

RMSC 4003 Statistical Modelling in Financial Markets

Chapter 0 - Course Outline

Agenda

1 Teachers

2 Course Content

3 Grading

4 Prerequisites

- Financial Risks
- Mathematics and Statistics
- Regression

- **Instructor:** Prof. YAU Chun Yip
 - Email : cyyau@sta.cuhk.edu.hk
 - Web : www.sta.cuhk.edu.hk/cyyau
 - Office: Lady Shaw Building 110
 - Office hours: Monday & Tuesday 3-5pm, or by appointment
- **Tutors:** Mr. Chen Jiaxu & Mr. Liu Chun Yu Parco & Mr. Sun Cho Fei
 - Email : chenjiaxu0469@link.cuhk.edu.hk & parcocyliu@link.cuhk.edu.hk & pascalsun@link.cuhk.edu.hk
 - Office: Lady Shaw Building G24, Lady Shaw Building G28 & Lady Shaw Building 101

Agenda

1 Teachers

2 Course Content

3 Grading

4 Prerequisites

- Financial Risks
- Mathematics and Statistics
- Regression

Topics covered

- ① Markowitz Portfolio Theory (2 weeks)
- ② Capital Asset Pricing Model (CAPM) (1.5 weeks)
- ③ Arbitrage Pricing Theory (APT) (1.5 weeks)
- ④ Midterm
- ⑤ Discrete Time Financial Model (extension of Binomial model) (1 week)
- ⑥ Continuous Time Financial Model (1 week)
- ⑦ Stochastic Calculus (2-3 weeks)
- ⑧ Parameter Estimation of Continuous Time Financial Models (if time allows)

Lecture Slides

- [Uploaded in the course website.](#)

www.sta.cuhk.edu.hk/cyyau/rmsc4003

Reference books

- [Lai, Tze Leung, and Xing, Haipeng \(2008\).](#)
STATISTICAL MODELS AND METHODS FOR FINANCIAL MARKETS.
Springer New York.
- [Mikosch, Thomas \(1998\).](#)
ELEMENTARY STOCHASTIC CALCULUS, WITH FINANCE IN VIEW.
World Scientific Publishing, Singapore.
- More references are given in each chapter of lecture slides

Agenda

- 1 Teachers
- 2 Course Content
- 3 Grading
- 4 Prerequisites
 - Financial Risks
 - Mathematics and Statistics
 - Regression

Grading

- Assignments (20%)
~5 assignments; upload to blackboard or dropbox at LSB 1/F
- Mid term exam (30%) 20/10 Tuesday
- Final exam (50%) TBA
- Absolute Grading System:

Overall points reach	Grade
90	A
80	A-
70	B+
60	B
50	B-
45	C+
40	C
Below 40	C-/D/F (case by case)

Agenda

1 Teachers

2 Course Content

3 Grading

4 Prerequisites

- Financial Risks
- Mathematics and Statistics
- Regression

Agenda

- 1 Teachers
- 2 Course Content
- 3 Grading
- 4 Prerequisites
 - Financial Risks
 - Mathematics and Statistics
 - Regression

Typology of Risks

- **Market:** movement in market price
 - include **basis risk:** mismatches in a hedged position
- **Credit:** change in credit quality, e.g. downgrade or default
- **Operational:** loss from inadequate systems, management, controls, human errors or fraud.
- **Liquidity:**
 - Funding liquidity: Unable to re-fund/rollover its debt
 - Trading liquidity: Lack of counter-party interest to trade
- **Legal:** being sued for e.g. misrepresentation of risk of products sold.
- **Model:** error in complex models e.g. in derivative valuation.

Famous Financial Pitfalls

- ❶ (1994) Orange County, California, USA
 - The county treasurer, Bob Citron, managed a \$ 7.5b portfolio
 - Heavy investment in **high yield (low grade) bonds**.
 - County went bankrupt when his huge bet on low interest rate reversed.
- ❷ (1995) Barings Bank (**British bank founded in 1762**)
 - Nick Lesson lost the entire equity of \$1.3b from **trading Nikkei stock index futures**, mainly.
- ❸ (1995) Japanese Yen Crisis
 - The value of USD was declining in the 1990s
 - Hedging using put options (sell USD for JPY) protects USD holders
 - Some firms hedged using (cheaper than vanilla) **down-and-out put** options, which voided below the knock-out of 85 JPY/USD.
 - USD drop was accelerated by its panic sale near the knock-out price 85.
 - Investment banks also drove the USD down to release their obligations.

Famous Financial Pitfalls

- 4 (1998) Peregrine Securities (HK investment company)
 - A front-desk credit derivative trader invested in questionable bonds, which defaulted in the Asian crisis. Peregrine went bankrupt and was acquired by BNP Paribas.
- 5 (1998) Long Term Capital Management (Hedge fund in USA)
 - Board of directors included Myron Scholes and Robert C. Merton
 - LTCM traded on discrepancies between fixed income & FX markets.
 - Their model suggested increasing risky bets during a global downturn.
 - The fund was liquidated and dissolved in early 2000.
- 6 (2008) Lehman Brothers (USA I-bank founded in 1847)
 - Heavily involved in subprime mortgage lending: high return & default risk
 - package those loans into mortgage-backed securities and sell to investors
 - subprime mortgage crisis $\Rightarrow$ loss of 75% of its stock value
 - Worsened by (a) unrealized takeover rumour, (b) mass exodus of clients, (c) severe asset downgrades by rating agencies
 - Largest bankruptcy in U.S. then ($\geq \$ 600\text{b}$ assets & $\geq 28\text{k}$ employees)

Agenda

- 1 Teachers
- 2 Course Content
- 3 Grading
- 4 Prerequisites
 - Financial Risks
 - **Mathematics and Statistics**
 - Regression

Background in mathematics and statistics

- Calculation of mean, variance, covariance (population & sample versions)
- Confidence interval and hypothesis testing
- Differentiation and Constraint Optimization
- Basic R programming
 - Download: <http://cran.r-project.org/bin/windows/base/>
 - Tutorial: <http://r-statistics.net/r-tutorial.html>

Statistics: Mean, Variance and Covariance

- Data: $\mathbf{x} = \{x_1, \dots, x_n\}$, $\mathbf{y} = \{y_1, \dots, y_n\}$
- Sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} (\sum_{i=1}^n x_i^2 - n\bar{x}^2)$
- Sample covariance $s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} (\mathbf{x} - \bar{x})^T (\mathbf{y} - \bar{y})$

Example 1

- X be the annual number of typhoon signals number 8 or above in HK
- Y be the annual number of black rainstorm signals in HK
- Samples (2013-2020): $\mathbf{x} = (3, 1, 1, 4, 10, 4, 1, 4)^T$ & $\mathbf{y} = (1, 2, 1, 1, 1, 0, 0, 2)^T$

Find the sample means, variances, and covariance of the above samples:

- Mean: $\bar{x} = (3 + 1 + 1 + 4 + 10 + 4 + 1 + 4)/8 = 28/8 = 3.5$
 $\bar{y} = (1 + 2 + 1 + 1 + 1 + 0 + 0 + 2)/8 = 8/8 = 1$
- Var.: $s_x^2 = [(3^2 + 1^2 + 1^2 + 4^2 + 10^2 + 4^2 + 1^2 + 4^2) - 8(3.5)^2] / 7 \approx 8.86$
 $s_y^2 = [(1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 2^2) - 8(1)^2] / 7 \approx 0.57$
- Cov.: $s_{xy} = \frac{(-0.5, -2.5, -2.5, 0.5, 6.5, 0.5, -2.5, 0.5)^T (0, 1, 0, 0, 0, -1, -1, 1)}{7} = 0$

Confidence Intervals and Hypothesis Testings

- Let $\mathbf{x} = (x_1, \dots, x_n)$ be from a population of **unknown** mean μ & variance σ^2 .
- Distributional result: $\frac{\bar{x} - \mu}{s/\sqrt{n}} \xrightarrow{d} N(0, 1)$ as $n \rightarrow \infty$
- The two-sided $100(1 - \alpha)\%$ confidence interval (CI) for μ is:

$$\left(\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right),$$

for large n , where $z_{\alpha/2}$ satisfies $P(N(0, 1) > z_{\alpha/2}) = \alpha/2$

Example 2

Let X be the annual number of typhoon signals number 8 or above in HK;
The mean & variance of $n = 60$ samples observed are $\bar{x} = 3.45$ & $s^2 = 3.4^2$.

- Given $z_{2.5\%} = 1.96$, construct 95% confidence interval for μ :

$$\text{C.I.} = \left(3.45 \pm 1.96 \frac{3.4}{\sqrt{60}} \right) \approx (3.45 \pm 0.86) = (2.59, 4.31)$$

- Test hypothesis $H_0 : \mu = 3$ vs $H_a : \mu \neq 3$
 $3 \in (2.59, 4.31) \Rightarrow$ do not reject H_0 at 5% significance level.

Probability: Mean and Variance

1 Probability density functions:

- Random variable $X \sim f_X(x)$:

$$\mathbb{P}(X \in (a, b)) = \int_a^b f_X(x) dx$$

- Random vector $\mathbf{X} = (X_1, \dots, X_n)^T \sim f_{\mathbf{X}}(x_1, \dots, x_n)$:

$$\mathbb{P}(\mathbf{X} \in A) = \int_A f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \cdots dx_n$$

2 Mean/Expectation & Covariance

- $E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$
- $E(\mathbf{X}) = (E(X_1), \dots, E(X_n))^T$
- $\text{Cov}(X_1, X_2) = E((X_1 - E(X_1))(X_2 - E(X_2)))$
- $\text{Cov}(\mathbf{X}, \mathbf{Y}) = (\text{Cov}(X_i, Y_j))_{1 \leq i, j \leq n}$; $\text{Var}(\mathbf{X}) = \text{Cov}(\mathbf{X}, \mathbf{X})$

$$3 \quad E\left(\sum_{t=1}^n a_t X_t\right) = \sum_{t=1}^n a_t \mu_t, \quad \text{where } \mu_t = E(X_t)$$

$$4 \quad \text{Cov}\left(\sum_{t=1}^n a_t X_t, \sum_{s=1}^n b_s X_s\right) = \sum_{t=1}^n \sum_{s=1}^n a_t b_s \text{Cov}(X_t, X_s)$$

Prerequisites

$$\textcircled{1} \quad \mathbb{E} \left(\sum_{t=1}^n a_t X_t \right) = \sum_{t=1}^n a_t \mu_t, \quad \text{where } \mu_t = \mathbb{E}(X_t)$$

$$\textcircled{2} \quad \text{Cov} \left(\sum_{t=1}^n a_t X_t, \sum_{s=1}^n b_s X_s \right) = \sum_{t=1}^n \sum_{s=1}^n a_t b_s \text{Cov}(X_t, X_s)$$

Vector version: $\mathbf{a} = (a_1, \dots, a_n)^T$, $\mathbf{X} = (X_1, \dots, X_n)^T$, $\Sigma = \text{Cov}(\mathbf{X}, \mathbf{X})$:

$$\textcircled{1} \quad \mathbb{E}(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \mathbb{E}(\mathbf{X}) = \mathbf{a}^T \boldsymbol{\mu}, \quad \text{where } \boldsymbol{\mu} := (\mathbb{E}(X_1), \dots, \mathbb{E}(X_n))^T$$

$$\textcircled{2} \quad \text{Cov}(\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{X}) = \mathbf{a}^T \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{b} = \mathbf{a}^T \Sigma \mathbf{b}$$

Prerequisites

- ① $E(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T E(\mathbf{X}) = \mathbf{a}^T \boldsymbol{\mu}$, where $\boldsymbol{\mu} := (E(X_1), \dots, E(X_n))^T$
- ② $\text{Cov}(\mathbf{a}^T \mathbf{X}, \mathbf{b}^T \mathbf{X}) = \mathbf{a}^T \text{Cov}(\mathbf{X}, \mathbf{X}) \mathbf{b} = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{b}$

Example 3

Two unfair coins $i = \{1, 2\}$, are tossed, whose event of a head is H_i , where

Coin 1: $\mathbb{P}(H_1) = E(1_{H_1}) = 0.6$, with a gain of $n_1 = \$10$ under H_1

Coin 2: $\mathbb{P}(H_2) = E(1_{H_2}) = 0.7$, with a gain of $n_2 = \$20$ under H_2

Find the expected total gain.

- Let $\mathbf{n} = (n_1, n_2)^T$, $\mathbf{1}_H = (1_{H_1}, 1_{H_2})^T$.
- $E(n_1 1_{H_1} + n_2 1_{H_2}) = E(\mathbf{n}^T \mathbf{1}_H) = \mathbf{n}^T E(\mathbf{1}_H) = (10, 20)(0.6, 0.7)^T = 20$

Example 4

Two investments of sizes \$ (8, 2)m are made into projects with profits $\mathbf{X} = (X_1, X_2)^T$

$$E(\mathbf{X}) = (10\%, 5\%), \text{Var}(X_1) = \text{Var}(X_2) = \sigma_2^2 = 0.25, \text{Cov}(X_1, X_2) = 0.125$$

Find the variance of total profit in (m \$)².

- Let $\mathbf{a} = (a_1, a_2)^T = (8, 2)^T$, $\boldsymbol{\Sigma} = \text{Var}(\mathbf{X})$
- $\text{Var}(\text{Total profit}) = \text{Var}(\mathbf{a}^T \mathbf{X}) = \text{Cov}(\mathbf{a}^T \mathbf{X}, \mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a}$
$$= \begin{pmatrix} 8 & 2 \end{pmatrix} \begin{pmatrix} 0.25 & 0.125 \\ 0.125 & 0.25 \end{pmatrix} \begin{pmatrix} 8 \\ 2 \end{pmatrix} = \begin{pmatrix} 2.25 & 1.5 \end{pmatrix} \begin{pmatrix} 8 \\ 2 \end{pmatrix} = 2.25 \times 8 + 1.5 \times 2 = 21$$

Definition 1

- Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a real-valued function with n arguments $x_1, \dots, x_n$.
- Differentiating a scalar $f(\mathbf{x}) = f(x_1, \dots, x_n)$ with respect to a vector $\mathbf{x} = (x_1, \dots, x_n)$:

$$\frac{df}{d\mathbf{x}} = \begin{pmatrix} \frac{df}{dx_1} \\ \vdots \\ \frac{df}{dx_n} \end{pmatrix}.$$

Example 5

Let $f(x_1, x_2, x_3) = x_1^2 + 2x_2x_3^3$, and $\mathbf{x} = (x_1, x_2, x_3)$.

$$\frac{df}{d\mathbf{x}} = \begin{pmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \\ \frac{df}{dx_3} \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_3^3 \\ 6x_2x_3^2 \end{pmatrix}.$$

Differentiation w.r.t. Vectors

Let $\mathbf{a}, \mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$. Also, let $\mathbf{u}, \mathbf{v} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be functions of $\mathbf{x}$:

$$\textcircled{1} \quad \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\textcircled{2} \quad \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}$$

Proof.

$$\textcircled{1} \quad y := \mathbf{a}^T \mathbf{x} = a_1 x_1 + \dots + a_n x_n. \quad \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \left(\frac{\partial y}{\partial x_1}, \dots, \frac{\partial y}{\partial x_n} \right)^T = (a_1, \dots, a_n)^T = \mathbf{a}$$

$\textcircled{2}$ Denote $\mathbf{A} = (a_{ij})$, whose i^{th} row vector is $\mathbf{r}_i$, & j^{th} column vector is $\mathbf{c}_j$.

$$\text{Denote } \mathbf{y} = (y_1, \dots, y_n)^T := \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial}{\partial \mathbf{x}} \sum_i \sum_j a_{ij} x_i x_j$$

$$\text{Consider } y_k = \frac{\partial}{\partial x_k} \sum_i \sum_j a_{ij} x_i x_j = \sum_j a_{kj} x_j + \sum_i a_{ik} x_i = \mathbf{r}_k \mathbf{x} + \mathbf{c}_k^T \mathbf{x}$$

$$\text{Putting together } y_k, \forall k \Rightarrow \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \mathbf{y} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x} \quad \square$$

Example 6

Show $\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$, where $\mathbf{x} = (x_1, x_2, x_3)^T$, $\mathbf{a} = (1, 2, 3)^T$.

$$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \left(\frac{\partial(x_1 + 2x_2 + 3x_3)}{\partial x_1}, \frac{\partial(x_1 + 2x_2 + 3x_3)}{\partial x_2}, \frac{\partial(x_1 + 2x_2 + 3x_3)}{\partial x_3} \right)^T = (1, 2, 3)^T = \mathbf{a}$$

Example 7

Show $\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{x}$, where $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{A} = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$.

$$\begin{aligned} \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} &= \frac{\partial}{\partial \mathbf{x}} \left[(x_1 + 2x_2, 3x_1 + 4x_2)(x_1, x_2)^T \right] \\ &= \frac{\partial}{\partial \mathbf{x}} (x_1^2 + 5x_1x_2 + 4x_2^2) \\ &= \begin{pmatrix} 2x_1 + 5x_2 \\ 5x_1 + 8x_2 \end{pmatrix} = \begin{pmatrix} 2 & 5 \\ 5 & 8 \end{pmatrix} \mathbf{x} = \left[\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix} + \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \right] \mathbf{x} \\ &= (\mathbf{A} + \mathbf{A}^T) \mathbf{x} \end{aligned}$$

Method of Lagrange Multipliers

- **Goal:** Minimize $f(\mathbf{x})$, subject to $(g_1(\mathbf{x}), \dots, g_K(\mathbf{x})) = \mathbf{c} \in \mathbb{R}^K$, with $\mathbf{x} \in \mathbb{R}^n$
- **Solution:**
 - 1 Define *Lagrangian* $L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{k=1}^K \lambda_k (g_k(\mathbf{x}) - c_k)$, with *Lagrange multipliers* λ_k .
 - 2 Set the gradient vector $\left(\frac{\partial L}{\partial x_1}, \dots, \frac{\partial L}{\partial x_n}, \frac{\partial L}{\partial \lambda_1}, \dots, \frac{\partial L}{\partial \lambda_K} \right)^T = \mathbf{0}$
 - 3 Solve the resulting system of $n + K$ equations.
- **Key Idea:** Setting $\frac{\partial L}{\partial \lambda_k} = 0$ automatically enforces the constraint $g_k(\mathbf{x}) = c_k$.

Example 8

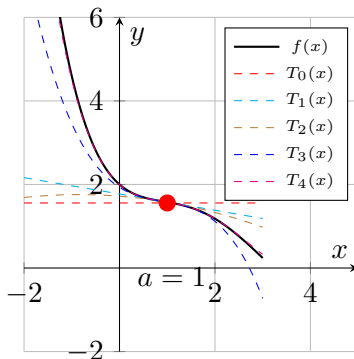
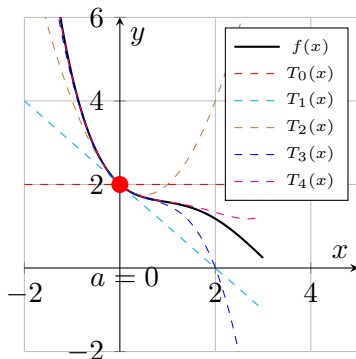
Minimize $x_1^2 + x_2^2 + x_3^2$ subject to $x_1 + x_2 + x_3 = 1$.

- 1 $L(x_1, x_2, x_3, \lambda) = \sum_{i=1}^n x_i^2 - \lambda(\sum_{i=1}^n x_i - 1)$
- 2 $\left(\frac{\partial L}{\partial x_1}, \frac{\partial L}{\partial x_2}, \frac{\partial L}{\partial x_3}, \frac{\partial L}{\partial \lambda} \right) = (2x_1 - \lambda, 2x_2 - \lambda, 2x_3 - \lambda, \sum_i x_i - 1) = \mathbf{0}$
- 3 On solving: $\lambda = 2x_i, \forall i \Rightarrow 1 = \sum_i x_i = \frac{3\lambda}{2} \Rightarrow \lambda = \frac{2}{3} \Rightarrow x_i = \frac{1}{3}, \forall i$

Taylor Series

Taylor Series is a **polynomial approximation** of function $f(x)$ around $x = a$:

$$f(x) = f(a) + f'(a)(x - a) + \frac{1}{2!}f''(a)(x - a)^2 + \frac{1}{3!}f^{(3)}(a)(x - a)^3 + \dots$$



Idea: Let $f(x) = f(a) + c_1(x - a) + c_2(x - a)^2 + \dots$. Find $\left. \frac{\partial^k f}{\partial x^k} \right|_{x=a}$ and solves for c_k 's.

Taylor Series (Two Dimensional Case)

A **polynomial approximation** of function $f(x, y)$ around $(x, y) = (a, b)$:

$$\begin{aligned} f(x, y) = & f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b) \\ & + \frac{1}{2!} f_{xx}(a, b)(x - a)^2 + f_{xy}(a, b)(x - a)(y - b) + \frac{1}{2!} f_{yy}(a, b)(y - b)^2 \\ & + \frac{1}{3!} f_{xxx}(a, b)(x - a)^3 + \frac{1}{2!} f_{xxy}(a, b)(x - a)^2(y - b) \\ & + \frac{1}{2!} f_{xyy}(a, b)(x - a)(y - b)^2 + \frac{1}{3!} f_{yyy}(a, b)(y - b)^3 \\ & + \dots \end{aligned} \tag{1}$$

Idea: Let $f(x, y) = f(a, b) + c_x(x - a) + c_y(y - b) + c_{xy}(x - a)(y - b) + \dots$

Find $\left. \frac{\partial^{j+k} f}{\partial x^j \partial y^k} \right|_{(x,y)=(a,b)}$ and solves for $c_{\alpha\beta}$'s.

- Heuristically, (1) can be expressed in a **differential form** as

$$df(x, y) = f_x dx + f_y dy + \frac{1}{2!} f_{xx}(x, y) dx^2 + f_{xy}(x, y) dx dy + \frac{1}{2!} f_{yy}(x, y) dy^2 + \dots$$

- In Chapter 6, we study the **Ito's lemma**, a stochastic version of Taylor Series:

$$df(t, W_t) = f_t(t_0, W_0) dt + f_w(t_0, W_0) dW_t + \frac{1}{2!} f_{ww}(t_0, W_0) dW_t^2$$

Agenda

- 1 Teachers
- 2 Course Content
- 3 Grading
- 4 Prerequisites**
 - Financial Risks
 - Mathematics and Statistics
 - Regression**

Regression

- Regression is a method for describing the **relationship between predictor variables** $X_1, \dots, X_p$ and the **response variable** Y as

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon,$$

where

- $\epsilon \sim N(0, \sigma^2)$ is an error term.
- $\beta_0, \beta_1, \dots, \beta_p, \sigma^2$ are unknown parameters to be determined
- Example: Predict GPA using IQ, hours of studies, ..., DSE scores

Subject	Y (GPA)	X_1 (IQ)	X_2 (Hours)	$\dots$ $\dots$	X_p (DSE score)
1	y_1	x_{11}	x_{21}	$\dots$	x_{p1}
2	y_2	x_{12}	x_{22}	$\dots$	x_{p2}
3	y_3	x_{13}	x_{23}	$\dots$	x_{p3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{1n}	x_{2n}	$\dots$	x_{pn}

Regression: Estimation

- Regression: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$
- Observed variables from a sample of size n :
 - $\mathbf{y} = (y_1, \dots, y_n)^T$
 - $\mathbf{x}_i = (x_{i1}, \dots, x_{in})^T$, $i = 1, \dots, p$.
 - $\mathbf{X} = (\mathbf{1} \ \mathbf{x}_1 \ \cdots \ \mathbf{x}_p) \in \mathbb{R}^{n \times (p+1)}$
- Least Squares Estimation for $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \arg \min_{\boldsymbol{\beta}} \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_{1j} - \cdots - \beta_p x_{pj})^2 = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y},\end{aligned}$$

by solving (noting $(\mathbf{X}\boldsymbol{\beta})^T \mathbf{y} = \mathbf{y}^T \mathbf{X}\boldsymbol{\beta}$ since both are scalar)

$$\frac{d}{d\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{d}{d\boldsymbol{\beta}} (\mathbf{y}^T \mathbf{y} - 2\mathbf{y}^T \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X}\boldsymbol{\beta}) = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = 0.$$

Regression: Implementation in R

```
lm(y~x1+x2+x3)
```

Example 9

- A sample of 8 RMSC graduates is randomly selected. Their information are recorded as follows
 - GPA=(3.57,3.43, 3.36, 3.63, 3.94, 3.85, 2.96, 4.0)
 - IQ=(110, 106, 100, 121, 130, 127, 136, 96)
 - Hours of studies per week = (10, 12, 8, 5, 16, 15, 3, 20)
- If your IQ is 150 and study 2 hours/week, what is your expected GPA?
- Analysis in R:

```
# Input data
```

```
GPA=c(3.57,3.43, 3.36, 3.63, 3.94, 3.85, 2.96, 4.0)
```

```
IQ=c(110, 106, 100, 121, 130, 127, 136, 96)
```

```
Hour=c(10, 12, 8, 5, 16, 15, 3, 20)
```

```
# Fitting
```

```
fit=lm(GPA~IQ+Hour)
```

```
# Prediction
```

```
newX=data.frame(IQ=150,Hour=2)
```

```
predict(fit,newX)
```

```
# Answer: GPA=3.193
```