FORECASTING

Levels of Data Measurement

- Nominal (Category) — Lowest level of measurement
- Ordinal
- Interval
- Ratio — Highest level of measurement

Nominal Level Data

Numbers are used to classify or categorize

Example: Employment Classification

- 1 for Educator
- 2 for Construction Worker
- 3 for Manufacturing Worker

Example: Ethnicity

- 1 for African-American
- 2 for Anglo-American
- 3 for Hispanic-American
- 4 for Oriental-American

Ordinal Level Data

Numbers are used to indicate rank or order

- Relative magnitude of numbers is meaningful
- Differences between numbers are not comparable

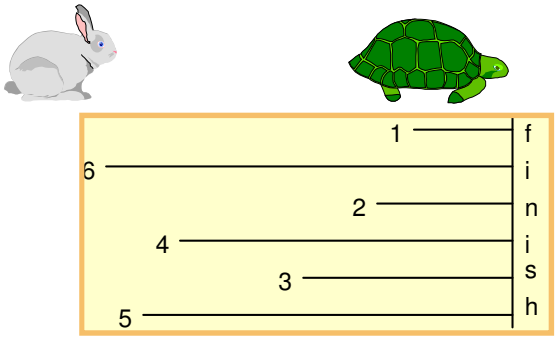
Example: Ranking productivity of employees

Example: Taste test ranking of three brands of soft drink

Example: Position within an organization

- 1 for President
- 2 for Vice President
- 3 for Plant Manager
- 4 for Department Supervisor
- 5 for Employee

Example of Ordinal Measurement



Ordinal Data

Faculty and staff should receive preferential treatment for parking space.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1	2	3	4	5

Interval Level Data

Distances between consecutive integers are equal

- Relative magnitude of numbers is meaningful
- Differences between numbers are comparable
- Location of origin, zero, is arbitrary
- Vertical intercept of unit of measure transform function is not zero

Example: Fahrenheit Temperature
F=32+(9/5)*C

Example: Calendar Time

Example: Monetary Units

Ratio Level Data

Highest level of measurement

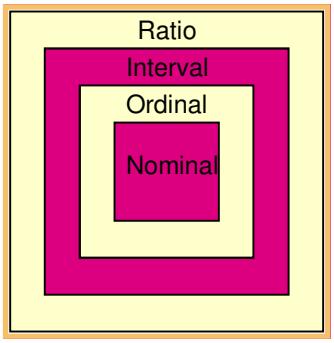
- Relative magnitude of numbers is meaningful
- Differences between numbers are comparable
- Location of origin, zero, is absolute (natural)
- Vertical intercept of unit of measure transform function is zero

Examples: Height, Weight, and Volume

Example: Monetary Variables, such as Profit and Loss, Revenues, and Expenses

Example: Financial ratios, such as P/E Ratio, Inventory Turnover, and Quick Ratio.

Usage Potential of Various Levels of Data



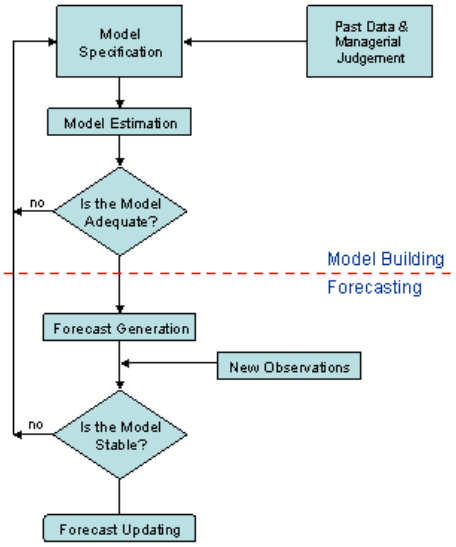
Data Level, Operations, and Statistical Methods

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction, Multiplication, and Division	Parametric
Ratio	All of the above	Parametric

APPROACH

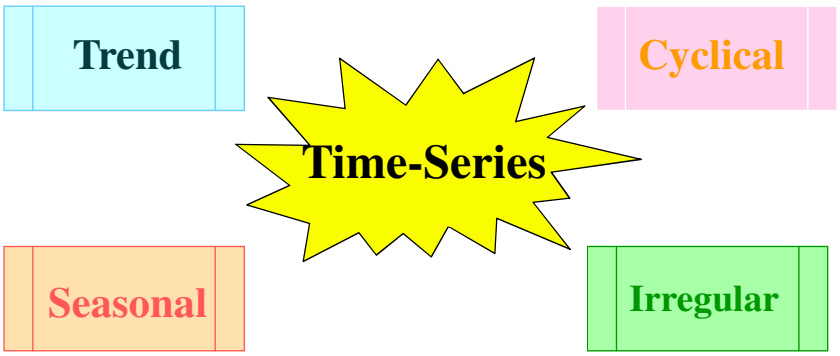
- extrapolation method : based on an inferred study of past general data behavior over time
- explanatory method : based on an analysis of factors which are believed to influence future values

Proses



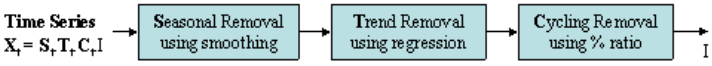
Forecasting System:
The Model-Building and The Forecasting Phases

Time-Series Components



Decomposition Analysis

$$X_t = S_t \,.\, T_t .\, C_t \,.\, I$$



The Three Signals Decomposition and Its Reversal Processes For Forecasting

- **Seasonal variation:** When a repetitive pattern is observed over some time horizon, the series is said to have seasonal behavior. Seasonal effects are usually associated with calendar or climatic changes. Seasonal variation is frequently tied to yearly cycles.
- **Trend:** A time series may be stationary or exhibit trend over time. Long-term trend is typically modeled as a linear, quadratic or exponential function
- **Cyclical variation:** An upturn or downturn not tied to seasonal variation. Usually results from changes in economic conditions.
- **Irregularities** : are any fluctuations not classified as one of the above.

SEASONAL

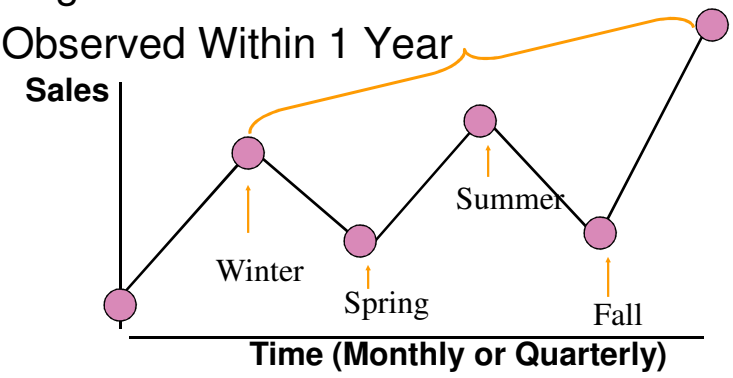
Seasonalities are regular fluctuations which are repeated from year to year with about the same timing and level of intensity. The first step of a times series decomposition is to remove seasonal effects in the data. Without deseasonalizing the data, we may, for example, incorrectly infer that recent increase patterns will continue indefinitely (i.e., a growth trend is present) when actually the increase is 'just because it is that time of the year' (i.e., due to regular seasonal peaks). To measure seasonal effects, we calculate a series of seasonal indexes. A practical and widely used method to compute these indexes is the ratio-to-moving-average approach. From such indexes, we may quantitatively measure how far above or below a given period stands in comparison tot he expected or 'business and usual' data period (the expected data are represented by a seasonal index of 100%, or 1.0).

Seasonal Component

Upward or Downward Swings

Regular Patterns

Observed Within 1 Year



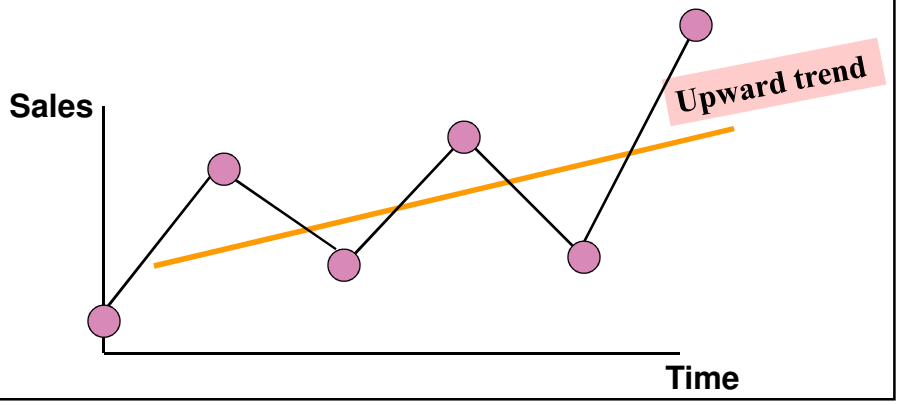
TREND

Trend is growth or decay that are the tendencies for data to increase or decrease fairly steadily over time. Using the deseasonalized data, we now wish to consider the growth trend as noted in our initial inspection of the time series. Measurement of the trend component is done by fitting a line or any other function. This fitted function is calculated by the method of least squares represents the overall trend of the data over time.

Trend Component

Overall Upward or Downward Movement

Data Taken Over a Period of Years



CYCLICAL

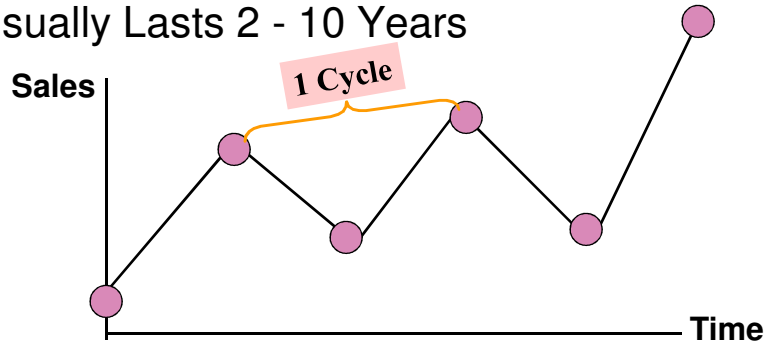
Cyclic oscillations are general up-and-down data changes due to changes e.g., in the overall economic environment (not caused by seasonal effects) such as recession-and-expansion. To measure how the general cycle affects data levels, we calculate a series of cyclic indexes. Theoretically, the deseasonalized data still contains trend, cyclic, and irregular components. Also, we believe predicted data levels using the trend equation do represent pure trend effects. Thus, it stands to reason that the ratio of these respective data values should provide an index which reflects cyclic and irregular components only. As the business cycle is usually longer than the seasonal cycle, it should be understood that cyclic analysis is not expected to be as accurate as a seasonal analysis. Due to the tremendous complexity of general economic factors on long term behavior, a general approximation of the cyclic factor is the more realistic aim. Thus, the specific sharp upturns and downturns are not so much the primary interest as the general tendency of the cyclic effect to gradually move in either direction. To study the general cyclic movement rather than precise cyclic changes (which may falsely indicate more accurately than is present under this situation), we 'smooth' out the cyclic plot by replacing each index calculation often with a centered 3-period moving average.

Cyclical Component

Upward or Downward Swings

May Vary in Length

Usually Lasts 2 - 10 Years



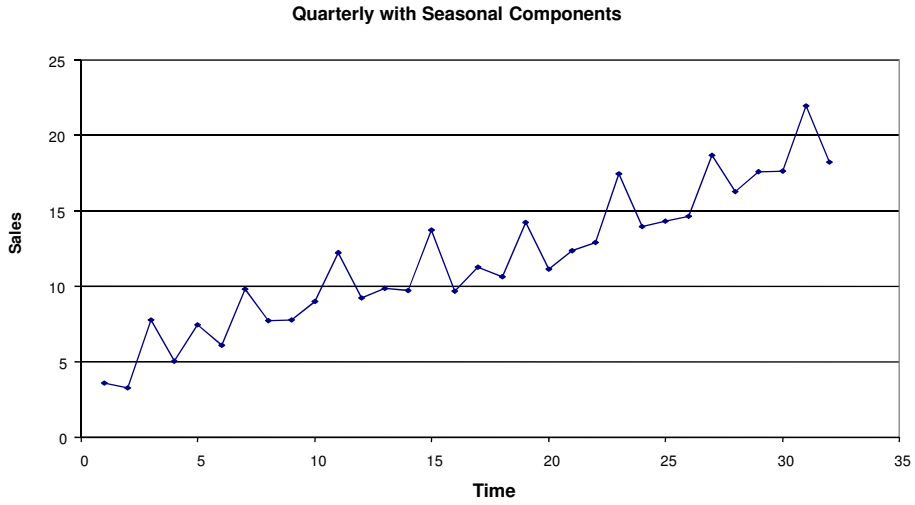
IRREGULAR

Irregularities (**I**) are any fluctuations not classified as one of the above. This component of the time series is unexplainable therefore it is unpredictable. Estimation of **I** can be expected only when its variance is not too large. *Otherwise, it is not possible to decompose the series*. If the magnitude of variation is large the projection for the future values will be inaccurate. The best one can do is to give a probabilistic interval for the future value give the probability of **I** is known.

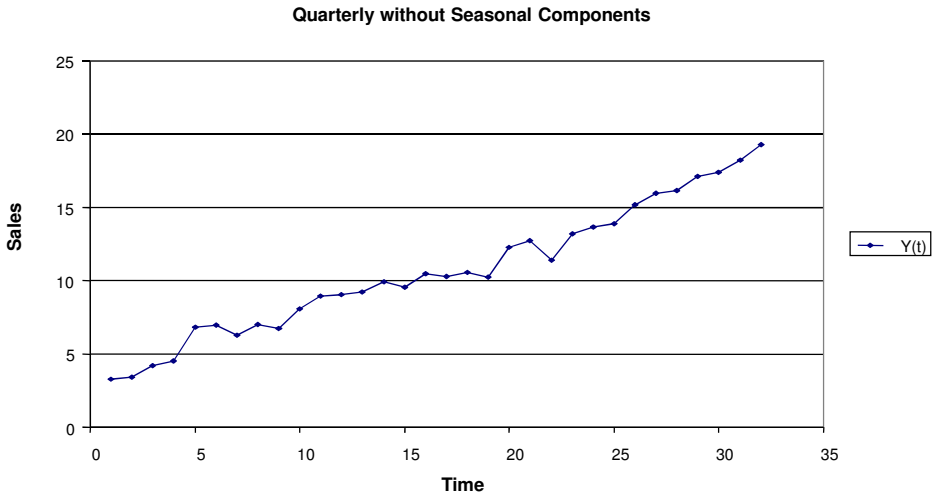
PROCEDURE

- Step 1: Compute the future trend level using the trend equation.
- Step 2: Multiply the trend level from Step 1 by the period seasonal index to include seasonal effects.
- Step 3: Multiply the result of Step 2 by the projected cyclic index to include cyclic effects and get the final forecast result.

E.g. Quarterly Retail Sales with Seasonal Components



E.g. Quarterly Retail Sales with Seasonal Components Removed



DATA

- Kondisi data dapat diperbandingkan (dari satu sumber, memperhatikan perbedaan hari dalam bulan, kenaikan harga, pertumbuhan jumlah penduduk, dll)
- Menghilangkan pengaruh outlier
- Perhatikan adanya kejadian khusus seperti wabah penyakit, pemilu, perang, bencana alam, dll

TREND

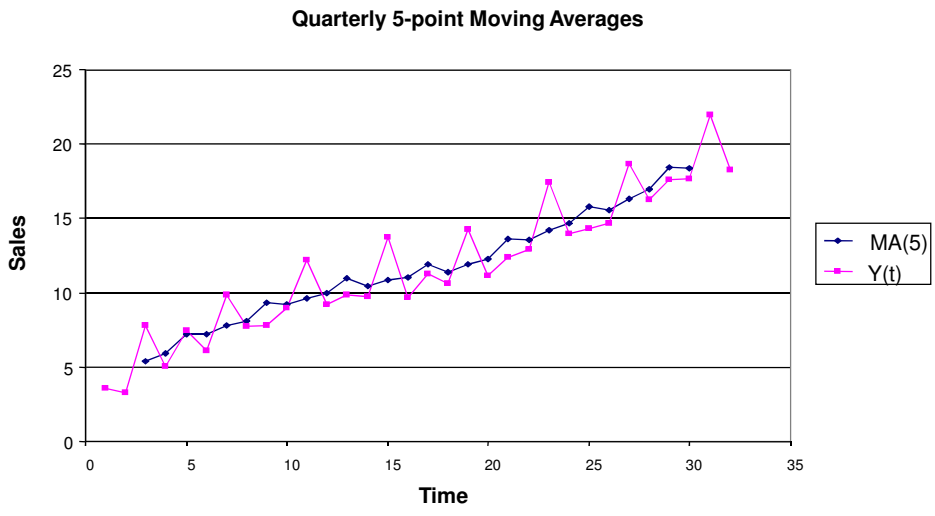
- Regresi Linier : $Y = a + bX + cX$
- Kuadratik : $Y = a + bX + cX^2$
- Eksponensial : $Y = ab^X$
- Gompertz : $Y = {}_{ab}c^X$
- Pearl-Reed : $Y = a/(1+e^{b+cX})$
- dll

SMOOTHING

- Moving Average
- Weighted Moving Average

- Untuk mengurangi pengaruh seasonal dan sebagian cyclical
- Mempunyai kelemahan karena awal data dan akhir data menjadi tidak dapat digunakan

E.g. 5-point Moving Averages of Quarterly Retail Sales



SEASONAL

- Metode Rata-Rata Sederhana :

  $S \approx [(T.S.C.I)/n] - T$ dimana n : jumlah tahun

- Metode Persentase dari Trend (Falkner's method) :

  $S \approx [(T.S.C.I)/T] / n$

- Metode Rasio terhadap rata-rata bergerak :

  $S \approx [(T.S.C.I)/(T.C)] / n$ dengan data diolah "moving average" terlebih dahulu

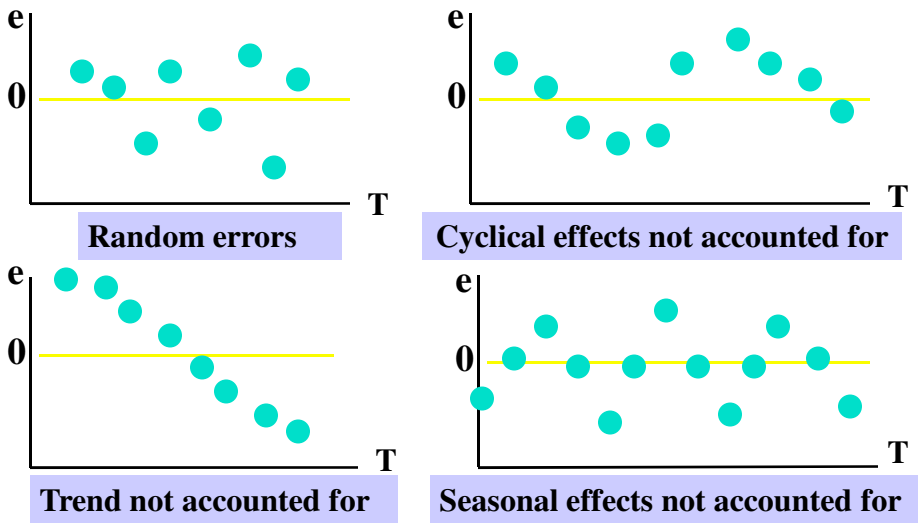
CYCLICAL

- Serupa dengan Seasonal

  $C.I \approx [(T.S.C.I)/(T.S)]$

- Sebaiknya menggunakan data bertahun-tahun
- Relatif sulit karena kemunculan dan lama perilakunya tidak dapat diperkirakan
- Sampai saat ini belum ada metode yang memuaskan

Residual Analysis



Autoregressive Modeling

Used for Forecasting

Takes Advantage of Autocorrelation

- 1st order - correlation between consecutive values
- 2nd order - correlation between values 2 periods apart

Autoregressive Model for *p*-th order:

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \cdots + A_p Y_{i-p} + \delta_i$$

δ_i: Random Error

Autoregressive Model: Example

The Office Concept Corp. has acquired a number of office units (in thousands of square feet) over the last 8 years. Develop the 2nd order Autoregressive model.

Year	Units
93	4
94	3
95	2
96	3
97	2
98	2
99	4
00	6

Autoregressive Model: Example Solution

Develop the 2nd order table

Use Excel to estimate a regression model

Excel Output

Year	Y_i	Y_{i-1}	Y_{i-2}
93	4	---	---
94	3	4	---
95	2	3	4
96	3	2	3
97	2	3	2
98	2	2	3
99	4	2	2
00	6	4	2

$$\hat{Y}_i = 3.5 + .8125Y_{i-1} - .9375Y_{i-2}$$

Autoregressive Model Example: Forecasting

Use the 2nd order model to forecast number of units for 2001:

$$\hat{Y}_i = 3.5 + .8125Y_{i-1} - .9375Y_{i-2}$$

$$\begin{aligned}\hat{Y}_{2001} &= 3.5 + .8125Y_{2000} - .9375Y_{1999} \\ &= 3.5 + .8125 \times 6 - .9375 \times 4 \\ &= 4.625\end{aligned}$$

Principal of Parsimony

Suppose 2 or More Models Provide Good Fit to Data

Select the Simplest Model!

Pitfalls Concerning Time-Series Forecasting

- Taking for Granted the Mechanism that Governs the Time Series Behavior in the Past will Still Hold in the Future
- Using Mechanical Extrapolation of the Trend to Forecast the Future Without Considering Personal Judgments, Business Experiences, Changing Technologies, Habits, etc.

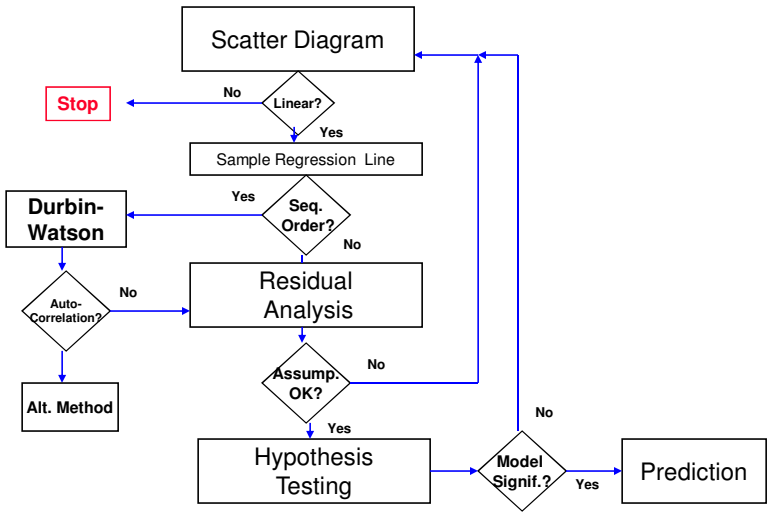
ANALISIS REGRESI

- Ditujukan untuk forecasting atau untuk mencari korelasi (asosiasi)
- Asosiasi tidak menunjukkan hubungan sebab akibat. Kejadian asosisasi dipicu oleh variabel lain (*lurking* variable) yang mungkin tidak tampak.
- Hubungan sebab-akibat boleh disimpulkan dari hasil eksperimen terkendali.

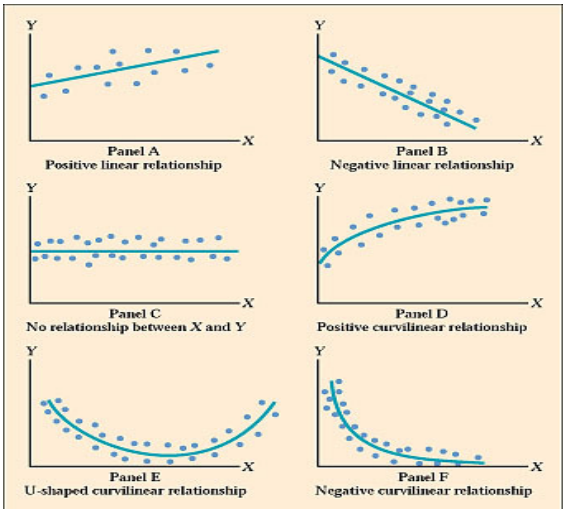
Purpose of Regression

- Predict values of a dependent (response) variable based on values of one or more independent (explanatory) variables
- Regression model is a statistical equation derived from a sample data set
- Valid over relevant range of independent variables

Linear Regression Flowchart



Types of Regression Models -- Scatter Diagram



Linear Regression Model

Relationship Between Variables Is a <u>Linear Function</u>

Y-Intercept

Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

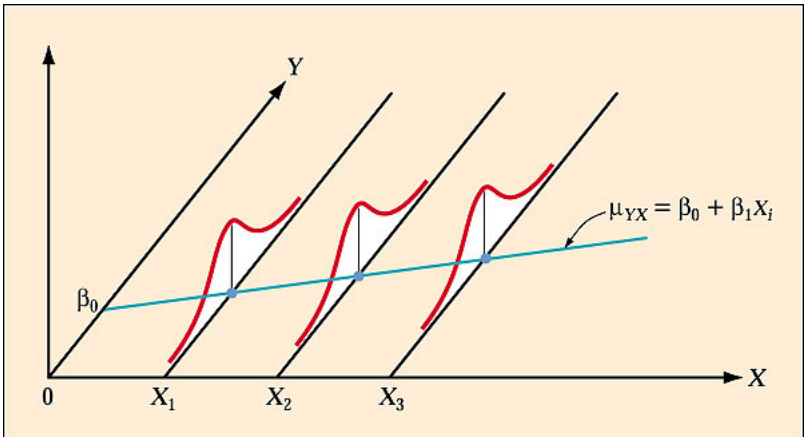
Dependent Variable

Independent Variable

Linear Regression Assumptions

1. Normality of Errors
   - Distribution of Errors around regression line is Normal
   - Evaluate using a histogram, normal probability plot, etc. of Studentized Residuals.
2. Homoscedasticity (Constant Variance)
   - Evaluate from observing residual plots -- desired outcome is constant range of variation
3. Independence of Errors
   - Evaluate from observing residual plots -- desired outcome is no pattern

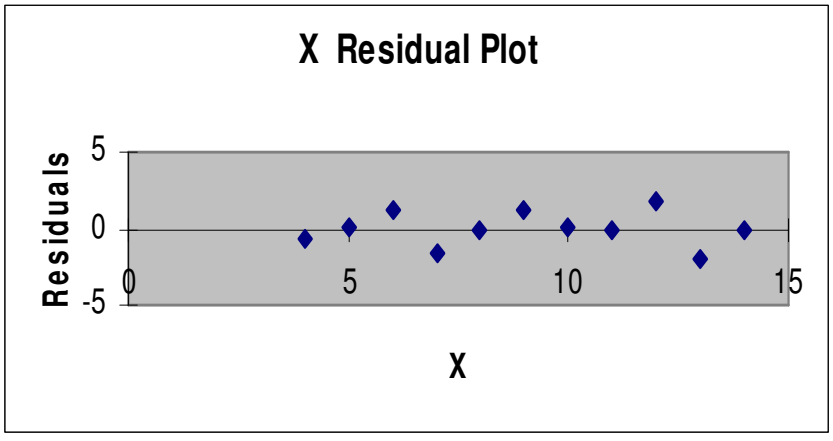
Linear Regression Assumptions



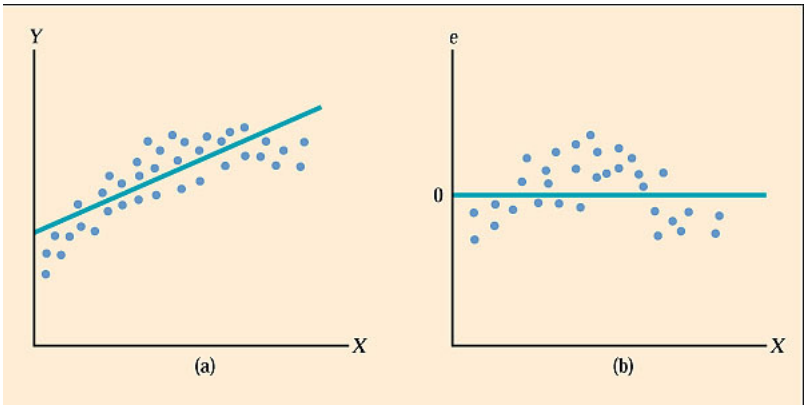
Residual Analysis

1. Graphical Analysis of Residuals ("errors")
   - Residuals = Difference between actual Y_i & predicted Y_i
   - Plot residuals vs. X_i values
2. Purpose
   - Examine functional form (linear vs. non-linear)
   - Evaluate violations of assumptions
     - Homoscedasticity
     - Independence of errors

Residual Analysis



Residual Analysis



Durbin-Watson Procedure

1. Used to Detect Autocorrelation
   - Residuals in one time period are related to residuals in another period
   - Violation of independence assumption
2. Durbin-Watson Test Statistic

Durbin Watson test

A test for serially correlated (or autocorrelated) residuals. One of the assumptions of regression analysis is that the residuals for consecutive observations are uncorrelated. If this is true, the expected value of the Durbin-Watson statistic is 2. Values less than 2 indicate positive autocorrelation, a common problem in time-series data. Values greater than 2 indicate negative autocorrelation.

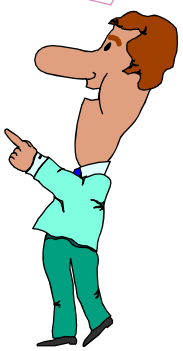
The d-statistic has values in the range [0,4]. Low values of d are in the region for positive autocorrelation. Values of d that tend towards 4 are in the region for negative autocorrelation.

Coefficient of Determination

Proportion of Variation 'Explained' by Relationship Between X & $\hat{Y}$

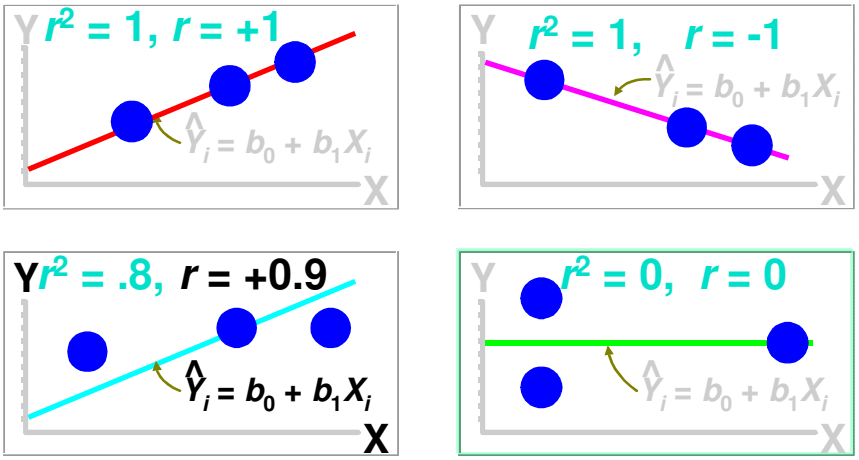
$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{SSR}}{\text{SST}}$$

$0 \le r^2 \le 1$



Ability of equation to "fit" the data

Coefficients of Determination (r^2) and Correlation (r)



KRITERIA R^2

Model	Variables	R Squared	Adjusted R Sq.
1	1	0.925	0.924
2	2	0.931	0.93
3	3	0.935	0.933
4	4	0.942	0.94
5	5	0.947	0.944
6	4	0.946	0.944

Model 6 lebih baik!

- Model semakin bagus bila nilai R^2 mendekati 1
- Bila membandingkan beberapa model yang memiliki perbedaan jumlah sampel atau beda jumlah prediktor maka digunakan Adjusted R^2

KRITERIA *t*

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Constant	82.677	0.626	0	132.176	0
	Fertility	-0.662	0.263	-0.119	-2.518	0.013
	Infant Mortality	-0.24	0.013	-0.863	-18.326	0

Model : female life expectancy = 82.677 - 0.662 fertility - 0.240 infant mort

infant mort. lebih penting!

- Tingkat penting variabel prediktor diukur dari uji t dengan nilai di bawah -2 atau di atas 2. Semakin jauh dari nilai-nilai tersebut, semakin penting sebuah prediktor
- Koefisien tidak menunjukkan tingkat penting suatu prediktor

STUDI KASUS

Perkiraan kualitas layanan dengan mengetahui jumlah pelanggan.

- Prediktor apa yang sebaiknya dipakai? Apakah konsisten untuk dibandingkan dari waktu ke waktu?
- bagaimana bentuk distribusi variable dependent? Apakah perlu ditransformasi?
- Bagaimana hasil analisis scatter plot? Apakah ada outlier? Apakah variansi stabil? Apakah linier?