THE CHINESE UNIVERSITY OF HONG KONG
Department of Mathematics
MMAT5330
Econometrics Principles and Data Analytics
Due Date: February 24, 2024 before 11:59 PM

Name: ________________________ Student ID.: ________________________

I declare that the assignment here submitted is original except for source material explicitly acknowledged, the piece of work, or a part of the piece of work has not been submitted for more than one purpose (i.e. to satisfy the requirements in two different courses) without declaration, and that the submitted soft copy with details listed in the "Submission Details" is identical to the hard copy, if any, which has been submitted. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained on the University website https://www.cuhk.edu.hk/policy/academichonesty/

It is also understood that assignments without a properly signed declaration by the student concerned will not be graded by the course teacher.

________________________________ ________________________________
Signature Date

General Regulations

- All assignments will be submitted and graded on Gradescope. You can view your grades and submit regrade requests there as well. For submitting your PDF homework on Gradescope, here are a few tips.

  Where is Gradescope?

  Do the following:

  1. Go to 2023R2 Econometric Principles and Data Analysis (MMAT5330)
  2. Choose Tools in the left-hand column
  3. Scroll down to the bottom of the page
  4. The green Gradescope icon will be there
- Late assignments will receive a grade of 0.
- Write your COMPLETE name and student ID number legibly on the cover sheet (otherwise we will not take any responsibility for your assignments). Please write your answers using a black or blue pen, NOT any other color or a pencil.

  For the declaration sheet:

  Either

  Use the attached file, sign and date the statement of Academic Honesty, convert it into a PDF and submit it with your homework assignments via Gradescope.

  Or

  Write your name on the first page of your submitted homework, and simply write out the sentence "I have read the university regulations."
- Write your solutions on A4 white paper or use an iPad or other similar device to present your answers and submit a digital form via Gradescope. Please do not use any colored paper and make sure that your written solutions are a suitable size (easily read). Please be aware that you can only use a ball-point pen to write your answers for any exams.
- Show all work for full credit. In most cases, a correct answer with no supporting work will NOT receive full credit. What you write down and how you write it are the most important means of your answers getting good marks on this homework. Neatness and organization are also essential.

Instructions for Homework 1

Please attempt to solve all the problems.

- Your solutions of problems 1 - 6 are to be submitted.
- Problems 7 - 10 are optional question.

Please be aware that the symbols we used are different from our lecture notes and lab exercises. Try to understand their meaning(s) and fit for the purpose of the following questions. You are free to use any computing tools to plot any figures and to assist you to answer all the required questions. There are also many theoretical questions that may be of interest for further investigations and studies.

[**Objective:**] The types of questions on this homework assignment are classified into three groups:

- Theoretical;
- Computational;
- Usage of Computing Software, namely `MATLAB` and/or **R** and/or **Python**.

The objective of this homework assignment is to understand:

- the concepts, applications and properties of probability and statistics;
- the concepts and applications of the simple linear regression model;
- the properties of least squares estimators;
- the applications of prediction, goodness-of-fit, and modeling issues.

Question 1. X and Y are discrete random variables with the following joint distribution:

		Value of Y				
		14	**22**	**30**	**40**	**65**
Value of X	**1**	0.02	0.05	0.10	0.03	0.01
	5	0.17	0.15	0.05	0.02	0.01
	8	0.02	0.03	0.15	0.10	0.09

That is, $P(X = 1, Y = 14) = 0.02$, and so forth.

(1) Calculate the probability distribution, mean, and variance of Y.

(2) Calculate the probability distribution, mean, and variance of Y given $X = 8$.

(3) Calculate the covariance and correlation between X and Y.

Question 2. Suppose $Y_1, Y_2, \cdots, Y_N$ is a random sample from a population with mean μ and variance σ^2. Rather than using all N observations, consider an easy estimator of μ that uses only the first two observations

$$Y^* = \frac{Y_1 + Y_2}{2}$$

(1) Show that Y^* is a linear estimator.

(2) Show that Y^* is an unbiased estimator.

(3) Find the variance of Y^*.

(4) Explain why the sample mean of all N observations is a better estimator than Y^*.

Question 3. Table 1 shows the mileage and the price of Japanese automobiles.

Japanese Car Brands	Vehicle	Mileage (mpg)	Price ($'000)
Mazda MPV V6	1	19	14.944
Nissan Van 4	2	19	14.799
Acura Legend V6	3	20	24.76
Mitsubishi Wagon 4	4	20	14.929
Nissan Axxess 4	5	20	13.949
Mitsubishi Sigma V6	6	21	17.879
Nissan Stanza 4	7	21	11.65
Mazda 929 V6 8	8	21	23.3
Nissan Maxima V6	9	22	17.899
Toyota Cressida	10	23	21.498
Nissan 240SX 4	11	24	13.249
Subaru Loyale 4	12	25	9.599
Mitsubishi Galant 4	13	25	10.989
Honda Prelude Si 4WS 4	14	27	13.945
Subaru XT 4	15	28	13.071
Mazda Protege 4	16	32	6.599
Honda Civic CRX Si 4	17	33	9.41
Subaru Justy 3	18	34	5.866
Toyota Tercel 4	19	35	6.488

TABLE 1. Mileage and Price of Japanese automobiles

The CSV data file for Question 3 can be obtained from the 2023R2 Econometric Principles and Data Analysis (MMAT5330) course on Blackboard, under the "Homework" folder.

Answer the following questions:

(1) Determine the following:

i Mean Mileage $$\overline{M} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

ii Median Mileage $$\text{median} = \begin{cases} \text{middle observation} & \text{if } n \text{ odd;} \\ \text{average of middle two observations} & \text{if } n \text{ even} \end{cases}$$

iii Mean Absolute Deviation $$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |M_i - \overline{M}|$$

iv Mean Squared Deviation $$\text{MSD} = \frac{1}{n}\sum_{i=1}^{n} (M_i - \overline{M})^2$$

v Sample Variance $$s_M^2 = \frac{1}{n-1}\sum_{i=1}^{n} (M_i - \overline{M})^2$$

vi Sample Standard deviation $$s_M = \sqrt{s_M^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (M_i - \overline{M})^2}$$

where the following table may be useful:

Japanese Car Brands	i	M_i	$(M_i - \overline{M})$	$\lvert M_i - \overline{M}\rvert$	$(M_i - \overline{M})^2$
Mazda MPV V6	1	19			
Nissan Van 4	2	19			
Acura Legend V6	3	20			
Mitsubishi Wagon 4	4	20			
Nissan Axxess 4	5	20			
Mitsubishi Sigma V6	6	21			
Nissan Stanza 4	7	21			
Mazda 929 V6 8	8	21			
Nissan Maxima V6	9	22			
Toyota Cressida	10	23			
Nissan 240SX 4	11	24			
Subaru Loyale 4	12	25			
Mitsubishi Galant 4	13	25			
Honda Prelude Si 4WS 4	14	27			
Subaru XT 4	15	28			
Mazda Protege 4	16	32			
Honda Civic CRX Si 4	17	33			
Subaru Justy 3	18	34			
Toyota Tercel 4	19	35			

(2) Determine the following:

i Covariance between M and P $$\text{Cov}_{MP} = \text{Cov}(M,P) = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{M})(P_i - \overline{P})$$

ii Correlation between M and P $$r_{MP} = \frac{\text{Cov}_{MP}}{s_P s_M} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{M})(P_i - \overline{P})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{M})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(P_i - \overline{P})^2}}$$

where the following table may be useful:

i	M_i	P_i	$M_i - \overline{M}$	$P_i - \overline{P}$	$(M_i - \overline{M})^2$	$(P_i - \overline{P})^2$	$(M_i - \overline{M})(P_i - \overline{P})$
1	19	14.944					
2	19	14.799					
3	20	24.76					
4	20	14.929					
5	20	13.949					
6	21	17.879					
7	21	11.65					
8	21	23.3					
9	22	17.899					
10	23	21.498					
11	24	13.249					
12	25	9.599					
13	25	10.989					
14	27	13.945					
15	28	13.071					
16	32	6.599					
17	33	9.41					
18	34	5.866					
19	35	6.488					

(3) Consider the following:

$$\widehat{\text{Price}} = a + b \times \text{mileage}$$

or

$$\widehat{P}_i = a + b \times M_i, \quad i = 1, \cdots, 19$$

Determine the following:

i Estimated Slope $$b = \frac{\sum_{i=1}^{n}(M_i - \overline{M})(P_i - \overline{P})}{\sum_{i=1}^{n}(M_i - \overline{M})^2}$$

ii Estimated Intercept $$a = \overline{P} - b\overline{M}$$

iii Estimated Coefficient of Determination $$R^2 = r^2_{P\widehat{P}} = \frac{\sum_{i=1}^{n}(\widehat{P}_i - \overline{P})^2}{\sum_{i=1}^{n}(P_i - \overline{P})^2}$$

iv Standard error of the regression (SER) $$s_{\widehat{u}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(P_i - \widehat{P}_i)^2}$$

where the following table may be useful:

i	M_i	P_i	$\widehat{P}_i$	$P_i - \overline{P}$	$P_i - \widehat{P}_i$	$\widehat{P}_i - \overline{P}$
1	19	14.944				
2	19	14.799				
3	20	24.76				
4	20	14.929				
5	20	13.949				
6	21	17.879				
7	21	11.65				
8	21	23.3				
9	22	17.899				
10	23	21.498				
11	24	13.249				
12	25	9.599				
13	25	10.989				
14	27	13.945				
15	28	13.071				
16	32	6.599				
17	33	9.41				
18	34	5.866				
19	35	6.488				

Please note that the formula for the estimated variance of the errors, i.e.,

$$s_{\widehat{u}}^2 = \widehat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}(P_i - \widehat{P}_i)^2$$

is not quite the same as the sample variance of the residuals. The sample variance would have a divisor of $n-1$, whereas here we have used $n-2$. The sum of squared errors is divided by the degrees of freedom, which can be defined as the number of data points minus the number of parameters estimated. In this case, we have estimated two parameters a and b, so the degrees of freedom are two less than the total number of observations.

Question 4. Suppose that a random sample of 200 twenty-year-old men is selected from a population and that these men's height and weight are recorded. A regression of weight on height yields

$$\widehat{Weight} = -99.41 + 3.94 \times Height,\ R^2 = 0.81,\ SER = 10.2$$

where *Weight* is measured in pounds, *Height* is measured in inches and SER stands for the standard error of the regression.

(1) What is the regression's weight prediction for someone who is 70 in. tall? 65 in tall? 74 in. tall?

(2) A man has a late growth spurt and grows 1.5 in. over the course of a year. What is the regression's prediction for the increase in this man's weight?

(3) Suppose that instead of measuring weight and height in pounds and inches these variables are measured in centimeters and kilograms. What are the regression estimates from this new centimeter-kilogram regression? (Give all results, estimated coefficients, R^2, and SER)

Question 5. A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam while others have 120 minutes.

Each students is randomly assigned one of the examination times based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i^{th} student ($0 \leq Y_i \leq 100$), let X_i denote the amount of time that the student has to complete the exam ($X_i = 90$ or 120), and consider the regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

(1) Explain what the term u_i represents. Why will different students have different values of u_i?

(2) Explain why $E(u_i|X_i) = 0$ for this regression model.

(3) **The Least Squares Assumptions**

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \cdots, n,$$

where

I The error term u_i has conditional mean zero given X_i: $E(u_i|X_i) = 0$;

II $(X_i, Y_i), i = 1, \cdots, n$, are independent and identically distributed (i.i.d.) draws from their joint distribution; and

III Large outliers are unlikely: X_i and Y_i have nonzero finite fourth moments.

Are the above assumptions satisfied? Explain.

(4) The estimated regression is $\widehat{Y_i} = 49 + 0.24X_i$.

(a) Compute the estimated regression's prediction for the average score of students given 90 minutes to complete the exam. Repeat for 120 minutes and 150 minutes.

(b) Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.

Question 6. Table 2 presents data on the aggregate consumption (Y, in billions of HK dollars) and disposable income X, also in billions of HK dollars) for a developing economy over a 12-year period from 2012 to 2023.

Year	i	Y_i	X_i
2012	1	102	114
2013	2	106	118
2014	3	108	126
2015	4	110	130
2016	5	122	136
2017	6	124	140
2018	7	128	148
2019	8	130	156
2020	9	142	160
2021	10	148	164
2022	11	150	170
2023	12	154	178

TABLE 2. Aggregate Consumption (Y) and Disposable Income (X)

The CSV data file for Question 6 can be obtained from the 2023R2 Econometric Principles and Data Analysis (MMAT5330) course on Blackboard, under the "Homework" folder. There is no need to submit two figures for this question. However, visualizing two figures and comparing their results could certainly be beneficial for you.

Answer the following questions:

(1) Draw a scatter diagram using the data and visually inspect whether there exists an approximate linear relationship between Y and X.

(2) Determine the simple regression equation for the consumption schedule in Table 2, using $\widehat{\beta}_1 = \dfrac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2}$ to find $\widehat{\beta}_1$ and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$ to find $\widehat{\beta}_0$.

(3) *(Optional)* Plot the simple regression line and illustrate the deviations of each Y_i from the corresponding $\widehat{Y}_i$.

(4) For the aggregate consumption-income observation in Table 2, use the results from (2) to compute

(a) $s^2_{\widehat{u}}$

(b) $s^2_{\widehat{\beta}_0}$

(c) $s^2_{\widehat{\beta}_1}$

(5) Use the results from (2) to compute r for the estimated consumption regression using the following methods:

(a) $\sqrt{R^2}$

(b) $r = \dfrac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$

(c) $r = \sqrt{\widehat{\beta}_1 \dfrac{\sum x_i y_i}{\sum y_i^2}}$

where $x_i = X_i - \overline{X}$, and $y_i = Y_i - \overline{Y}$.

(6) Derive the following:

(a) Starting with $\widehat{\beta}_1 = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$, derive the equation for $\widehat{\beta}_1$ in deviation form for the case where $\overline{X} = \overline{Y} = 0$.

(b) Determine the value of $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$ when $\overline{X} = \overline{Y} = 0$.

(c) *(Optional)* Use the results of (a) and (b) to plot the simple regression line on a graph where the variables are measured as deviations from their respective means. Compare this simple regression line with the simple regression line plotted in (3) on the same graph.

Question 7. *(Optional)* A household has weekly income of \$2,000. The mean weekly expenditure for households with this income is

$$E(Y|X = \$2,000) = \mu_{Y|X=\$2,000} = \$200,$$

and expenditures exhibit variance

$$\mathrm{Var}(Y|X = \$2,000) = \sigma^2_{Y|X=\$2,000} = 100.$$

Answer the following questions:

(a) Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$180 and \$215 on food in a week. Include a sketch with your solution.

(b) Find the probability that a household with this income spends more than \$250 on food in a week. Include a sketch with your solution.

(c) Find the probability in part (a) if the variance of weekly expenditures is

$$\mathrm{Var}(Y|X = \$2,000) = \sigma^2_{Y|X=\$2,000} = 81.$$

(d) Find the probability in part (b) if the variance of weekly expenditures is

$$\mathrm{Var}(Y|X = \$2,000) = \sigma^2_{Y|X=\$2,000} = 81.$$

Question 8. *(Optional)* Let X, Y, and V be random variables, let μ_X and σ^2_X be the mean and variance of X, let σ_{XY} be the covariance between X and Y (and so forth for the other variables), and let a, b, and c be constants. Show that Properties (1) through (7) follow from the definitions of the mean, variance, and covariance:

(1) $E(a + bX + cY) = a + b\mu_X + c\mu_Y$

(2) $\mathrm{Var}(a + bY) = b^2\sigma^2_Y$

(3) $\mathrm{Var}(aX + bY) = a^2\sigma^2_X + 2ab\sigma_{XY} + b^2\sigma^2_Y$

(4) $E(Y^2) = \sigma^2_Y + \mu^2_Y$

(5) $\mathrm{Cov}(a + bX + cV, Y) = b\sigma_{XY} + c\sigma_{VY}$

(6) $E(XY) = \sigma_{XY} + \mu_X\mu_Y$

(7) $|\mathrm{Corr}(X, Y)| \leq 1$ (or $|r_{XY}| \leq 1$) and $|\sigma_{XY}| \leq \sqrt{\sigma^2_X\sigma^2_Y}$ (correlation inequality)

Question 9. *(Optional)* X is a Bernoulli random variable with $P(X = 1) = 0.99$, Y is distributed $N(0, 1)$, W is distributed $N(0, 100)$, and X, Y, and W are independent. Let

$$S = XY + (1 - X)W.$$

(That is, $S = Y$ when $X = 1$, and $S = W$ when $X = 0$.)

Answer the following questions:

(a) Show that $E(Y^2) = 1$ and $E(W^2) = 100$.

(b) Show that $E(Y^3) = 0$ and $E(W^3) = 0$. (Hint: What is the skewness for a symmetric distribution?)

(c) Show that $E(Y^4) = 3$ and $E(W^4) = 3 \times 100^2$. (Hint: Use the fact that the kurtosis is 3 for a normal distribution.)

(d) Derive $E(S), E(S^2), E(S^3)$ and $E(S^4)$. (Hint: Use the law of iterated expectations conditioning on $X = 0$ and $X = 1$.)

(e) Derive the skewness and kurtosis for S.

Question 10. *(Optional)* X is a random variable with moments $E(X), E(X^2), E(X^3)$, and so forth. Answer the following questions:

(a) Show that $E(X-\mu)^3 = E(X^3) - 3[E(X^2)][E(X)] + 2[E(X)]^3$.

(b) Show that $E(X-\mu)^4 = E(X^4) - 4[E(X)][E(X^3)] + 6[E(X)]^2[E(X^2)] - 3[E(X)]^4$.