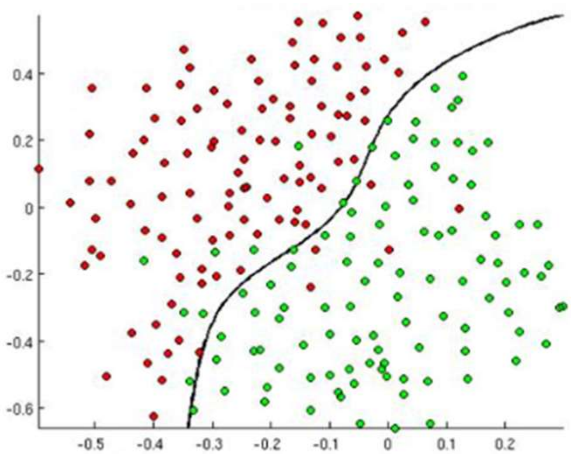
Regression Analysis

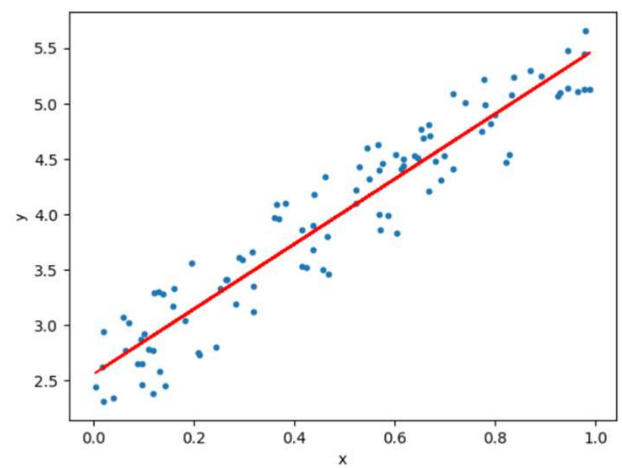
Linear Regression

Classification vs. Regression

- Classification predicts a discrete value/class label
- Regression is a type of machine learning that predicts a continuous value.



e.g., Spam filtering, Image Classification



e.g., a house's [Area, Age] ($\mathbf{x}$) vs. its Price(y)

- Is there any correlation between the observations $(\boldsymbol{X}_i, y_i)$

Recall Covariance from PCA

- Covariance tells us about the amount of dependency between two variables

$$\mathbf{cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- ❑ $\mathbf{cov}(X, Y) > 0$ → X and Y are positively related
- ❑ $\mathbf{cov}(X, Y) < 0$ → X and Y are inversely related
- ❑ $\mathbf{cov}(X, Y) = 0$ → X and Y are independent

Correlation Coefficient

- Assuming a linear relationship between the variables, the relative strength between them can be observed

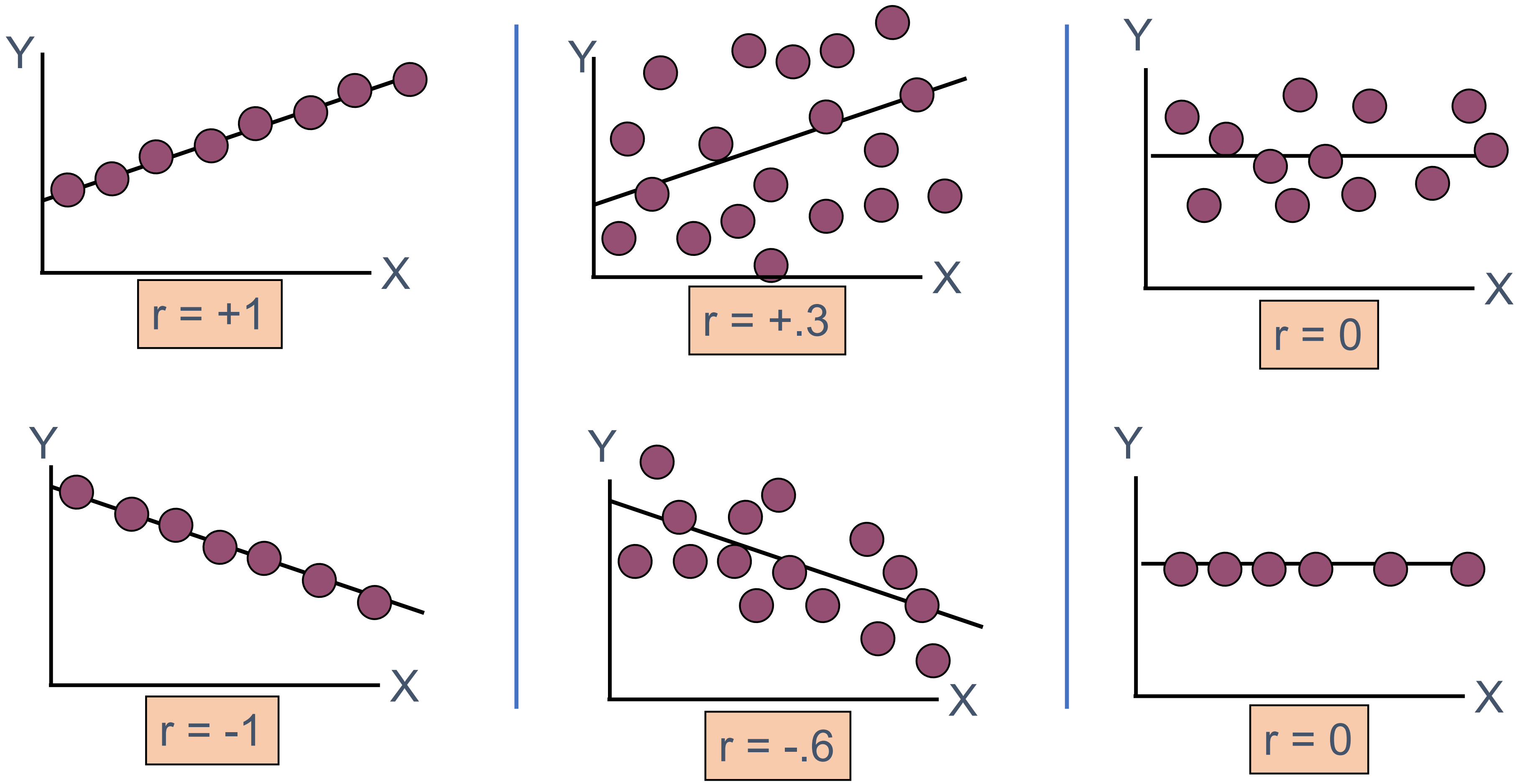
- Pearson's Correlation Coefficient is standardized covariance ranging between -1 and 1, and is unitless

$$\boldsymbol{r = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}\, X}\,\sqrt{\mathrm{var}\, Y}}}$$

- $\boldsymbol{r = 1}$: Perfect positive linear correlation
- $\boldsymbol{r = -1}$: Perfect negative linear correlation
- $\boldsymbol{r = 0}$: No linear correlation

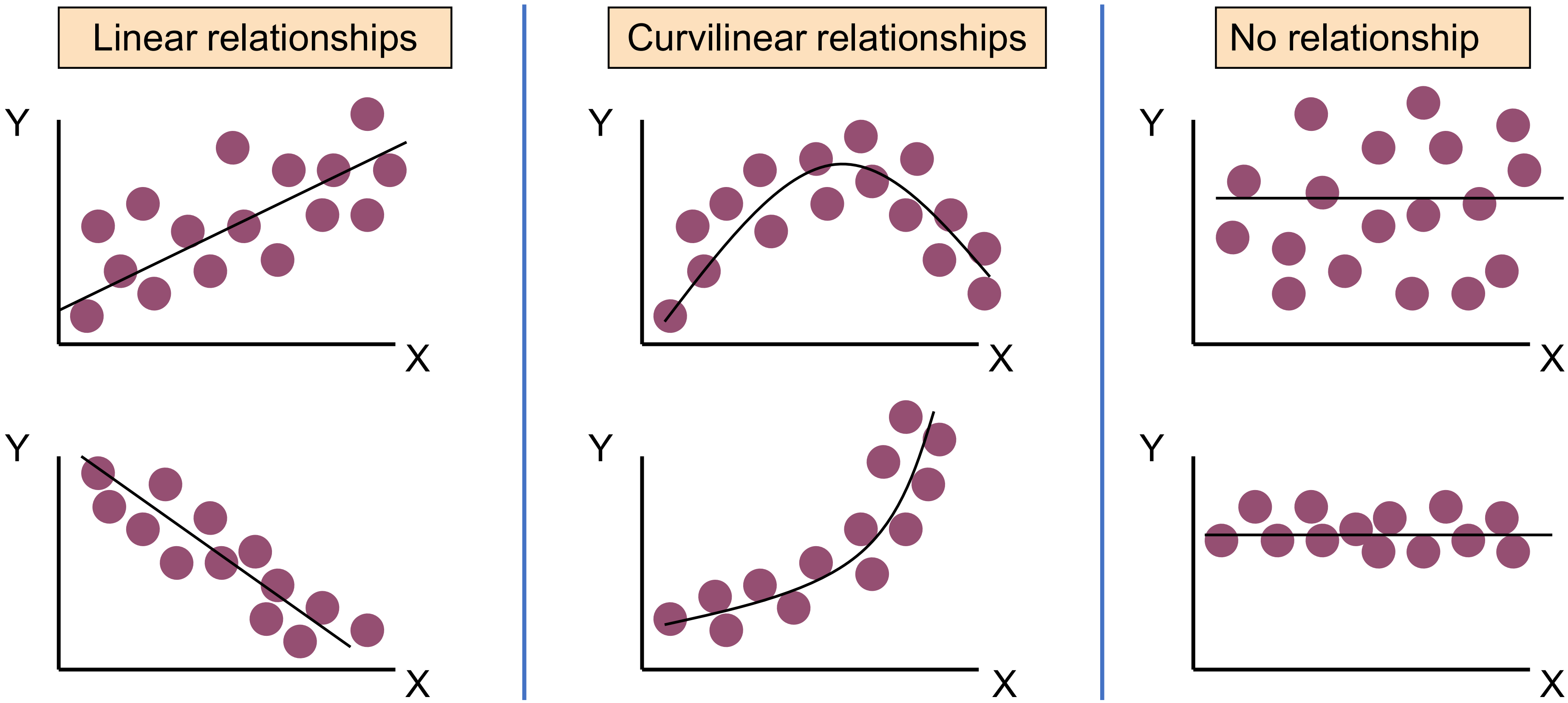
Note: Correlation does not imply causation. Even if two variables are correlated, it does not necessarily mean that changes in one variable cause changes in the other.

Scatter Plots of Data with Various Correlation Coefficients



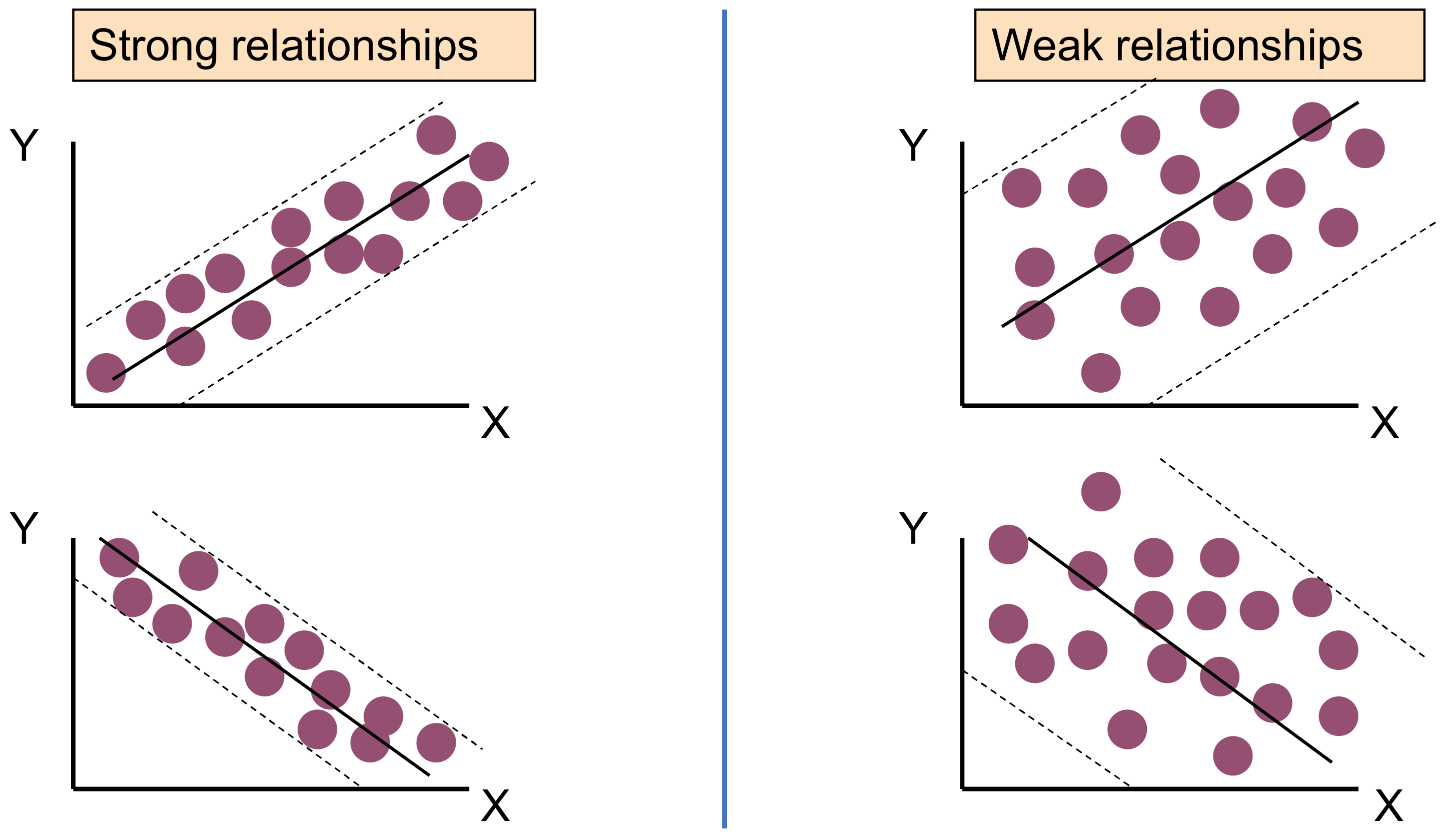
▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Regression Analysis

- The two variables (x_i, y_i) are treated as equals in correlation
- Regression analysis is a statistical method that helps us to analyse and understand the relationship between two or more variables of interest
- **Dependent Variable:** This is the variable that we are trying to forecast (**y**).
- **Independent Variable:** These are the factors that influence the analysis and provide us with information regarding the relationship of the variables with the target variable (**x**).

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	12

Linear Regression

- Linear Regression is a predictive model used for finding the ***linear*** relationship between a dependent variable and one or more independent variables

$$Y = m\,X + b,$$

- Y = dependent variable,
- X = independent variable
- m = slope (or Gradient, determines change in Y, per unit change in X),
- b = Y-intercept

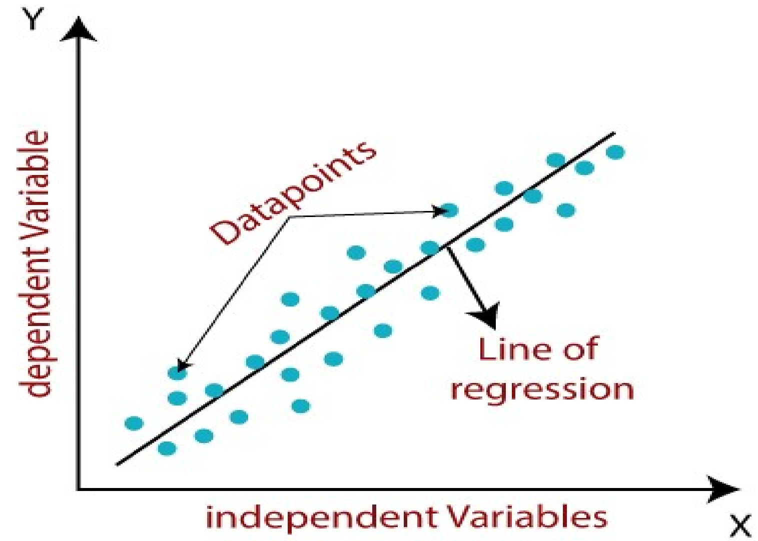


Image Source: Link

Linear Model

- A model is linear, when it is linear in its parameters: $\frac{\partial y}{\partial \alpha_i}$ is independent of $\alpha_i' s$

- $\boldsymbol{y = \alpha_0 + \alpha_1 X}$ **Linear**

- $\boldsymbol{y = \alpha_0 + \alpha_1{}^2 X}$ **Non-Linear**

- $\boldsymbol{y = \alpha_0 + \alpha_1 e^{\alpha_2 X}}$ **Non-Linear**

Modelling the dependent and independent variables

- Simple Linear Regression: $\boldsymbol{y = \alpha_0 + \alpha_1 X + \epsilon}$

- Multiple Linear Regression: $\boldsymbol{y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n + \epsilon}$

- Polynomial Regression: $\boldsymbol{y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_n X^n + \epsilon}$

✓ ε reflects the stochastic nature of the relationship between y and X indicating that such a relationship is not exact in nature

Example:

- The income and education of a person are related, with on an average basis a higher level of education providing a higher income:

$$\textbf{income} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} \times \textbf{education} + \boldsymbol{\epsilon}$$

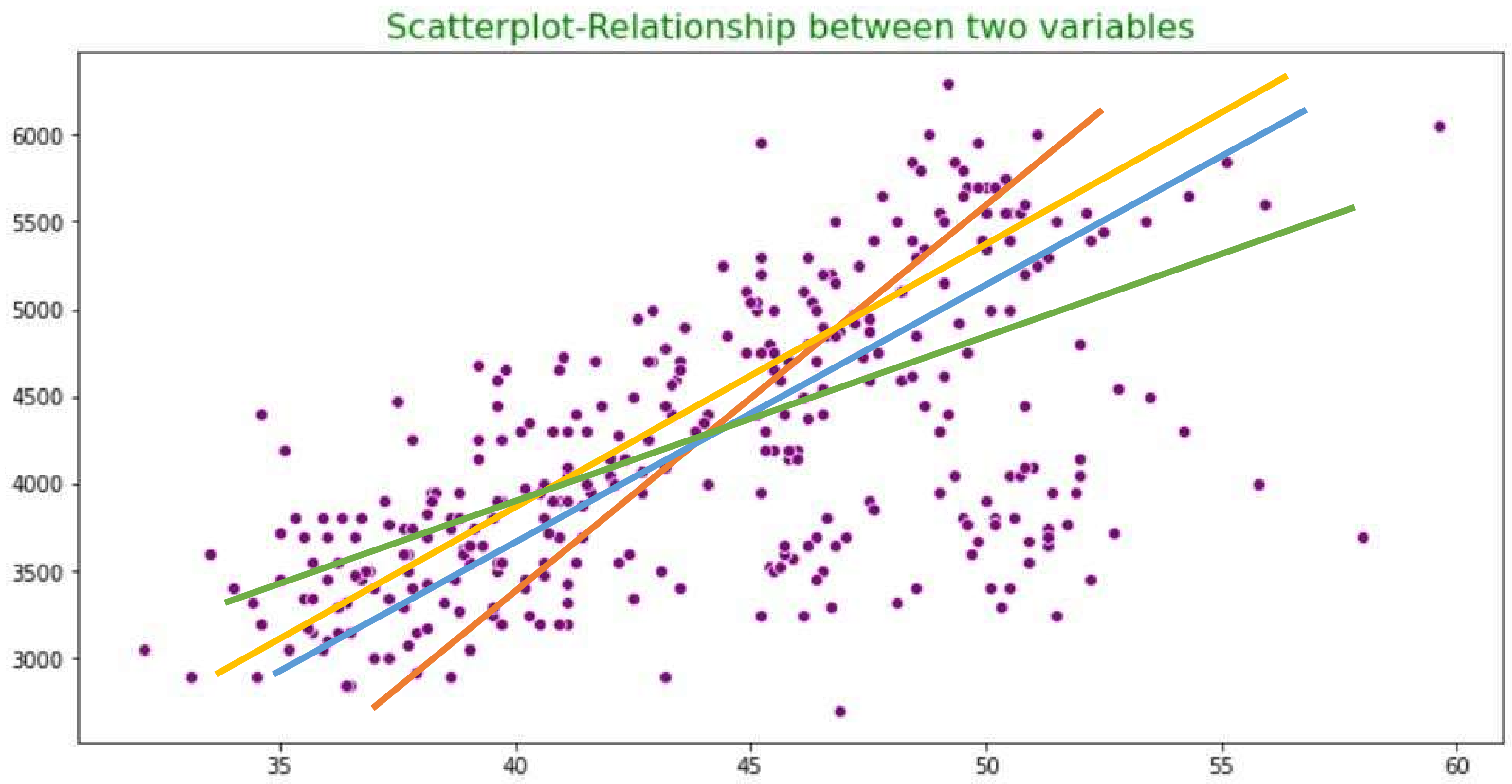
- We neglected the fact that most people have higher income when they are older than when they are young, regardless of education

$$\textbf{income} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} \times \textbf{education} + \boldsymbol{\alpha_2} \times \boldsymbol{age} + \boldsymbol{\epsilon}$$

- Let's say that the income tends to rise less rapidly in the later earning years than in early years.

$$\textbf{income} = \boldsymbol{\alpha_0} + \boldsymbol{\alpha_1} \times \textbf{education} + \boldsymbol{\alpha_2} \times \textbf{age} + \boldsymbol{\alpha_3} \times \textbf{age}^2 + \boldsymbol{\epsilon}$$

How a Linear Regression Model Works

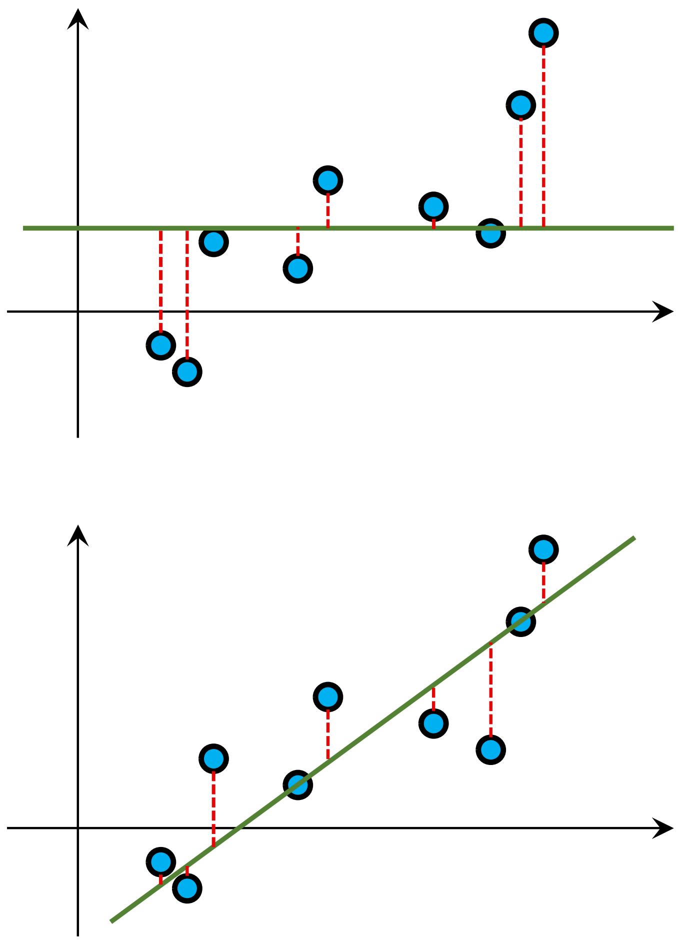


The Linear Regression Problem

- Given a set of samples, (x_i, y_i), find $\alpha_{0,1}$ that minimizes:

$$\sum_i (\alpha_1 x_i + \alpha_0 - y_i)^2$$

- Minimum squared error (MSE) criterion
- Simple Case: $y_i = \alpha_0$
- Generic Case (one independent variable): $y_i = \alpha_1 x_i + \alpha_0$
- Maximum Likelihood Estimate of α_0 and α_1.

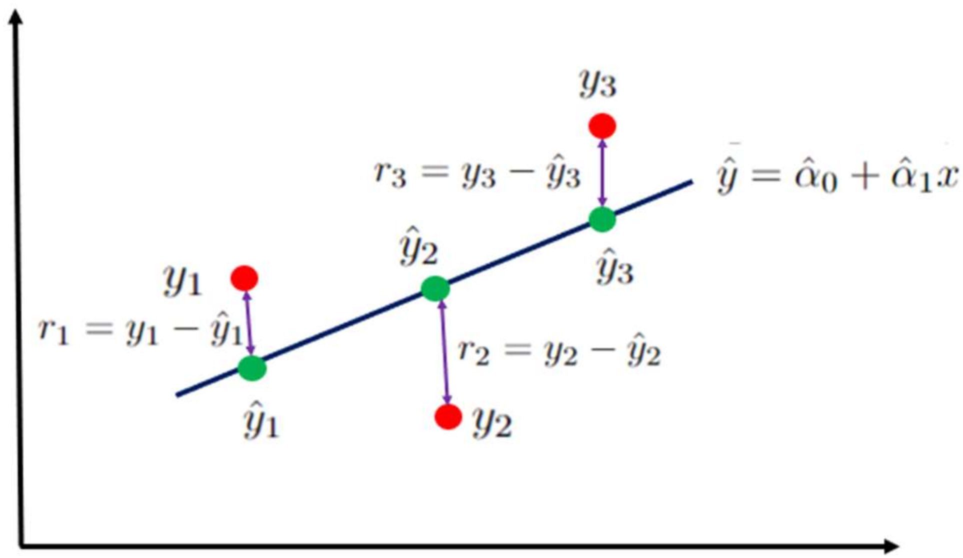


Least Square Estimation

- The best fit line is drawn across a scatter plot of data points in order to represent a relationship between those data points

- The line for which the error between the predicted values and the observed values is minimum is called the best fit line

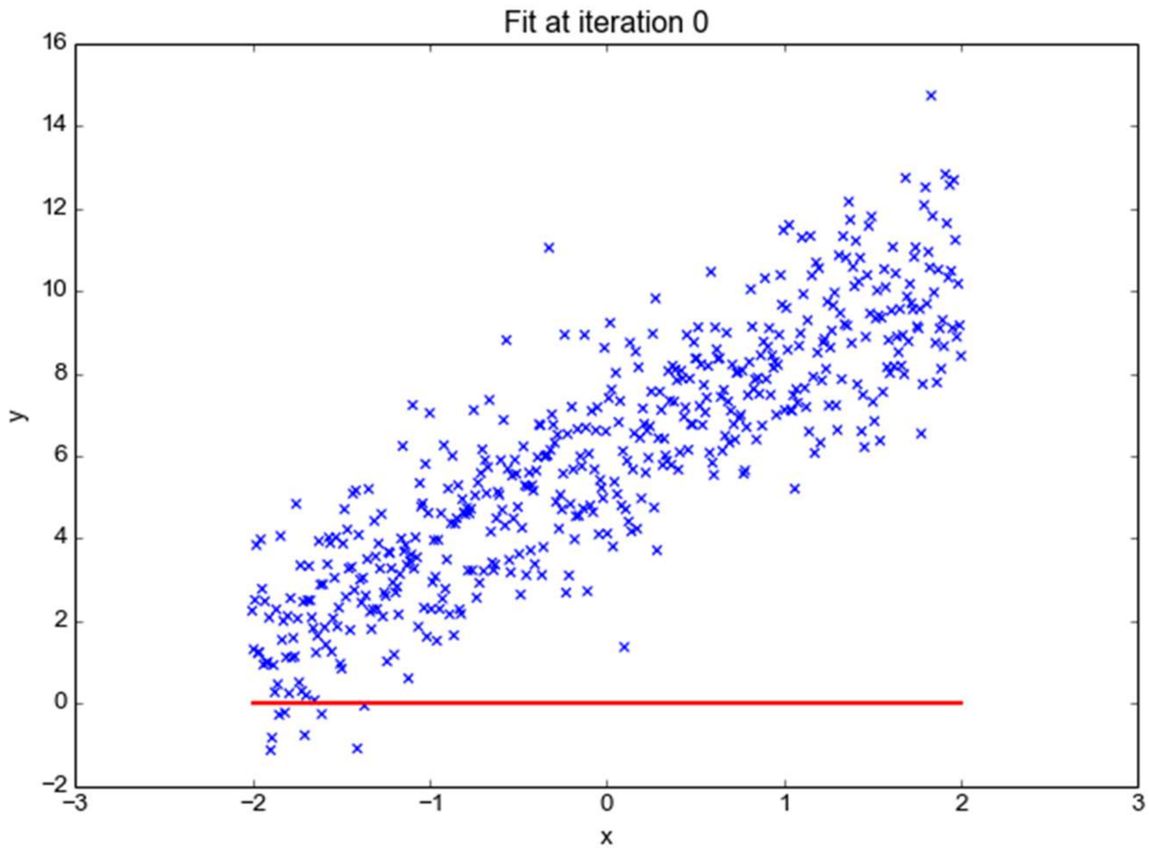
- The Loss function expresses how far off the mark our computed output is and is used to determine the error between the output of our algorithms and the given data.



$$\boldsymbol{Residual} = (y_{actual} - y_{predicted}) = \epsilon$$

✓ **Goal : Minimizing the loss**

How to gradually choose the best line



Linear Regression: by Matrix Inverse

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$ call it: $\boldsymbol{Dw} = \boldsymbol{y}$.

- Can we compute: $\boldsymbol{w} = \boldsymbol{D}^{-1}\boldsymbol{y}$?
- $\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{w} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{y}$
- $\boldsymbol{w} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{D}^{\dagger}\boldsymbol{y}$, where

$\boldsymbol{D}^{\dagger} = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}$ is called the pseudo-inverse of D.

Linear Regression: by Gradient Descent

1. Randomly initialize $\boldsymbol{w} = [\boldsymbol{\alpha_0}, \boldsymbol{\alpha_1}]$; call it: $\mathbf{w}^0$
2. Compute the gradient of the error function E at $\mathbf{w}$: $\nabla E = \frac{\partial \mathrm{E}}{\partial \mathbf{w}}$
3. $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta \nabla E$
4. Repeat steps 2 and 3 until convergence

$$\mathrm{E} = \frac{1}{2} \sum_i (\boldsymbol{\alpha_1} x_i + \boldsymbol{\alpha_0} - y_i)^2$$

$$\nabla E = \begin{bmatrix} \sum_i (\boldsymbol{\alpha_1} x_i + \boldsymbol{\alpha_0} - y_i) \\ \sum_i x_i \, (\boldsymbol{\alpha_1} x_i + \boldsymbol{\alpha_0} - y_i) \end{bmatrix}$$

- Convergence: when $||\eta \nabla E||$ becomes small

Understanding Data with Linear Regression

- Data Scientists use linear regression to understand the relationship between variables in a dataset.

- Assume that we have a line fit between weight and height of a set of students.
  - We need to understand how much the weight of a person is dependent on their height.
  - Fit a line between height (x_i) and weight (y_i): Find the best α_0 and α_1.
  - The Sum of Squared residual errors to fit (SS_{fit}) is: $\sum_i (\alpha_1 x_i + \alpha_0 - y_i)^2$
  - The Sum of Squared residual errors to mean (SS_{mean}) is: $\sum_i (y_i - \bar{y})^2$

Understanding Data with Linear Regression

- The Sum of Squared residual errors to mean (SS_{mean}) is: $\sum_i (y_i - \bar{y})^2$
- The variation around mean: $Var_{mean} = \frac{1}{n} \sum_i (y_i - \bar{y})^2 = \frac{1}{n} SS_{mean}$
- The Sum of Squared residual errors to fit (SS_{fit}) is: $\sum_i (\alpha_1 x_i + \alpha_0 - y_i)^2$
- The variation around fit: $Var_{fit} = \frac{1}{n} \sum_i (\alpha_1 x_i + \alpha_0 - y_i)^2 = \frac{1}{n} SS_{fit}$
- Var_{fit} will be smaller than Var_{mean} , and we can define:
- $\mathbf{R^2} = \frac{\mathbf{Var_{mean} - Var_{fit}}}{\mathbf{Var_{mean}}} = \frac{SS_{mean} - S_{\ fit}}{SS_{mean}}$

Understanding R^2

- R^2 is a statistical measure that assesses the proportion of the variance in the dependent variable that is explained by the independent variables in a multiple regression model.

$$\mathbf{R^2 = 1 -} \frac{\mathrm{SS_{fit}}}{\mathrm{SS_{mean}}}$$

1. $\mathbf{R^2 = 0}$: Model does not explain any of the variability in the dependent variable
2. $\mathbf{R^2 = 1}$: Model explains all the variability in the dependent variable
3. $\mathbf{0 < R^2 < 1}$: The proportion of variability in the dependent variable explained by the model

- If $\mathbf{R^2}$ **=0.6**, we say that the height can explain 60% of the variation from mean.

× However it does not indicate the quality of the model's predictions or the significance of individual predictors.

Adjusted R^2

- R^2 tends to increase as more independent variables are added, even if they do not contribute meaningfully to the model
- Adjusted R^2 is the variation of Y that is explained by the set of independent variables selected, adjusted for the number of independent variables (k) and the sample size (n)
- Penalizing models with more predictors unless those predictors significantly improve the fit

$$\textbf{Adjusted } \boldsymbol{R^2 = 1 - \left(\frac{(1 - R^2).(n-1)}{n-k-1}\right)}$$

Assumptions for Linear Regression

- **Linearity**: the Y variable is linearly related to the value of the X variable.
    - The change in the mean of the dependent variable is constant for any change in the independent variable, $\boldsymbol{E(Y|X) = \alpha_0 + \alpha_1 X}$

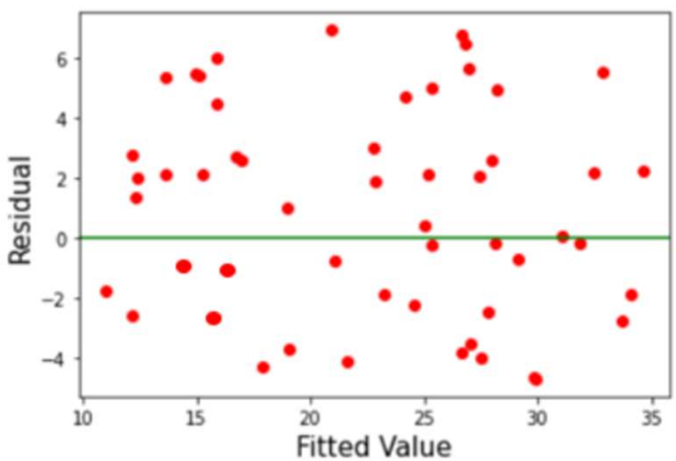
- **No Perfect Multicollinearity**: Multicollinearity indicates that there is a high correlation between independent variables in the dataset.
    - A high level of correlation between the independent variables limits our model's interpretability ability.
    - Example: Using both total number of rooms and number of bedrooms as independent variables in same model
    - Examine the correlation matrix

Assumptions

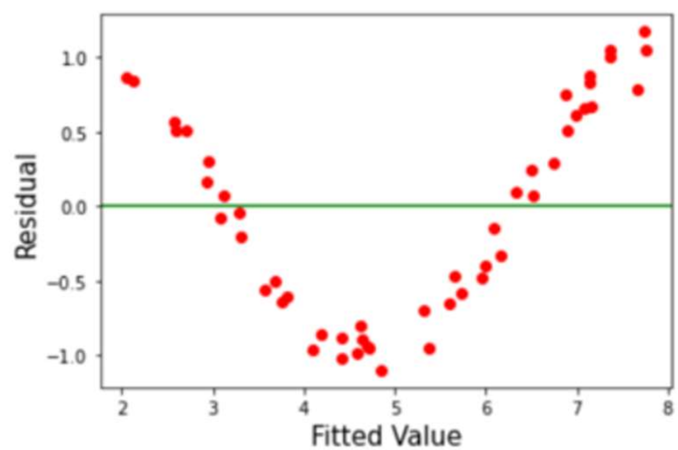
- **Normality of Residuals**: Residuals are assumed to be normally distributed, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
  - Create a histogram or a Q-Q (quantile-quantile) plot of the residuals, where points fall approximately along a straight line, when normally distributed.

- **Independence of Error** - The error (residual) is independent for each value of X
  - Residual plots, where residuals are plotted against the predicted values or X, can reveal patterns or trends that might indicate a lack of independence

Residual Plot with no pattern

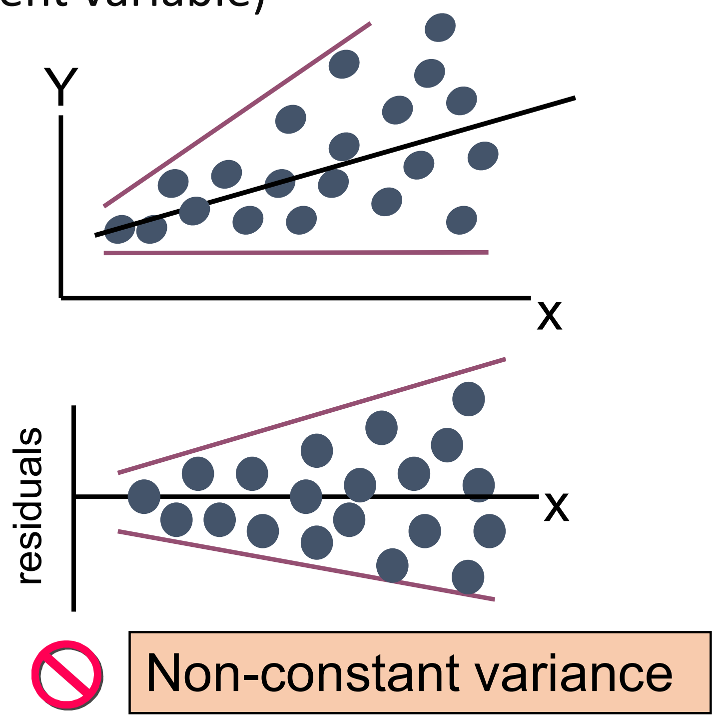


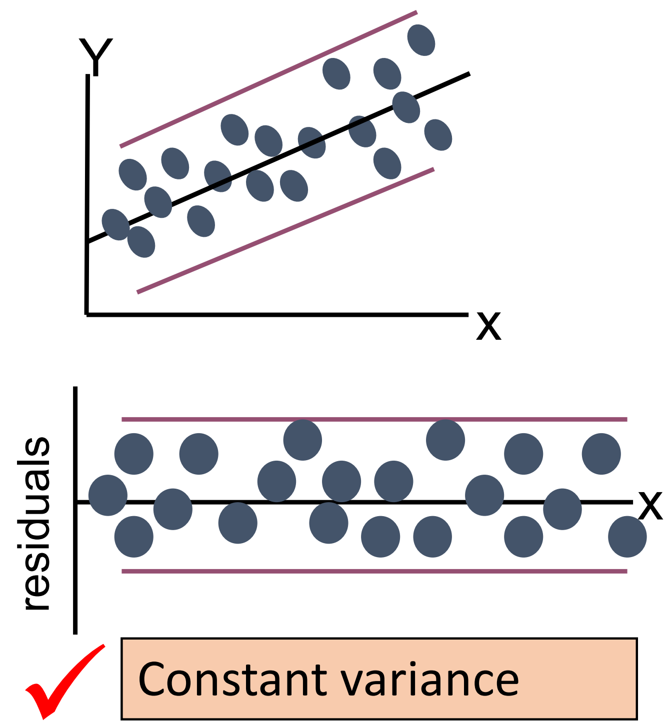
Residual Plot with a distinct pattern



Residual Analysis for Homoscedasticity

- **Homoscedasticity** - The variation around the line of regression be constant for all values of X. (the spread or dispersion of your data points remains constant across different levels of an independent variable)





▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Comments on Linear Regression

- Least squared linear regression fits a line to a pair of observations in a dataset.
- What if we have more that one independent variable?
  - One can fit a plane or hyperplane
  - The more parameters we have, the better is R^2.
  - Adjusted R^2 : scaled with the number of dimensions.
- Multiple algorithms to find the line fit.
- One can use other models of noise (not squared residual error)

Polynomial Regression

- It is a type of linear regression where the relationship between the independent variable X and the dependent variable y is modeled as an n-degree polynomial.

- Second order polynomial in one variable: $\boldsymbol{y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \varepsilon}$

- Second order polynomial in two variables:

$$\boldsymbol{y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_{11} X_1^2 + \alpha_{22} X_2^2 + \alpha_{12} X_1 X_2 + \varepsilon}$$

- Polynomial models are useful in situations where curvilinear effects are present in the true response function.

- Polynomial models are also useful as approximating functions to unknown and possible very complex nonlinear relationship.

Polynomial Regression Shape

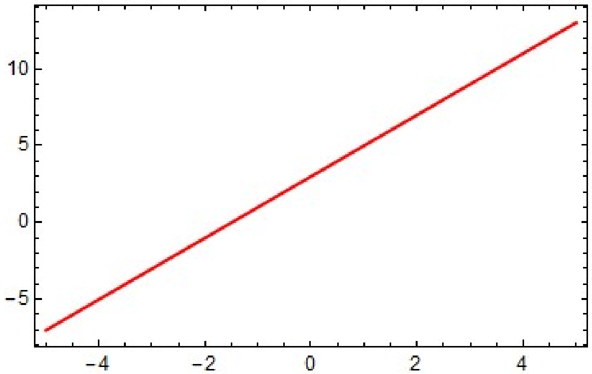
- The highest order determines the overall shape of the relationship:

$$y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_N X^N + \varepsilon$$

Linear

$y = \alpha_0 + \alpha_1 X$

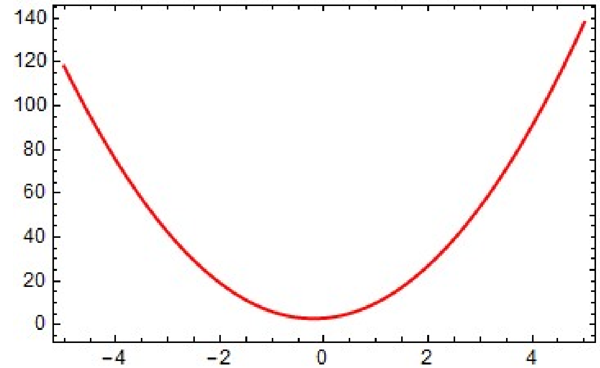
Zero bends



Quadratic

$y = \alpha_0 + \alpha_1 X + \alpha_2 X^2$

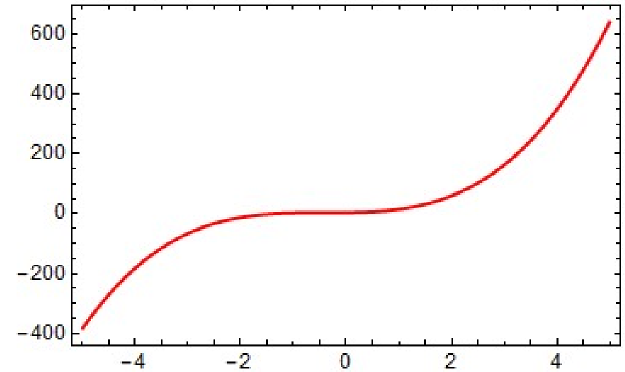
One bend



Cubic

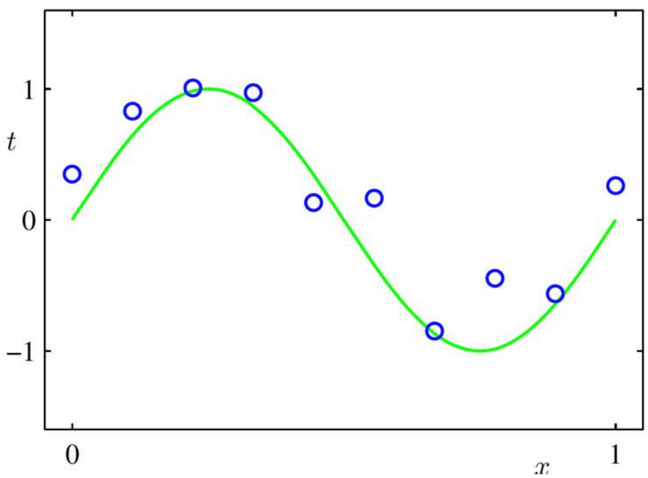
$y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3$

Two bends



Note: There is one less bend than the highest order in the polynomial model

Polynomial Regression: Example



A set of data points captured from a noisy variant of a sine function: $\sin(2\pi x)$.