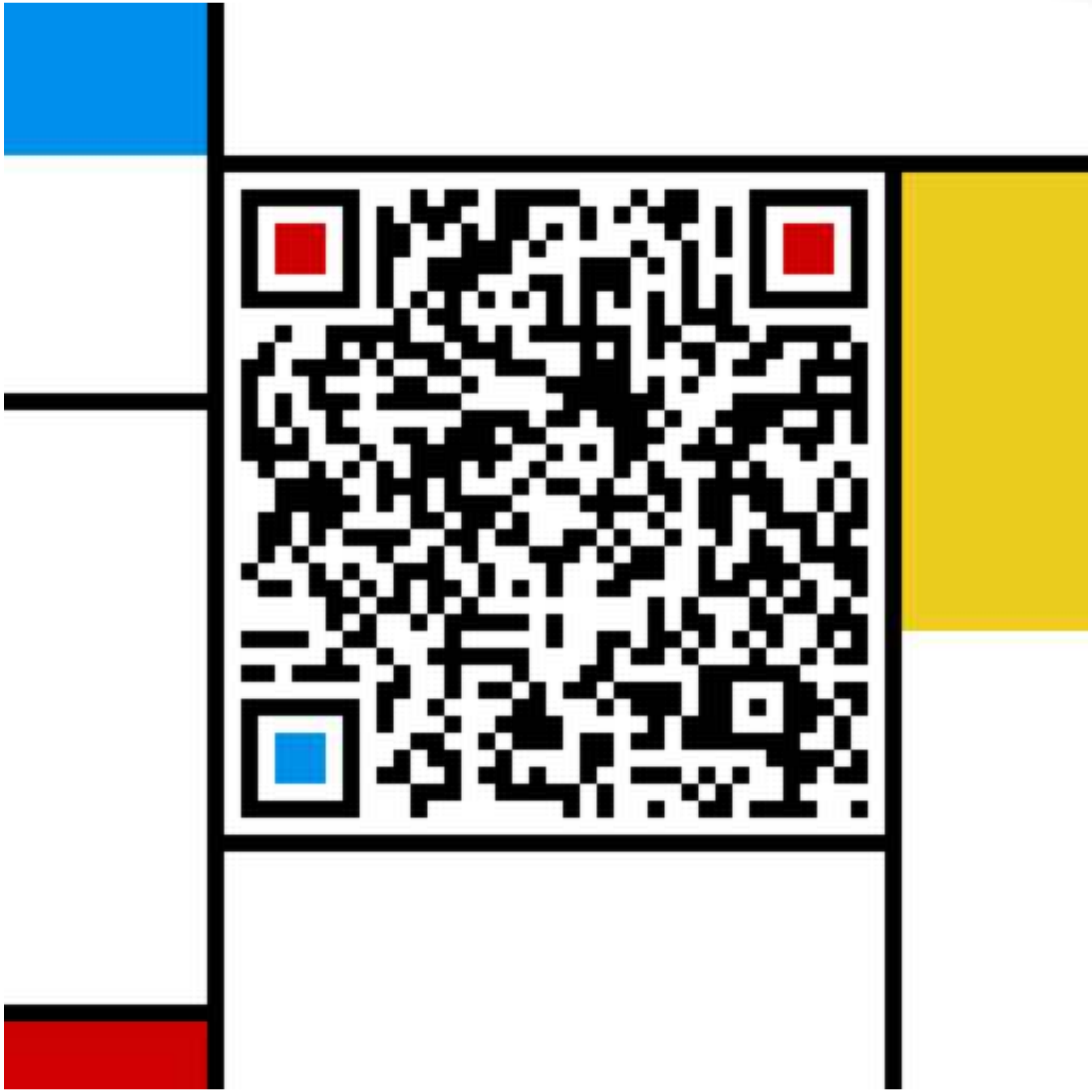
北京师范大学
香港浸会大学 联合国际学院

BEIJING NORMAL UNIVERSITY · HONG KONG BAPTIST UNIVERSITY
UNITED INTERNATIONAL COLLEGE

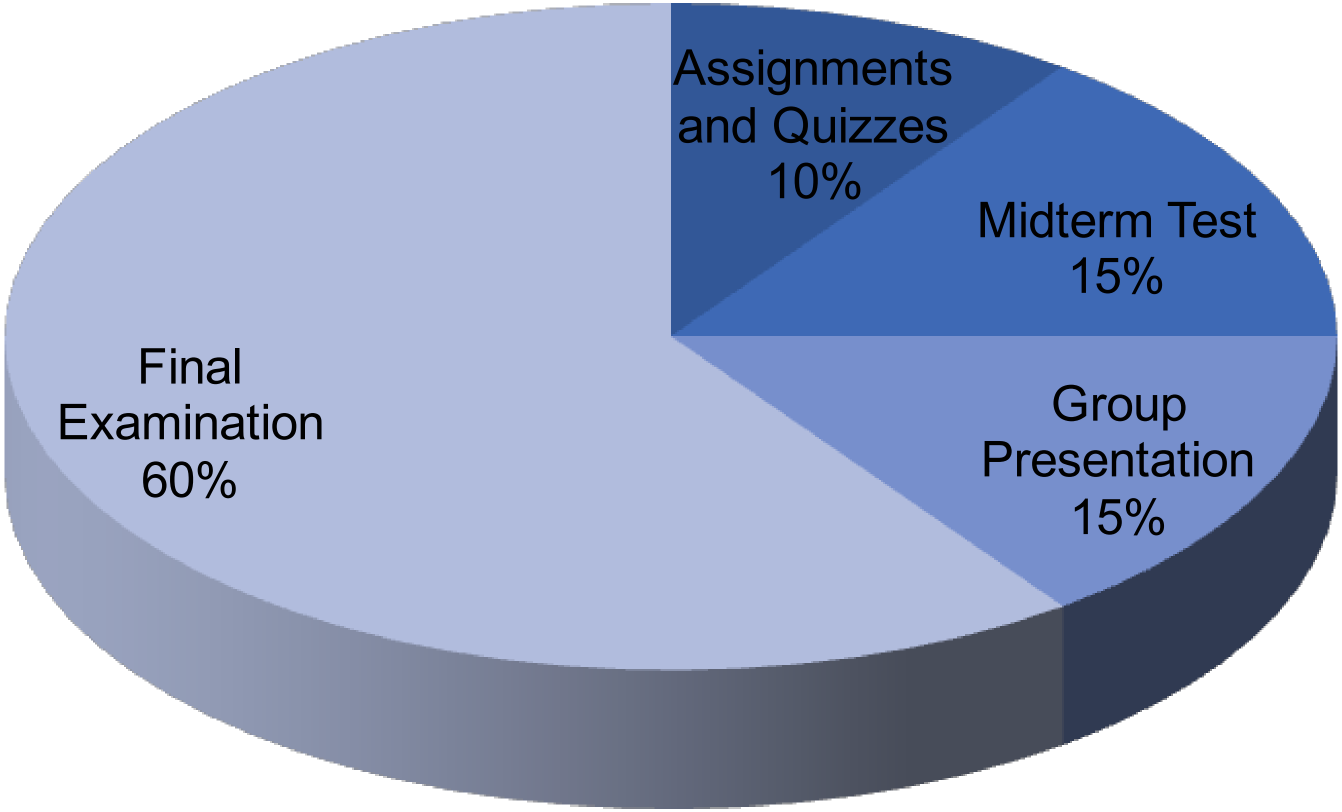
STAT2013 Regression Analysis

Semester 1, 2018-2019
Room: T3-401-R1
Instructor: Ye, Huajun Terry
TA：Jenna Otto
Email:hjye@uic.edu.hk
Jotto@uic.edu.hk (TA)



Course Code & Title: STAT2013 Regression Analysis	
CILO	**Upon successful completion of the course, students should be able to:**
CILO 1	Explain the basic concepts and ideas of linear regression models.
CILO 2	Describe and evaluate basic concepts and methods in various regression models.
CILO 3	Apply regression models to solve real world problems.

Percentage


Assignments and Quizzes 10%
Midterm Test 15%
Group Presentation 15%
Final Examination 60%

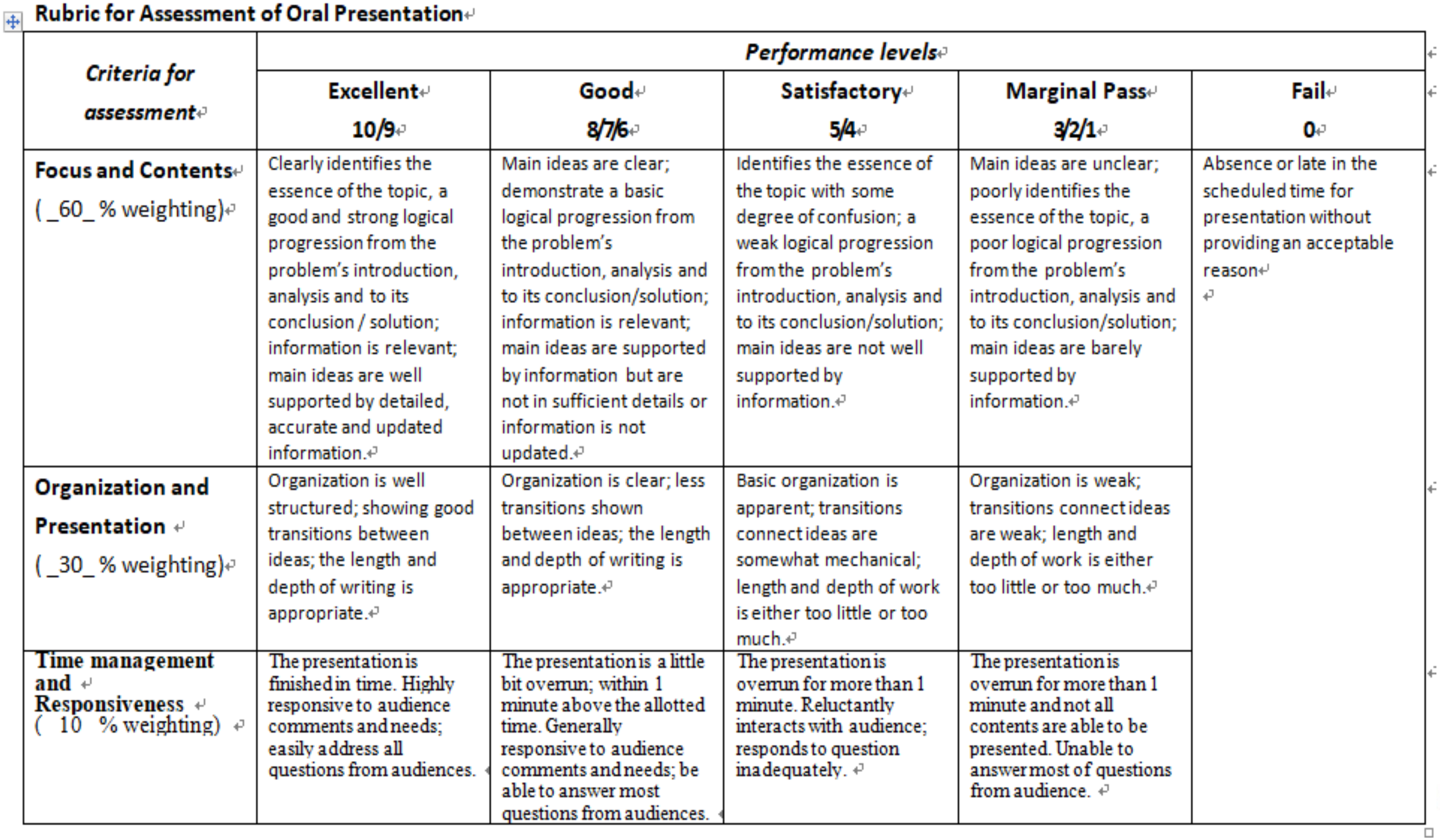
CRA (Criterion-Referenced Assessment)

- Adoption of the Criterion-Referenced Assessment (CRA) for evaluating students' performance

- Syllabus

- CRA model is directly compatible with the OBTL philosophy.

Sample of Rubric



Rubric for Assessment of Oral Presentation

Criteria for assessment	*Performance levels*				
	Excellent **10/9**	**Good** **8/7/6**	**Satisfactory** **5/4**	**Marginal Pass** **3/2/1**	**Fail** **0**
Focus and Contents (_60_ % weighting)	Clearly identifies the essence of the topic, a good and strong logical progression from the problem's introduction, analysis and to its conclusion / solution; information is relevant; main ideas are well supported by detailed, accurate and updated information.	Main ideas are clear; demonstrate a basic logical progression from the problem's introduction, analysis and to its conclusion/solution; information is relevant; main ideas are supported by information but are not in sufficient details or information is not updated.	Identifies the essence of the topic with some degree of confusion; a weak logical progression from the problem's introduction, analysis and to its conclusion/solution; main ideas are not well supported by information.	Main ideas are unclear; poorly identifies the essence of the topic, a poor logical progression from the problem's introduction, analysis and to its conclusion/solution; main ideas are barely supported by information.	Absence or late in the scheduled time for presentation without providing an acceptable reason
Organization and Presentation (_30_ % weighting)	Organization is well structured; showing good transitions between ideas; the length and depth of writing is appropriate.	Organization is clear; less transitions shown between ideas; the length and depth of writing is appropriate.	Basic organization is apparent; transitions connect ideas are somewhat mechanical; length and depth of work is either too little or too much.	Organization is weak; transitions connect ideas are weak; length and depth of work is either too little or too much.	
Time management and Responsiveness (10 % weighting)	The presentation is finished in time. Highly responsive to audience comments and needs; easily address all questions from audiences.	The presentation is a little bit overrun; within 1 minute above the allotted time. Generally responsive to audience comments and needs; be able to answer most questions from audiences.	The presentation is overrun for more than 1 minute. Reluctantly interacts with audience; responds to question inadequately.	The presentation is overrun for more than 1 minute and not all contents are able to be presented. Unable to answer most of questions from audience.	

Oral Presentation/Project Report

- Oral Presentation Period: Dec.10-Dec 13

  Deadline for project report: Dec 13, 2018

- 3 members for one team

- Study rubric for oral presentation before presentation

- Submit your team project report

- The report should include some parts as follows.
  - Abstract
  - Introduction including your problem from your life, data set, graphical display.
  - Methodology
  - Data analysis
  - The conclusions

Grade System

Letter Grade	Academic Performance
A	Excellent
A-	Excellent
B+	Good
B	Good
B-	Good
C+	Satisfactory
C	Satisfactory
D	Marginal Pass
F	Fail

Some notices on this Course

- Assignments must be handed in before the deadline. After the deadline, we refuse to accept your assignments!

- For the final examination, we can not tell you the score before the AR inform the official results. If you have any question on the score, you can check the marked sheet via AR.

Textbook: Myers, R.H. (1990)

Classical and Modern Regression with Applications, 2nd Edition

Duxbury Press, Belmont, California. 2968 citation

海外优秀数学类教材系列丛书
影印版
Classical and Modern Regression with Applications (Second Edition)
经典和现代回归分析及其应用（第2版）
RAYMOND H. MYERS
高等教育出版社
Higher Education Press

Reference books

A second course in Statistics: REGISSION ANALYSIS

sixth edition

William Mendenhall & Terry Sincich , Pearson Education

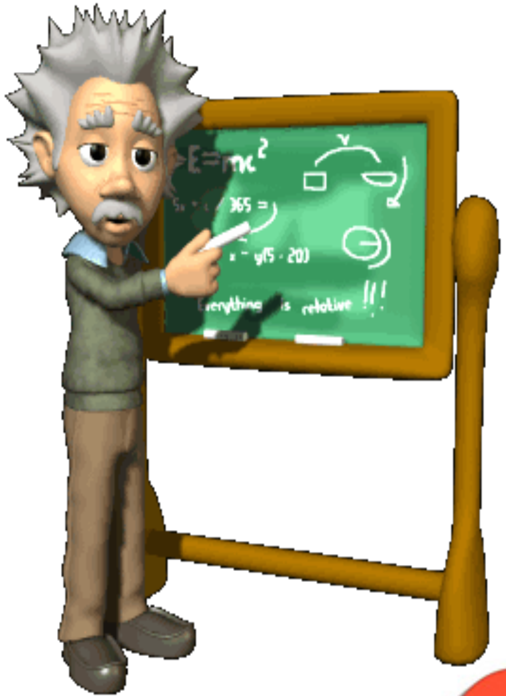
王松桂，陈敏，陈立萍，线性统计模型，高等教育出版社，1999.

方开泰，陈敏，统计学中的矩阵代数, 高等教育出版社, 2013.

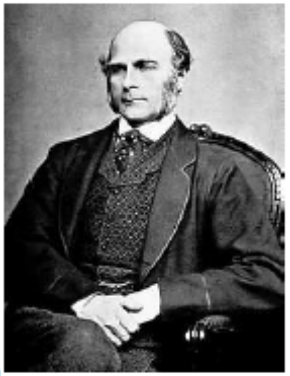
Chapter 1
Introduction

In this Chapter, we will know some important concepts.

- Linear Regression Models
- The Methods of Estimation
- Hypothesis Testing
- Using R



Section 1.1 Linear Regression Models



Regression Analysis

- 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).
- For Galton, regression had only this biological meaning,-but his work was later extended by Udny Yule and Karl Pearson_to a more general statistical context.

Regression Analysis

- Statistics **regression analysis** includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.
- **Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.**

Distinguished Statisticians in History!

Sir R. A. Fisher

1890-1962

Karl Pearson

1857-1936

There are two kinds of relationships among variables:

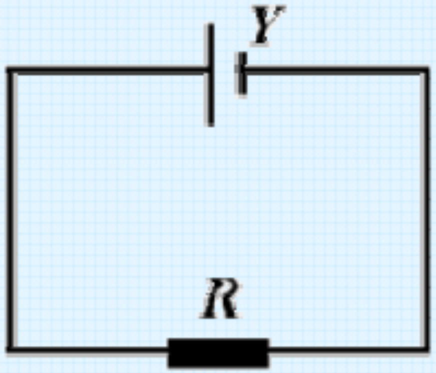
1) Determinate Relation :

$$Y = RX$$

where Y: Voltage

R: Resistance

X: Current



When R of a resistor is fixed, given X, Y can be exactly calculated.

2) Correlation Relation :

Height **VS** Weight

Income **VS** Intelligence Quotient (IQ)

Example 1.1

The House Price Case

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

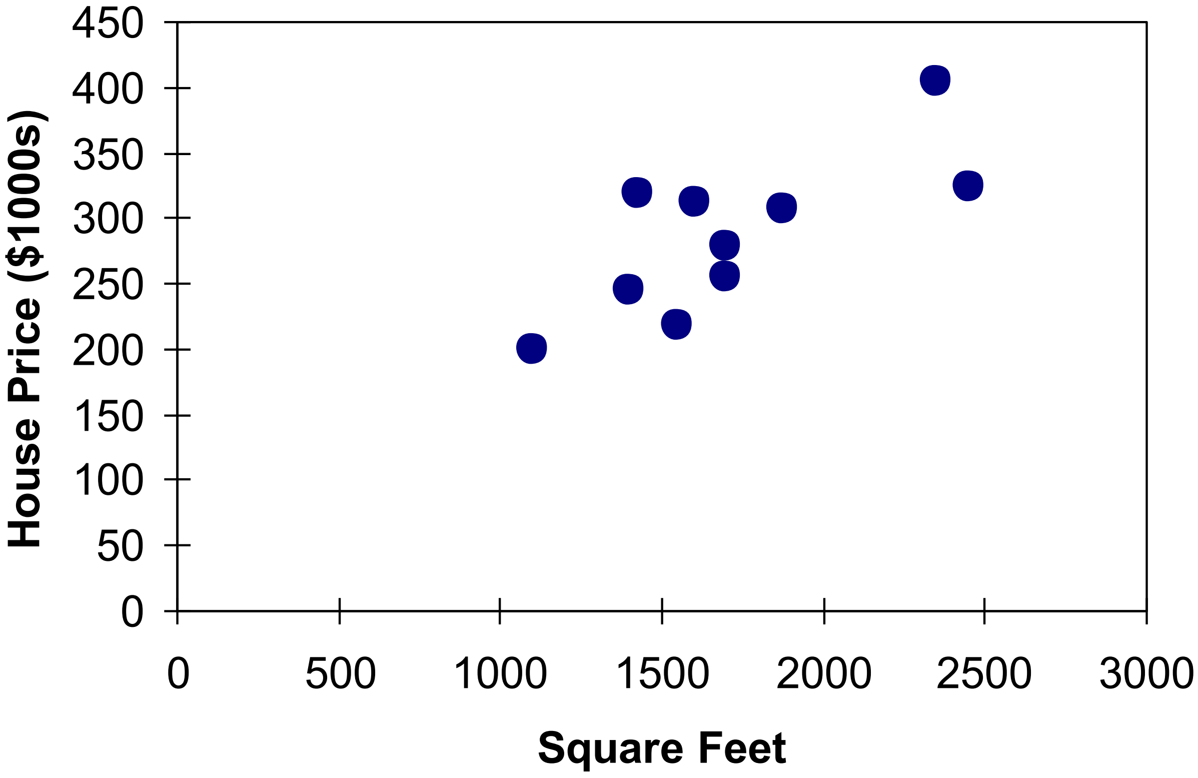
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Question:

- Is there any correlation between Y and X?
- Can we predict Y by the X-value?

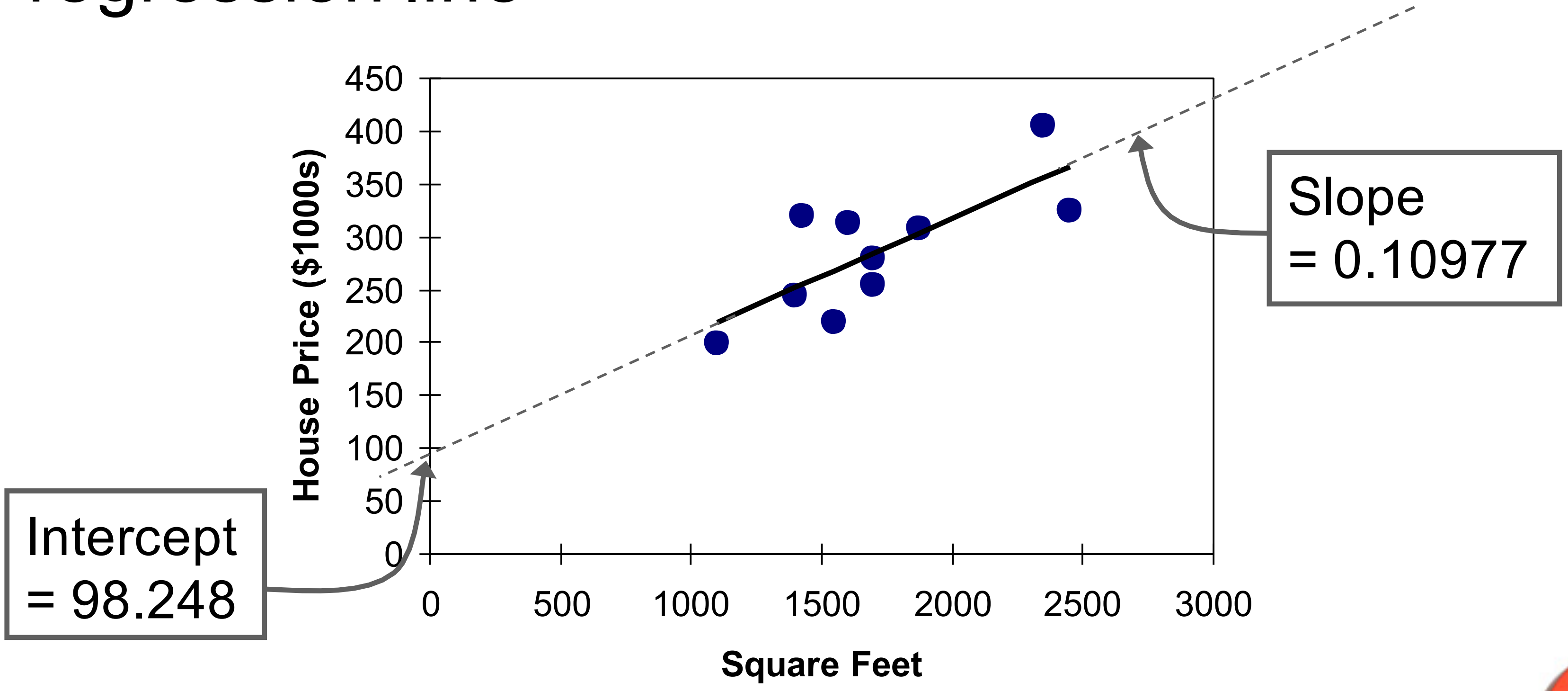
Graphical Presentation

- House price model: scatter plot



Graphical Presentation

- House price model: scatter plot and regression line



Slope
= 0.10977

Intercept
= 98.248

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

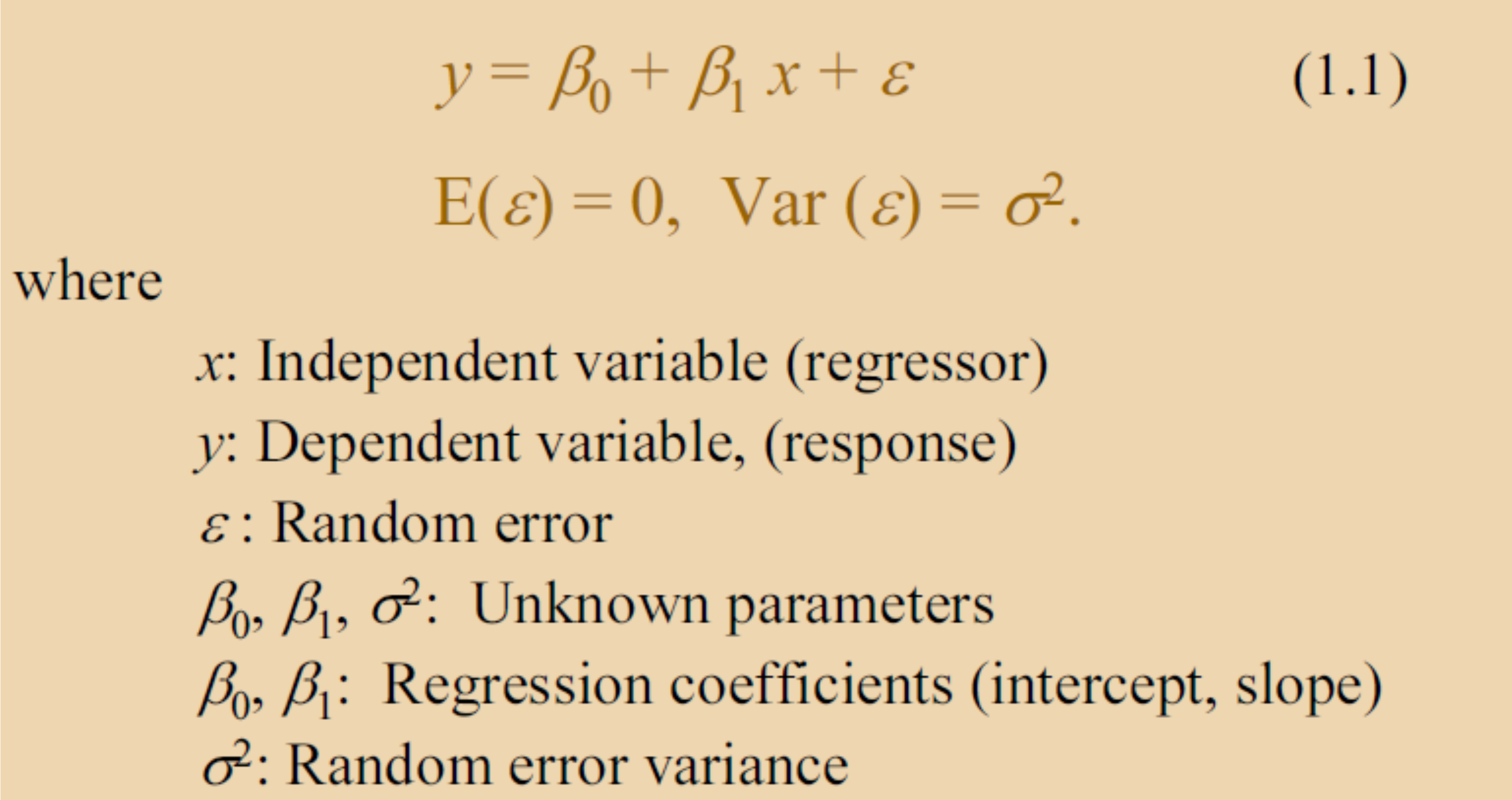
Let (X_i, Y_i), $i = 1, \ldots, n$ be n pairs of observations.

- Plot the points by ***Scatter Plot***;
- Fit a ***smooth curve*** through the points such that the points are as 'close' to the curve as possible.

The following characteristics in this kind of problems are:

- Y cannot be exactly determined by X;
- Y can be estimated by X in the statistical sense;
- There are random errors.

For Example 1.1, we want to fit the data by a model of



$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1.1}$$

$$E(\varepsilon) = 0, \quad \text{Var}(\varepsilon) = \sigma^2.$$

where

x: Independent variable (regressor)

y: Dependent variable, (response)

ε : Random error

$\beta_0, \beta_1, \sigma^2$: Unknown parameters

β_0, β_1: Regression coefficients (intercept, slope)

σ^2: Random error variance

Section 1.2 Methods of Estimation

Observation: $(x_1, y_1), \cdots, (x_n, y_n)$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \cdots, n \qquad (1.2)$$

$$\varepsilon_1, \cdots, \varepsilon_n \text{ are i.i.d.,} \quad \mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

We want to estimate β_0, β_1 and σ^2.

- If we know the distribution of ε, the *maximum likelihood estimation (MLE)* is available. Otherwise, the *least squares estimates (LSE) and other estimates* are considered.

$e_i = y_i - \hat{y}_i$: residual

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: estimated value of y_i

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1 respectively.

→ We want to minimize all e_i's in a certain sense.

L_1 - norm estimate

$$Q_1 = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \left| y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right| \qquad (1.3)$$

Least squares estimate

$$Q_2 = \sum_{i=1}^{n} e_i^{2} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \qquad (1.4)$$

Robust estimate

$$Q_r = \sum_{i=1}^{n} f(e_i), \qquad f(x) = \begin{cases} x^2, & |x| \le k \\ k^2, & |x| > k \end{cases} \quad , \quad k \text{ given.}$$

$$(1.5)$$

Least squares

- The least-squares method was first described by **Carl F. Gauss** around 1794 motivated by prediction the future location of the newly discovered asteroid Ceres (**小行星谷神星**).
- The idea of least-squares analysis was also independently formulated by the Frenchman Adrien-Marie Legender in 1805 and the American Robert Adrain in 1808 motivated by the study of theory of error.

LSE (Least Squares Estimation)

Recall

$$Q_2 = \sum_{i=1}^{n} e_i{}^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

For minimizing $Q = Q_2$ with respect to β_0 and β_1, take derivatives:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (1.6)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \qquad (1.7)$$

$$\boldsymbol{S}_{\mathrm{XX}} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$\boldsymbol{S}_{\mathrm{XY}} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$\Rightarrow \begin{cases} b_1 = \mathrm{S}_{\mathrm{XY}} / \mathrm{S}_{\mathrm{XX}} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases} \qquad (1.10)$$

Interpretation of the Intercept, b_0

$$\boxed{\widehat{\text{house price}} = \boxed{98.24833} + 0.10977\,(\text{square feet})}$$

- b_0 is the estimated average value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
  - Here, no houses had 0 square feet, so $\boxed{b_0 = 98.24833}$ just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, $\boxed{b_1 = .10977}$ tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Predictions using Regression Analysis

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

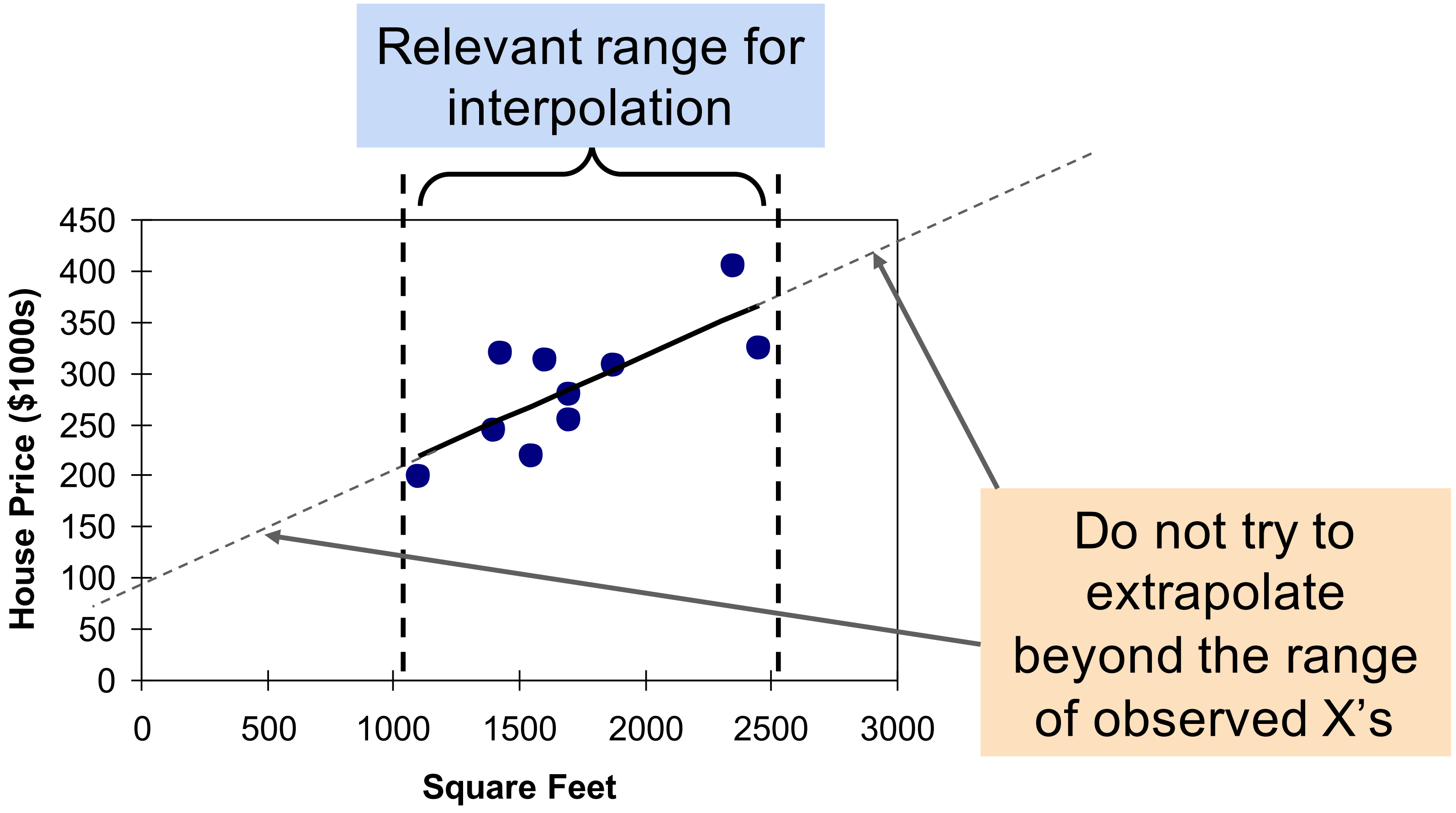
$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Interpolation vs. Extrapolation

- When using a regression model for prediction, only predict within the relevant range of data



Relevant range for interpolation

Do not try to extrapolate beyond the range of observed X's

The Least Square Point Estimates

Estimation/prediction equation

$$\hat{y} = b_0 + b_1 x$$

Least squares point estimate of the slope β_1

$$b_1 = \frac{S_{xy}}{S_{xx}} \qquad S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \qquad S_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

Least squares point estimate of the y-intercept β_0

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \bar{y} = \frac{\sum y_i}{n} \qquad \bar{x} = \frac{\sum x_i}{n}$$

Model Assumptions

1. **Mean of Zero**
   At any given value of x, the population of potential error term values has a mean equal to zero
2. **Constant Variance Assumption**
   At any given value of x, the population of potential error term values has a variance that does not depend on the value of x
3. **Normality Assumption**
   At any given value of x, the population of potential error term values has a normal distribution
4. **Independence Assumption**
   Any one value of the error term ε is statistically independent of any other value of ε

Example 1.2.

Forecasting Sales for a Clothing Store

- The sales for Sunflowers, a chain of apparel stores for women, have increased during the past 12 years as the chain expanded the number of stores open.
- Until now, Sunflowers senior managers selected sites based on subjective factors such as the availability of a good lease or the perception that a location seemed ideal for an apparel stores.
- As the new director of planning, you need to develop a systematic approach to selecting new sites that will allow Sunflowers to make better-informed decisions for opening additional stores.

Example 1.2.

Forecasting Sales for a Clothing Store

- This plan must be able to forecast annual sales for all potential stores under consideration.
- You believe that the size of the store significantly contributes to the success of a store and you want to use this relationship in the decision-making process.
- How can you use statistics so that you can forecast the annual sales of a proposed store based on the size of that store?

Data Set

Store	Square Feet	Annual Sales (in Millions of Dollars)	Store	Square Feet	Annual Sales (in Millions of Dollars)
1	1.7	3.7	8	1.1	2.7
2	1.6	3.9	9	3.2	5.5
3	2.8	6.7	10	1.5	2.9
4	5.6	9.5	11	5.2	10.7
5	1.3	3.4	12	4.6	7.6
6	2.2	5.6	13	5.8	11.8
7	1.3	3.7	14	3.0	4.1

Least Squares Method

$$\overline{x}=\frac{40.9}{14}=2.92143,\overline{y}=\frac{81.8}{14}=5.842857$$

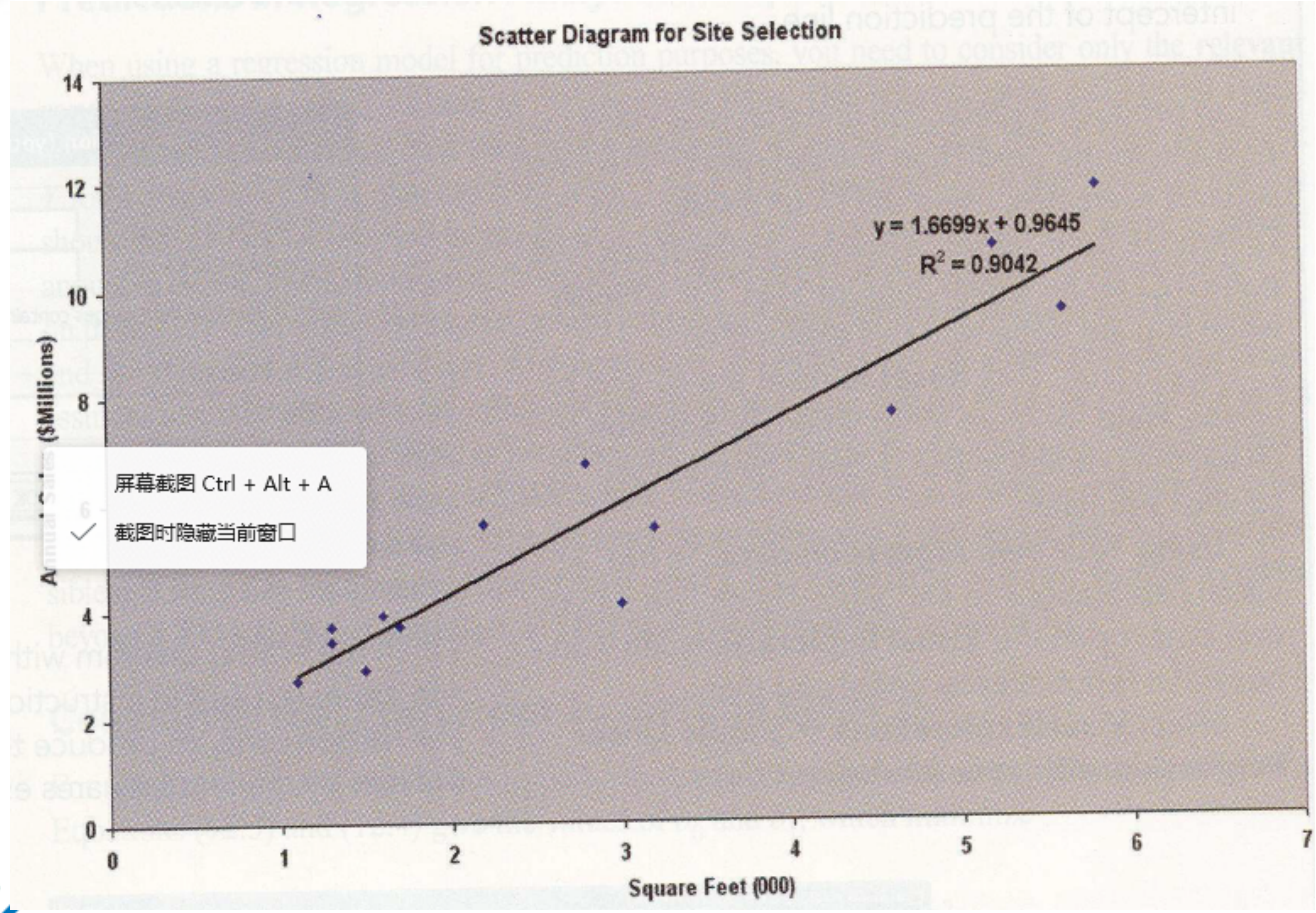
$$SS_{xy}=63.32715$$

$$SS_{x}=37.92358$$

$$b_1=\frac{63.32715}{37.92358}=1.66986\approx 1.67$$

$$b_0=\overline{y}-b_1\overline{x}=5.842857-1.67\times 2.92143=0.964478$$

$$\hat{y}=0.964478+1.67x$$


Scatter Diagram for Site Selection
y = 1.6699x + 0.9645
R² = 0.9042
Annual Sales ($Millions)
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
Square Feet (000)
屏幕截图 Ctrl + Alt + A
✓ 截图时隐藏当前窗口

Section 1.3 Hypothesis Testing

Several questions may raise from the fitted regression line:

1. **Does *x* truly influence *y*?**
2. **Is there an adequate fit of the data to the model?**
3. **Will the model adequately predict response?**

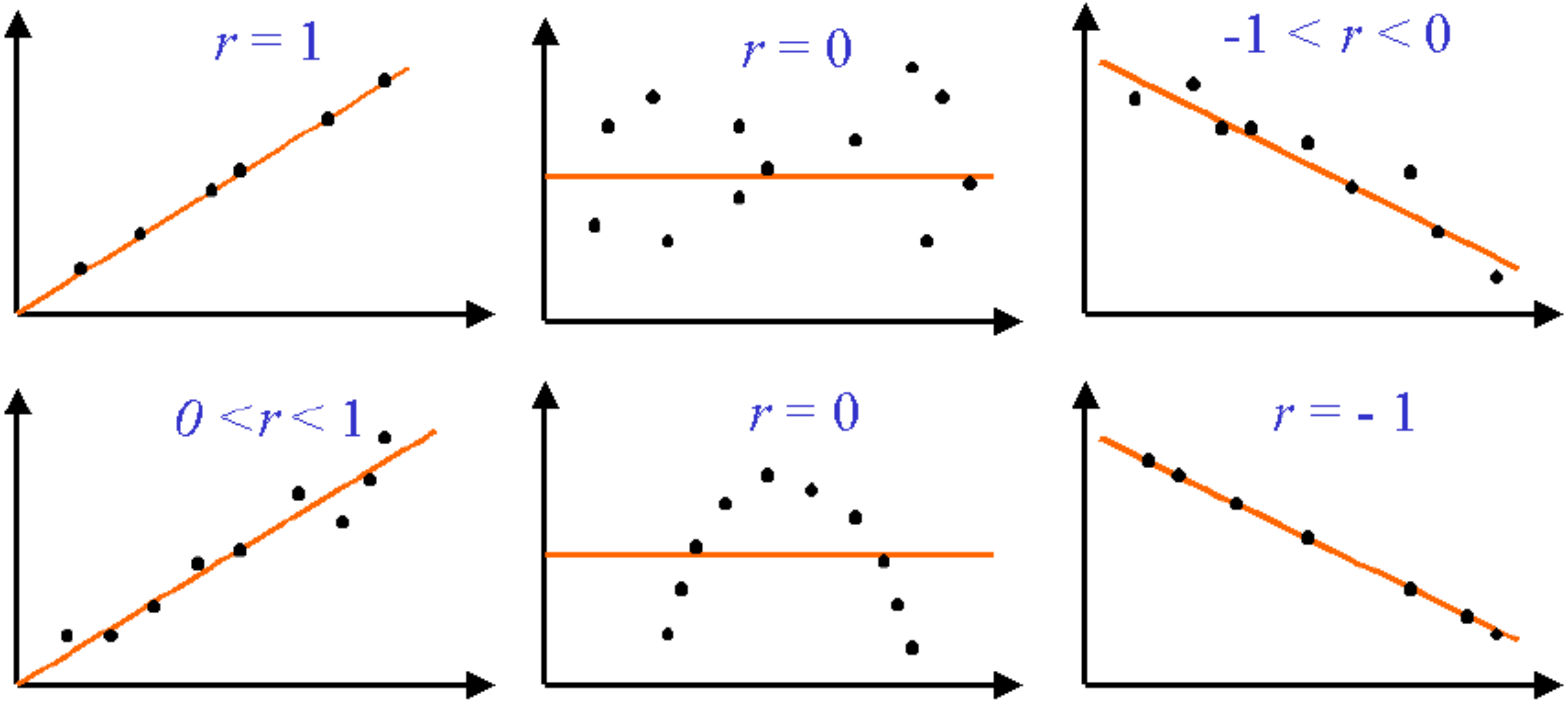
Hypothesis: $H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$

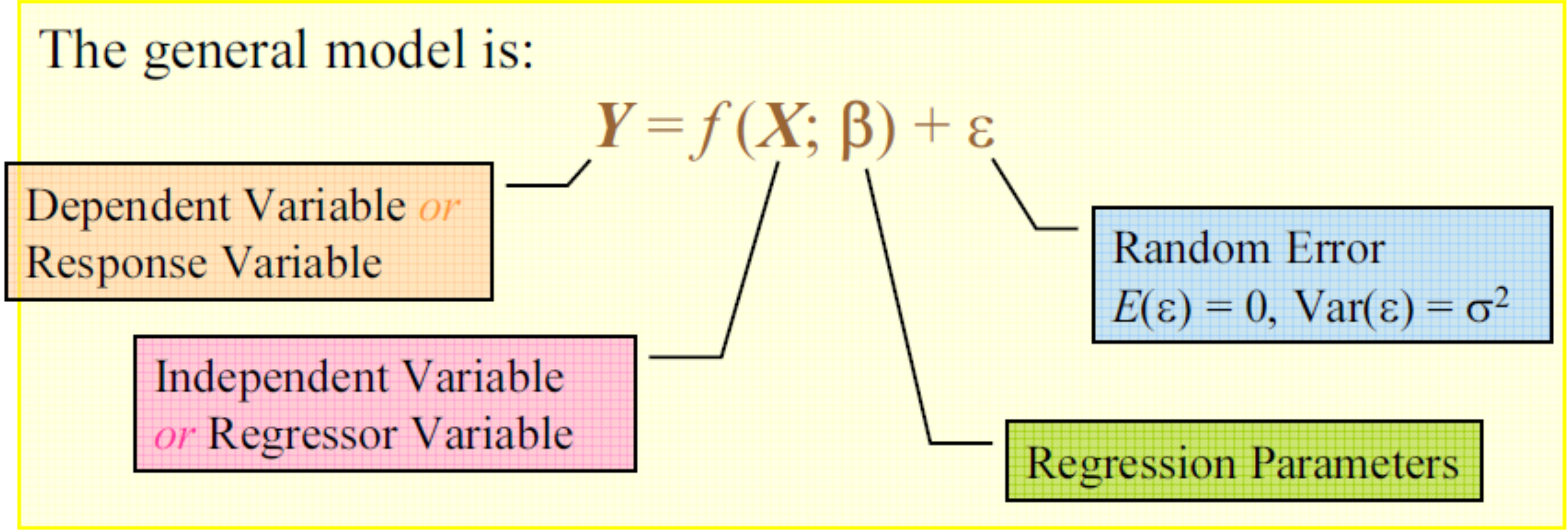
If H_0 is true,

$E(y) = \beta_0$, and **the regressor variable does not influence *y*.**

There are several ways to test this hypothesis:

1. **Correlation Coefficient**
2. **ANOVA (analysis of variance)**
3. **Coefficient of Determinate (*R*2)**


$r = 1$
$r = 0$
$-1 < r < 0$
$0 < r < 1$
$r = 0$
$r = -1$



The general model is:

$$Y = f(X; \beta) + \varepsilon$$

Dependent Variable *or* Response Variable

Independent Variable *or* Regressor Variable

Regression Parameters

Random Error
$E(\varepsilon) = 0, \ \mathrm{Var}(\varepsilon) = \sigma^2$

The simple model is:

Linear Model

$$\boldsymbol{Y} = \beta_0 + \beta_1\boldsymbol{X} + \varepsilon$$

Given n observations (X_i, Y_i), $i = 1, \ldots, n$, the model will be:

$$\boldsymbol{Y_i} = \beta_0 + \beta_1\boldsymbol{X_i} + \varepsilon_{\boldsymbol{i}}$$

Classifications of Regression Models

A. According to the number of regressors

- One regressor variable model;
- Multiple regression model.

B. According to the function *f*:

- Linear regression;
- Nonlinear regression.

C. According to the method of estimation parameters

- Least squares estimation;
- Maximum likelihood estimation;
- Robust estimation;
- Nonparametric regression.

D. Data analysis:

- Estimation of β and σ^2;
- Testing hypothesis (on β or σ^2);
- Selection of variables;
- Statistical Diagnostics

Remarks on Regression Analysis:

- Long history in statistics;
- Most popular methods with wide applications
- Most active branch of statistics;
- The basis of many useful method, such as:
  - time series;
  - principal component analysis;
  - canonical analysis;
  - factor analysis;
  - structure models, etc.

Remarks on Regression Analysis:

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for **robust regression**, regression involving correlated responses such as time series and growth curves, regression in which the predictor or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, **nonparametric regression, Bayesian** methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Section 1.4 Using R

R Language

- R is free. Open source and development
- Flexible
  - Extensible
  - Interoperability with other languages: **C, Java**
- Strengths
  - statistical and numerical methods
  - High quality visualization and graphics tools
  - Effective, extensible user interface
- Supports the creation, testing, and distribution of software and data modules: **packages**.

Website

- R www.r-project.org
  - software;
  - documentation;
  - RNews.

Installing R

http://cran.r-project.org

Data can either be typed into R or imported into the program. If you type the data, then use the `c(·)` function to create a vector or the `matrix(·,nrow,ncol)` function to generate a matrix of size $(nrow \times ncol)$. For example,

```
>c(1,2.3,3.45,0.56)
```

produces

$$\begin{pmatrix} 1 \\ 2.3 \\ 3.45 \\ 0.56 \end{pmatrix}$$

and

```
>matrix(c(1,2,3,4,5,6,7,8),4,2)
```

creates the matrix

$$\begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}.$$

Importing data can be done with either the `scan(·)` or `read` functions. Assume you had a file called regression.csv on your desktop. Then

```
>scan("c:\Desktop\regression.csv")
```

would import the file **as a vector**. So, if the file contained 4 variables or 25 observations each, then the `scan` function would call the data in as a vector of 100 observations. You would then have to convert the vector into the correct (4×2) matrix.

Alternatively, the `read.table(·)` and `read.csv(·)` functions call the data files into R in matrix form. i.e. 4 columns in the data file will remain 4 columns when imported. Saving the data as a *comma separated* (.csv) file and using the `read.csv(·)` function is strongly recommended.

```
>read.csv("c:\Desktop\regression.csv",header=TRUE)
```

calls the data into R and treats the first row as the column headings.

There are 3 ways to seek help from R. They are

1. `help.search(·)`
2. `help(·)`
3. Use `?`

If you do not know the function you need, then you can use the first help tool. Say, we wanted to look for a function to do linear regression, but we don't know what functions exist in R. Typing

If you do not know the function you need, then you can use the first help tool. Say, we wanted to look for a function to do linear regression, but we don't know what functions exist in R. Typing

```
>help.search("linear model")
```

will open up a screen that lists functions related to linear regression. Looking through the list, you should find `lm(·)` and `glm(·)`. The functions will be very useful for this course. Click on one of them and scroll down the information screen. This screen provides useful information about the function. It starts with possible function inputs, then tells you the outputs provided, and towards the bottom, it should provide a list of functions that you should *See also*. *Examples* are also often provided at the end.

Many basic statistical concepts have logical function names. Here is a list of many simple concepts or figures

```
mean()
median()
mode()
range()
min()
max()
sd()
var()
cor()
cov()
IQR()
hist()
plot()
qqline()
abline()
points()
lines()
boxplot()
stem()
persp()
quantile()
```