Data Analytics

Course Taught at IIFT

Session 6: Regression Analysis

Dr. Tanujit Chakraborty

www.ctanujit.org

Today’s Topics……

- Regression Analysis
  - Simple Linear Regression
  - Multiple Linear Regression
  - Stepwise Regression
  - Non-Linear Regression Analysis

Relationship Analysis

- **Example: Wage Data**

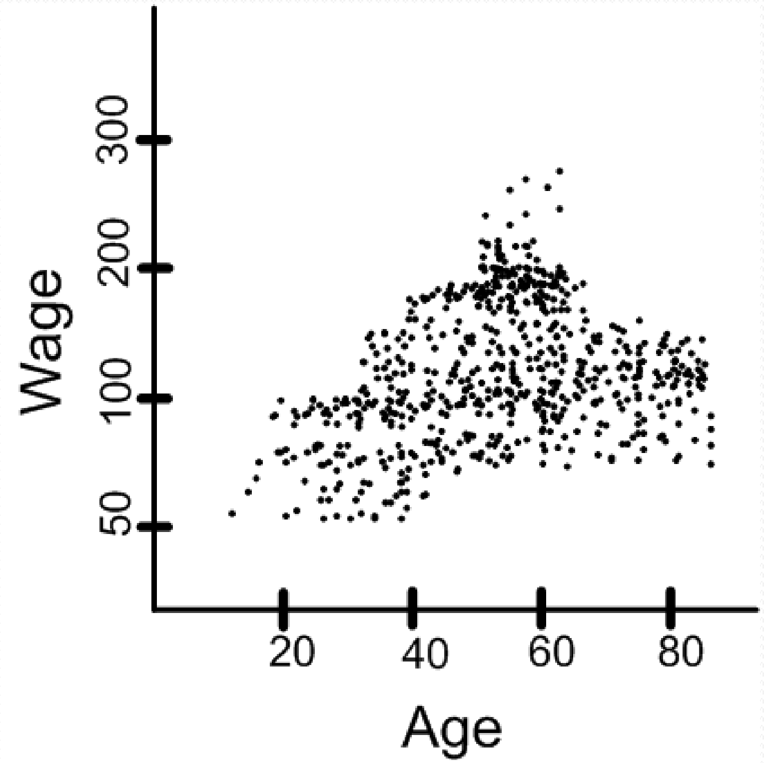
  A large data regarding the wages for a group of employees from the eastern region of India is given.

  In particular, we wish to understand the following relationships:

  - *Employee's age and wage:* How wages vary with ages?
  - *Calendar year and wage:* How wages vary with time?
  - *Employee's age and education:* Whether wages are anyway related with employees' education levels?

Relationship Analysis

- Example: Wage Data
  - Case I. Wage versus Age
    - From the data set, we have a graphical representations, which is as follows:

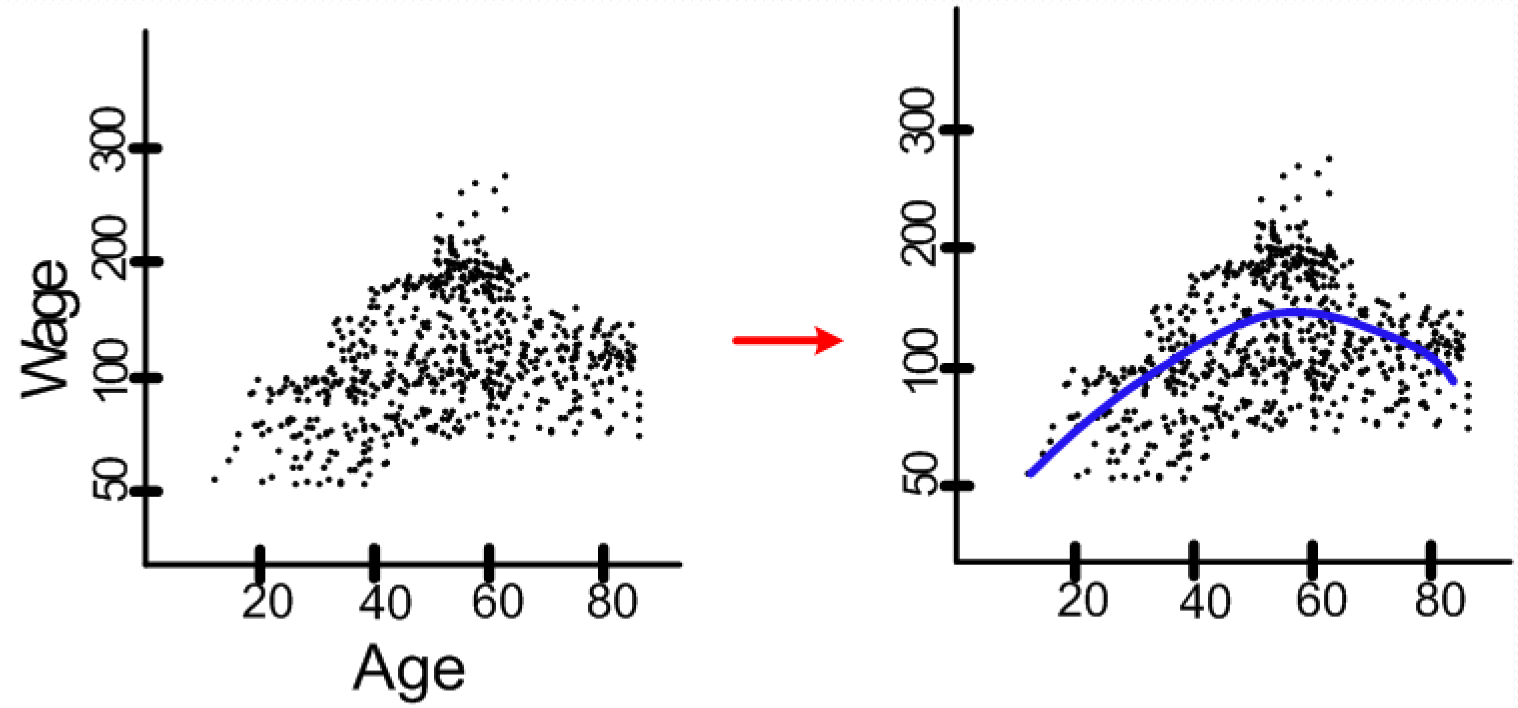


?

How wages vary with ages?

Relationship Analysis

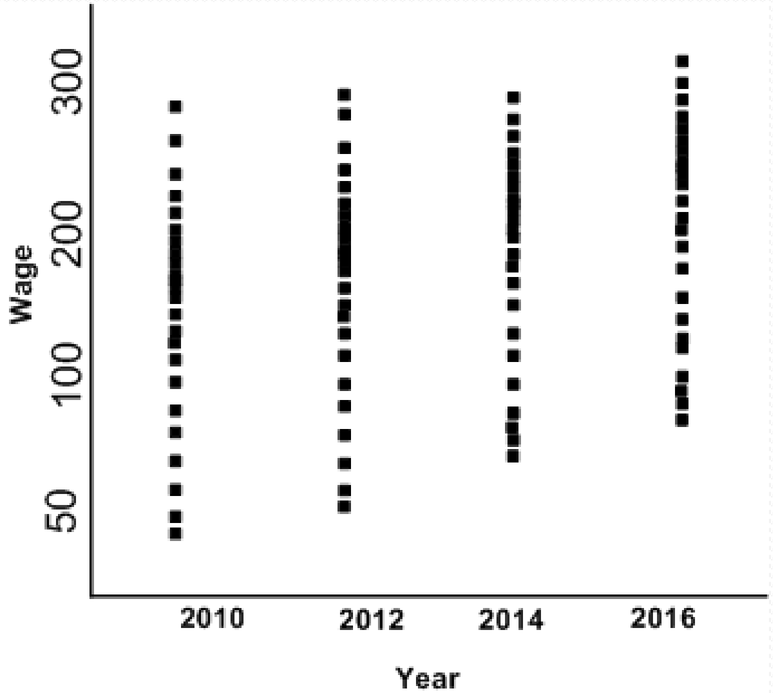
- Example: Wage Data
  - *Employee's age and wage:* How wages vary with ages?



Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

Relationship Analysis

- Example: Wage Data
  - Case II. Wage versus Year
    - From the data set, we have a graphical representations, which is as follows:

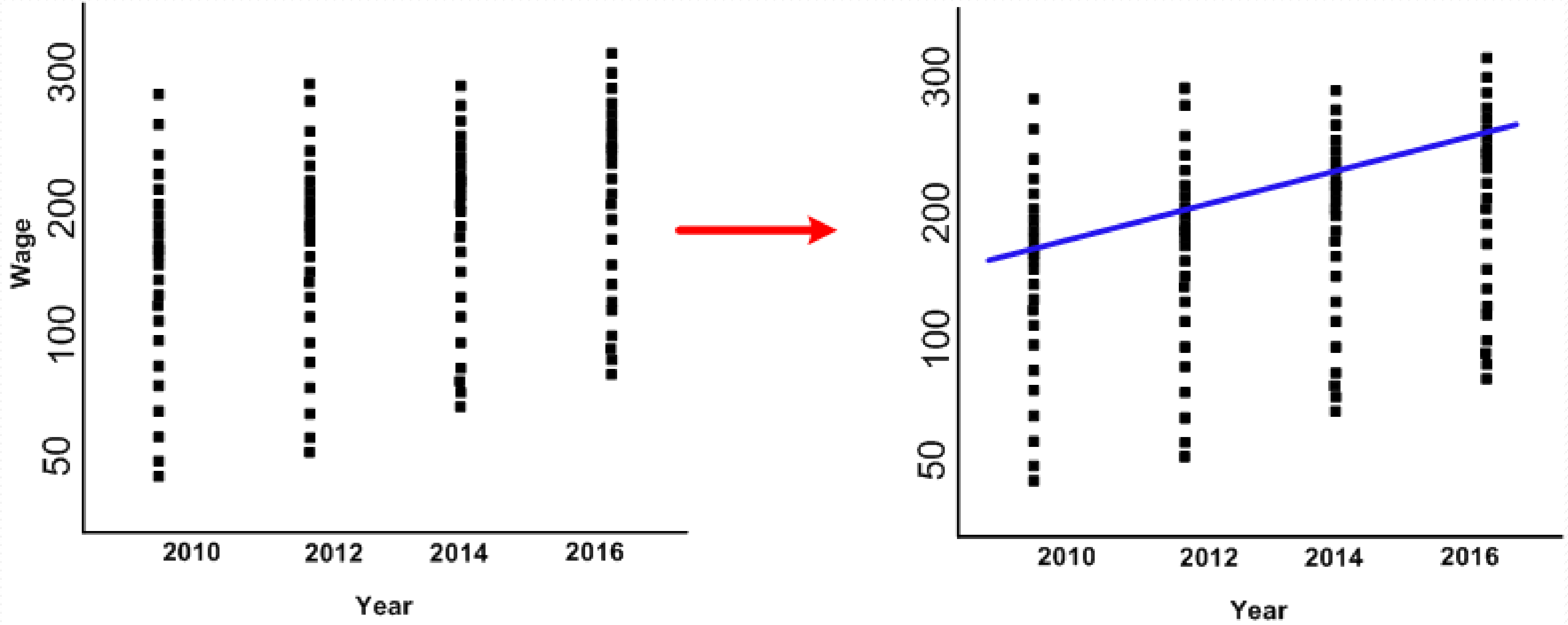


?

How wages vary with time?

Relationship Analysis

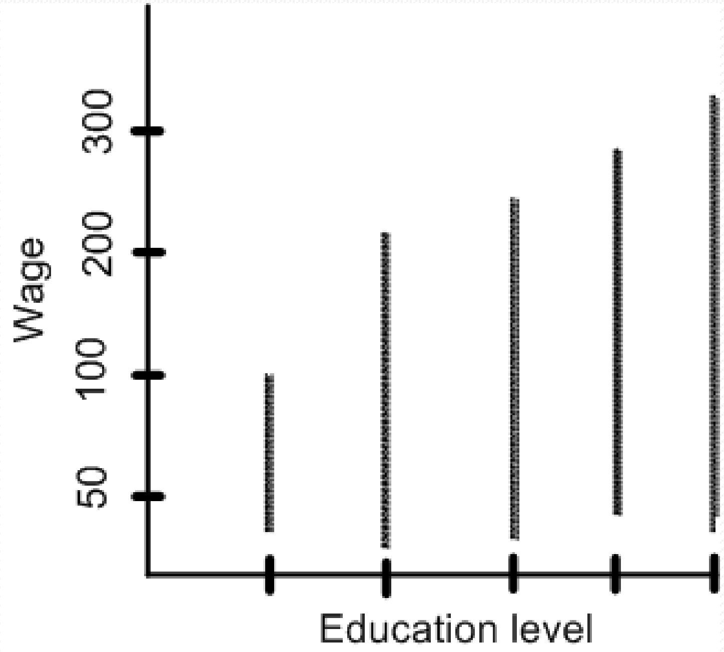
- Example: Wage Data
  - *Wage and calendar year:* How wages vary with years?



Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

Relationship Analysis

- Example: Wage Data
  - Case III. Wage versus Education
    - From the data set, we have a graphical representations, which is as follows:

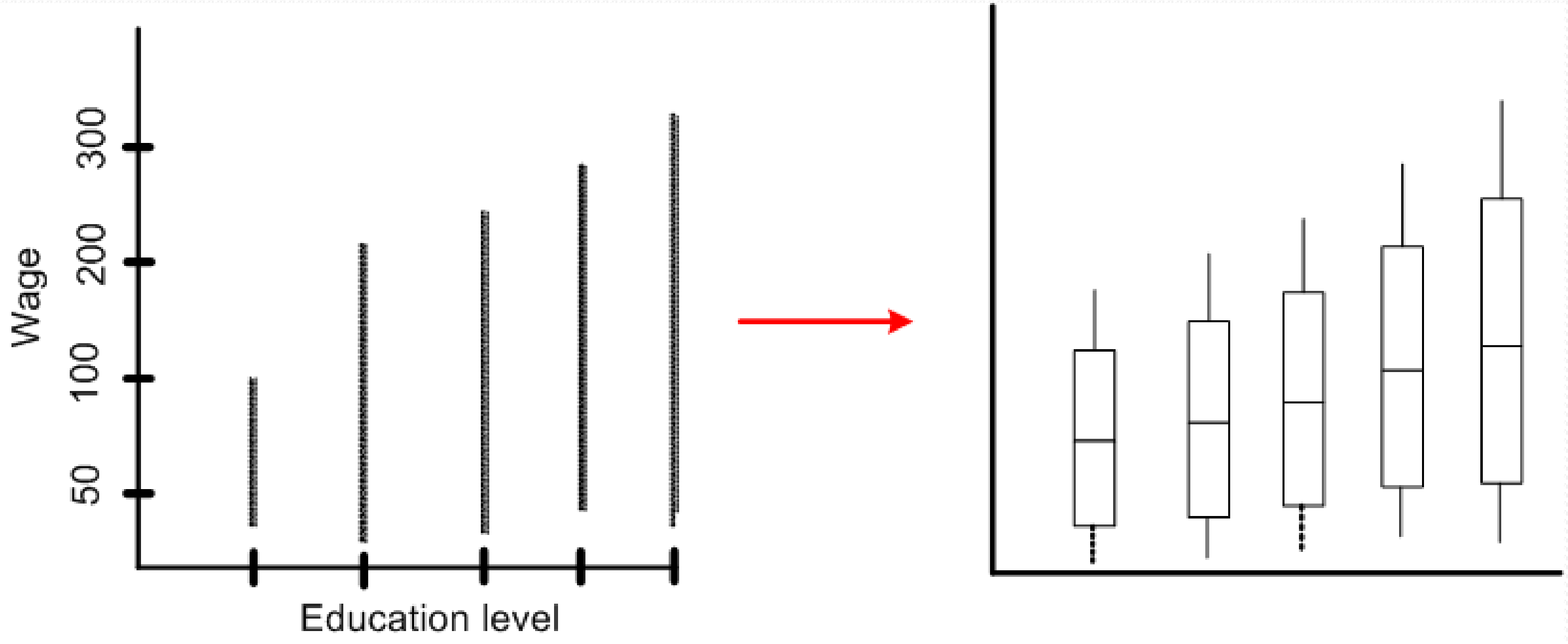


?

Whether wages are related with education?

Relationship Analysis

- Example: Wage Data
  - *Wage and education level:* Whether wages vary with employees' education levels?



Interpretation: On the average, wage increases with the level of education.

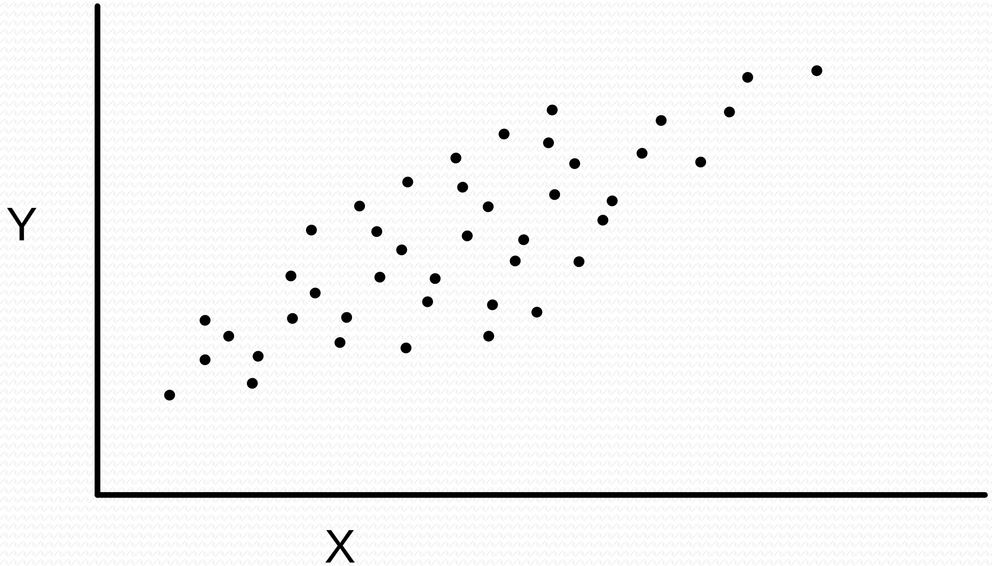
Relationship Analysis

Given an employee’s wage can we predict his age?

Whether wage has any association with both year and education level?

etc….

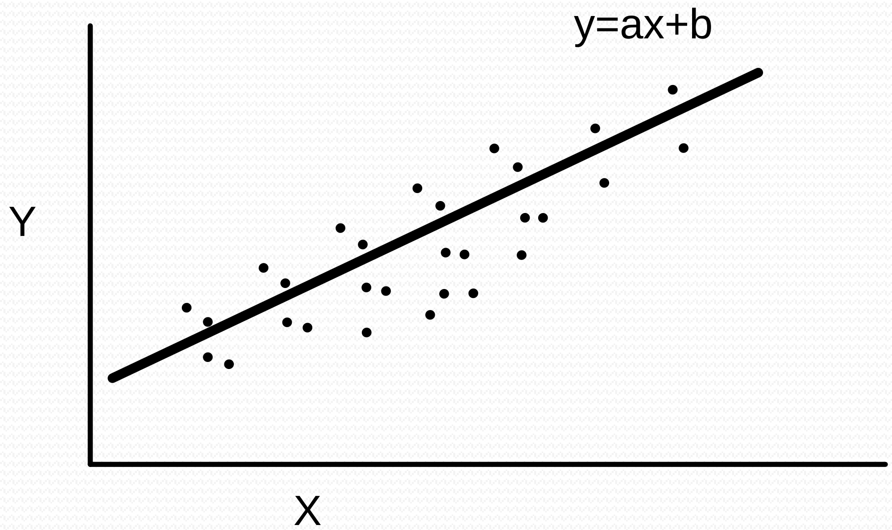
Question for You!



Suppose there are countably infinite points in the $XY\ plane$. We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Solution:



Just decide the values of **a** and **b**

(as if storing one point's data only!)

Note: Here, the trick was to find a relationship among all the points.

Measures of Relationship

Univariate population: The population consisting of only one variable.

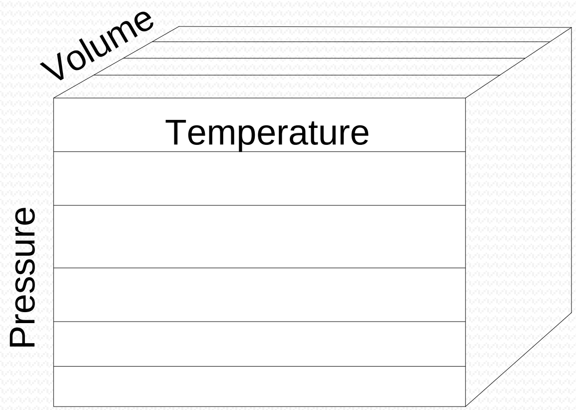
$Temperature$	20	30	21	18	23	45	52

Here, statistical measures are suffice to find a relationship.

Bivariate population: Here, the data happen to be on two variables.

$Pressure$	1	1.1	… …	0.8
$Temperature$	35	41		29

Multivariate population: If the data happen to be one more than two variable.



Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?

If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.

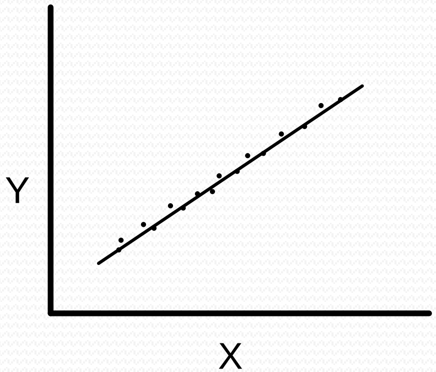
- **Correlation Analysis**
- **Regression Analysis**

Regression Analysis

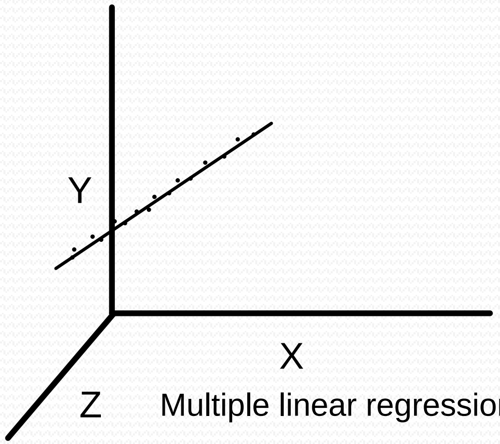
Regression Analysis

- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

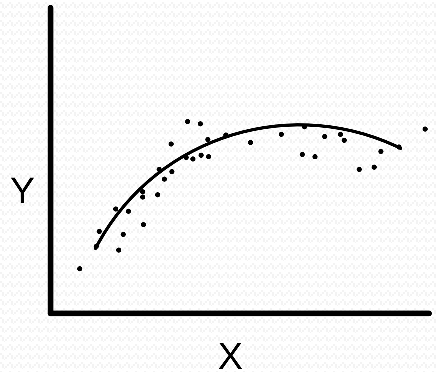
- **Classification of Regression Analysis Models**
  - Linear regression models
    1. Simple linear regression
    2. Multiple linear regression
  - Non-linear regression models



Simple linear regression

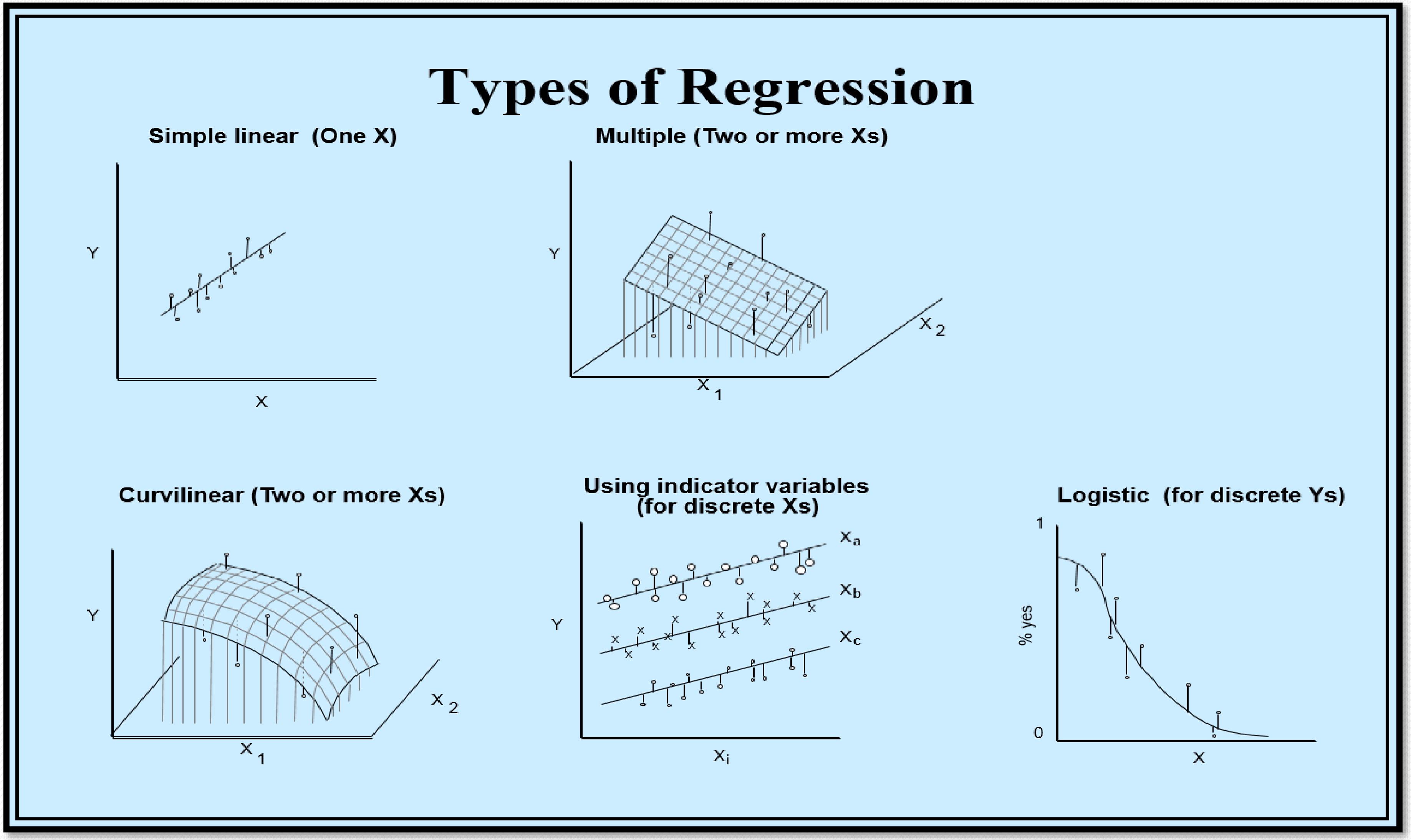


Multiple linear regression



Non-linear regression

Regression Analysis



Earlier Developments of Regression

- The earliest form of regression was the method of least squares, which was published by Legendre in 1805 and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations; the orbits of bodies about the Sun. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem.

- The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

- Galton’s work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be.

Galton Board

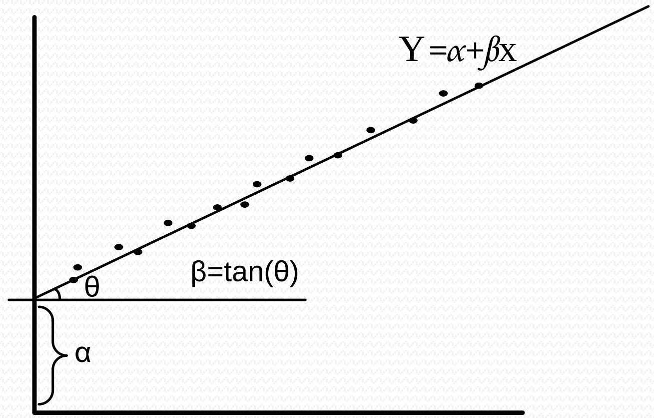
- Sir Francis Galton, Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.

- It specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.

Simple Linear Regression Model

In simple linear regression, we have only two variables:

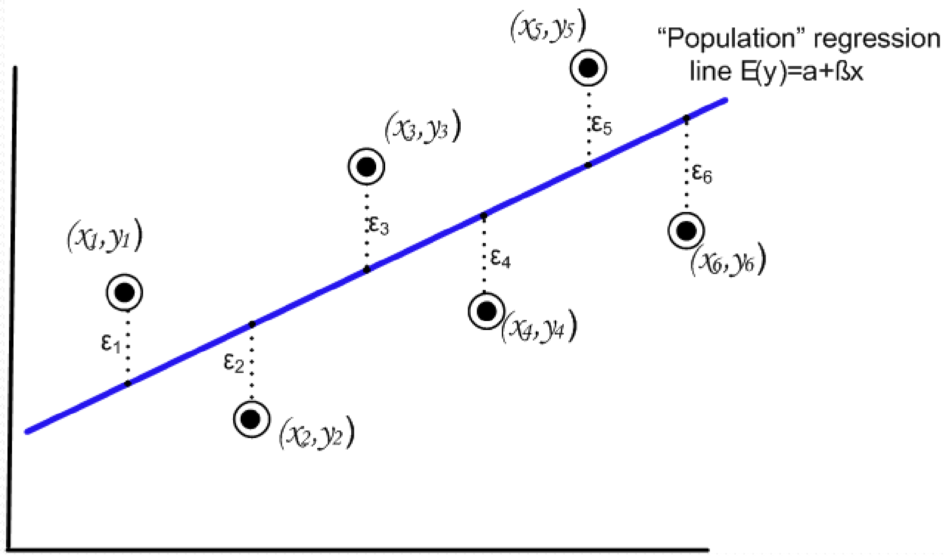
- Dependent variable (also called Response), usually denoted as Y.
- Independent variable (alternatively called Regressor), usually denoted as x.
- A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$



Note:

- There are infinite number of lines (and hence $\alpha_s \ and \ \beta_s$)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Regression Analysis



Given the set $[(x_i, y_i), i = 1,2, \ldots \ldots, n]$ of data involving n pairs of (x, y) values, our objective is to find "true" or population regression line such that $Y = \alpha + \beta x + \in$

Here, $\in$ is a random variable with $E(\in) = 0$ and $var(\in) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Note:

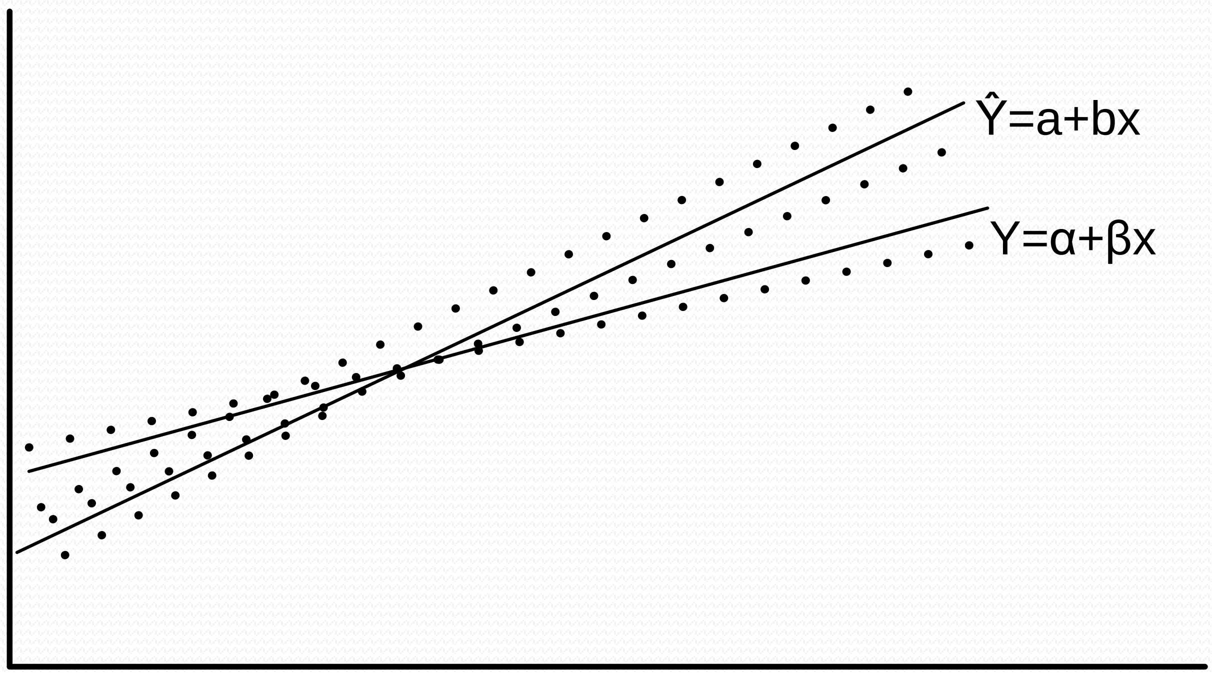
- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the data.

True versus Fitted Regression Line

- The task in regression analysis is to estimate the regression coefficients α and β.
- Suppose, we denote the estimates a for α and b for β. Then the fitted regression line is

$$\hat{Y} = a + bx$$

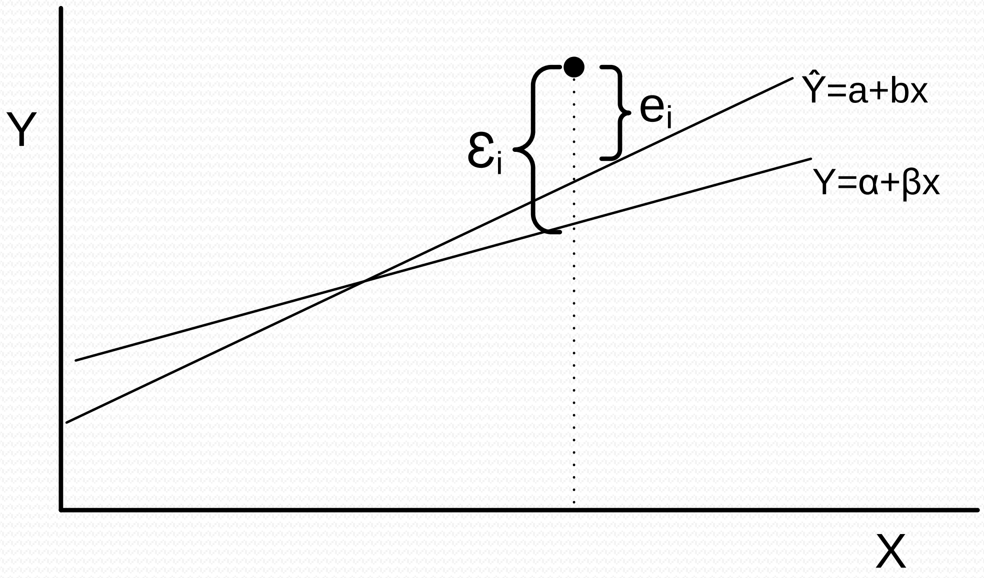
where $\hat{Y}$ is the predicted or fitted value.



Least Square Method to estimate α and β

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\widehat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \widehat{Y}_i,\ i = 1,2,3, \ldots \ldots, n$$



Least Square method

- The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as SSE

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

- We are to minimize the value of SSE and hence to determine the parameters of a and b.
- Differentiating SSE with respect to a and b, we have

$$\frac{\partial(SSE)}{\partial a} = -2\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2\sum_{i=1}^{n}(y_i - a - bx_i).x_i$$

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$

$$\frac{\partial(SSE)}{\partial b} = 0$$

Least Square method to estimate α and β

- Thus, we set

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- These two equations can be solved to determine the values of a and b, and it can be calculated that

$$\boxed{\begin{gathered} b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\[2ex] a = \overline{y} - b\overline{x} \end{gathered}}$$

R^2 : Measure of Quality of Fit

- A quantity R^2, is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.
- We have $SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$
- It signifies the **variability due to error**.
- Now, let us define the total corrected sum of squares, defined as

$$\text{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

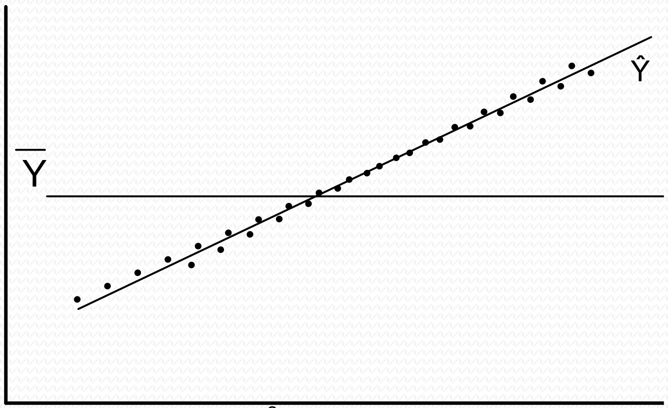
- SST represents the variation in the response values. The R^2 is

$$R^2 = 1 - \frac{SSE}{SST}$$

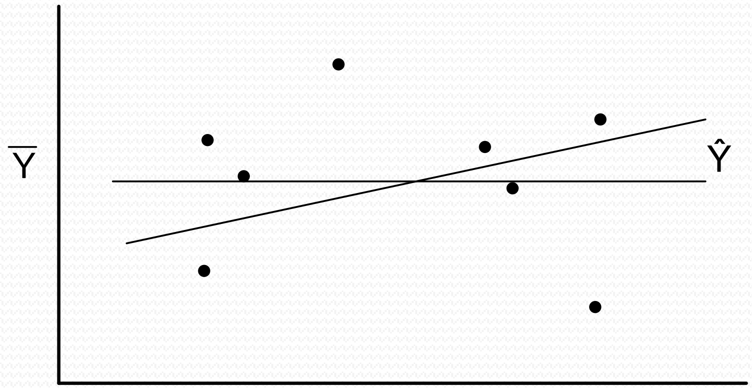
Note:

- If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)

R^2 : Measure of Quality of Fit



$R^2 \approx 1.0$ (Very good fit)



$R^2 \approx 0$ (Very poor fit)

Adjusted R^2

- The above formula for R^2 does not take into account the loss of degrees of freedom from the introduction of the additional explanatory variables in the function. The inclusion of additional explanatory variables in the function can never reduce the coefficient of multiple determination and will usually raise it.
- We introduce adjusted R^2 to compare the goodness of fit of two regression equations with different degrees of freedom. The formula for adjusted R^2 is

$$\bar{R}^2 = 1 - \Sigma(e^2/(n\text{-}K\text{-}1))/\ (\Sigma y^2/(n\text{-}1)).$$

Or

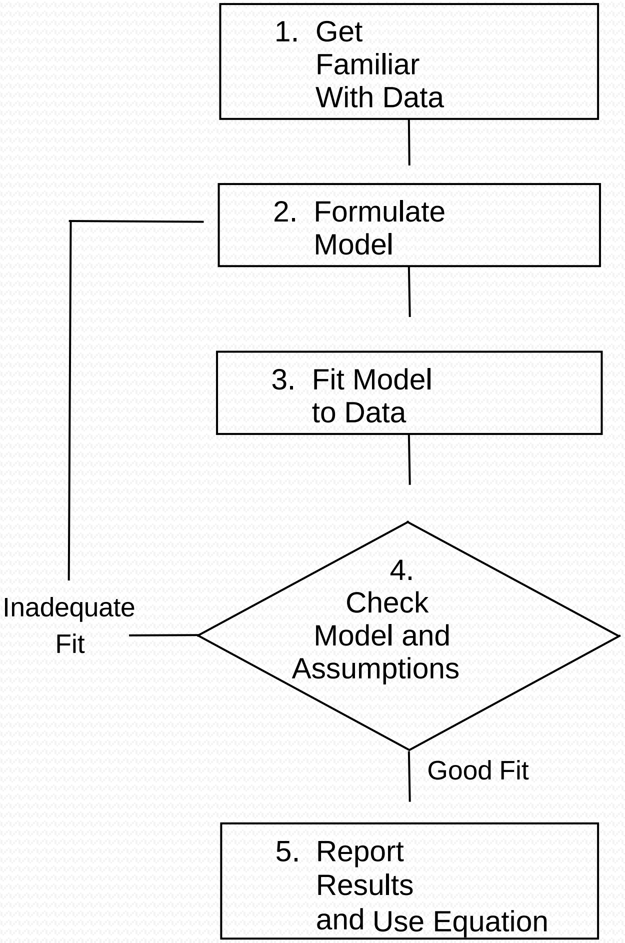
$$\bar{R}^2 = 1 - (1\text{-}R^2)(n\text{-}1)/(n\text{-}K\text{-}1).$$

For large n the value of $\bar{R}^2$ and R^2 remains almost same. For small sample, $\bar{R}^2$ will be much less than R^2 especially for large number of regressors and it may even take negative value.

Review: Interpreting Output for Regression

Name	Definition	Range	Meaning
P-value for slope	Probability that the slope is significant (different from zero)	0 to 1	If less than .05, the slope is significant (different from zero) and X is linearly related to Y.
r	Correlation coefficient	-1 to +1	Indicates the strength of a linear relationship. Numbers near zero indicate no linear relationship.
R-Square (R-sq)	Percent of explained variation $= r^2$	0 to 100%	% of variation in the Y-values explained by the linear relationship with X.
s	Standard deviation of the residuals (unexplained variation)	0 to ∞	Indicates how much the typical observed value differs from the fitted value, in units of the original data.
Residual	= Observed Y – Predicted Y	$-\infty$ to $+\infty$	Residuals are assumed to be random, and Normal with a mean of zero (represent common cause variation).
Standardized Residual	$= \dfrac{\text{residual}}{\text{standard deviation}}$	About –3 to about +3	If the absolute value of a standardized residual is > 3, then it's an unusual observation. Investigate it.
Influential Observation	An observation whose X-value has a large influence on the values of the coefficients (the regression line)	$-\infty$ to $+\infty$	View them on a plot to decide whether you will keep them or drop them from the regression analysis.

Five Step Regression Procedure: Overview



- Look at plots
- Look at descriptive statistics

- Linear or curvilinear?
- One X or more Xs?
- Transform?
- Discrete X, discrete Y?

- Do the regression

- Look at residuals plots
- Look at unusual observations
- Look at R-Sq
- Look at P-values for b

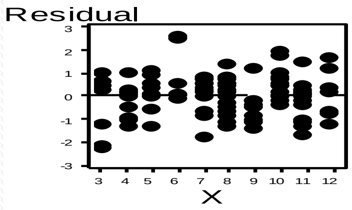
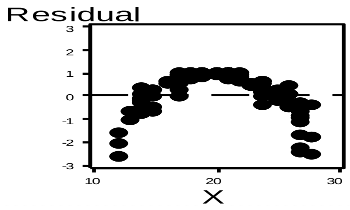
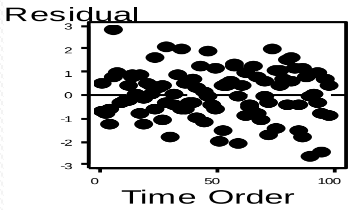
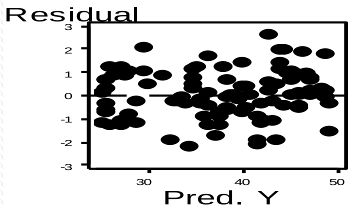
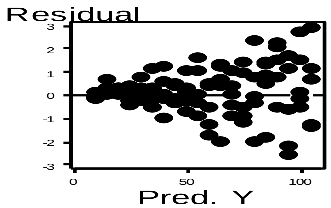
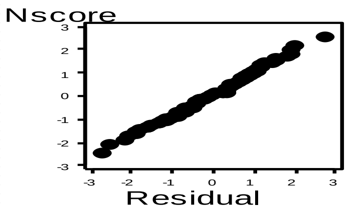
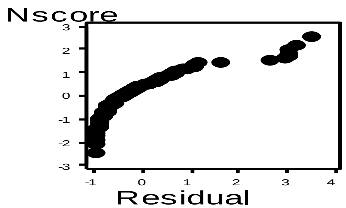
- Make predictions for X-values of interest

Assumptions Revisited

- **In regression analysis:**
  - All assumptions are made about the **residuals**
  - **No assumptions are made for X or Y**
  - **Residuals** need to be:
    - Bell-shaped (normal)
    - Stable (over time)*
    - Random
    - Unrelated (we think X *is* related to Y) Plot your data before doing regression analysis
  - **Residuals need to show certain properties for regression to work properly**
  - **The regression equation can be used to predict (or possibly manage) output data from input/process data**

Checking Assumptions About Residuals

Residuals plots must be checked to ensure the assumptions hold; otherwise, the regression equation may be incorrect or misleading.

Residuals Plot	Good	Bad	Meaning / Actions
1. Residuals vs Each X *Used to check that the residuals are not related to the Xs*			The relationship between X & Y is not a straight line, but a curve. Try a transformation on X, Y, or both. Or use X^2 in a multiple regression.
2. Time Plot of Residuals *Used to check for stability over time*			*Any* pattern visible over time means another factor, related to time, influences Y. Try to discover it and include it in a multiple regression.
3. Residuals vs Predicted Y (Fits) *Used to check that they are constant over the range of Ys*			This fan shape means the variation increases as Y gets larger (it's not constant). Try a square root, log, or inverse transformation on Y.
4. Normal Probability Plot of Residuals *Used to check that residuals are Normal*			The residuals are not Normal. Try a transformation on X or Y or both.

Multiple Linear Regression

- When more than one variable are independent variable, then the regression can be estimated as a multiple regression model
- When this model is linear in coefficients, it is called multiple linear regression model
- If k-independent variables $x_1, x_2, x_3 \ldots\ldots\ldots\ldots, x_k$ are associated, the multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \epsilon$$

- And the estimated response is obtained as

$$\hat{y}_i = \mathrm{b}_0 + \mathrm{b}_1 x_1 + \mathrm{b}_2 x_2 + \mathrm{b}_k x_k$$

Multiple Linear Regression

Estimating the coefficients

Let the data points given to us is

$$(x_{1i},\ x_{2i},\ \ x_{3i}, \dots \dots \dots \dots \dots \dots ., x_{ki}, y_i) \quad i = 1,2, \dots \dots \dots, n, \qquad n > k$$

where y_i is the observed response to the values $x_{1i},\ x_{2i},\ \ x_{3i}, \dots \dots \dots \dots \dots \dots ., x_{ki}$ of k independent variables $x_1,\ x_2,\ \ x_3, \dots \dots \dots \dots \dots \dots ., x_k$.

Thus,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \epsilon_i$$

and

$$\hat{y}_i = \mathrm{b}_0 + \mathrm{b}_1 x_{1i} + \mathrm{b}_2 x_{2i} + \mathrm{b}_k x_{ki} + \mathrm{e}_i$$

where ϵ_i and e_i are the random error and residual error, respectively associated with true response y_i and fitted response $\hat{y}_i$.

Using the concept of **Least Square Method** to estimate $b_0, b_1, b_2, \dots, b_k$, we minimize the expression

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Multiple Linear Regression

- Differentiating SSE in turn with respect to $b_0, b_1, b_2, \ldots, b_k$ and equating to zero, we generate the set of (*k+1*) normal estimation equations for multiple linear regression.

$$\mathrm{nb}_0 + \mathrm{b}_1 \sum_{i=1}^{n} x_{1i} + \mathrm{b}_2 \sum_{i=1}^{n} x_{2i} + \ldots + \mathrm{b}_k \sum_{i=1}^{n} x_{\mathrm{k}i} = \sum_{i=1}^{n} y_i$$

$$\mathrm{b}_0 \sum_{i=1}^{n} x_{1i} + \mathrm{b}_1 \sum_{i=1}^{n} x_{1\mathrm{i}}^2 + \mathrm{b}_2 \sum_{i=1}^{n} x_{1i}.x_{2i} + \ldots + \mathrm{b}_k \sum_{i=1}^{n} x_{1i}.x_{\mathrm{k}i} = \sum_{i=1}^{n} x_i.y_i$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\mathrm{b}_0 \sum_{i=1}^{n} x_{\mathrm{k}i} + \mathrm{b}_1 \sum_{i=1}^{n} x_{\mathrm{k}i}.x_{1i} + \mathrm{b}_2 \sum_{i=1}^{n} x_{\mathrm{k}i}.x_{2i} + \ldots + \mathrm{b}_k \sum_{i=1}^{n} \mathrm{x}_{ki}^2 = \sum_{i=1}^{n} x_i.y_i$$

- The system of linear equations can be solved for $b_0, b_1, \ldots, b_k$ by any appropriate method for solving system of linear equations.
- Hence, the multiple linear regression model can be built.

Testing the Significance of the Regression Coefficients

Model: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon$

Hypothesis: The x variables are not relevant to y.

H_0 : $\boxed{\beta_1 = 0 \text{ and } \beta_2 = 0 \text{ and } ... \beta_K = 0}$

H_1 : At least one coefficient is not zero.

Set α level to 0.05 as usual.

Rejection region: In principle, values of coefficients that are far from zero

Rejection region for purposes of the test: Large R^2

Test procedure: Compute $\boxed{F = \dfrac{R^2 / K}{(1 - R^2)/(N-K-1)}}$

Reject H_0 if F is large. Critical value depends on K and N-K-1.

(F is not the square of any t statistic if K > 1.)

Degrees of Freedom for the F statistic are K and N-K-1

Multiple Linear Regression : Dealing with multi-collinearity

- Many predictor variables or independent variables 'X_1, X_2, ...X_k' (e.g.: gender, height) and a response variable or dependent variable 'Y' (e.g.: weight).

$$\text{The regression equation is}$$

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

$$\text{where, } \hat{Y} = \text{Predicted value of } Y$$

$$a = \text{Intercept (the predicted value of } Y \text{ when all } X_i = 0)$$

$$b_j = \text{Slope of the line (the amount of difference in } Y \text{ associated}$$

$$\text{with a 1-unit difference in } X_j) : j = 1, 2, \ldots, k$$

- One of the assumption of model accuracy is that X's are not correlated. But this may not be true always. The multi-collinearity can be checked by Variance Inflation Factor (VIF).

Variance Inflation Factor (VIF)

- The variance inflation factor (VIF) is used to detect whether one predictor has a strong linear association with the remaining predictors (the presence of multi-collinearity among the predictors).
- VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated (multi-collinear).
- VIF = 1 indicates no relation; VIF > 1, otherwise. The largest VIF among all predictors is often used as an indicator of severe multi-collinearity.
- Montgomery and Peck suggest that when VIF is greater than 5-10, then the regression coefficients are poorly estimated.
- We should consider the options to break up the multi-collinearity: collecting additional data, deleting predictors, using different predictors, or an alternative to least square regression.
- If VIF is greater than 5, we should consider the options to break up the multi-collinearity: collecting additional data, deleting predictors, using different predictors, or an alternative to least square regression.

Checking Multi-collinearity

- Matrix plot : A matrix plot is a two-dimensional matrix of individual plots. Matrix plots are good for, among other things, seeing the two-variable relationships among a number of variables all at once.
- Correlation Coefficient :

❖ A measure of the relationship between variables.

❖ The most commonly used coefficient is Pearson Product-Moment Correlation Coefficient (measure of linear relationship denoted by 'r').

❖ 'r' lies between -1 and +1. r = 0 means no correlation.

❖ If one variable tends to increase as the other decreases, the correlation coefficient 'r' is negative. Conversely, if the two variables tend to increase together the correlation coefficient 'r' is positive.

For a two-tailed test of the correlation:

$H_0: r = 0$ versus $H_1: r \neq 0$ where 'r' is the correlation between a pair of variables.

STEPWISE REGRESSION

- **Many predictor variables or independent variables 'X_1, X_2, ...X_k' (e.g.: gender, height)** and a response variable or dependent variable 'Y' (e.g.: weight).

- It begins by selecting the single independent variable (entire set of predictors) that is the 'best' predictor which maximizes R^2. Then it adds (eliminates) variables in sequential manner, in order of importance and at each step it increases R^2.

- When you choose the stepwise method, you can enter a starting set of predictor variables in Predictors in initial model. These variables are removed if their p-values are greater than the Alpha to enter value. If you want keep variables in the model regardless of their p-values, enter them in Predictors to include in every model in the main dialog box.

BEST SUBSETS REGRESSION

- **Many predictor variables or independent variables '$X_1, X_2, \ldots X_k$' (e.g.: gender, height)** and a response variable or dependent variable 'Y' (e.g.: weight).

- It generates regression models using the maximum R^2 criterion by first examining all one-predictor regression models and then selecting the two-predictor models giving the largest R^2. It examines all two-predictor models, selects the two models with the largest R^2, and displays information on these two models. This process continues until the model contains all predictors.
- $Cp = (SSEp / MSEm) - (n-2p)$: where SSEp is SSE for the best model with 'p' parameter and MSEm is the mean square error for the model with all 'm' predictors.
- We look for models where Cp is small and is also close to p, the number of parameters in the model.

Non Linear Regression Model

- When the regression equation is in terms of *r*-degree, *r*>1, then it is called nonlinear regression model. When more than one independent variables are there, then it is called Multiple Non linear Regression model. Also, alternatively termed as polynomial regression model. In general, it takes the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon$$

- The estimated response is obtained as

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r$$

Solving for Polynomial Regression Model

Given that (x_i, y_i); i = 1,2,…,n are n pairs of observations. Each observations would satisfy the equations:

$$\mathrm{y}_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon_i$$

and $$\hat{\mathrm{y}}_i = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r + e_i$$

where, r is the degree of polynomial

ϵ_i = is the i^{th} random error

e_i = is the i^{th} residual error

Note: The number of observations, n, must be at least as large as r+1, the number of parameters to be estimated.

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2$, …, $x_n = x^r$. Thus, the equation assumes the form:

$$\mathrm{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_r x^r + \epsilon_i$$

$$\hat{\mathrm{y}}_i = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + e_i$$

This model then can be solved using the procedure followed for multiple linear regression model.

References

- The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Any question?

You may also send your question(s) at ctanujit@gmail.com