## **LINEAR REGRESSION - TASK 4**



|  |  |
| --- | --- |
| **NAME** | ABDUL MUIZZ, MUHAMMAD ABDULLAH |
| **REGISTRATION NUMBER** | FA23-BCS-117, FA23-BCS-108 |
| **DATE OF SUBMISSION** | 3RD OCTOBER 2025 |
| **MODERATOR** | SIR UMAR NAUMAN |

**Table of Contents:**

[**LINEAR REGRESSION - TASK 4** 1](#_Toc210395115)

[**Abstract:** 3](#_Toc210395116)

[**Part 1: Simple Linear Regression (Canada per Capita Income)** 4](#_Toc210395117)

[**Problem Statement:** 4](#_Toc210395118)

[**Dataset Description:** 5](#_Toc210395119)

[**Methodology:** 6](#_Toc210395120)

[**Data Analysis and Visualization:** 7](#_Toc210395121)

[**Model Implementation:** 8](#_Toc210395122)

[**Result and Analysis:** 9](#_Toc210395123)

[**Part 2: Multiple Linear Regression (Hiring Dataset)** 10](#_Toc210395124)

[**Problem Statement:** 10](#_Toc210395125)

[**Dataset Description:** 11](#_Toc210395126)

[**Dataset Preprocessing:** 11](#_Toc210395127)

[**Methodology:** 12](#_Toc210395128)

[**Model Implementation:** 13](#_Toc210395129)

[**Result and Analysis:** 14](#_Toc210395130)

[**List of Figures:** 15](#_Toc210395131)

[**Conclusion:** 16](#_Toc210395132)

[**References:** 18](#_Toc210395133)

## **Abstract:**

This project explores the application of linear regression models for predictive analysis on two distinct datasets. First, a **simple linear regression** model is developed to predict Canada's per capita income based on the year. This model demonstrates a foundational understanding of regression and achieves an R² score of **0.7738**, indicating a reasonably good fit. Second, a **multiple linear regression** model is implemented to predict a candidate's salary using features like experience, test score, and interview score. After thorough data cleaning and preprocessing, this model achieves a very high R² score of **0.9497**, showcasing its strong predictive power. The project covers key machine learning steps including data loading, preprocessing, model training, evaluation, and prediction.

## **Part 1: Simple Linear Regression (Canada per Capita Income)**

## **Problem Statement:**

The primary objective of this task is to build a simple linear regression model that can accurately predict the per capita income of Canada for a given year. The model will be trained on historical data from 1970 to 2016 and then used to make a prediction for the year 2020.

## **Dataset Description:**

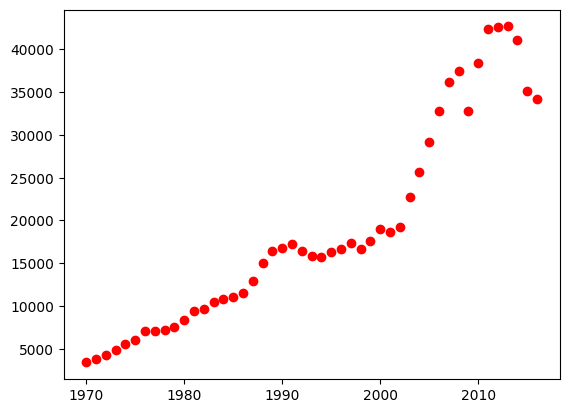
* **Source:** canada\_per\_capita\_income.csv
* **Total Records:** 47 entries
* **Features:**
  + year: The year (1970-2016).
  + per capita income (US$): The per capita income for that year.
* **Data Quality:** The dataset is clean with no missing values.

## **Methodology:**

* **Data Loading:** The dataset was loaded using the pandas library.
* **Data Splitting:** The data was split into training and testing sets with a **1/3** test size ratio to evaluate the model's performance on unseen data.
* **Model Training:** A simple linear regression model from scikit-learn was trained on the training data.

## **Data Analysis and Visualization:**

A scatter plot of the dataset (Figure 1) was created to visualize the relationship between the year and per capita income. The plot shows a clear positive linear trend, suggesting that linear regression is a suitable model for this data.



**Figure 1:** Scatter plot of Year vs. Per Capita Income

## **Model Implementation:**

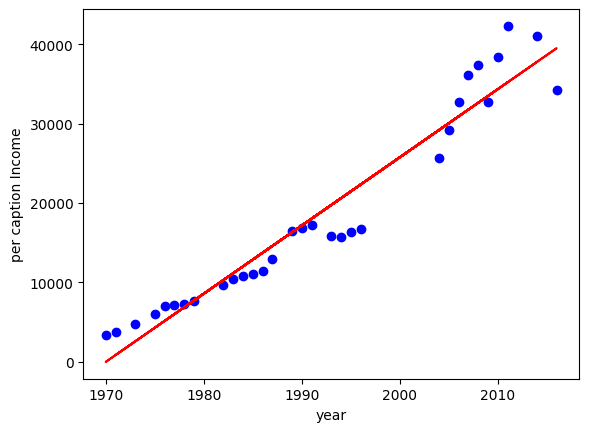
A LinearRegression model was instantiated and trained using the .fit() method on the training data (x\_train, y\_train). The trained model was then used to make predictions.

## **Result and Analysis:**

 **Model Evaluation:** The model's performance was evaluated on the test set, achieving an R-squared (R²) score of **0.7738**. This indicates that approximately 77.4% of the variance in per capita income can be explained by the year.

 **Prediction:** The model was used to predict the per capita income for the year 2020. **Prediction for 2020:** **$42,960.58**

 **Model Visualization:** The regression line was plotted against the training data to visualize the model's fit (Figure 2). The line effectively captures the upward trend of the data points.



**Figure 2:** Linear Regression Model Fit on Training Data

## **Part 2: Multiple Linear Regression (Hiring Dataset)**

## **Problem Statement:**

The goal is to build a multiple linear regression model to predict the salary of a potential employee based on their years of experience, test score (out of 10), and interview score (out of 10). This involves handling missing data and converting categorical features into a numerical format.

## **Dataset Description:**

* **Source:** hiring.csv
* **Total Records:** 8 entries
* **Features:**
  + experience: Years of experience (contains numerical, text, and NaN values).
  + test\_score(out of 10): Score in a technical test (contains NaN values).
  + interview\_score(out of 10): Score in an interview.
  + salary($): The final salary offered (target variable).
* **Data Quality:** The dataset contains missing values (NaN) and non-numerical data in the 'experience' column, requiring preprocessing.

## **Dataset Preprocessing:**

Several steps were taken to clean and prepare the data for modeling:

1. **Handling Missing Values:**
   * The missing experience values were filled with **0**, assuming no experience.
   * The missing test\_score was filled with **8.0**, which is the median/mode value.
2. **Feature Encoding:**
   * The text-based numbers in the experience column (e.g., 'five', 'two') were converted to their corresponding integer values (e.g., 5, 2).

## **Methodology:**

1. **Data Loading & Cleaning:** The dataset was loaded and preprocessed as described above.
2. **Data Splitting:** The cleaned data was split into an 80% training set and a 20% testing set.
3. **Model Training:** A multiple linear regression model was trained on the three input features (experience, test\_score, interview\_score) to predict salary($).

## **Model Implementation:**

A LinearRegression model from scikit-learn was trained on the preprocessed training data. This model learns a linear equation of the form:

salary=m1​×experience+m2​×test\_score+m3​×interview\_score+c

## **Result and Analysis:**

* **Model Evaluation:** The model performed exceptionally well on the test data, achieving an R-squared (R²) score of **0.9497**. This signifies that nearly 95% of the variability in salary can be explained by the combination of experience, test score, and interview score.
* **Predictions:** The model was used to predict salaries for two hypothetical candidates:

|  |  |  |  |
| --- | --- | --- | --- |
| **Experience (yrs)** | **Test Score (/10)** | **Interview Score** | **Predicted Salary ($)** |
| 2 | 9 | 6 | **$54,565.22** |
| 12 | 10 | 10 | **$90,869.57** |

## **List of Figures:**

1. **Figure 1:** Scatter plot of year vs per capita income
2. **Figure 2:** Linear Regression model fit on training data

## **Conclusion:**

* **Key Findings:**
* Both simple and multiple linear regression models were successfully implemented.
* The simple linear regression model for Canada's per capita income showed a good fit (**R² = 0.77**), confirming the strong linear relationship over time.
* The multiple linear regression model for salary prediction demonstrated a very strong fit (**R² = 0.95**), proving its effectiveness after appropriate data preprocessing.
* **Model Strengths:**
* **Simplicity and Interpretability:** Linear regression is easy to understand and the coefficients can be interpreted to understand feature importance.
* **Fast Performance:** The model is computationally efficient and trains quickly, even on larger datasets.
* **Limitations:**
* **Linearity Assumption:** The model assumes a linear relationship between features and the target, which may not hold true for all datasets.
* **Preprocessing Impact:** The salary prediction model's accuracy is highly dependent on the chosen methods for handling missing data (e.g., filling with 0 or the median).
* **Future Work:**
* For the income dataset, a polynomial regression could be explored to capture any non-linear trends.
* For the hiring dataset, more advanced imputation techniques could be used for missing values.
* The models could be compared against other regression algorithms like Decision Trees or Support Vector Regression.

## **References:**

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine
2. Scikit-learn Documentation: Linear Regression. (n.d.). Retrieved from scikit-learn.org
3. Pandas Documentation. (n.d.). Retrieved from pandas.pydata.org