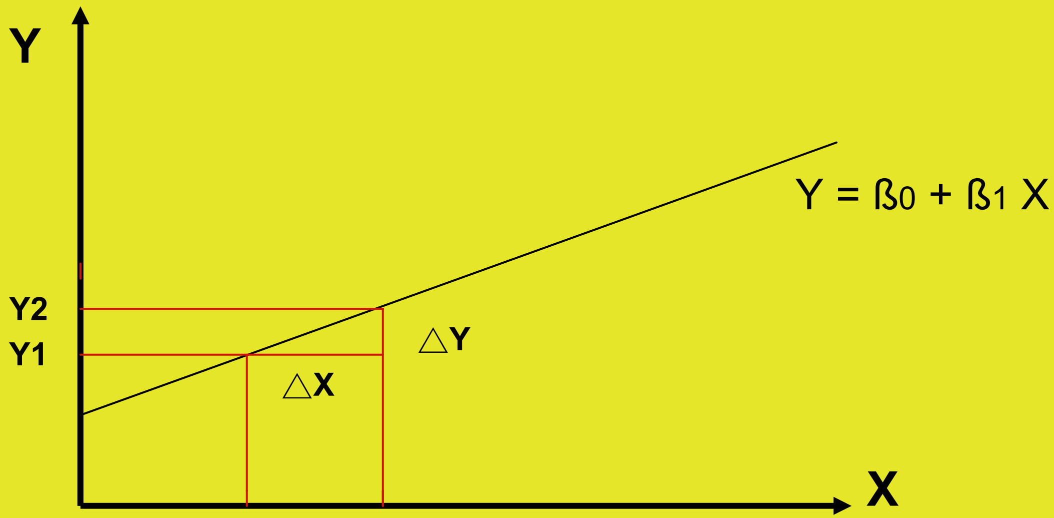
Statistical Data Mining



Regression Analysis (part 1)

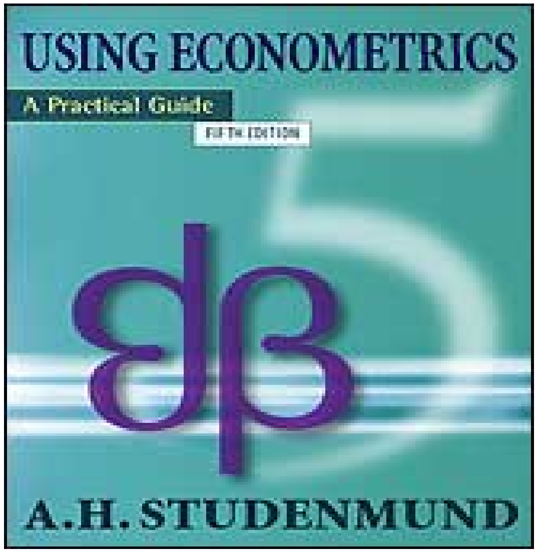
Professor Dr. Gholamreza Nakhaeizadeh

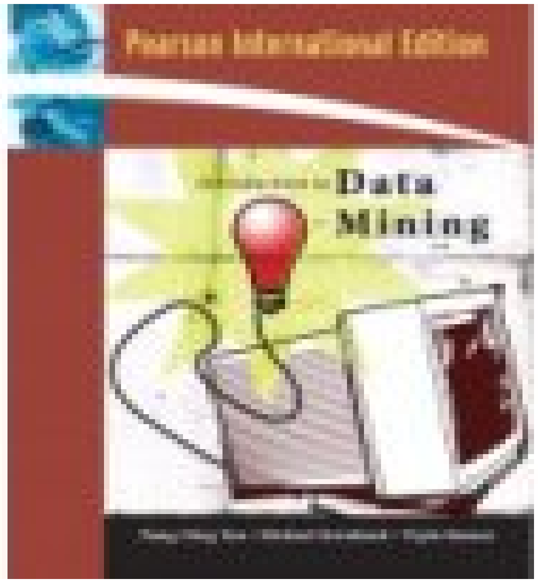
Content

Regression Analysis (Part 1)

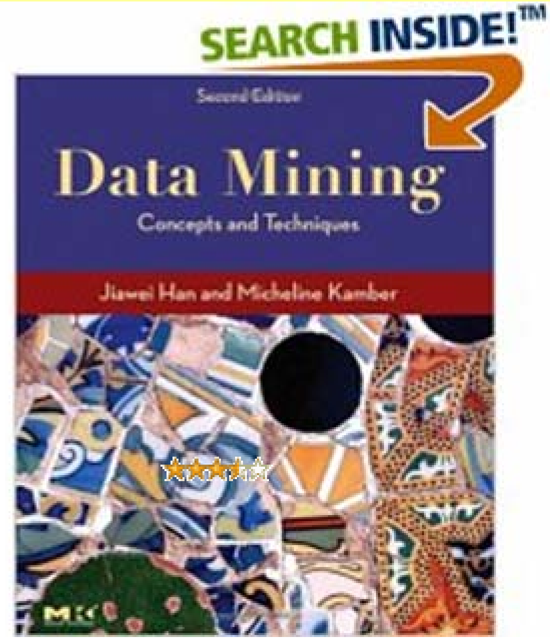
- Literature used
- Regression analysis
- introduction
- Simple linear regression
- Stochastic aspects in Simple linear regression
- Multivariate linear regression
- Matrix representation
- OLS-Estimators in Simple linear regression
- Overall fit of the estimated regression
- Coefficient of Determination
- Simple Correlation Coefficient

Literatur used (1)



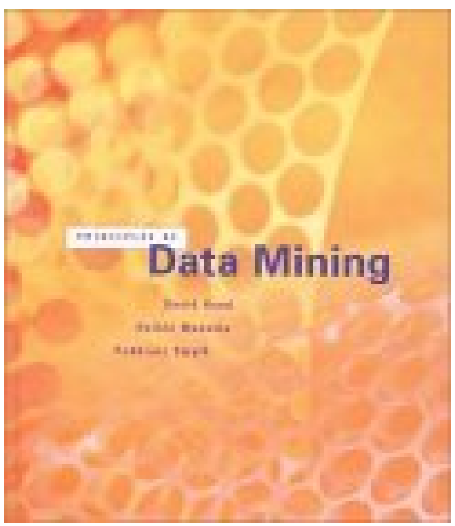


Pang-Ning Tan,
Michael Steinbach,
Vipin Kumar



Jiawei Han and
Micheline Kamber

Schneeweiss:
***Ökonometrie.* 1990**
Physica-Verlag



Principles of Data Mining
David J. Hand, Heikki Mannila,
Padhraic Smyth

Literature Used (2)

http://www2.chass.ncsu.edu/garson/PA765/regress.htm
http://www.statsoft.com/textbook/stmulreg.html
http://www.statsoft.com/textbook/glosfra.html?glosm.html&1

Regression Analysis

Data Mining Algorithms

Machine Learning

- Rule Based Induction
- Decision Trees
- Neural Networks
- Conceptional clustering
- ....

Statistics

- Discriminant Analysis
- Cluster Analysis
- Regression Analysis
- Logistic Regression Analysis
- ....

Database Technology

- Association Rules
- ....

Regression Analysis

Introduction

- **Tools for prediction and causal analysis based on Supervised Learning**

- Regression function, **y = f (X),** maps a set of attributes **X** known also as **exogenous, independent or explanatory variables**

  into an output **y** known also as **endogenous dependent , response, or target variable** by learning from the tuples observed for X and y

Regression Analysis

Introduction

- The aim is to use the input data to perform the **best estimation** for **y** with **minimum error**
- **Time Series and Cross-Section aspects regarding prediction**
- Endogenous variable must be **continuous-valued** but the exogenous variables can be **nominal or continuous**
- Estimation of parameters and their Significant tests are based on **statistical methods**

Regression and Artificial Neural Networks

Regression Analysis

Introduction

Estimation Method

- Error Function: Sum of **squared errors**

$$\sum_i [\, y_i - f(X_i) \,]^2$$

- Estimation on **the training data**, assessment on the **test data** or **validation data**

- In **Stepwise** regression **backward** and **forward** possible (like pruning in DT)

Regression and Artificial Neural Networks

Regression Analysis **Introduction**

Examples of applications

- Prediction of the family consumption using other indicators like, income, price, family size, living place
- Prediction of stock market index by applying other economic indicators
- Prediction of the air temperature based on other atmospheric factors
- Trend Prediction

Regression Analysis

Single-Equation Linear Models

$$Y = \overbrace{\beta_0 + \beta_1}\, X \qquad (1)$$

ßs are the coefficients
ß0 : Constant or intercept
$X=0 \longrightarrow Y = \beta_0$

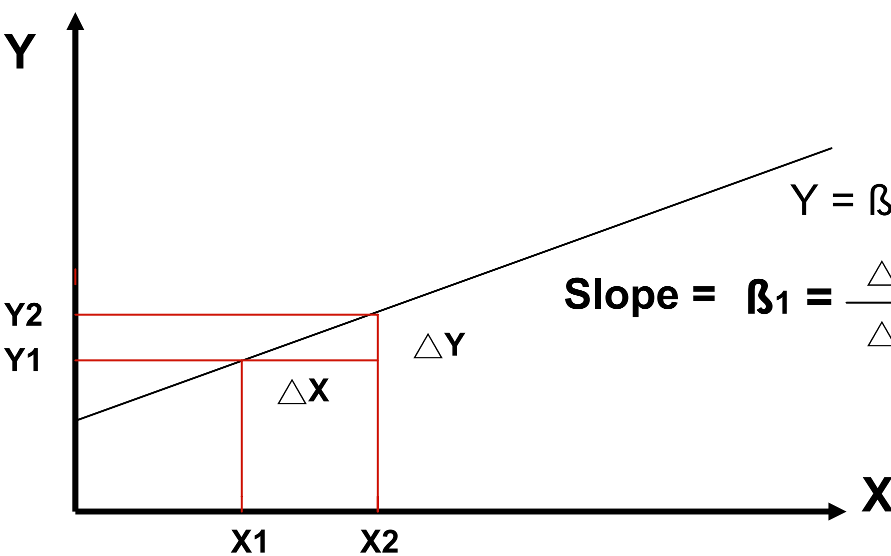
β_1 : slope coefficient

X increases by one unit → Y increases by β_1

Making nonlinear equations linear

$Y = \beta_0 + \beta_1 X^2$

$Z = X^2 \qquad Y = \beta_0 + \beta_1 Z$



$Y = \beta_0 + \beta_1 X$

Slope = $\beta_1 = \dfrac{\triangle Y}{\triangle X}$

Regression Analysis

The stochastic Error Term

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{deterministic component}} + \epsilon \longrightarrow \text{Stochastic error term} \qquad (2)$$

Stochastic error term must be preset, because

- All relevant explanatory variables are not considered
- Measurement error
- Misspecification of functional form
- …..

$$E(\epsilon \mid X) = 0 \qquad (3)$$

$$E(Y \mid X) = \beta_0 + \beta_1 X \qquad (4)$$

Regression Analysis

Consideration of the observations

$Y_i = ß_0 + ß_1 X_i + \epsilon_i$ (i = 1, 2, 3,……, n) n : Number of observations

Y_i : the ith observation of the dependent variable

X_i : the ith observation of the independent variable

ϵ_i : the ith *observation* of the stochastic error term

$$\left.\begin{aligned} Y_1 &= ß_0 + ß_1 X_1 + \epsilon_1 \\ Y_2 &= ß_0 + ß_1 X_2 + \epsilon_2 \\ &\dots\dots\dots\dots\dots\dots \\ Y_n &= ß_0 + ß_1 X_n + \epsilon_n \end{aligned}\right\}$$

The coefficients $ß_0$ and $ß_1$ do not change from observation to observation

Regression Analysis

General Case: Multivariate Regression Equation

$$Y_i = ß_0 + ß_1 X_{1i} + ß_2 X_{2i} + \ldots + ß_m X_{mi} + \epsilon_i \qquad (5)$$

One unit increase in the independent variable X_k

Change in the dependent variable Y is equal to $ß_k$, holding constant the other independent variables

Regression and Artificial Neural Networks

Regression Analysis

Multivariate Linear Regression

$$Y_i = ß_0 + ß_1 X_{1i} + ß_2 X_{2i} + \dots + ß_m X_{mi}$$

Observations

$$\begin{array}{cccccccc} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1m} & & Y_1 \\ \dots & & & & & & & \dots \\ X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{im} & & Y_i \\ \dots & & & & & & & \dots \\ X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{nm} & & Y_n \end{array}$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1m} \\ 1 & X_{21} & X_{22} & \dots & X_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nm} \end{pmatrix}$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{pmatrix} \qquad \mathbf{ß} = \begin{pmatrix} ß_0 \\ ß_1 \\ . \\ . \\ ß_m \end{pmatrix}$$

Matrix notation

$$\mathbf{Y = X\,ß}$$

Regression Analysis

Ordinary Least Square

In the regression equation $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

The parameters β_0 and β_1 are unknown
they can be estimated by using the observations of Y and X

$\hat{\beta}_0$ and $\hat{\beta}_1$: Estimates of β_0 and β_1

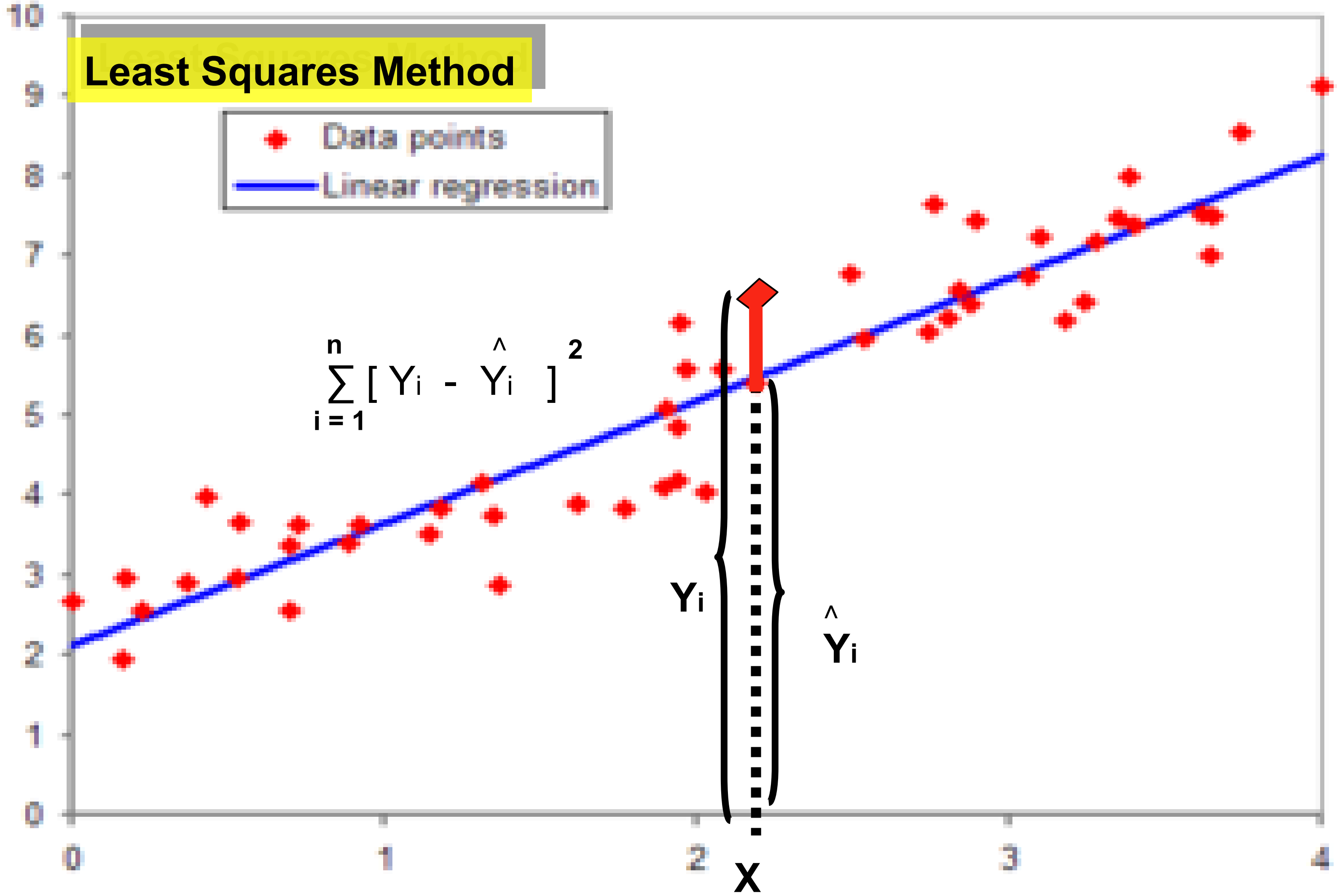
$\hat{Y}_i$: Estimate of Y_i

and $\hat{\epsilon}_i = Y_i - \hat{Y}_i$: residual

OLS: Determine $\hat{\beta}_0$ and $\hat{\beta}_1$ so that is minimized $\sum_{i=1}^{n} \hat{\epsilon}_i^2$

OLS is relatively easy and OLS – estimates have useful characteristics

Least Squares Method



Regression Analysis

OLS-estimates for single-equation linear model

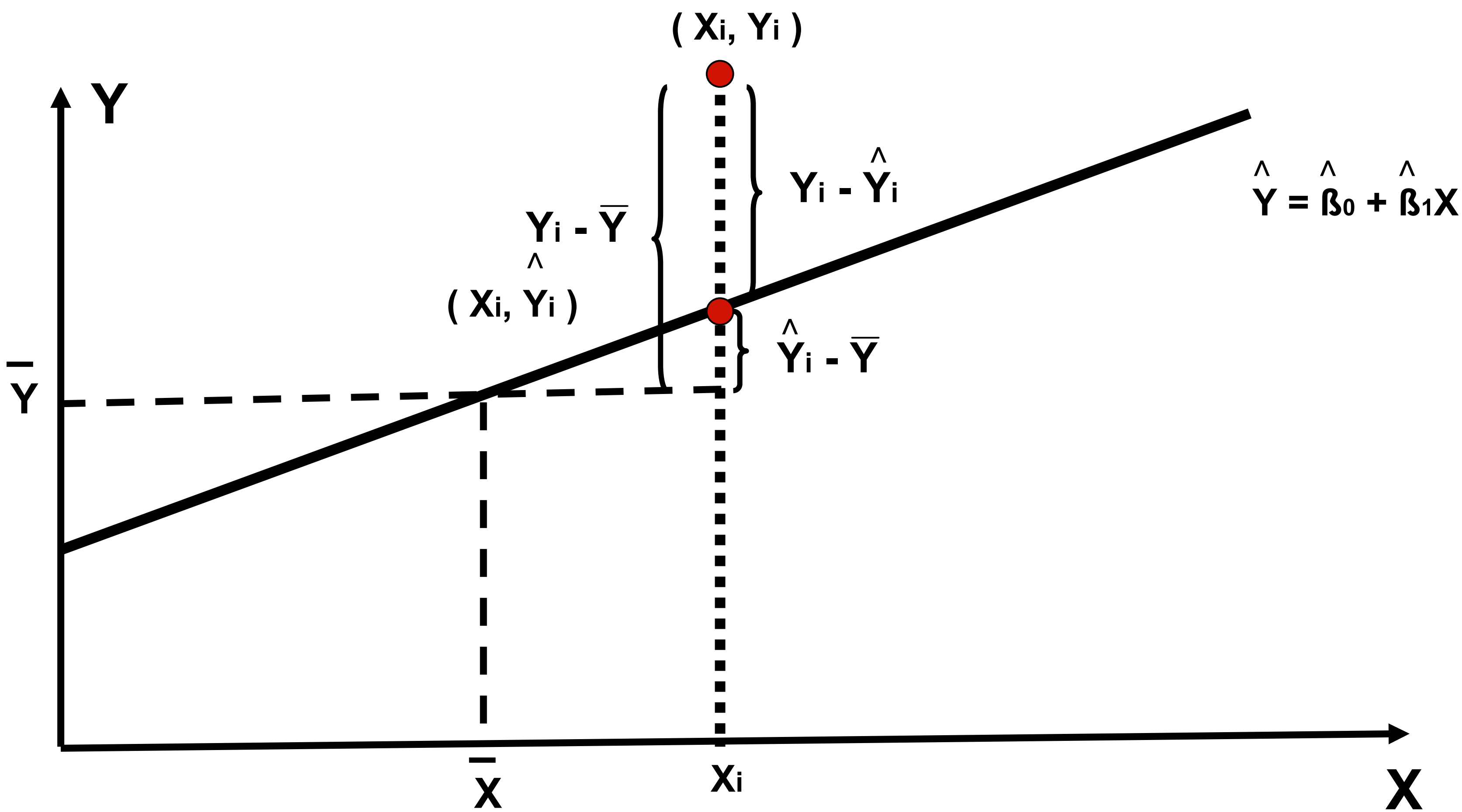
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad , \quad \hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \qquad (6)$$

with

$$\left.\begin{aligned} \bar{X} &= 1/n \sum_{i=1}^{n} X_i \\ \bar{Y} &= 1/n \sum_{i=1}^{n} Y_i \end{aligned}\right\}$$

Regression Analysis

Overall fit of the estimated regression **Decomposition of Variance**



Regression Analysis

Overall fit of the estimated regression

Decomposition of Variance

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \quad (7)$$

Total Sum of Squares
(TSS)

Residual Sum of Squares
(RSS)

Explained Sum of Squares
(ESS)

TSS = ESS + RSS

Smaller RSS to TSS $\longrightarrow$ better the estimated regression fits the data

Regression Analysis

Overall fit of the estimated regression | **Coefficient of Determination**

TSS = ESS + RSS

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (Y_i - \hat{Y})^2}{\sum (Y_i - \bar{Y})^2} \qquad (8)$$

From (7) and (8) $\quad 0 \leq R^2 \leq 1 \qquad (9)$

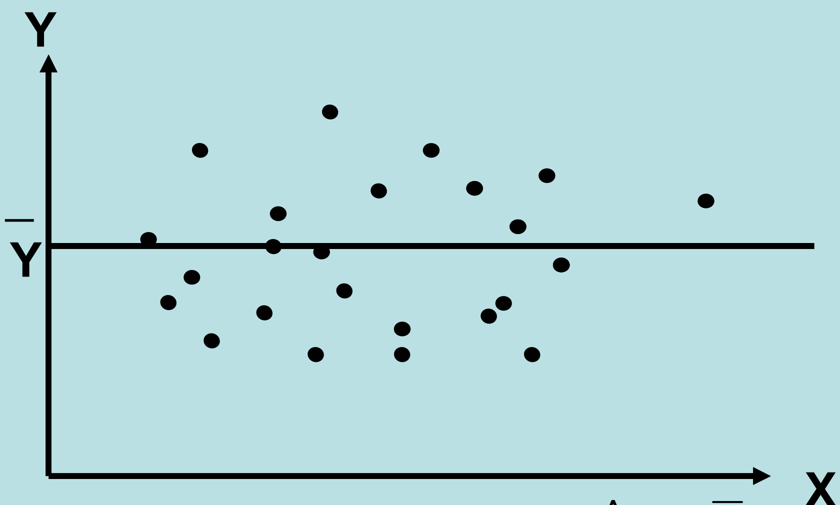
Value of R^2 close to one $\longrightarrow$ excellent overall fit

Value of R^2 close to zero $\longrightarrow$ very poor fit

Regression Analysis

Overall fit of the estimated regression

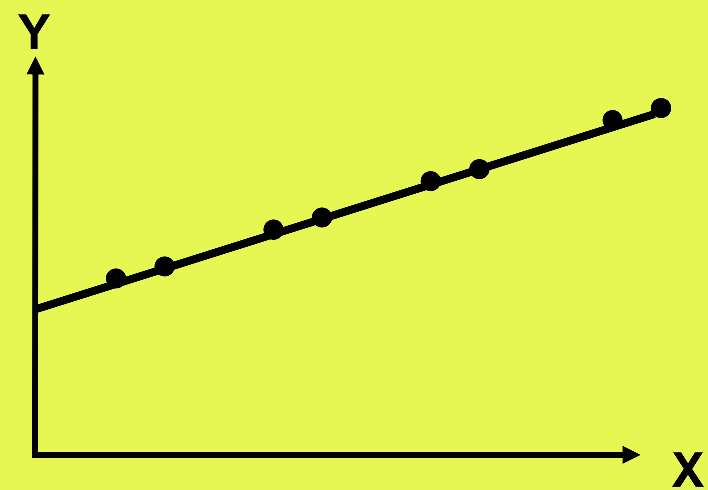
Two extreme cases



Estimated Regression : $\hat{Y} = \bar{Y}$

$R^2 = 0$ see (8)

X and Y are not related

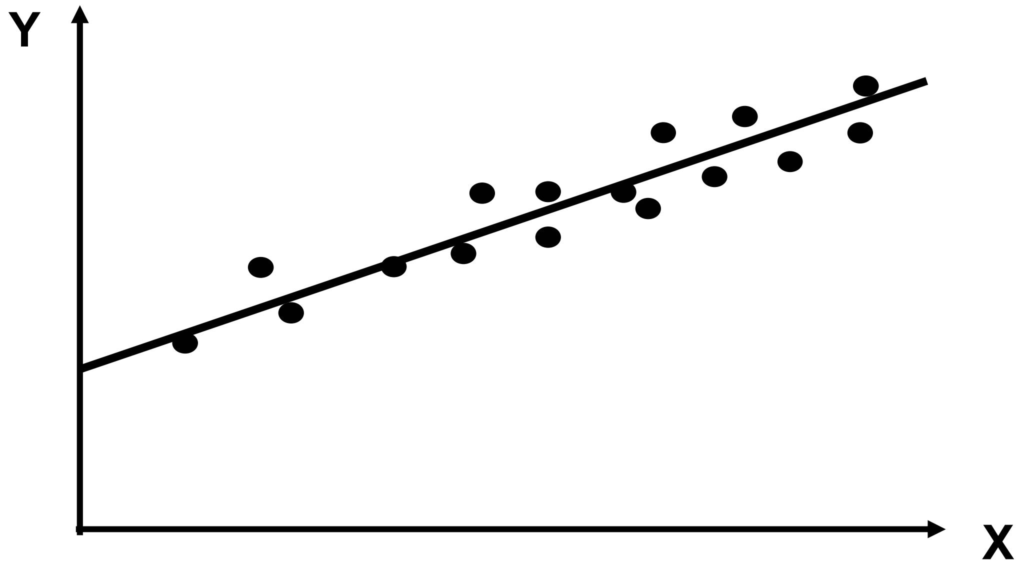


$R^2 = 1$

all the observations are on the regression line

Regression Analysis

Overall fit of the estimated regression | **Coefficient of Determination**



R^2 : very close to one
very good fit

Regression Analysis

Overall fit of the estimated regression

Adjusted Coefficient of Determination

R^2 is biased to the number of independent variables

More independent variables $\longrightarrow$ higher R^2

Solution: Adjusted R^2

$$\bar{R}^2 = 1 - \frac{\sum (Y_i - \hat{Y})^2 / (n - k - 1)}{\sum (Y_i - \bar{Y})^2 / (n - 1)}$$

k : Number of independent variables

- Normally, $\bar{R}^2$ is used to compare the goodness of fit of regression equations with different numbers of independent variables
- $\bar{R}^2$ is not a percent but an index

Regression Analysis

Overall fit of the estimated regression | **Simple Correlation Coefficient**

$$r_{X,Y} = \frac{\sum [(X_i - \overline{X})(Y_i - \overline{Y})]}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}} \qquad (10)$$

$$-1 \leq r \leq +1$$

X and Y are perfectly positively correlated, then r = +1

X and Y are perfectly negatively correlated, then r = -1

X and Y are totally uncorrelated, then r = 0

Regression Analysis

Simple linear regression model

Model assumptions

$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ for $i = 1, 2, \ldots n$

Assumption 1:

$E(\varepsilon_i \mid x) = 0$ for $i = 1, 2, \ldots n$

Assumption 2:

$V(\varepsilon_i \mid x) = \sigma^2$ for $i = 1, 2, \ldots n$ $\rightarrow$

Variance of ϵ_i is constant for $i=1,2,\ldots n$

Homoscedasticity

(Heteroscedasticity)

Assumption 3:

$E(\varepsilon_i \varepsilon_j \mid x) = 0$ for $i \neq j$ $i, j = 1, 2, \ldots n$

$\downarrow$

ε_i and ε_j are not correlated

Regression Analysis

Simple linear regression model

Model assumptions

Assumption 4:

$$\text{Sample-Var}(x) = S^2(x) = 1/n \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$$

und

$$\lim_{n \to \infty} \overline{X^2} < \infty \qquad \overline{X^2} = 1/n \sum_{i=1}^{n} X_i^2$$

und

$$\lim_{n \to \infty} S^2(X) > 0$$

Aussumption 5:
The explanatory variables must be linearly independent

no collinearity or multicollinearity

Under these 5 assumptions the OLS-Estimators are **Best Linear Unbiased Esimator (BLUE).**
It means that they are the efficient ones amongst the set of unbiased linear estimators

Assumption 6: (not always necessary)
For given x the error term ε is normally distributed