A REPORT FOR STATISTICS IN DATA ANALYTICS. CA1 - (35%)

Investigating Multiple Linear Regression Models Best Suited for Weekly Earnings in the US

ADEOLA DEBORAH ADENIJI - x23104201
MSCDAD_ A - School of Computing
National College of Ireland, Dublin, Ireland
x23104201@student.ncirl.ie

Abstract*—This report investigates a wide range of intricate variables that affect a crucial economic metric: weekly earnings for individuals in the US. With data from over 1.3 million individuals between the ages of 17 and 65, this research provides an excellent sample of the working-age population in the U.S. The investigation uses advanced statistical techniques, specifically linear regression analysis, to identify the variables that have a major impact on the variances in weekly earnings. Differentiating between the dependent variable in this case, weekly earnings—and independent variables, such as demographic and employment-related variables. The success of this model will be evaluated using the R-squared statistics, which measures the percentage of weekly earnings variance that can be attributed to the identified independent variables. The greater the R-squared value, the more capable the model is of describing capability.*

Index Terms: *Exploration Data Analysis (EDA), Simple and Multiple Linear Regression Model , Least Square Method.*

I. INTRODUCTION

GENERAL FOUNDATIONS OF THE METHODS APPLIED

Understanding the factors that can influence the weekly earnings in the United States is crucial for well-informed decision-making in domains such as economic policy, wage inequality analysis, and human resource management [1]. This research makes use of data analytics and statistics to determine which linear regression model best explains and forecasts weekly wages in the US across a range of occupations and demographics. For this purpose, a sample dataset was provided called "microwage.csv" and Some milestones called phases has to be reached to complete an in-depth investigation.

Beginning with phase one which is the called "***Data Collection and Description***" phase. This phase [2, p. 4] is important in other to understand the data , to make well informed decisions and also decreasing the possibilities of errors during the investigation. By utilizing the given source file "*microwage"* in this case, an insight is deduced to distinction between the dependent variable and multiple independent variables.

The second phase called "***Exploration Data Analysis"*** (EDA). This is a crucial phase in the investigation process. It provides more in-depth understanding of data patterns, anomalies as well as visualizing through graphical representation [3, p. 2]. Using metrics such as descriptive statistics analysis that describes the central tendency, spread, and distribution that investigates the distribution of each of the individual variables. Additionally, employed analysis such as data cleaning, which addresses outliers, inconsistent data, , missing values in the dataset, etc., variable relationships, which aid in analyzing the connections between each independent variables and the dependent variable, and visualization, which assists in producing informative diagrams to comprehend the distribution of variables, identify potential relationships, and investigate potential transformations. All these would help wen building the model.

In phase three, the focus moved to building a "***Simple and Multiple Linear Regression Model***" to investigate the best suited model for the weekly earns in the US. This model uses an assumption based a linear relation [4] to predict performance based on several variables outcomes. Applying model selections, transformation, and model diagnostics checks (e.g., Normality of errors, Homoscedasticity, and Multicollinearity). This stage aims to establish the relationship between independent variables and the dependent variable "*Weekly wage*" based on the Exploratory Data Analysis (EDA) insights.

The fourth and the final stage called "***The Model Evaluation and Interpretation Phase***". [5] This stage focus on understanding the model performance and it predictions from the analysis of phase 4. It provide insights on the directions and strength of the independent and dependent variables. An assessment made on the model performance called “***Model Fitting***” using metrics like ‘***R-squared, Adjusted R-squared”, and Residual Plots***’ [4]. Furthermore, a “Model interpretation” coefficient of the final model is inducted to understand the direction and strength of the relationships between each independent variable and the weekly earnings. By following these phases, the need to find the best statistical test for the regression model that might be used to estimate US weekly wages and provide insightful information for making decisions.

II. DATA COLLECTION AND DESCRIPTION - (*PHASE ONE*)

The initial phase of this report deals with the collection and understanding of the sourced dataset with. The investigation into the factors that influence weekly earnings in the US begins with a dataset named “microwage.csv”. The dataset contains just over 1.3 million working people across several industries and region within the US, representing about 140 million working people within the age group of about 17 to 65 years. The dataset is divided into the set following categories that contain a variable range of characteristics that have an impact on weekly earnings.

Geographical Data: This contains information such as “region”- an integer encoding of the census region stored in an additional file called “regions.csv”, “statefip”- an IT standard encoding for US states stored in an additional file called “fips.csv”, “metaread”- a code for the metropolitan area stored in an additional file called “metareas.csv”, “puma”- a public use micro area that divides states into areas of at least 100,000 inhabitants. The geographical datasets may not be relevant for the ongoing investigation analysis as they are independent variables and are just stated for record purposes, but they are also included when loading datasets.

Individual Data: This category includes information such as “Age”- as the age of a person in years (from 17 to 65), “edyrs”- as the number of years in education, “female”- as the binary encoding of a legal gender (0=male, 1=female), ”race_ nonwhite”- as the binary encoding of race (0=White Caucasian, 1=other), and ”perwt”- which indicates the number of people across the whole US that the individual represents. The individual data variables a considered and treated as independent variables because they represent characteristics of individual observations in the dataset and are used to support or predict the changes in the dependent variable - “weekly earnings.”

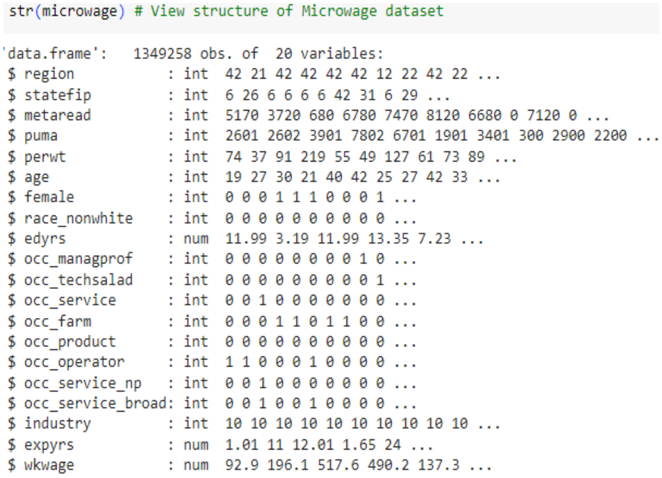
Professional Data: This category involves information such as “Industry”- for industries where each person works using the encoding of industries (1990) standard, “expyrs”- for the numbers of potential years of experience on the job, mutually exclusive job classification which includes data such as “Occ_managprof”- as 1=managerial or professional role, “Occ_techsalad”- as 1=technical and salaried, “occ_service”- as 1=service industry, “occ_farm”- as 1=farming, ” occ_operator” – as 1=machine operator, “occ_product”- as 1=general production and some additional job classification under the professional data are included such as “occ_service_broad”- as 1=new, broader sense of service and “occ_service_np”- as 1=non-professional services. The professional data which is an independent variable provides more insights into the professional backgrounds, experience, and industry codes of individuals in the US in line with the datasets.

Dependent variable: Weekly Wage “*wkwage*”- represents the average weekly wages, accounting for weeks worked, and the number of hours worked each week of each individual. Analyzing this variable with the independent variables will provide insights into factors that influence the investigations, and also to develop a model for explaining and possibly predicting earnings.

Independent Variable: Taking into account the numerical variables contained in the dataset such as Age as “*age*”, Years in Education as “*edyrs*”, People across US as “*perwt*”, Years of Experience as “*expyrs*”. The predictor “*wkwage*” uses the independent variables through a linear equation that will then quantify the strength and the direction of their relationship.

The next step involves utilizing a statistical analysis software “**R**” programming language on “**Jupyter notebook**” environment to conduct the investigation. By using this tool, an effective analysis is made possible. The main data collection is made by loading the “microwage” dataset with the read.csv function to populate the environment with the data.

To get a preliminary view of the dataset, the “*head*” and “*tail*” functions is than used. This is to understand the overall structure and to get an idea of the variable ranges by examining the first and last tenth rows of the dataset. Additionally, the "*names*" function is utilized to view all the column names in the dataset responsible for the behavior of the investigation. While the "*str*" function aids in understanding the structure and data types of the data. The dataset contains a total of “**1,349,258 rows**” and “**20 columns**” across the US. Also the below fig.1 present the over view structure including all numerical and categorical variables of the sample data. For data cleaning, exploration, and subsequent analysis, this information is essential to note.



```
str(microwage) # View structure of Microwage dataset
```

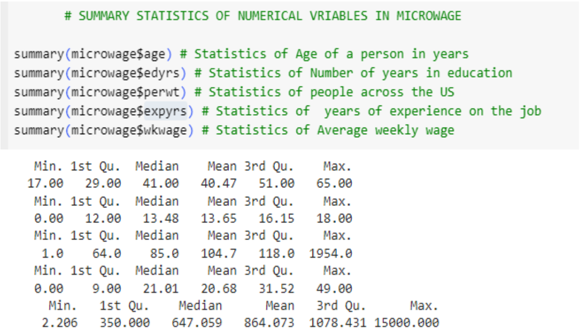
```
'data.frame':   1349258 obs. of  20 variables:
 $ region           : int  42 21 42 42 42 42 12 22 42 22 ...
 $ statefip         : int  6 26 6 6 6 6 42 31 6 29 ...
 $ metaread         : int  5170 3720 680 6780 7470 8120 6680 0 7120 0 ...
 $ puma             : int  2601 2602 3901 7802 6701 1901 3401 300 2900 2200 ...
 $ perwt            : int  74 37 91 219 55 49 127 61 73 89 ...
 $ age              : int  19 27 30 21 40 42 25 27 42 33 ...
 $ female           : int  0 0 0 1 1 1 0 0 0 1 ...
 $ race_nonwhite    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ edyrs            : num  11.99 3.19 11.99 13.35 7.23 ...
 $ occ_managprof    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ occ_techsalad    : int  0 0 0 0 0 0 0 0 0 1 ...
 $ occ_service      : int  0 0 1 0 0 0 0 0 0 0 ...
 $ occ_farm         : int  0 0 0 1 1 0 1 1 0 0 ...
 $ occ_product      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ occ_operator     : int  1 1 0 0 0 1 0 0 0 0 ...
 $ occ_service_np   : int  0 0 1 0 0 0 0 0 0 0 ...
 $ occ_service_broad: int  0 0 1 0 0 1 0 0 0 0 ...
 $ industry         : int  10 10 10 10 10 10 10 10 10 10 ...
 $ expyrs           : num  1.01 11 12.01 1.65 24 ...
 $ wkwage           : num  92.9 196.1 517.6 490.2 137.3 ...
```

Fig.1. **Summary structure of microwage dataset.**

III. EXPLORATION DATA ANALYSIS ‘EDA’ - (*PHASE TWO*)

Following the collection and description of the sample dataset, the next phase is a very crucial and important phase for this investigation called the *Exploratory Data Analysis*. This phase helps to gain deeper understanding of the dataset.

Begun with a “**Descriptive Statistical Analysis**”, by summarizing the numerical variables using the ‘*summary*()’ function. The summary statistics provided valuable insight into the central tendency (*mean*, and *median*), spread (*quartiles*) and the potential outliers (*minimum* and *maximum*) for the numerical variables. The '*age'* variable analysis, which ranges from 17 to 65 years old, closely matches the working-age population that the study is aimed at, having a central tendency that indicates a median age of 41 years, which is marginally younger than the 40.47-year-old mean. The middle 50% of the data are captured by the interquartile range (IQR), which spans 22 years from 29 to 51. This suggests a modest skew towards a younger population within the sample. For the distribution in *'edyrs'* variable, which has a range of 0 to 18 years. This range includes the average length of time spent in formal education; the median value of 13.48 years indicates that most people in the sample have finished their higher education. The distribution in *'perwt'*, which measures how many people in the US each person represents, has a broad range from 1 to 1954 when analyzed. The sample's distribution across various demographic groups and geographic regions in the US varies significantly, as indicated by this variance. Next, the ‘*expyrs*’ variable represents years of work experience. Here, the range of 0 to 49 years. With a mean experience of about 20.68 years, the workforce in the sample appears to be moderately experienced, and a large part of the individuals have accumulated extensive expertise in their respective industries. There is a significant difference in the values of the dependent variable, "*Weekly Wage*," which ranges from 2.21 to 15,000. This wide range suggests that there is a sizable income gap in the population. A strong right-skewed distribution is confirmed by the notable disparity between the mean of 849.024 and median of 647.06 earnings, which highlights the fact that most workers earn less than the average income. Having an IQR of 647.05, ranging from 350.00 to 1078.43 highlights the frequency of lower incomes even further. These insights serve as a basis of the exploratory data analysis, which shows the economic and demographic traits of the working population in the dataset. Below in fig. 2. shows the summary of the above descriptive statistics.



```
      # SUMMARY STATISTICS OF NUMERICAL VRIABLES IN MICROWAGE

summary(microwage$age) # Statistics of Age of a person in years
summary(microwage$edyrs) # Statistics of Number of years in education
summary(microwage$perwt) # Statistics of people across the US
summary(microwage$expyrs) # Statistics of  years of experience on the job
summary(microwage$wkwage) # Statistics of Average weekly wage
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17.00   29.00   41.00   40.47   51.00   65.00 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   12.00   13.48   13.65   16.15   18.00 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0    64.0    85.0   104.7   118.0  1954.0 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    9.00   21.01   20.68   31.52   49.00 
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
    2.206   350.000   647.059   864.073  1078.431 15000.000 
```

Fig.2. **Descriptive Statistical Analysis.**

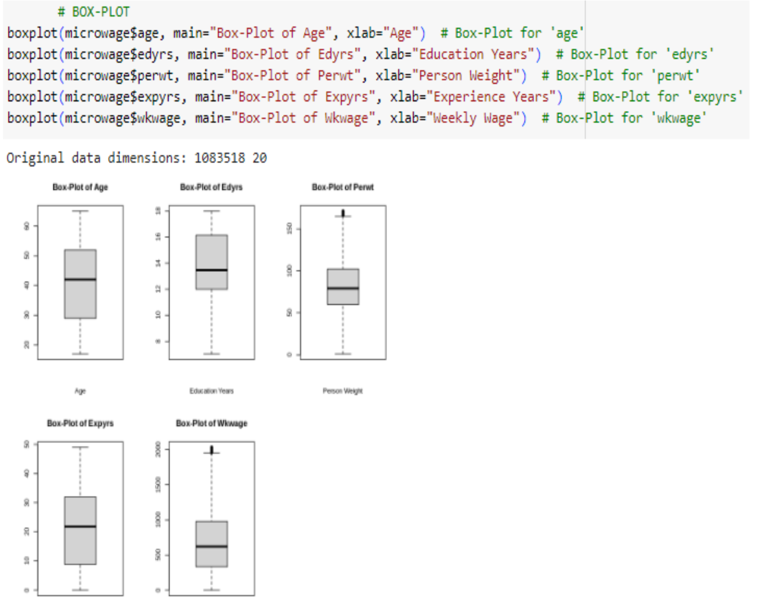
Categorial Variables as Factors and in Tables: The focus turned to the categorical variables in the dataset using the “*as.factor*()” function. These variables were converted into factors to allow for proper statistical analysis. The distribution of each category was then understood by creating frequency tables. The tables may reveal, for example, which gender dominates the dataset or whether any occupations are noticeably underrepresented.

Unique Values: Moreover, “*length(unique*()” was used to determine the total number of unique values in each category. As a result, it is easier to identify categories with sparse observation counts, which may need to be paired with related categories for a more comprehensive analysis. Having a sum length of nine unique variables in “*region*” while the rest variable been binary summed up to just two unique variables.

Missing Values: To verify the accuracy and dependability of the data, a comprehensive check was also performed for missing values in the numerical variables. Finding and addressing missing values is imperative for maintaining the accuracy of the analysis and coming to relevant conclusions. There were no missing values discovered during the check, as each summed up to 0.

Visualization of Dependent and Independent Variables:

Box-Plots were created to detect outliers and anomalies for the numerical variables . In doing so, ‘*age* and *expyrs*’ showed **0** anomalies, while significant outliers of **13661, 131995,** and **71072** in ‘*edyrs*’, ‘*perwt*’, and "*wkwage*" respectively were found. The dataset was then shrunk to **1,140,514** rows as a result of the removal of outliers and still retaining and 20 columns. Through using such measures, it is ensured that extreme data points won't adversely affect the succeeding analyses, producing more trustworthy and consistent findings. Fig.3. shows the image of the boxplot after removing the outliers.



```
      # BOX-PLOT
boxplot(microwage$age, main="Box-Plot of Age", xlab="Age")  # Box-Plot for 'age'
boxplot(microwage$edyrs, main="Box-Plot of Edyrs", xlab="Education Years")  # Box-Plot for 'edyrs'
boxplot(microwage$perwt, main="Box-Plot of Perwt", xlab="Person Weight")  # Box-Plot for 'perwt'
boxplot(microwage$expyrs, main="Box-Plot of Expyrs", xlab="Experience Years")  # Box-Plot for 'expyrs'
boxplot(microwage$wkwage, main="Box-Plot of Wkwage", xlab="Weekly Wage")  # Box-Plot for 'wkwage'
```

```
Original data dimensions: 1083518 20
```

Fig.3. **Box-Plot After Removing Outliers.**

To obtain a deeper understanding of the dependent and independent variables, as well as the relationships between them, visualizations were carried out as part of the EDA phase. In the beginning of the visualization, the distributions of numerical variables were represented using **Histograms**. The histogram for ‘*age and expyrs*’ showed a fairly normal distribution, while ‘*edyrs*’ histogram suggest that there was peak at a certain level of education and made it to be right skewed. For ‘*perwt*’, it showed a large portion of population on the lower left side and fewer people on the right side making it left skewed, and lastly the ‘*wkwage’* histogram is a right-skewed distribution as with fewer high earners on the left side. Then, for the purpose of ensuring that the variables were distributed appropriately and to help normalize skewed distributions, **transformations** such the “**logarithmic** transformation and **square** functions” were implemented.

Furthermore, **Bar-Plots** were created to illustrate the distributions of independent categorical variables and to provide insight into the relative frequency of various categories within each variable. Showing the frequencies in ‘*region*’, ‘*statefip*’, ‘*female*’ etc. was represented.

Visualization Relationship Between Dependent And Independent:

Scatter-Plots were then used to investigate the correlations between independent and dependent variables. Potential correlations and patterns between variables can be found with the help of these visualizations. Initially, the plots were made without lines to show the relationship between the variables in an easy-to-understand manner, it was then recreated with lines to better show trends and make it easier to find any underlying patterns or correlations.

Lastly for the visualization, a **Heatmap** was used to create a correlation matrix, providing a thorough understanding of the correlations between the variables. To evaluate the direction and strength of correlations between variables using this visualization technique, which helped to guide future modeling efforts and provide insightful information about potential multicollinearity.

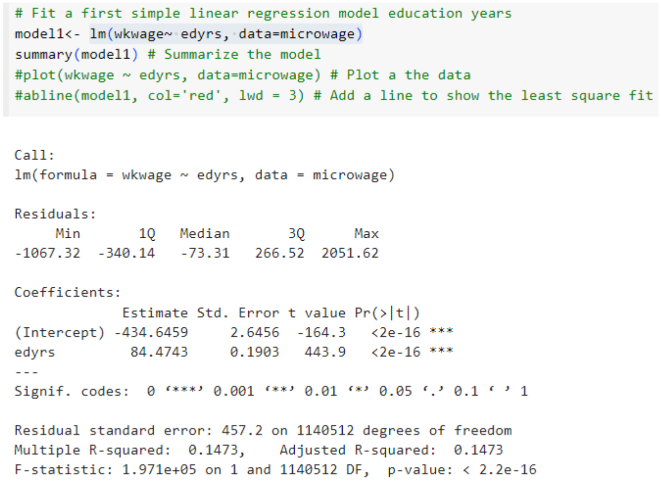
By highlighting patterns, trends, and correlations between variables, these visualizations help us better comprehend the dataset. By carrying out an in-depth visual investigation, one may create the foundation for advanced investigations that lead to the next stage of the analytics journey, which uses modeling approaches and finally yields insightful information about weekly wages in the United States.

IV. MODEL BUILDIG - (*PHASE THREE*)

Moving on from EDA, series of regression analysis would be carried out to find the best possible model for the weekly earnings in the US. The regression analysis would be applied using the **Simple and Multiple Linear Regression.** Starting off in building the model an investigation on simple linear regression was the first model implemented, due to its simplicity, and interpretability. It provided a straightforward approach for analyzing the relationship between independent and dependent variables. Although it assumes a linear relationship, it serves as a foundation for more complex analysis and is especially helpful when examining the first relationships in the sample data ‘*microwage*’.

This model first applied numerical variables that affect weekly wages, it included independent variables, such as ‘*age*’, ‘*expyrs*’, ‘*perwt*’, ‘*edyrs*’ ‘*region*’, ‘*industry*’, ‘*female*’, and ‘*race_nonwhite*’. The model ran eight times to find the best result of the adjusted R square yet proved unsuccessful.

The first model (**Model1**: wkwage ~ edyrs) explored the relationship between years of education and weekly wages. Although p-value < 2.2e-16 indicates a positive statistical significance, the adjusted R-squared value was a lowly 0.1473. It implies that the variance in weekly wages is explained by roughly around 14.7% of cases. On a large dataset of 1,140,512 observations, the residual standard error of 457.2 suggests a significant degree of unexplained variation. Fig.4. shows the first model created using simple linear regression.



```
# Fit a first simple linear regression model education years
model1<- lm(wkwage~ edyrs, data=microwage)
summary(model1) # Summarize the model
#plot(wkwage ~ edyrs, data=microwage) # Plot a the data
#abline(model1, col='red', lwd = 3) # Add a line to show the least square fit
```

```
Call:
lm(formula = wkwage ~ edyrs, data = microwage)

Residuals:
     Min       1Q   Median       3Q      Max 
-1067.32  -340.14   -73.31   266.52  2051.62 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -434.6459     2.6456  -164.3   <2e-16 ***
edyrs         84.4743     0.1903   443.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.2 on 1140512 degrees of freedom
Multiple R-squared:  0.1473,	Adjusted R-squared:  0.1473 
F-statistic: 1.971e+05 on 1 and 1140512 DF,  p-value: < 2.2e-16
```

Fig.4. **First Simple Linear Regression Model.**

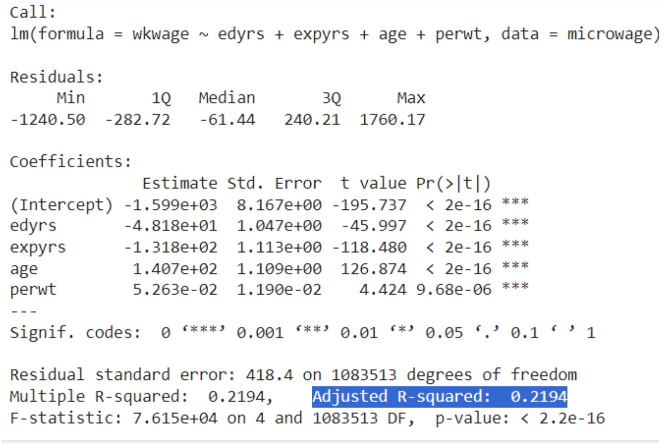
Likewise, **Model 2, 3**, and **4** evaluated between ‘*perwt*’, ‘*age*’, and ‘*expyrs’*, respectively and the relationships with ‘*wkwage*’. The models all obtained statistical significance (p-value < 2.2e-16); however, their adjusted R-squared values, which ranging from 0.08847 to values of 0.05 were significantly unsatisfactory.

Despite producing statistically significant evaluations, these basic linear regression models had little explanatory power when it came to weekly wage changes. This limitation is emphasized by the adjusted R-squared values, which range from a meager 3.6% for gender in **model 7** to 19% for industry **model** 6. The impact of industry on weekly wages, for example, was examined in model 6. Even though it was the simplest model with the highest adjusted R-squared (0.1898), industry alone appears to account for just around 19% of the variation in weekly wages.

These above models findings imply that the complex relationships among variables influencing weekly earnings are more complicated than one independent variable could explain. Most likely, several of these variables affect weekly wages, hence a broader approach is required to for it complexity.

Multiple Linear Regression: The use of this regression model is to have a more thorough grasp of the variables influencing weekly wages. With this technique, the use of several explanatory variables can be applied at once, which could result in a model that is more precise and helpful. Compared to the simple linear model examined, an anticipation for a more complete model that explains a greater output of the variation in ‘wkwage’ by including multiple independent variables at once. This will give a better idea of the variables that affect the weekly wages. During the investigation, the results of the multiple linear regression models progressively increased the explanatory power of the weekly wage. In this section, the use of all available both numerical and categorical were considered in other to get the best result.

The first model (**Multi _Model1:** lm (wkwage~ edyrs + expyrs + age + perwt) explored the relationship between years in education, experience, age and person in the US with the predictor weekly wage. The adjusted R-square value of 0.2194 points around 21.94% of variation in the weekly wage as was seen. This was slightly improved to 0.2226, indicates 22.26% difference by the inclusion of region in the second model and yet still proved unsatisfactory. Fig. 5 shows the summary result of the first multiple regression model.



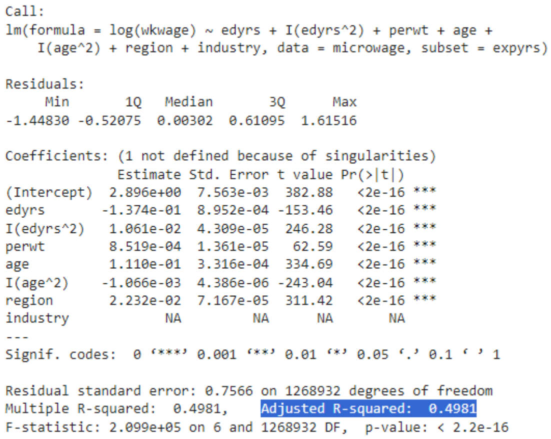
```
Call:
lm(formula = wkwage ~ edyrs + expyrs + age + perwt, data = microwage)

Residuals:
     Min       1Q   Median       3Q      Max 
-1240.50  -282.72   -61.44   240.21  1760.17 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept) -1.599e+03  8.167e+00 -195.737  < 2e-16 ***
edyrs       -4.818e+01  1.047e+00  -45.997  < 2e-16 ***
expyrs      -1.318e+02  1.113e+00 -118.480  < 2e-16 ***
age          1.407e+02  1.109e+00  126.874  < 2e-16 ***
perwt        5.263e-02  1.190e-02    4.424 9.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 418.4 on 1083513 degrees of freedom
Multiple R-squared:  0.2194,	Adjusted R-squared:  0.2194 
F-statistic: 7.615e+04 on 4 and 1083513 DF,  p-value: < 2.2e-16
```

Fig.5. First Multiple Linear Regression Model with the dependent and independent variables.

The third model included ‘*female*’ variable, which generated a significant rise in the adjusted R-squared to 0.2706 at 27.06%, this suggest a larger percentage of wages variability explained by the model. Subsequently, the fourth model improved it adjusted R-square to 0.2738 when ‘*female* and *region*’ were considered. Furthermore by adding industry variable to the fifth model resulted to a significant increase in explanatory power, with an adjusted R-squared value of 0.3636, indicating a considerable impact of industry on wage variation. This trend persisted when the sixth model polynomial terms for age and education were included. These factors captured non-linear effect and yielded an adjusted R-square of 0.5058, which explained more than half of the wage’s variation.

Even though the sixth model had a very high proportion of variance of about 50.58% adjusted R-square, the linear regression analysis's diagnostic plots provided several negative signs of the model's weakness. The Residuals vs. Fitted and Scale-Location plots demonstrated a non-linearity in the residuals and perhaps **Heteroscedasticity**. The residuals could not have a normal distribution, according to the Q-Q plot, especially around the tail. Furthermore, the Residuals vs. Leverage plot indicates the existence of significant outliers, notably observation of 31320, which may have an outsized impact on the model's predictions. These diagnostics highlight the necessity to improve the model and this made model6 not a best fit for the investigation.

The above analysis of the sixth model made an opening for the seventh model which applied **logarithmic transformation** to the dependent variable by using the ‘*log*’ function to model6. This led to an improved **Homoscedasticity** in the model. The residuals variance became the same at all levels of the independent variables. The **Cook's Distance** plot validated the robustness of the model result and confirmed the effectiveness of the log transformation on weekly wage, with no data points having a disproportionate effect on the seventh model. Fig.6. shows the model increment last seen in Fig.5. while Fig.7. shows the regression diagnostic plot of model7.



```
Call:
lm(formula = log(wkwage) ~ edyrs + I(edyrs^2) + perwt + age + 
    I(age^2) + region + industry, data = microwage, subset = expyrs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.44830 -0.52075  0.00302  0.61095  1.61516 

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.896e+00  7.563e-03  382.88   <2e-16 ***
edyrs       -1.374e-01  8.952e-04 -153.46   <2e-16 ***
I(edyrs^2)   1.061e-02  4.309e-05  246.28   <2e-16 ***
perwt        8.519e-04  1.361e-05   62.59   <2e-16 ***
age          1.110e-01  3.316e-04  334.69   <2e-16 ***
I(age^2)    -1.066e-03  4.386e-06 -243.04   <2e-16 ***
region       2.232e-02  7.167e-05  311.42   <2e-16 ***
industry            NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7566 on 1268932 degrees of freedom
Multiple R-squared:  0.4981,	Adjusted R-squared:  0.4981 
F-statistic: 2.099e+05 on 6 and 1268932 DF,  p-value: < 2.2e-16
```

Fig.6. Model7:Transformation with Multiple Regression

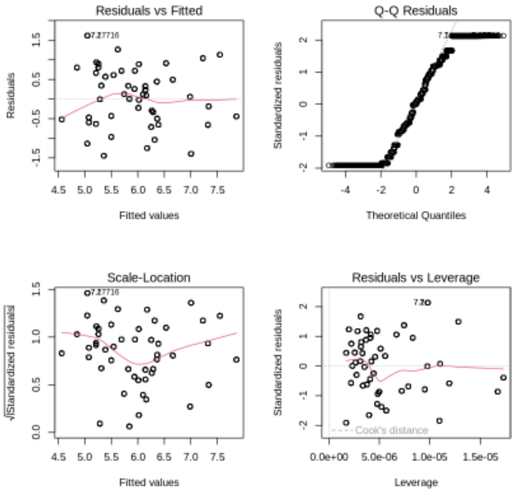


Fig.7. Regression Diagnostic Plot for Model7

V. MODEL EVALUATION AND INTERPRETATION - (*PHASE FOUR*)

Evaluation: Corresponding to the objective outlined at the beginning of this report "*Investigating Multiple Linear Regression Models, Best Suited for Weekly Earnings in the US*," the final version of the model has undergone an in-depth investigation. To achieve the goal, rigorous assessments were conducted for an accomplished model to be made possible. The model takes into consideration several predictive factors mentioned above. with the use of transformations, outlier handling, and other methodology to find the best possible outcome for the objective.

Coupled with the large amount of data (more than 1.3 million observations), the residual standard error of 0.7566 made reliable coefficient estimations feasible. The logarithm of weekly wages has a variability that can be described by the predictors of the model to the extent that the Multiple R-squared value of 0.4981 is over 50%. Based on the statistical significance of the model (an F-statistic of over 200,000), this is a significant improvement over simpler models.

Interpretation: The interpretation of the perimeters of the final model uncovers the complex dynamics in the U.S. The education coefficients show a non-linear relationship between education years and weekly wages; earnings grow at a declining rate as education years increase. The positive correlation between age and its squared term points to an influence on earnings that increases initially before declining; this is consistent with the normal career trajectory, which shows a slowdown in earnings growth as one gets closer to retirement. The wages vary between regions, as indicated by the positive coefficient for the "region" variable.

In summary, the final model offers a robust and statistically significant representation of the variables affecting weekly wages in the US, together with detailed rationalizations of the effects of age, experience, education, and regional variations. It continues to be a vital analytical model for evaluating and forecasting labor market weekly wages changes.

VI. CONCLUSION

The investigation discussed in this report successfully advanced through the phases of data collecting, exploratory data analysis, building a model, and comprehensive evaluation to create a multiple linear regression model that is suited analyzing weekly wages in the labor market in the United States. The study methodically revealed the complex effects of professional and demographic factors on earnings by including a large dataset that included over 1.3 million individuals. The accuracy of the model was greatly increased during the Model Building phase by applying a logarithmic transformation to the dependent variable, resulting in an adjusted R-squared of 0.4981. This shows that the model is resilient because it explains almost half of the variance in weekly earnings.

Finally, this report's incrementally approach has produced a statistically significant and insightful model that explains close to 50% of the earnings variation. Notwithstanding the advantages of the model, the unexplained variation points to the possibility that other factors could provide more insights to contribute patterns and indicate directions for further investigation.

References

[1] L. M. a. J. Bivens, "Identifying the policy levers generating wage suppression and wage inequality," Economic Policy Institute., Washington, DC, 2021.

[2] H. Taherdoost, "Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection," International Journal of Academic Research in, Switzerland, 2022.

[3] J. D. S. D. C. M. a. Y. C. Matthieu Komorowski, "Exploratory Data Analysis," ResearchGate, London, 2016.

[4] D. H. M. a. A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine," Interdisciplinary Publishing Academia, 2020.

[5] F. S. M. R. a. S. U. K. David M. Raisuddin, "A Review of Evaluation Metrics for Linear Regression Models," Journal of Applied Research and Technology, 2018.