

Assignment 2

Instructions:

- There are 3 questions in this assignment, complete them all.
- There are 10 datasets, your lecturer will assign one dataset to you. Marks will be deducted for the wrong dataset used.
- Present your solutions in a Word document. Indicate the question number.
- State the dataset you are using at the beginning of the report. e.g. “2. Condo Tamp Paris Ris H1 price”
- Save the file as **x_name_admin.doc**, where x is your dataset number, name is your name, and admin is your admission number.
- Submit your report, code and video to Brightspace. The report should not exceed 13 pages.
- For the Python code, paste it into the report as instructed in the questions. In additional, place the code for Q1, Q2 and Q3 in one Jupyter notebook and submit it as well. Indicate clearly which question number the code is for.
- An oral presentation may be required at your tutor’s discretion.
- Submit the *Declaration of Academic Integrity* before submitting your assignment.
- Use the following template to acknowledge the use of any AI Tools.

Name of AI tool	< For example, ChatGPT >
Input prompt	< Insert the question that you asked ChatGPT >
Date generated	< Insert the date that ChatGPT response was generated, since ChatGPT is an evolving technology >
Output generated	< Insert the response verbatim from ChatGPT >
Impact on submission	< Briefly explain which part of your submitted work was ChatGPT’s response applied >

Introduction

Given a set of data points with at least one predictor and one continuous response variable, we want to construct a linear model to predict the response. This is the aim of **Linear Regression**, which is a supervised learning technique.

In the context of this assignment, data on life expectancies for the Singapore resident population from 2018 to 2020 are [extracted](#). The data file is zipped together with this document. The following table lists two of the variables in the file and their descriptions:

Variable	Description
<i>Age_x</i>	Age x (years)
<i>Expectation of life at age x</i>	The length of time (in years) that an individual can expect to live at age x.

The response variable is *Expectation of life at age x*, and the predictor is *Age_x*.

Simple Linear Regression (SLR)

We will first build a SLR model using *Age_x* as the predictor to predict *Expectation of life at age x*.

In SLR notations, let:

x_i = predictor value of the i -th data point

y_i = actual response value of the i -th data point

\hat{y}_i = predicted response value of the i -th data point based on model

Thus, $\hat{y}_i = a + bx_i$, where values of a (intercept) and b (slope) are to be determined.

The squared-error of the i th prediction is $e_i^2 = (y_i - \hat{y}_i)^2$. Errors (also known as residuals) are squared to remove the signs, so that errors of opposite signs do not cancel out each other, giving the false impression of small aggregated errors.

Then, we define **Error function** as the mean of squared-error (of the whole data set):

$$\begin{aligned}
 E(a, b) &= \frac{1}{n} \sum_{i=1}^n e_i^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\
 &= \frac{1}{n} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2] \dots (1)
 \end{aligned}$$

We want to find the values of a and b such that the Error function is **minimised**.

The resultant equation $\hat{y} = a + bx$ will give the best-fit line that passes through the data points.

MODEL 1: SLR with intercept a fixed $\Rightarrow \hat{y}_i = bx_i$ (30 marks – 10 marks for each sub-qn)

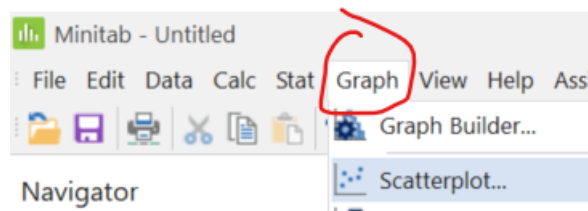
Revision on SLR: <https://youtu.be/HoqXask9cN8>

We will first build a SLR model to predict *Expectation of life at age x* (y) using *Age_ x* (x) as the predictor.

To get things started, we will simplify the regression line of $\hat{y} = a + bx$ to $\hat{y} = bx$ by setting $a = 0$, this means that \hat{y} is a constant multiple of x with $a = 0$. In this SLR model 1, we will only need to determine the slope, b , in $\hat{y} = bx$.

Note that the Expectation of life at age x is believed to be negatively correlated to Age_ x . Hence, the regression line of $\hat{y} = bx$ may not be a suitable regression line for this dataset. A more practical model will be $\hat{y} = a + bx$ which you will be building in Model 2. We simplify the problem at this stage to prepare us to build Model 2.

1(a) Use Minitab to plot the scatter plot of *Expectation of life at age x* and *Age_ x* with the regression line in it, paste the scatter plot to your report. Write down the equation of the regression line obtained from Minitab, in the context of your dataset, e.g. *sale of ice-cream = 2.31 temperature*.



(Hint: This question we want a regression line of $\hat{y} = a + bx$ with $a = 0$, to do that in Minitab, look for “Fit Intercept” in “Data View” in the Scatterplot options, then uncheck it)

=====

In this assignment, we are going to learn how to obtain the regression line from scratch by using the gradient descent algorithm, which will involve differentiation.

We are going to use the univariate gradient descent algorithm to obtain the value of the slope, b , which is reported by Minitab in Q1(a).

The video in Brightspace: Week 14 – Video 4 (see pic) explains the gradient descent algorithm.

Video	Section	Content
Video 1	8.5	Local Extrema
	8.5.1	Example 7 (First Derivative Test)
Video 2	8.5.2	Second Derivative Test
		Example 8
Video 3	8.5.3	Global Extrema
		Examples 9 & 10
Video 4	8.6	Optimisation
		Example on Univariate Gradient Descent

This YouTube video explains the error function (Equation 1) in page 2 and its derivative. Do watch until 8:55 min only: <https://youtu.be/sDv4f4s2SB8>

Note:

- (i) There is a slight difference in the loss function in the YouTube video, and in page 2, Equation 1, there is a division by n in Equation 1. In all your calculations for this assignment, stick to Equation 1:

$$\frac{1}{n} [(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2]$$

- (ii) In the video, it explains the concept of obtaining the value of the intercept. That can be applied to calculate the value of the slope in Q1(b)

1(b) Express Error function $E(b)$ in terms of b only since $a=0$, indicate the value of n in your Error function. Hence, derive $E'(b)$.

1(c) Use univariate gradient descent algorithm to find the value of b for which $E(b)$ is at its minimum.

1c(i) Write your Python code in a single cell and copy-paste (do NOT take screenshots) your **code** and **output** into your report. Refer to the sample output below as an example.

Sample output: Number of iterations is xx
 The local minimum occurs when b is xx
 Minimum error is xx

1c(ii) Write down the equation of the regression line you have obtained from gradient descent in the context of your dataset, e.g. $\widehat{\text{sale of ice-cream}} = 2.31 \text{ temperature}$.

Hint: Adjust the learning rate to get the minimum error. You may continue to watch this video from 8:55 to 15:40 min to enhance your understanding of the algorithm of gradient descent: <https://youtu.be/sDv4f4s2SB8?t=534>

MODEL 2: SLR $\Rightarrow \hat{y}_i = a + bx_i$

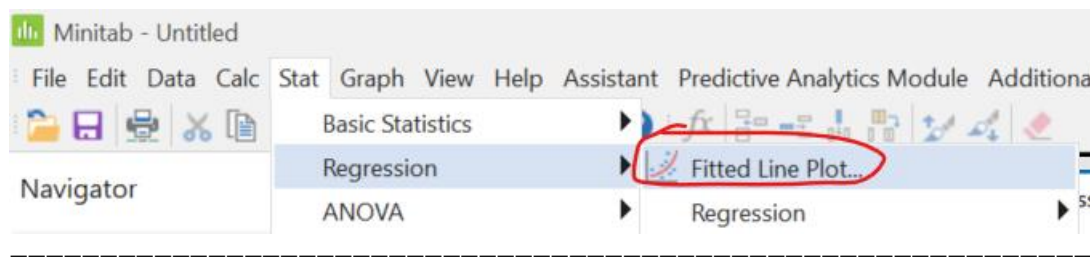
(40 marks – 10 marks each sub-qn)

Now we apply the SLR model where **both intercept a and slope b are to be determined**, when predicting *Expectation of life at age x* (y) using *Age_x* (x) as the predictor.

2(a) Use Minitab to plot the scatter plot of *Expectation of life at age x* and *Age_x* with the regression line in it, paste the scatter plot to your report. Your plot should include the equation of the regression line obtained from Minitab, in the context of your dataset, e.g.

$$\widehat{\text{sale of ice-cream}} = 10.9 + 2.31 \text{ temperature}$$

Use the regression line to estimate the value of the response variable with a suitable predictor value of your choice.



We are going to use gradient descent to obtain the regression line, $\hat{y} = a + bx$, from scratch. As there are 2 variables involved, we will need to apply partial derivative to the error function. You can start watching videos in week 16 folder in Brightspace to learn partial derivatives.

You may want to watch this YouTube video from 15:43 till the end, for the implementation of gradient descent involving two variables: <https://youtu.be/sDv4f4s2SB8?t=943>

2(b) Express Error function $E(a, b)$ in terms of a and b . Hence, derive $E_a(a, b)$ and $E_b(a, b)$.

2(c) Use gradient descent algorithm **to find the values of a and b** for which $E(a, b)$ is at its minimum.

2c(i) Write your Python code in a single cell and copy-paste (do NOT take screenshots) your **code** and **output** into your report, refer to the sample output below as an example.

Sample output: Number of iterations is xx
 The local minimum occurs when b is xx
 Minimum error is xx

2c(ii) Write down the equation of the regression line you have obtained from gradient descent in the context of your dataset, e.g. $\widehat{\text{sale of ice-cream}} = 10.9 + 2.31 \text{ temperature}$.

2(d) Describe the changes and decisions you made on the parameters for your solution to reach convergence.

MODEL 3: MLR $\Rightarrow \hat{y}_i = a + bx_i + cw_i$

(30 marks – 10 marks each sub-qn)

We can extend the SLR model to include more predictors. A linear regression model with more than one predictor is called **Multiple Linear Regression (MLR)** model.

Apply the MLR model where intercept a , and slopes b and c are to be determined, when predicting *Expectation of life at age x* (y) using *Age_x* (x) and w as the predictors.

You decide on a suitable variable w based on the context of your dataset.

Explain your Model 3 in a video. Your face must be visible for the entire duration of the recording.

Your video should include the following:

3(a) Data Collection

- Pick at least 10 records (rows) from your existing dataset, then insert data for the field w . The inserted data can come from the existing data given (if any), your own measurements, some randomly assigned values, etc.
- Explain your data collection procedure.
- Copy and paste the records you used to create Model 3 into your report.

3(b) Implementation: Error Function

- Obtain the expressions for the error function and any related functions needed for the gradient descent algorithm.
- Explain how the above expressions are derived mathematically.
- Explain how the above expressions are used in the gradient descent algorithm to find the values of a , b and c in the regression line, $\hat{y} = a + bx + cw$

3(c) Implementation: Coding and Verification

- Code the gradient descent algorithm in Python to find the values of a , b and c
- Explain your code.
- Run your code with the data inserted to demonstrate how you obtain the regression line.
- Explain and demonstrate how you verify that your regression line for Model 3 is correct.
- Copy-paste (do NOT take screenshots) your code and output into your report and include it in the .ipynp file.

Summary of files to be submitted:

1. Word document
2. Video file for Q3
3. One .ipynp file that contains the codes for Q1, Q2 and Q3. Label the question number clearly in your code.

Marks will be awarded for clear explanations.

Your video should not exceed 4 minutes. Marks will be deducted for video exceeding 4 minutes.

END OF ASSIGNMENT