Assignment 2

Instructions:

- Type your answers in the spaces provided in this Word document. Your submission should not exceed 11 pages, including this page.
- Submit the *Declaration of Academic Integrity* before submitting your assignment.

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 is collected from 35 staff employed in a pharmaceutical company. The data can be found in the file *salary.xlsx*. The following table lists the variables used in the file and their descriptions:

Variable	**Description**
salary	Salary in dollars per hour earned by staff
years	Number of years staff has been with company
gender	1 = male, 0 = female

The response variable is *salary*, and the predictors are *years* and *gender*.

Use *pandas.read_excel* to extract the data from *salary.xlsx* into a dataframe.

Simple Linear Regression (SLR)

We will first build a SLR model using *years* as the predictor to predict *salary*.

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 sum of squared-error (of the whole data set):

$$E(a,b) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

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

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 $\Rightarrow \hat{y}_i = 50 + bx_i$ (25 marks)

We will first build a SLR model to predict *salary* (*y*) using *years* (*x*) as the predictor.
Suppose any new staff who is employed by the company will earn a minimal salary of $50/hour. This means that when $x = 0$, $\hat{y} = a = 50$. Then, in the SLR model, we will only need to determine slope *b*.

(a) Express Error function $E(b)$ in terms of b only. Hence, derive $E'(b)$.

(b) Use univariate gradient descent algorithm to find the value of b for which $E(b)$ is at its minimum. Write your Python code in a single cell and copy-paste your code below.

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

(d) Describe your MODEL 1 by filling the information below.

Final MODEL 1 equation is:

Minimum value of Error function is:

Number of iterations ran to reach convergence:

MODEL 2: SLR $\Rightarrow$ $\hat{y}_i = a + bx_i$ (25 marks)

Now we apply the SLR model where both intercept a and slope b are to be determined, when predicting *salary* (y) using *years* (x) as the predictor.

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

(b) Use gradient descent algorithm to find the values of a and b for which $E(a, b)$ is at its minimum. Write your Python code in a single cell and copy-paste your code below.

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

(d) Describe your MODEL 2 by filling the information below.

Final MODEL 2 equation is:

Minimum value of Error function is:

Number of iterations ran to reach convergence:

MODEL 3: MLR $\Rightarrow$ $\boldsymbol{\hat{y}_i = a + bx_i + cg_i}$ (25 marks)

We can extend the SLR model to include more predictors. A linear regression model with more than 1 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 *salary* (y) using *years* (x) and *gender* (g) as the predictors.

(a) Explain how gradient descent algorithm can be extended for MODEL 3.

(b) Use gradient descent algorithm to find the values of a, b and c for which Error function is at its minimum. Write your Python code in a single cell and copy-paste your code below.

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

(d) Describe your MODEL 3 by filling the information below.

Final MODEL 3 equation is:

Minimum value of Error function is:

Number of iterations ran to reach convergence:

Conclusion (25 marks)

(a) Using Python (or other software), in a single figure, plot the data points (scatterplot) together with the linear lines representing the three models. Insert the figure below.

Note:

- The categorical variable *gender* can be represented in a bivariate scatterplot as legend (typically in colour).
- MODEL 3 equation can be written as two separate equations, one representing male and one representing female.

(b) Compare the 3 models. Which model will you use to predict salary in this context?