To find the best suited multiple linear regression model for the weekly earnings in the US.

Sneha Ramesh Dharne
National College Of Ireland
Dublin, Ireland
x23195703@student.ncirl.ie

***Abstract*—**

This study primarily examines and applies the Multiple Linear Regression Model on the provided dataset, "microwage.csv," which is a Comma Separated Values file. The goal of the investigation is to determine which multiple linear regression model is most appropriate for US weekly wages. utilizing the model of multiple linear regression.

Additionally, this research looks at the variables that are regarded as independent, such as '*region*', '*statefip*', '*metaread*', '*puma*', '*age*', *edyrs*', '*female*', '*race_nonwhite*', '*perwt*', '*industry*','*expyrs*', 'mutually exclusive job classification', 'additional job classification' etc; as well as the one considered dependent variable - '*wkwage*'. As part of the exploratory data analysis it investigates the distribution of each of the individual variables. It uses suitable visualisation to decide about possible transformations. Further using visualisation and suitable statistical tests to investigate the pairwise relations between each of the independent variables and the dependent variable.

Additionally, in order to evaluate and investigate the performance of the Multiple Linear Regression model, metrics such as the Mean Squared Error (MSE) and R-Squared Value have been generated. The Python code utilized for this analysis is run in the Jupiter Notebook. Numerous Python libraries and modules, such as NumPy, Matplotlib, sklearn, pandas, and others, are utilized in the code to carry out various statistical tests, plot visualizations, analyze multiple linear regression models, and manipulate data from the provided dataset, in that order.

Keywords—
Multiple Linear Regression, Independent variables, Dependent variables, Visualizations

I. INTRODUCTION

A kind of statistical method called multiple linear regression makes use of two or more independent variables to forecast or ascertain their relationship. Multiple regression refers to the regression between the response variable and the predictors when there are more than one predictor variable, such as two, three, or n predictor variables (where n is not equal to 1). Multiple linear regression describes the relationship between the predictors and the response variable.

In the following report, multiple linear regression will be used. The "'microwage " dataset is been provided on the moodle, which is a csv file.

A. This report's study uses multiple linear regression to look at the several aspects that may impact house prices. In this examination, the considered independent variables are 'region', 'statefip', 'metaread', 'puma', 'age', edyrs', 'female', 'race_nonwhite', 'perwt', 'industry','expyrs', 'mutually exclusive job classification', 'additional job classification' etc; as well as the one considered dependent variable - 'wkwage'. Maintaining the Integrity of the Specifications

II.

A variety of visualizations, including the histogram, scatterplot, coefficient plot, box plot, and others, have been used to show and plot the analysis. Calculating multiple metrics, such as the Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared value, is another aspect of the analysis.

Regarding the code, several Python libraries and modules, including pandas, numpy, seaborn, matplotlib.pyplot, sklearn.model_selection, sklearn.linear_model, sklearn.metrics, scipy.stats, statsmodels.stats.stattools, and statsmodels.stats.outliers_influence, are utilized for dataframe creation, multiple mathematical calculations, plotting and animated visualizations, and performing regression on the dataset.

II. METHODOLOGY

1) Data Collection :

- • The moodle hosted the dataset that was under consideration. The dataset, which is a Comma Separated Values (csv) file, is named "microwage.csv."
- From the given dataset, a set of independent variables like '*region*', '*statefip*', '*metaread*', '*puma*', '*age*', *edyrs*', '*female*', '*race_nonwhite*', '*perwt*', '*industry*','*expyrs*',

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

'mutually exclusive job classification', 'additional job classification' etc; as well as the one considered dependent variable - '*wkwage*'.

2) Data Cleaning :

The "microwage.csv" dataset that was supplied was already clean and prepared for analysis. However, identifying and eliminating the outliers was essential to obtaining a smooth analytical outcome. Consequently, the necessary Python modules and libraries are used to handle and remove the outliers.

3) Data Splitting :

- Essentially, the methodology's phase involves splitting the regarded or provided dataset (here, the microwave.csv' dataset) into training and testing sets in order to assess the model's performance on the unobserved data.
    - Using the 'train_test_split' function, the provided dataset is divided into training and testing sets for this study. The test_size that is being considered is 0.2.

4) Multiple Linear Regression Modeling :

Typically, there were two steps involved in this methodological step: the model's specification and fitting. Consequently, the definition or examination of the independent and dependent variables is carried out in accordance with the exploratory data analysis in the FIRST phase, which is the "Specification of the Model" section. In contrast, the multiple regression model is trained using the considered training set by modifying the coefficients in the SECOND phase, which is the "Fitting of the Model" component. [1] This is done in order to reduce or eliminate the discrepancy between the projected and actual values, accordingly.

Now, according to this analysis, the multiple linear model is been fitted or applied by using the function 'LinearRegression()' and 'model.fit' from the python module 'sklearn.linear_model' respectively.

5) Model Evaluation :

This process involved two main tasks once more: first, the testing set was evaluated, and then the residuals were studied or analyzed. The model's performance is examined in the FIRST section in order to evaluate or research its generalizability. In the 'Residual Analysis' section of the second part, it is determined if the residuals are randomly distributed around zero. If they are, this indicates that the regression model that was chosen is appropriately fitted and that it is also capturing all of the significant patterns found in the dataset. Additionally, if it is determined that the residuals are not randomly distributed around zero, the chosen regression

Relative to this analysis, residuals have been computed using the formula. The difference between the expected and actual numbers is the formula used to determine the residuals. In this investigation, the abolute standardized residuals have also been computed using the NumPy method 'np.abs'.

6) Multicollinearity Check :

- In essence, multicollinearity check is defined as a regression problem that arises when two or more predictor variables in the selected dataset have a high degree of correlation with one another. This ultimately impacts how the regression model's (in this case, multiple linear regression model) results are interpreted.

7) **Exploratory Data Analysis (EDA):**

- This phase involves performing descriptive statistics, which entails analyzing and studying all variables—both independent and dependent—to comprehend the distribution and average properties of the data.
- Additionally, a variety of visualization techniques, like as the histogram, boxplot, scatterplot, coefficient plot, and others, are employed to examine and evaluate the patterns and relationships between the variables under consideration.

8) Interpretation of the results :

- The regression model has been interpreted, calculated, and analyzed in this investigation. Numerous metrics, including mean squared error, R-squared value, and mean absolute error, have been successfully calculated. Analysis of each independent variable's coefficients has been completed.

III.EXPLORATORY DATA ANALYSIS

1) Descriptive statistics :

The process of performing descriptive statistics involves examining and analyzing all variables—both independent and dependent—in order to ascertain the distribution and average characteristics of the data.

Furthermore, a range of visual aids, including the histogram, boxplot, scatterplot, coefficient plot, and others, are utilized to scrutinize and assess the correlations and patterns among the variables in question.

A. The interpretation of the findings and the analysis of the coefficients are also completed in this analysis. Next, the coefficient values of the independent variables under consideration are calculated. In this case, the study also includes the computation of other metrics, such as the R-squared value and the Mean Squared Error (MSE) and Mean Absolute Error (MAE).

```
print(df.describe())
```

	region	statefip	metaread	puma	perwt
count	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06
mean	2.782701e+01	2.784883e+01	3.392802e+03	2.210346e+03	1.046662e+02
std	1.033126e+01	1.588875e+01	2.966956e+03	2.391580e+03	7.446740e+01
min	1.100000e+01	1.000000e+00	0.000000e+00	1.000000e+02	1.000000e+00
25%	2.100000e+01	1.300000e+01	0.000000e+00	8.000000e+02	6.400000e+01
50%	3.100000e+01	2.700000e+01	3.120000e+03	1.800000e+03	8.500000e+01
75%	3.300000e+01	4.100000e+01	5.960000e+03	3.202000e+03	1.180000e+02
max	4.200000e+01	5.600000e+01	9.360000e+03	7.777700e+04	1.954000e+03

	age	female	race_nonwhite	edyrs	occ_managprof
count	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06
mean	4.046613e+01	4.989468e-01	1.602207e-01	1.365140e+01	3.188441e-01
std	1.319065e+01	4.999991e-01	3.668107e-01	2.617561e+00	4.660287e-01
min	1.700000e+01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.900000e+01	0.000000e+00	0.000000e+00	1.200000e+01	0.000000e+00
50%	4.100000e+01	0.000000e+00	0.000000e+00	1.348000e+01	0.000000e+00
75%	5.100000e+01	1.000000e+00	0.000000e+00	1.615000e+01	1.000000e+00
max	6.500000e+01	1.000000e+00	1.000000e+00	1.800000e+01	1.000000e+00

	occ_techsalad	occ_service	occ_farm	occ_product	occ_operator
count	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06	1.349258e+06
mean	2.980875e-01	1.644971e-01	8.651422e-03	9.235669e-02	1.175631e-01
std	4.574183e-01	3.707262e-01	9.260983e-02	2.895289e-01	3.220902e-01
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
75%	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
max	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00

Fig.1 - Descriptive Statistics

A. Plotting of a Scatter Plot :

Scatterplots are a sort of visual aid that are mostly used to determine the relationship between two chosen independent variables in a chosen dataset (in this example, the housing.csv dataset file). Every data point in the chosen dataset is represented by a dot in the scatter plot.

For more detailed understanding of the concept of the Scatterplot, refer the figure below (Fig. 2).

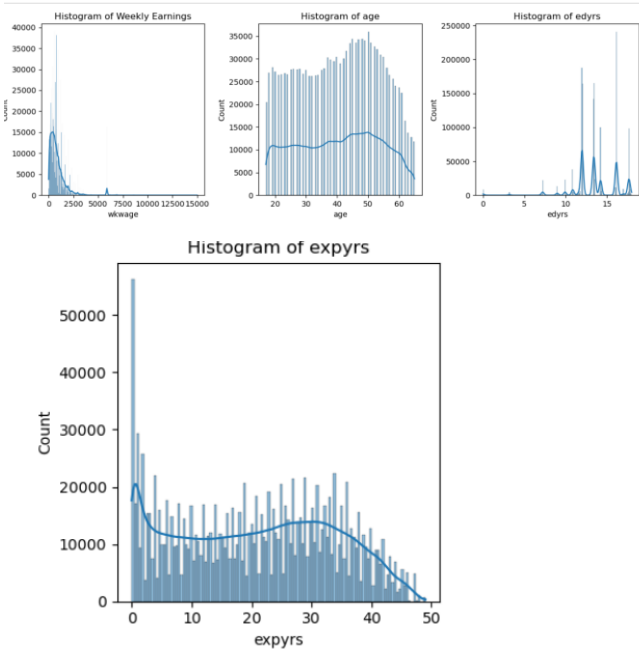


Fig. 2 histograms to visualize the distributions of the dependent and independent variables

Microwage.csv is the name of Moodle. As indicated by the file's extension, it is a Comma Separated Values file. Following a thorough examination of the dataset, the dependent variable—that is, the target variable—and independent variables are chosen in order to carry out the analysis (in this example, the Multiple Linear Regression Model analysis).

Data Cleaning: The provided data is made clean by eliminating any outliers from the provided dataset. We start by identifying the outliers. Subsequently, those anomalies are removed with the use of several Python routines. This was carried out to ensure that the chosen regression model operated cleanly, smoothly, and well.

Data Splitting: In this study, the provided dataset is divided into training and testing sets, respectively, using the function "train_test_split."Test_size is 0.2 in this case.

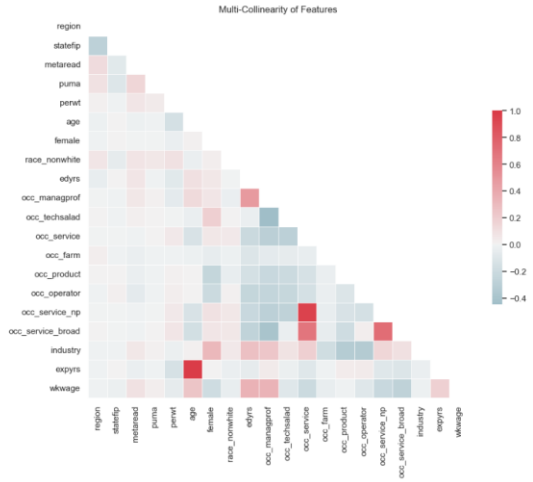
IV. MODELLING

This step consisted of the Model's Specification and the Model Fitting process, in essence. Thus, we can state that the first component, which is the "Model Specification" part of the Modelling, defines or takes into account the dependent and independent variables in accordance with the exploratory data analysis (EDA). On the other hand, in the second step, referred to as "Model Fitting," the multiple regression model is trained using the traiData Splitting: In this study, the provided dataset is divide d into training and testing sets, respectively, using the functio n "train_test_split."Test_size is 0.2 in this case

'LinearRegression()' and'model.fit' from the Python module'sklearn.linear_model' are used to fit and apply the multiple linear regression model, respectively, based on the analysis results.

```
In [23]: # using R squared value to check how much prediction is good and how much is not good
         from sklearn.metrics import r2_score
         r2_score(y_train, y_pred_train) # portion of variance for dependet variable - wkwage can
Out[23]: 0.28038764670033856
```

In this analysis of the multiple linear regression, the correlation matrix has also been computed.
PLEASE VIEW THE FIGURES BELOW FOR A CLEARER UNDERSTANDING OF THE COMPUTED CORRELATION MATRIX.



VIII. EVALUATION

THIS STEP CONSISTS OF TWO STEPS: THE TESTING SET WAS EVALUATED FIRST, AND THE RESIDUALS WERE EXAMINED OR EXAMINED IN ORDER TO DO THE ANALYSIS. IN THE FIRST

SECTION, THE MODEL'S PERFORMANCE IS BEING EXAMINED IN ORDER TO ASSESS ITS OVERALL PERFORMANCE. IN THE SECOND SECTION, THE RESIDUALS ARE EXAMINED TO DETERMINE WHETHER OR NOT THEY WERE DISTRIBUTED RANDOMLY ABOUT ZERO. IF SO, THIS MEANS THAT THE SELECTED REGRESSION MODEL (IN THIS CASE, MULTIPLE LINEAR REGRESSION MODEL) IS FITTING APPROPRIATELY AND HAS CAPTURED ALL THE NECESSARY PATTERNS AVAILABLE IN THE GIVEN DATASET. THE REGRESSION MODEL IS FAILING TO EXTRACT THE NECESSARY DATA FROM THE PROVIDED DATASET IF THE RESIDUALS ARE NOT BEING DISTRIBUTED RANDOMLY AROUND ZERO RESPECTIVELY.

REFERENCES

[1] Performance Analysis of Machine Learning Algorithm on Cloud Platforms: AWS vs Azure vs GCP | SpringerLink

[2] ORGANON: SOUTHWEST OREGON - PDF Free Download (businessdocbox.com)

[3]ubir.buffalo.edu/xmlui/bitstream/handle/10477/79900/Chauhan_buffalo_0656M_16362.pdf?sequence=3