**ML Assignment 2 - Linear Regression (Simple and Multiple) [40 pts]**

Regression is an analytical model that can identify the linear relationships (correlation) of variables. Simple or Multiple Linear regression can be used as an effective tool in Machine Learning. There is an important difference between classification and regression problems. Fundamentally, classification is about predicting a label and regression is about predicting a quantity. The Dependent Variable (DV) is typically a continuous variable.

This assignment consists of two parts:

* Part 1 - Create a Simple Linear Regression (SLR), a single predictor variable predictive model using birthweight data provided and discuss your findings
* Part 2 - Create a Multiple Linear Regression (MLR), a multiple predictor variable predictive model using birthweight data provided and discuss your findings

**Source Dataset:**

The data Birthweight\_reduced.csv will be used for creating and testing the two models i) Simple Linear Regression, ii) Multiple Linear Regression. This (real) dataset contains information on newborn babies and their parents. It contains all continuous except for one variable “LowBirthWeight” which is categorical. Hence this dataset is most useful for correlation and regression. The attribute metadata is available in the below table. Birthweight is the dependent variable.

|  |  |  |
| --- | --- | --- |
| **Name** | **Variable** | **Data type** |
| **ID** | Baby number |  |
| **length** | Length of baby (inches) | integer |
| **Birthweight** | Weight of baby (lbs) | floating |
| **headcirumference** | Head Circumference | integer |
| **Gestation** | Gestation (weeks) |  |
| **smoker** | Mother smokes 1 = smoker 0 = non-smoker | Binary |
| **motherage** | Maternal age | integer |
| **mnocig** | Number of cigarettes smoked per day by mother | integer |
| **mheight** | Mothers height (inches) | integer |
| **mppwt** | Mothers pre-pregnancy weight (lbs) | integer |
| **fage** | Father's age | integer |
| **fedyrs** | Father’s years in education | integer |
| **fnocig** | Number of cigarettes smoked per day by father | integer |
| **fheight** | Father's height (inches) | integer |
| **lowbwt** | Low birth weight, 0 = No and 1 = yes | integer |
| **mage35** | Mother over 35, 0 = No and 1 = yes | Binary |
| **LowBirthWeight** | Low or normal birth weight | Category |

**PART A (SLR) [20 pts]**

**Research question:**

1. Check if baby birthweight is dependent upon mother’s pre-pregnancy weight
2. Check if the baby length is dependent upon mother's height

Create two separate Models Using the Python Jupyter Notebook to investigate the above two research questions as follows.

Step 1: import all our standard libraries. Also include seaborn

Import seaborn as sns

1. Load the CSV file into a dataframe df
2. Explore the top 5 rows of this dataframe
3. How many rows and columns does it have?
4. See all the columns this dataset has

Step 2: Using seaborn visualize the pairplots between 'headcirumference', 'length', 'Birthweight', 'mppwt', 'mheight'

Comment of what kind of relationship do you see from the plots?

Step 3: compute the correlations of all these variables. Then show a heatmap of the correlations.

Step 4: You will now develop two SLR Models to answer to the two research questions:

1. Research question# 1, configure the attribute “mppwt” as predictor(independent) variable and “Birthweight” outcome (dependent) variable.
2. Research question# 2, configure the attribute “mheight” as predictor(independent) variable and “Length” outcome (dependent) variable

Step 4: You will repeat the following steps to first check relationship between “mppwt” and “Birthweight” and later between “mheight” and “length” of the baby.

Step 5: Constructing the training and testing sets for the 2 SLR.

1. Put only the “mppwt” data in X

X = df['mppwt'].values.reshape(-1, 1)

Make the y matrix for “Birthweight”.

y = df['Birthweight'].values

Check to make sure your X and y matrix are good.

2. import necessary libraries for slit test and train sets. Make the testing set 20%.

3. Create and fit your Linear Regression models.

4. Plot the linear regression line on training set and testing set. Use plt.plot function.

4. Print the intercepts and the coefficients. What do you find? Explain.

Repeat step 5 for the second research question.

**PART B [20 pts]**

**Research question:**

In this part, we like to investigate which other variables are significant in predicting the birthweight? Also, check how close your prediction is to actual values?

Step 1: Some data cleanup may be necessary. Let us go ahead and drop the features “id

”, “LowBirthWeight”, and “lowbwt”.

Verify that these columns are now gone.

Step 2: Setup your X matrix with all the independent variables left and your target of prediction is “Birthweight” which goes to y matrix.

Step 3: Split your data set into training and testing with a split ratio of 75:25.

Step 4: After importing the necessary Linear Regression libraries, create and fit your model.

Step 5: Print your intercept value. What do you find here?

Step 6: Print all your coefficients. You will see this as an array.

Step 7: Make your predictions. Remember you provide X\_test data now.

Step 8: Check your prediction value with actual real values.

Do a scatter plot to see how they align.

Step 9: from sklearn import metrics

Show the values MAE, MSE, RMSE.

Step 10: Save and upload your Jupyter Notebook.

What do you conclude from this multiple regression analysis?