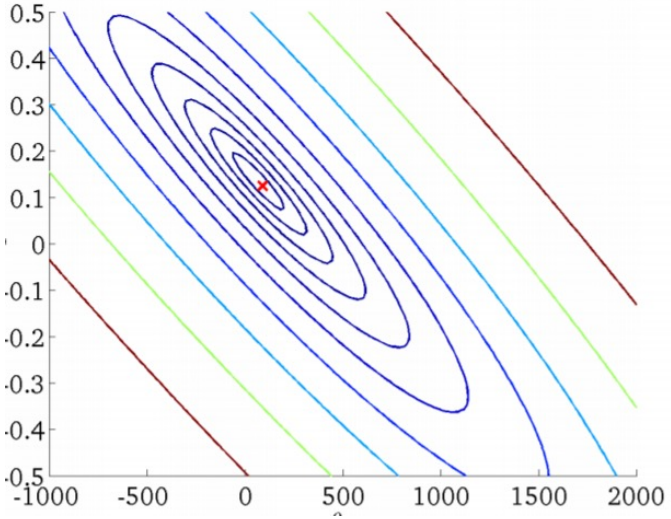
- Slope of the regression line i.e $\hat{w}_1$ can be interpreted as – with increase of 1 unit in x, the predicted value of y is estimated to increase by 1.6699 units
- Negative coefficients indicate the negative relation among the Predictor and Response variables.
- The intercept represents the predicted value of y when x equals 0

Optimizing the Cost Function

- RSS is a convex function
- Optimization problem = minimizing(RSS)
- Rate of change of RSS becomes/approaches zero at min(RSS)
- Rate of change of RSS is nothing but Gradient of RSS w.r.t parameters.







Optimizing the Cost Function (cont…)

- We can find solution (i.e w's of line with minimum cost) in different approaches.
  1. Setting Gradient(RSS) = 0 and solving (closed form solution)
  2. Gradient descent (Iterative approach)

Find Gradient

RSS = $\sum_{i=1}^{n}\left(y_i - (h(x_i)\right)^2 = \sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$

Gradient(RSS) = $\nabla\, RSS(w_0, w_1) = \frac{\partial}{\partial \boldsymbol{w}}(RSS)$

$\frac{\partial}{\partial w_0}(\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2) = \sum_{i=1}^{n}\frac{\partial}{\partial w_0}((y_i - (w_0 + w_1 x_i))^2)$

$= -2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)]$

$\frac{\partial}{\partial w_1}(\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2) = -2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)]x_i$

$$\nabla\, RSS(w_0, w_1) = \begin{matrix} -2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)] \\ -2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)]x_i \end{matrix}$$

Approach 1

Setting Gradient of RSS to zero

$\nabla RSS(w_0, w_1) = 0$

$-2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)] = 0 \quad (1)$

$-2\sum_{i=1}^{n}[y_i - (w_0 + w_1 x_i)]x_i = 0 \ (2)$

From (1) on solving for w_0, we get $w_0 = \bar{y} - w_1\bar{x}$

And Solving (2) after substituting the values of w_0, we get, $w_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}$

Or equivalently $w_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n}}$

Closed form Normal solution $\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{y}$

Approach 2
A Simple Prediction Workout

- We know that 1 km = 0.62137 * 1 Mile (Ground truth)
- Can we learn this formula by looking at some data that we have in our hand?
- Let's see
- Assume that we have the following data (Training data)

Sl.No	Kms	Miles
1	100	62.137
2	0	0

- In practice we have many more observations

All We Know….

- Miles = f(kms)

- Data (Observations)
- There exists a linear relationship between input and output

i.e miles = C * kms ------ (1)

Predict, Compare, Learn

- First assume that C = 0.5 . Substituting in (1)

  100 kms = 50 Miles (≠ 62.137)

  62.137 – 50 = 12.137 → Error

- Let's now increase C to 0.6

  100 kms = 60 Miles (≠ 62.137)

  Error = 2.137 (Better than Previous guess)

- Further increase C to 0.7

  100 kms = 70 Miles (≠ 62.137)

  Error = -7.863

Predict, Compare, Learn

- Let's now make C to 0.61

  100 kms = 61 Miles (≠ 62.137)

  Error = 1.137

  ------- and so on till we are okay with the error or till no further change in the error.

Gradient Descent (cont..)

While !(converge){

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla \boldsymbol{w}_t$$

$$t = t + 1$$

}

∇w is the partial derivative of the cost function w.r.t w i.e $\nabla \boldsymbol{w} = \frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial w}$

η is a small constant called as Learning rate

Gradient Descent

Iterative approach

While not converged update

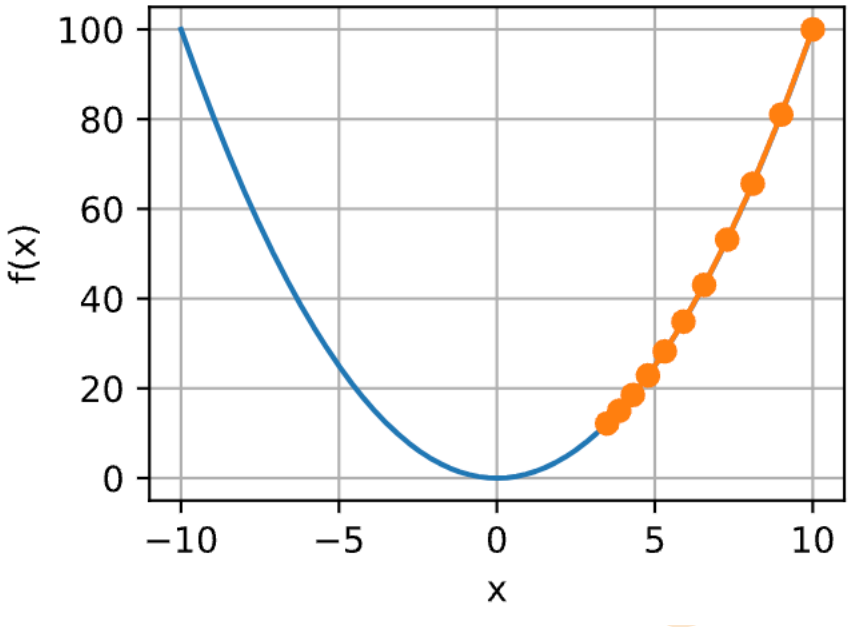
$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} - \eta \nabla RSS(\boldsymbol{w})^{t}$$

$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t - \eta(-2\sum_{i=1}^{n}[y_i - (w_0^t + w_1^t x_i)]) \\ w_1^t - \eta(-2\sum_{i=1}^{n}[y_i - (w_0^t + w_1^t x_i)]x_i) \end{bmatrix}$$

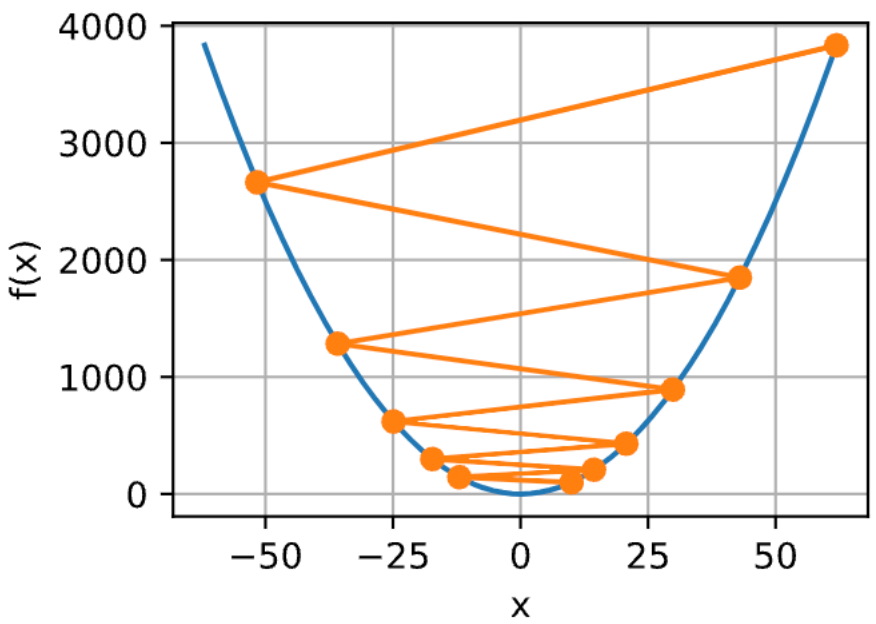
$$\begin{bmatrix} w_0^{t+1} \\ w_1^{t+1} \end{bmatrix} = \begin{bmatrix} w_0^t \\ w_1^t \end{bmatrix} + 2\eta \begin{bmatrix} \sum_{i=1}^{n}[y_i - (w_0^t + w_1^t x_i)] \\ \sum_{i=1}^{n}[y_i - (w_0^t + w_1^t x_i)]x_i \end{bmatrix}$$

Learning Rate Impact on Solution Convergence

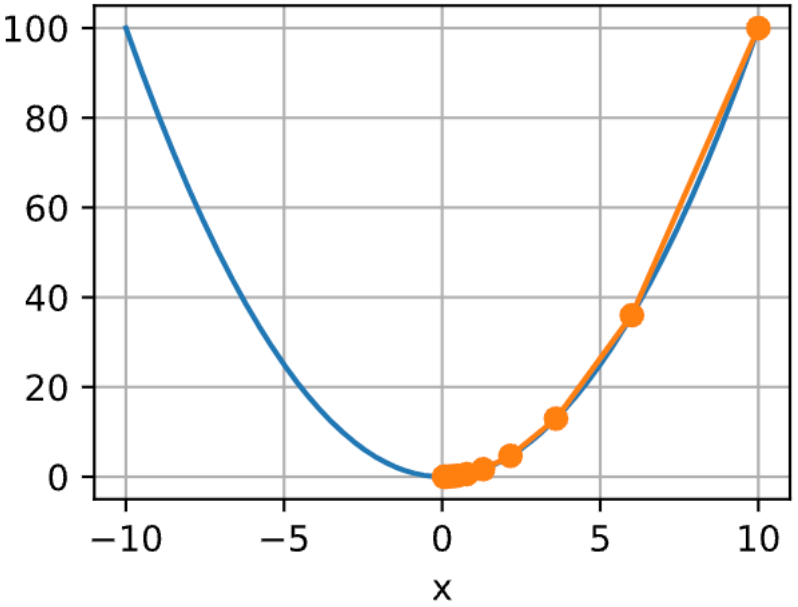
η = 0.05



η = 1.1



η = 0.2



Prediction Using More Than One Predictor

What should be the selling price of a used car based on the following features

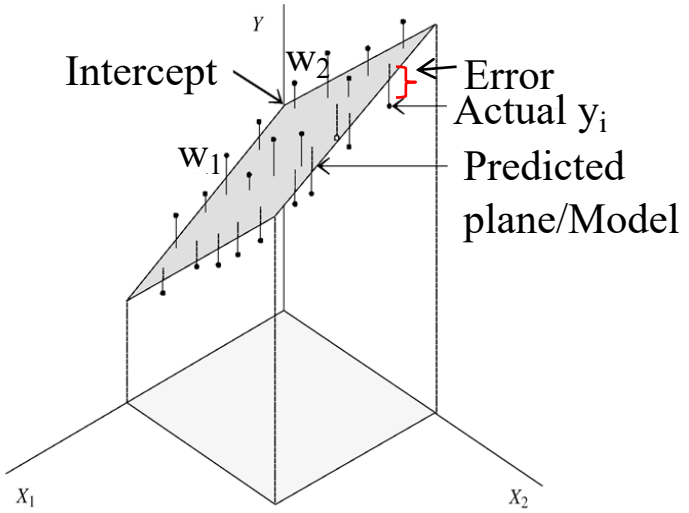
- Color
- Age
- Model
- Mileage
- Distance travelled

Multiple Linear Regression

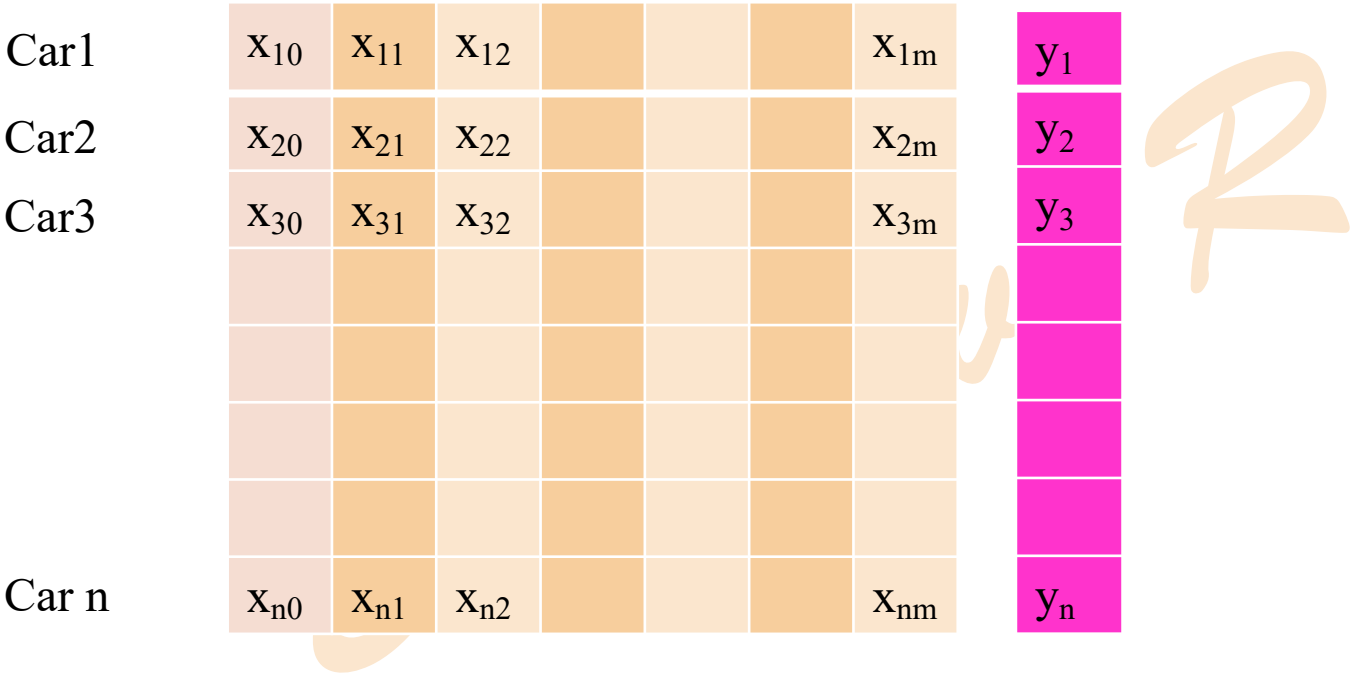
Hypothesis function $h(x) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m$

A Multiple Regression model with n predictor variables fits a regression plane in *(n+1)* dimensional space.

Ex: A multiple regression model with two explanatory variables fits a regression plane in *3*-dimensional space



Data


Car1
Car2
Car3
Car n
x_{10} x_{11} x_{12} x_{1m}
x_{20} x_{21} x_{22} x_{2m}
x_{30} x_{31} x_{32} x_{3m}
x_{n0} x_{n1} x_{n2} x_{nm}
y_1
y_2
y_3
y_n

MLR Predicted Model and Interpretation

$y = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m$ where $x_0 = 1$

- The intercept represents the predicted value of y when all x are set to 0.
- Slope for input feature x_i i.e w_i predicts the change in y per unit x_i holding all other input features constant.

Single Observation -Vector Notation

$$y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im} + \epsilon_i$$

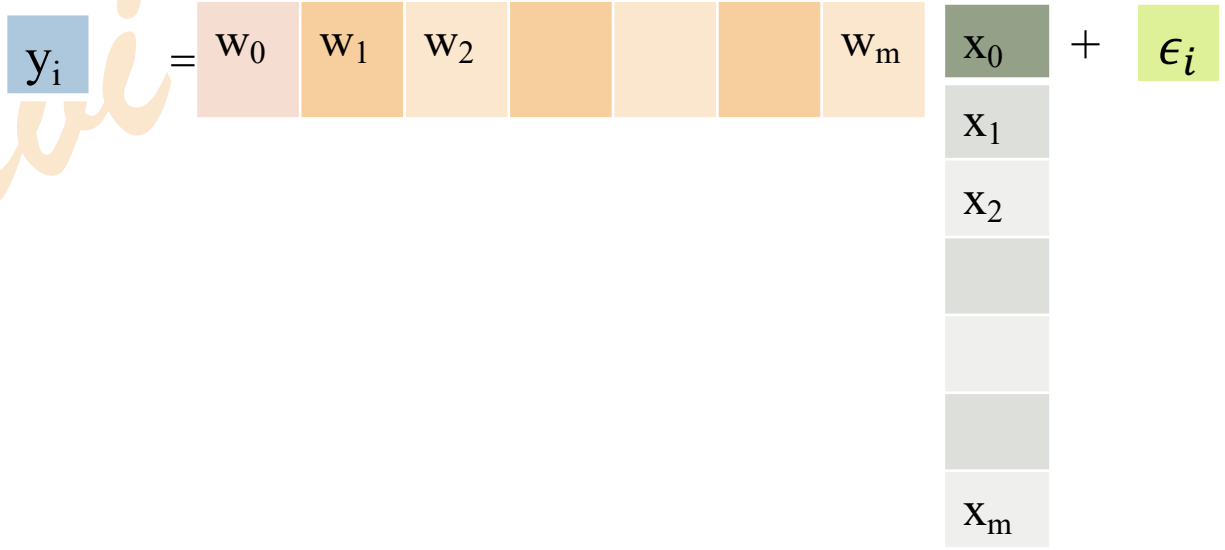
Introducing a dummy variable $x_{i0} = 1$

$$y_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im} + \epsilon_i$$

$$y_i = \sum_{j=0}^{m} w_j x_{ij} + \epsilon_i$$

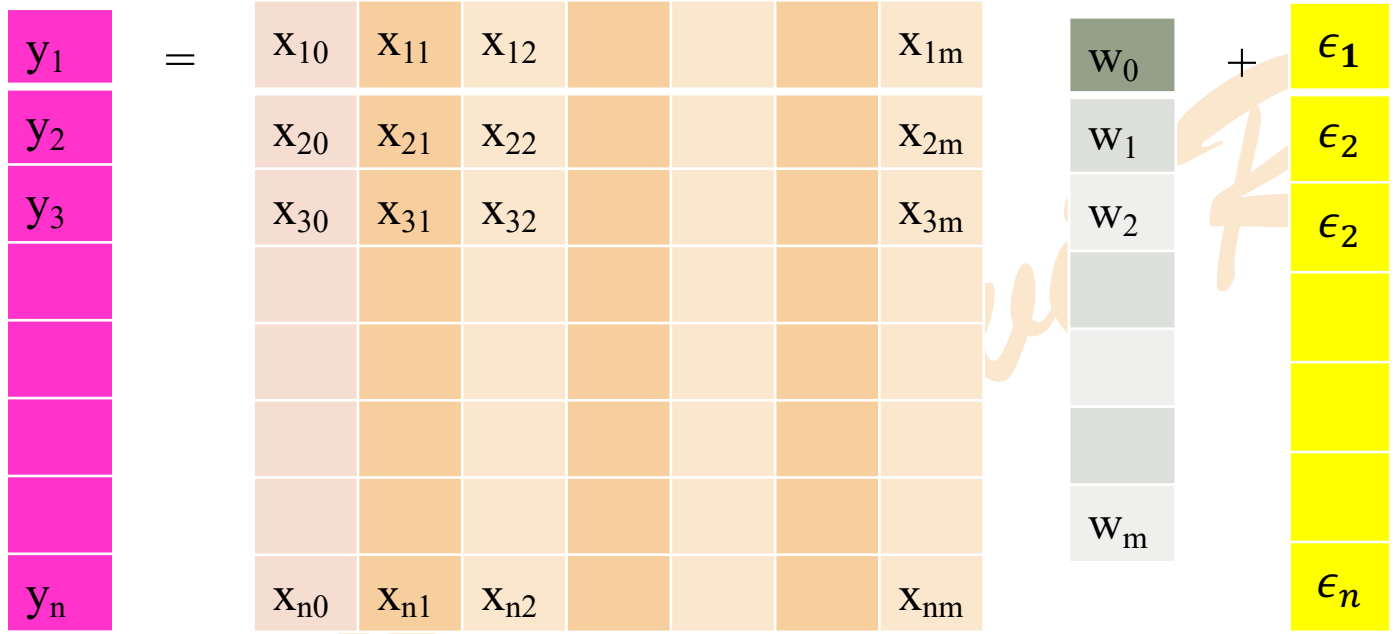
$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i$$

Equivalently we can also write $y_i = \boldsymbol{x}_i^T \boldsymbol{w} + \epsilon_i$



All Observations - Matrix Notation

n observations with m attributes



$\mathbf{y}$ = X $\mathbf{W}$ + ϵ

$$\boldsymbol{y} = X\boldsymbol{W} + \boldsymbol{\epsilon}$$

MLR Cost Function - RSS

RSS for Simple Linear Regression is given by

$$\text{RSS} = \sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

We will extend it to MLR

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{n}\left(y_i - (w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_m x_{im})\right)^2$$

$$= \sum_{i=1}^{n}(y_i - (\boldsymbol{w}^T \boldsymbol{x}_i))^2$$

$$RSS(\boldsymbol{w}) = (Y - X\boldsymbol{w})^T (Y - X\boldsymbol{w})$$

Gradient of RSS

$$\nabla RSS(\boldsymbol{w}) = \nabla[(Y - X\boldsymbol{w})^T(Y - X\boldsymbol{w})]$$

$$= -2X^T(Y - X\boldsymbol{W})$$

Approach1

Set Gradient = 0

$\nabla RSS(\boldsymbol{w}) = -2X^T(Y - X\boldsymbol{W}) = 0$

$X^T(Y - X\boldsymbol{W}) = 0$

$X^T\boldsymbol{Y} - X^TX\boldsymbol{W} = 0$

$X^TX\boldsymbol{W} = X^T\boldsymbol{Y}$

Multiplying both sides by $(X^TX)^{-1}$

$(X^TX)^{-1}X^TX\boldsymbol{W} = (X^TX)^{-1}X^T\boldsymbol{Y}$

But we know that $(X^TX)^{-1}X^TX = I$, the Identity matrix

$\boldsymbol{W} = (X^TX)^{-1}X^T\boldsymbol{Y}$ **Closed form solution**

Points to Remember – Closed form Solution

We can use closed form solution only if we can find the inverse of $X^T X$.

- $X^T X$ is a m x m matrix where m is the number of features.
- $X^T X$ is invertible only if the number of linearly independent instances n > m.
- Complexity of closed form solution is $O(n^3)$. Hence for huge data sets closed form solution is computationally very expensive

Approach 2 – Gradient Descent

While not converged {

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} - \eta \nabla RSS(\boldsymbol{w})^{t}$$

$t = t+1$

}

Substituting Gradient of RSS

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} - \eta\left(-2X^{T}(Y - X\boldsymbol{W}^{t})\right) = \boldsymbol{w}^{t} + 2\eta X^{T}(Y - X\boldsymbol{W}^{t})$$

While not converged {

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^{t} + 2\eta X^{T}\left(Y - \hat{Y}(\boldsymbol{w}^{t})\right)$$

$t = t+1$

}

Single Feature Update

Let's consider weight update for j[th] feature

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{n}(y_i - (w_0x_{i0} + w_1x_{i1} + w_2x_{i2} + \cdots + w_mx_{im}))^2$$

Differentiating partially w.r.t w_j

$$\frac{\partial}{\partial w_j}(RSS) = \sum_{i=1}^{n} 2\ (y_i - w_0x_{i0} - w_1x_{i1} - w_2x_{i2} - \cdots - w_mx_{im})(-x_{ij})$$

$$= -2\sum_{i=1}^{n} x_{ij}\left(y_i - \boldsymbol{x_i^T w}\right)$$

Single Feature Update (cont…)

Now Update of j[th] feature weight

$$w_j^{t+1} = w_j^t - \eta\left(-2\sum_{i=0}^{n} x_{ij}\left(y_i - \boldsymbol{x_i^T w^t}\right)\right)$$

$$w_j^{t+1} = w_j^t + 2\eta\left(\sum_{i=1}^{n} x_{ij}\left(y_i - \boldsymbol{\hat{y}_i}(\boldsymbol{w^t})\right)\right)$$

Algorithm

Inititialize $\mathbf{w}^{(1)} = 0$ (or random values)

While $\left\|\nabla RSS(\boldsymbol{w}^{(\boldsymbol{t})})\right\| > \varepsilon$ {

For j = 0 to m

$$\text{Partial[j]} = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \hat{\boldsymbol{y}}_{\boldsymbol{i}}(\boldsymbol{w}^{t})\right)$$

$$w_j^{t+1} = w_j^t - \eta * Partial[j]$$

t = t+1

}

Performance Metrics - Quality of Fit

Any regression model can be evaluated with the following metric

MAE – Mean Absolute Error

$$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{y_i} - \hat{y}_i))^2$$

RMSE – Root Mean Squared

RSS – Residual Sum Squares = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- RSS is Scale variant statistic

Metric - Quality of Fit (cont …)

R^2 (Coefficient of determination) - represents the proportion of variance (of y) that has been explained by the independent variables in the model.

- R^2 is scale invariant statistic

$$R^2 = 1 - \frac{Residual\ Sum\ Squares}{Total\ Sum\ Squares} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Drawback - It either remains the same or increases with the addition of new independent variables.
- The value of R-squared does not decrease even when redundant variables are added.

Metric - Quality of Fit (cont …)

Adjusted R-squared statistic – It takes into account the number of independent variables used for predicting the target variable.

- We can determine whether adding new variables to the model actually increases the model fit.

$$Adjusted\ R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-m-1}\right)$$

n – Number of observations

m – Number of independent variables/features

- If R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.
- On the other hand, if on adding the new independent variable results in significant increase in R-squared value, then the Adjusted R-squared value will also increase.

Multicollinearity

- Multicollinearity refers to a situation in multiple linear regression where two or more independent variables are highly correlated.
- Multicollinearity causes several issues in the regression analysis like
  - Instability of Coefficient Estimates
  - Inflated Standard Errors of regression coefficients.
  - Interpretation Difficulties
- Correlation matrix gives collinearity between any two pairs of variables only it fails to capture multicollinearity.
- Variance Inflation Factor (VIF) can be used to identify multicollinearity and address the problem.

Variance Inflation Factor (VIF)

- VIF is a measure of multicollinearity among independent variables in a multiple regression model.
- VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity.
- $VIF = \frac{1}{1 - R^2}$
- VIF starts at 1 and has no upper limit.
- VIF = 1, no correlation between the independent variable and the other variables.
- VIF exceeding 5 or 10 indicates high multicollinearity between this independent variable and the others.

Variance Inflation Factor (cont..)

- Fit the Initial Model: Fit a multiple linear regression model with all the potential predictors.
- Calculate VIF: Compute the VIF for each predictor.
- Identify High VIF: Identify predictors with high VIF values (commonly, a VIF > 10 is considered problematic).
- Remove High VIF Predictors: Remove the predictor with the highest VIF value.
- Refit the Model: Refit the model without the removed predictor.
- Recalculate VIF: Recalculate the VIF values for the remaining predictors.
- Repeat: Repeat the process until all VIF values are below the threshold (commonly 10).

Example

Consider a multiple regression problem with Three features x1, x2, x3 and the target feature y.

To calculate the VIF for each predictor:

For $X1$:

1. **Regress $X1$ on $X2$ and $X3$:**

   After performing this regression, let's assume we get $R^2 = 0.96$.

2. **Calculate VIF for $X1$:**

$$\text{VIF}_{X1} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.96} = \frac{1}{0.04} = 25$$

For $X2$:

1. **Regress $X2$ on $X1$ and $X3$:**

   Suppose $R^2 = 0.97$.

2. **Calculate VIF for $X2$:**

$$\text{VIF}_{X2} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.97} = \frac{1}{0.03} = 33.33$$

For $X3$:

1. **Regress $X3$ on $X1$ and $X2$:**

   Suppose $R^2 = 0.98$.

2. **Calculate VIF for $X3$:**

$$\text{VIF}_{X3} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.98} = \frac{1}{0.02} = 50$$

3. Address Multicollinearity by Removing Variables

From the VIF calculations:

- **VIF for $X1$**: 25
- **VIF for $X2$**: 33.33
- **VIF for $X3$**: 50

All VIF values are high, but $X3$ has the highest VIF. We'll start by removing the predictor with the highest VIF.

Remove $X3$ and Recalculate VIFs for Remaining Predictors:

1. **Remove $X3$ from the model:**

   The remaining predictors are $X1$ and $X2$.

2. **Recalculate VIF for $X1$ and $X2$ without $X3$:**

   - **Regress $X1$ on $X2$:**
     Suppose $R^2 = 0.95$.

$$\text{VIF}_{X1} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.95} = \frac{1}{0.05} = 20$$

   - **Regress $X2$ on $X1$:**
     Suppose $R^2 = 0.94$.

$$\text{VIF}_{X2} = \frac{1}{1 - R^2} = \frac{1}{1 - 0.94} = \frac{1}{0.06} = 16.67$$

4. Interpret Results

- **New VIFs:**
  - **VIF for $X1$**: 20 (still high but reduced)
  - **VIF for $X2$**: 16.67 (also high but reduced)

Even after removing $X3$, the VIF values are still high, which indicates some multicollinearity remains. If necessary, consider removing $X1$ or $X2$ next or using alternative methods like PCA or Ridge Regression for better model stability.