Regression Analysis

This presentation includes…

- Regression Analysis
  - Simple Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression Analysis
- Auto-Regression Analysis

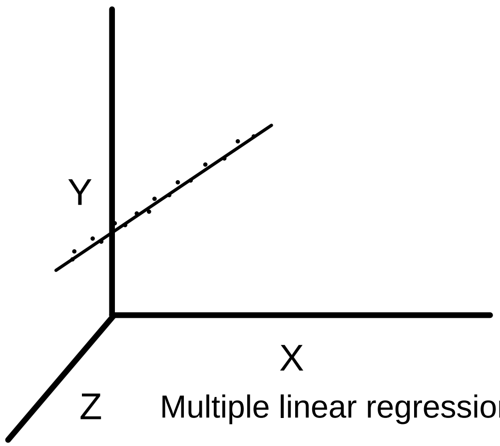
Regression Analysis

Regression Analysis

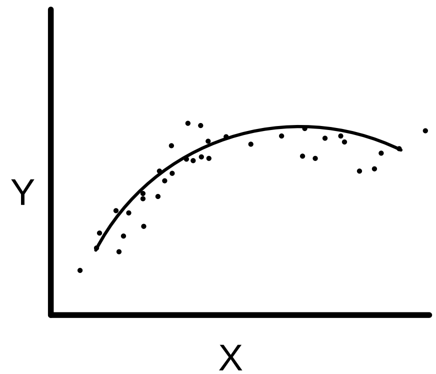
- The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of dependent variable, given the values of independent variables.

- **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

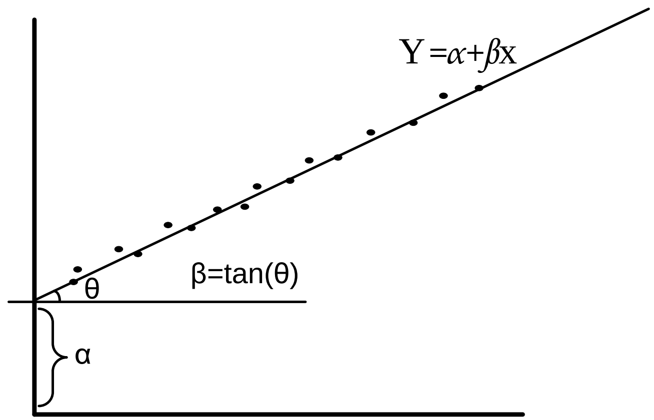


Non-linear regression

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 $Y = \alpha + \beta x$

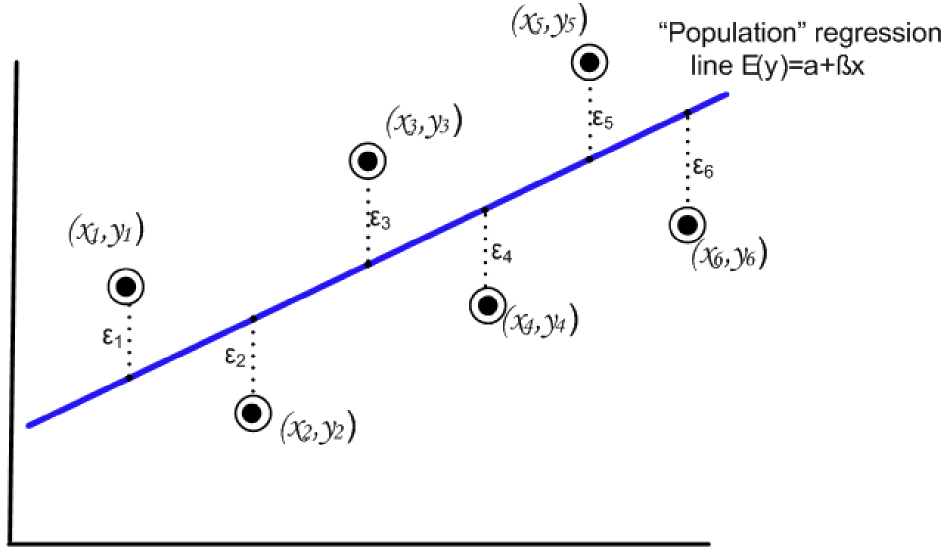


Note:

- There are infinite number of lines (and hence α_s and β_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:

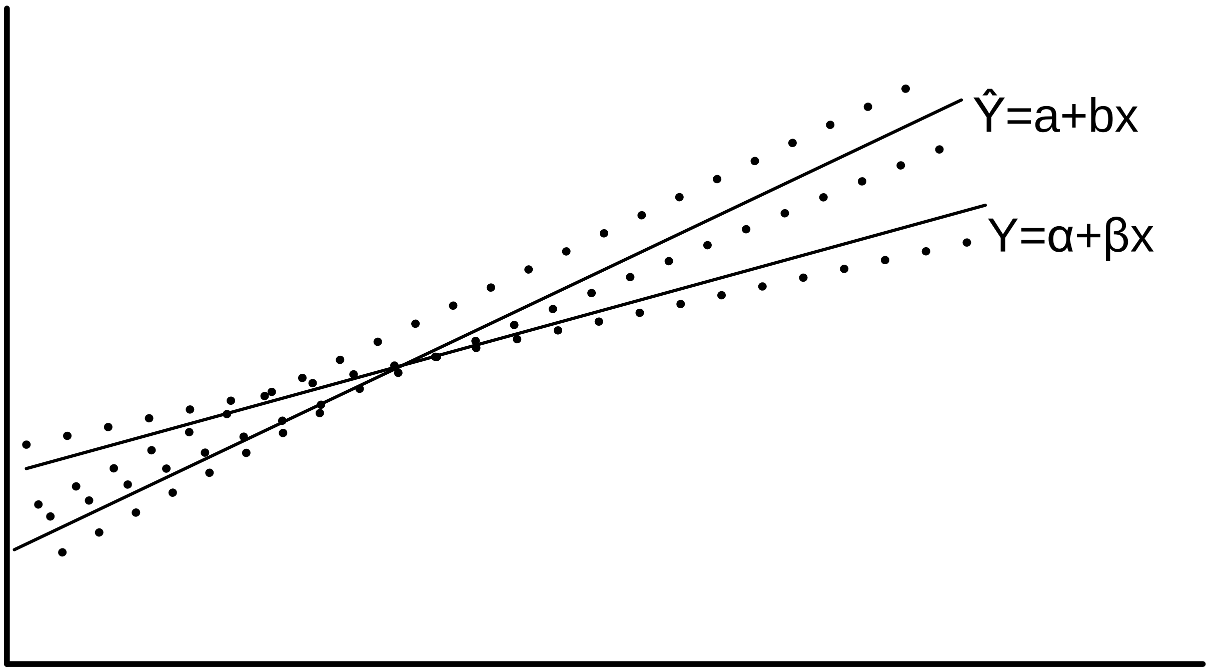
- $E(\in) = 0$ implies that at a specific x, the y values are distributed around the "true" regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- α and β are called **regression coefficients**.
- α and β values are to be estimated from the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

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

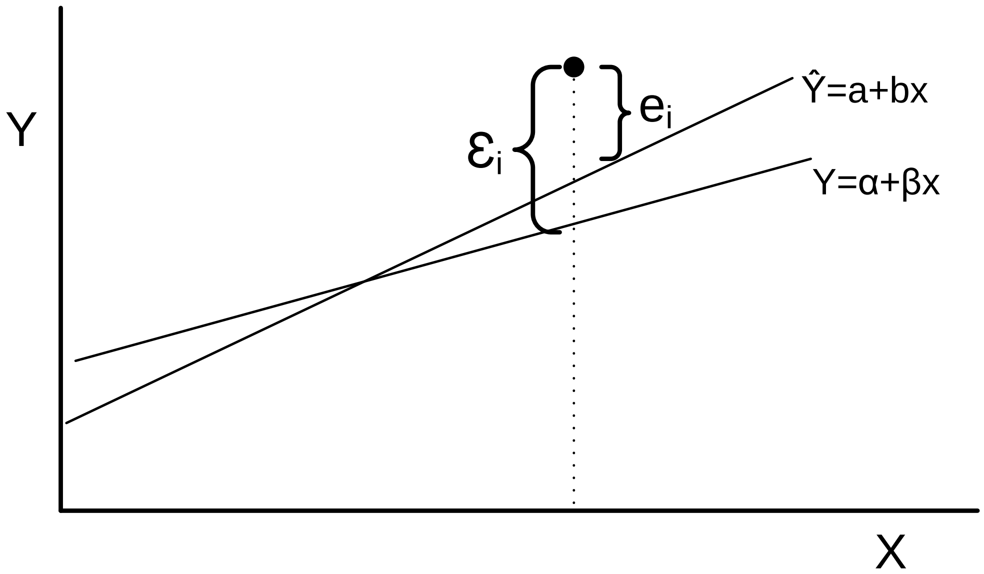
where $\hat{Y}$ is the predicted or fitted value.



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 $\widehat{\mathrm{Y}} = a + bx$. Thus, i^{th} residual is

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



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

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

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

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

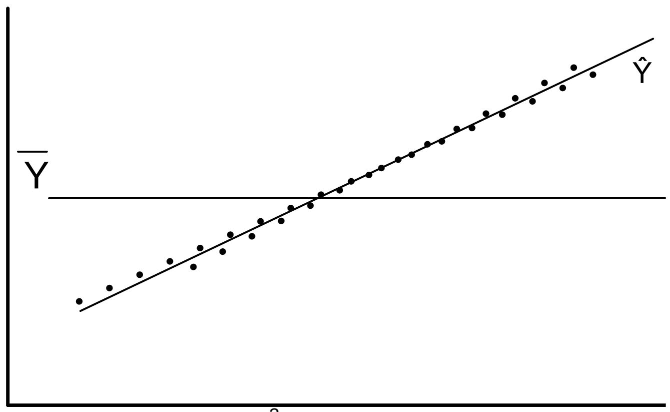
$$R^2 = 1 - \frac{SSE}{SST}$$

Note:

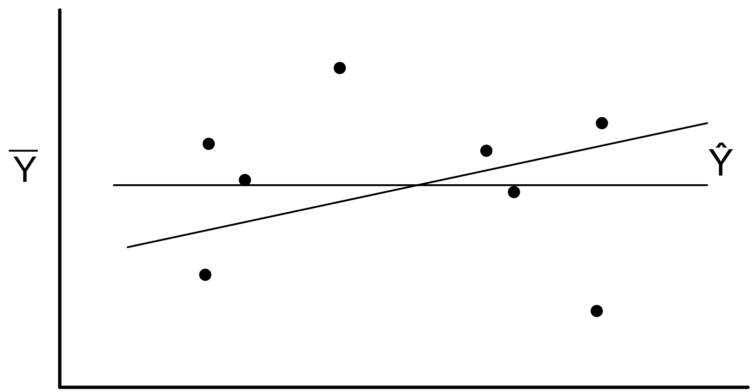
- If fit is perfect, all residuals are zero and thus R^2 = 1.0 (very good fit)

- 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)

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

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

- 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.

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

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

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

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

- 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.

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:

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

and

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

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:

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

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

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

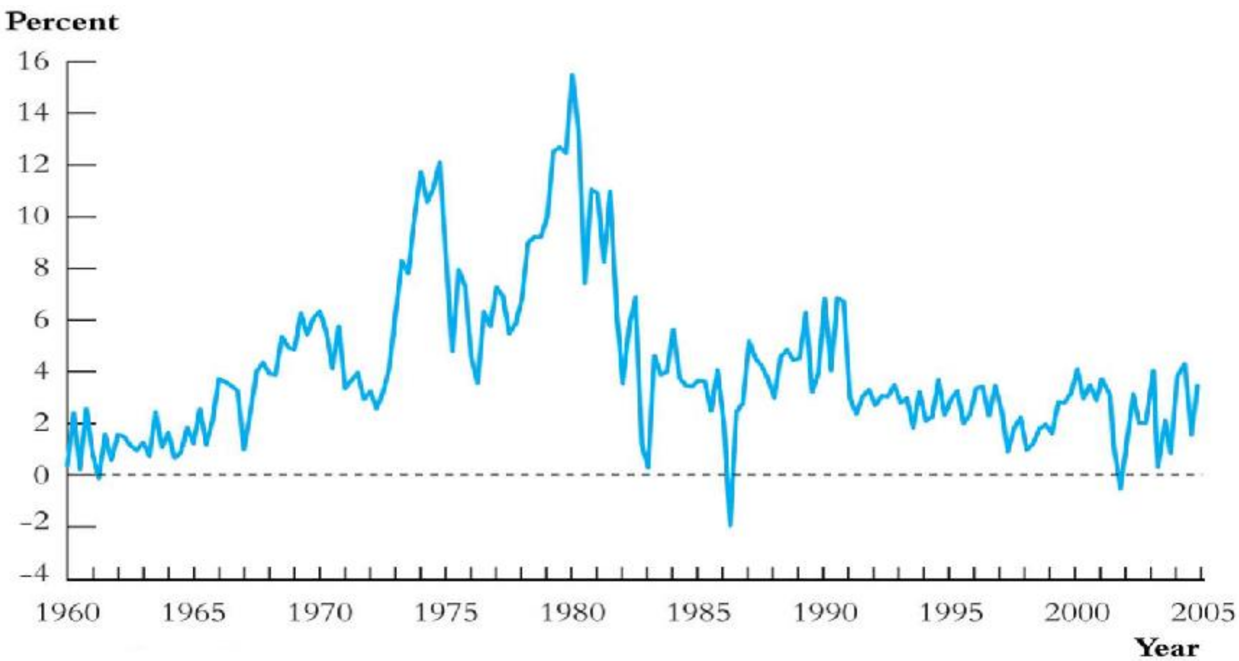
Auto-Regression Analysis

Auto Regression Analysis

- Regression analysis for time-ordered data is known as Auto-Regression Analysis

- Time series data are data collected on the same observational unit at multiple time periods

Example: Indian rate of price inflation

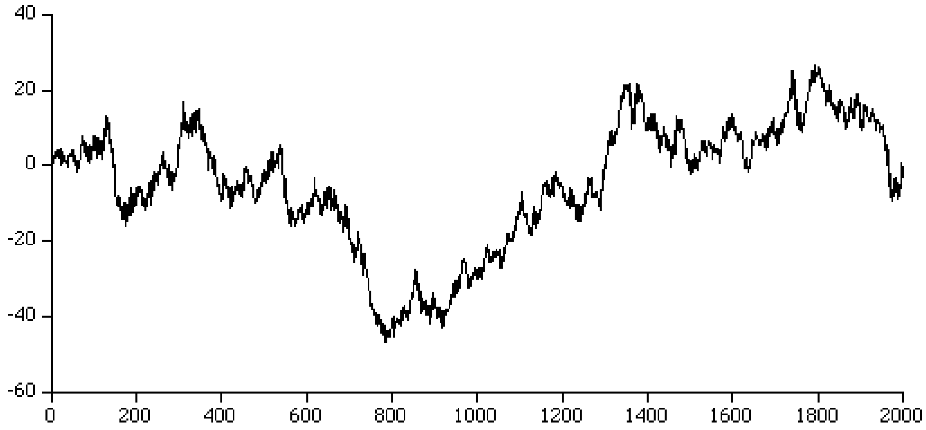


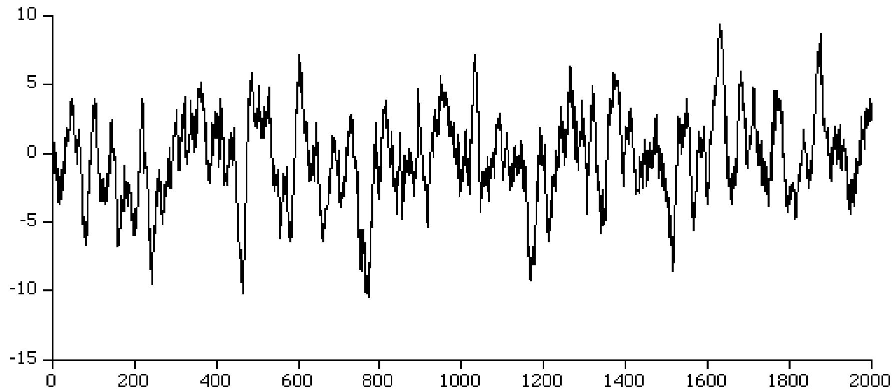
Auto Regression Analysis

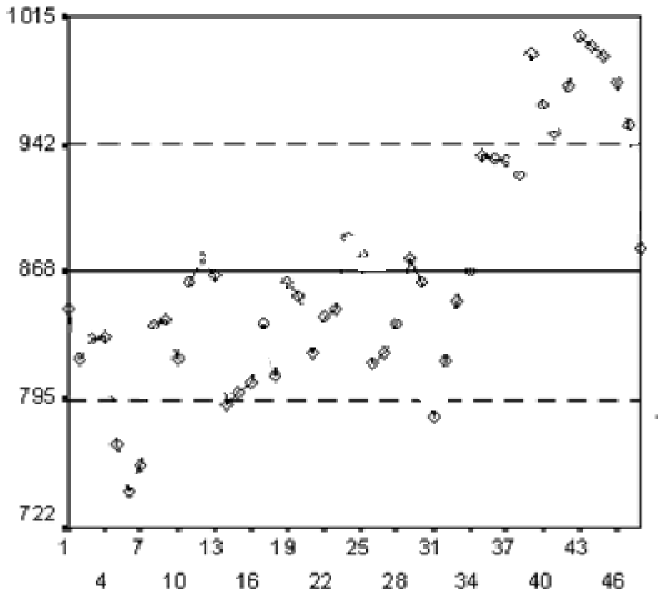
- **Examples:** Which of the following is a time-series data?
  - Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
  - Yen/$, pound/$ and Euro/$ exchange rates (daily data for 1 year = 365 observations)
  - Cigarette consumption per capita in a state, by years
  - Rainfall data over a year
  - Sales of tea from a tea shop in a season

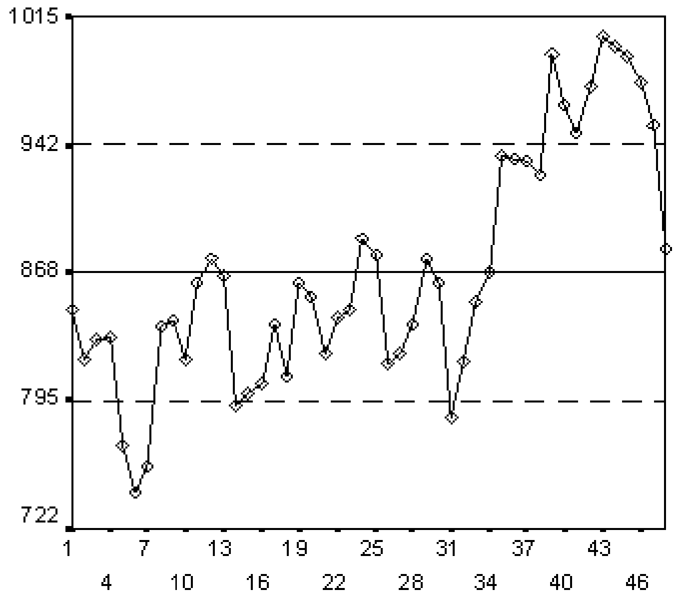
Auto Regression Analysis

- **Examples:** Which of the following graph is due to time-series data?









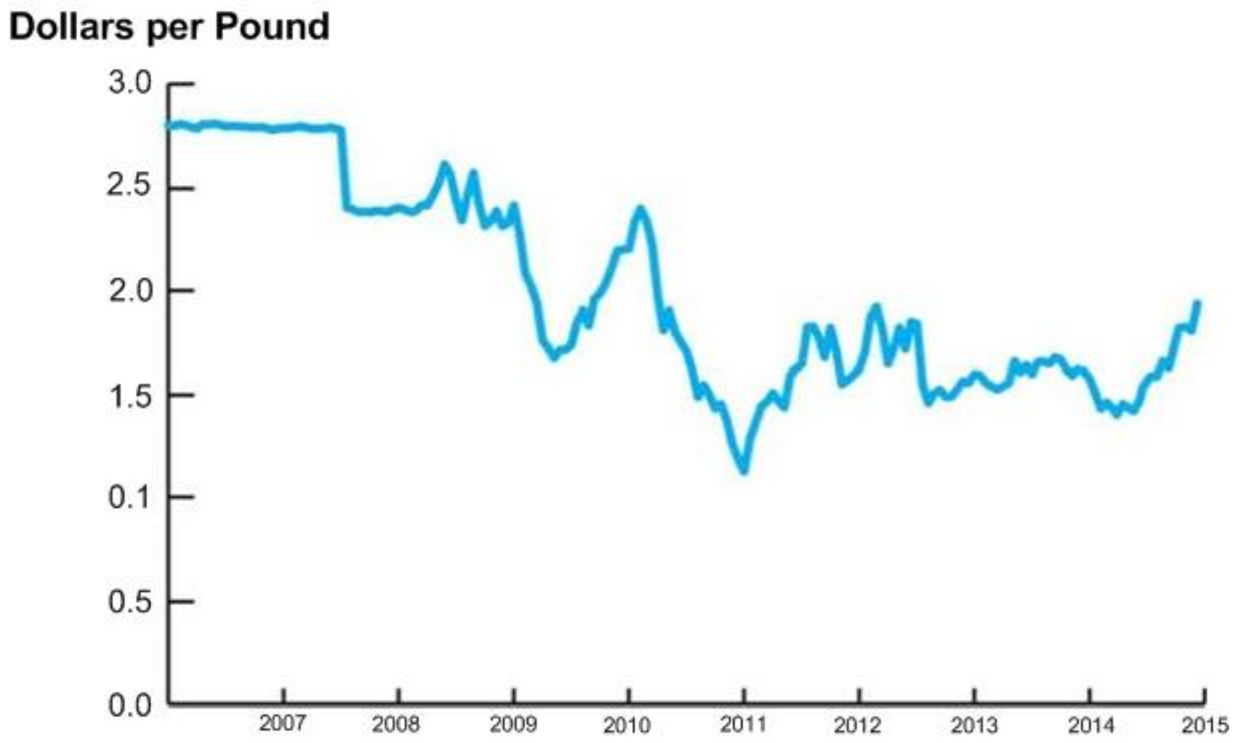
Use of Time Series Data

- To develop forecast model
  - What will the rate of inflation be next year?
- To estimate dynamic causal effects
  - If the rate of interest increases the interest rate now, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?
  - What is the effect over time on electronics good consumption of a hike in the excise duty?
- Time dependent analysis
  - Rates of inflation and unemployment in the country can be observed only over time!

Modeling with Time Series Data

- Correlation over time
  - Serial correlation, also called autocorrelation
  - Calculating standard error
- To estimate dynamic causal effects
  - Under which dynamic effects can be estimated?
  - How to estimate?
- Forecasting model
  - Forecasting model build on regression model

Auto-Regression Model for Forecasting



- Can we predict the tend at a time say 2017?

Some Notations and Concepts

- Y_t = Value of Y in a period t
- Data set [Y_1, Y_2, … Y_{T-1}, Y_T]: T observations on the time series random variable Y
- **Assumptions**
  - We consider only consecutive, evenly spaced observations
    - For example, monthly, 2000-2015, no missing months
  - A time series Y_t is **stationary** if its probability distribution does not change over time, that is, if the joint distribution of (Y_{i+1}, Y_{i+2}, …, Y_{i+T}) does not depend on i.
    - Stationary property implies that history is relevant. In other words, Stationary requires the future to be like the past (in a probabilistic sense).
    - Auto Regression analysis assumes that Y_t is stationary.

Some Notations and Concepts

- There are four ways to have the time series data for AutoRegression analysis
  - **Lag:** The first lag of Y_t is Y_{t-1}, its j-th lag is Y_{t-j}
  - **Difference:** The first difference of a series, Y_t is its change between period t and $t-1$, that is, $y_t = Y_t - Y_{t-1}$
  - **Log difference:** $y_t = \log(Y_t) - \log(Y_{t-1})$
  - **Percentage:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

Some Notations and Concepts

- **Autocorrelation**
  - The correlation of a series with its own lagged values is called autocorrelation (also called serial correlation)

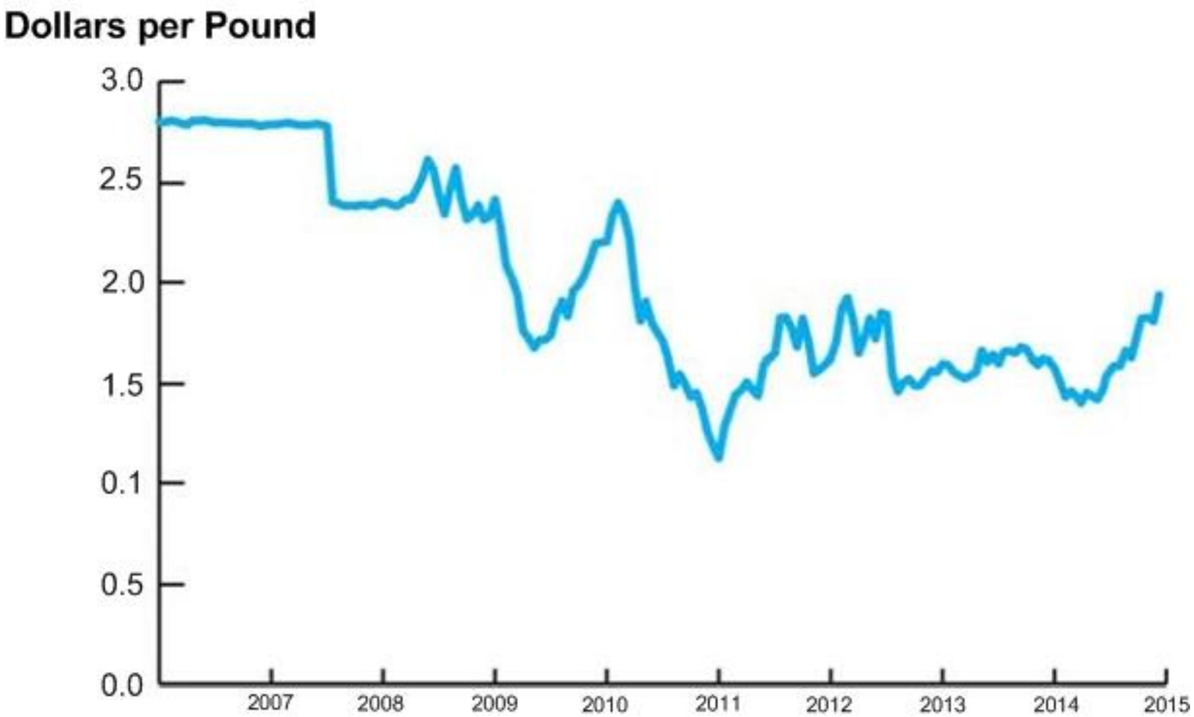
Definition 7.4: ***j*-th Autocorrelation**

The j-th autocorrelation, denoted by ρ_j is defined as

$$\rho_j = \frac{COV(Y_t, Y_{t-j})}{\sqrt{\sigma_{Y_t}\sigma_{Y_{t-j}}}}$$

where, $COV(Y_t, Y_{t-j})$ is the ***j*−th autocovariance**

Some Notations and Concepts



- For the given data, say $\rho_1 = 0.84$
  - This implies that the Dollars per Pound is highly serially correlated
- Similarly, we can determine $\rho_2, \rho_3 \ldots$ etc., and hence different regression analyses

Auto-Regression Model for Forecatsing

- A natural starting point for forecasting model is to use past values of Y, that is, Y_{t-1}, Y_{t-2}, … to predict Y_t
- An autoregression is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the **order** of autoregression
  - In first order autoregression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
  - In p-th order autoregression (denoted as AR(p)), Y_t is regressed against, Y_{t-1}, Y_{t-2}, …,Y_{t-p}

p-th Order AutoRegression Model

Definition 7.5: ***p*-th AutoRegression Model**

In general, the p-th order autoregression model is defined as

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t$$

where, $\beta_0, \beta_1, \ldots, \beta_p$ is called autoregression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gausian white noise

- For example, AR(1) is $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$
- The task in AR analysis is to derive the "best" values for β_i $i = 0, 1, \ldots, p$ given a time series Y_t.

Computing AR Coefficients

- A number of techniques known for computing the AR coefficients
- The most common method is called **Least Squares Method** (LSM)
- The LSM is based upon the Yule-Walker equations

$$\begin{bmatrix} 1 & r_1 & r_2 & r_3 & r_4 & \cdots\cdots & r_{p-2} & r_{p-1} \\ r_1 & 1 & r_1 & r_2 & r_3 & \cdots\cdots & r_{p-3} & r_{p-2} \\ r_2 & r_1 & 1 & r_1 & r_2 & \cdots\cdots & r_{p-4} & r_{p-3} \\ r_3 & r_2 & r_1 & 1 & r_2 & \cdots\cdots & r_{p-5} & r_{p-4} \\ . & . & . & . & . & \cdots\cdots & . & . \\ . & . & . & . & . & \cdots\cdots & . & . \\ . & . & . & . & . & \cdots\cdots & . & . \\ . & . & . & . & . & \cdots\cdots & . & . \\ . & . & . & . & . & \cdots\cdots & . & . \\ r_{p-1} & r_{p-2} & r_{p-3} & r_{p-4} & r_{p-5} & \cdots\cdots & r_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ . \\ . \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ . \\ . \\ . \\ . \\ . \\ r_{p-1} \\ r_p \end{bmatrix}$$

- Here, r_i (i = 1, 2 , 3, …, p-1) denotes the i-th auto correlation coefficient.
- β_0 can be chosen empirically, usually taken as zero.

Reference

- The detail material related to this lecture can be found in

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