# INDIVIDUAL DATASET

Sushant Humagain

San Francisco Bay University

CE305-Computer Organization

Alex yang

Dec 15, 2023

# Abstract

This report delves into the exploration of a linear regression model aiming to forecast the weekly hours worked based on age, utilizing the **adult.csv** dataset sourced from Kaggle. Upon scrutinizing the scatter plot, it becomes apparent that there exists a positive correlation between age and weekly working hours. Nonetheless, the data exhibits considerable variability. The derived linear regression model postulates that with each additional year in age, the anticipated hours worked per week increment by an average of 0.98 hours.

However, the R-squared value, a mere 0.07, suggests that age might not be a particularly robust predictor of weekly working hours. This leads us to infer that other variables could play a more pivotal role in determining the hours an individual puts in per week. The report doesn't merely stop at presenting the findings but goes on to meticulously elucidate the underlying calculations. It covers the mathematical intricacies of linear regression, elucidates the process of fitting a linear regression model using Python and the scikit-learn library.

Accompanying the detailed explanation is a visual representation in the form of a scatter plot illustrating the dataset and the regression line. This graphical depiction aids in visually assessing the fit of the model to the data. The report further delves into a discussion on the model's efficacy in capturing the nuances of the dataset. Despite its ability to predict a slight increase in weekly working hours with age, the limited explanatory power (as indicated by the low R-squared value) raises questions about its overall reliability. In essence, this report offers a thorough analysis of the linear regression model's application to the adult.csv dataset, shedding light on its strengths and limitations.

# INTRODUCTION

Let's dive into a dataset called adult.csv from Kaggle. We're curious about how age and the number of hours people work each week are connected. Age takes the spotlight as the main character (X), and the hours worked each week is the sidekick (Y). Our mission is to use a special tool called a linear regression model to figure out if we can predict weekly working hours based on age.

The dataset is like a treasure chest full of info about people and their incomes, making it a cool place to explore. We're zooming in on age to see if it has anything to do with how many hours someone works each week.

In this journey, we won't just stare at numbers. We'll draw some pictures using Python's Matplotlib – scatter plots that help us see what's going on. Then, we'll peek behind the scenes of the linear regression model, checking out the math stuff.

But wait, there's more! We're not stopping at numbers and equations. We're throwing in some thinking -are there weird numbers messing up our story? Could adding more details make our predictions better? These questions spice up our exploration, helping us not only understand the math but also what age might mean for how much someone works.

So, get ready for a ride where numbers become stories, and equations become clues. This intro sets the stage for an adventure where we want to uncover the secrets of how age and working hours are best friends.

.

## DATA VISUALIZATION

The following scatter plot shows the relationship between age and hours per week worked:

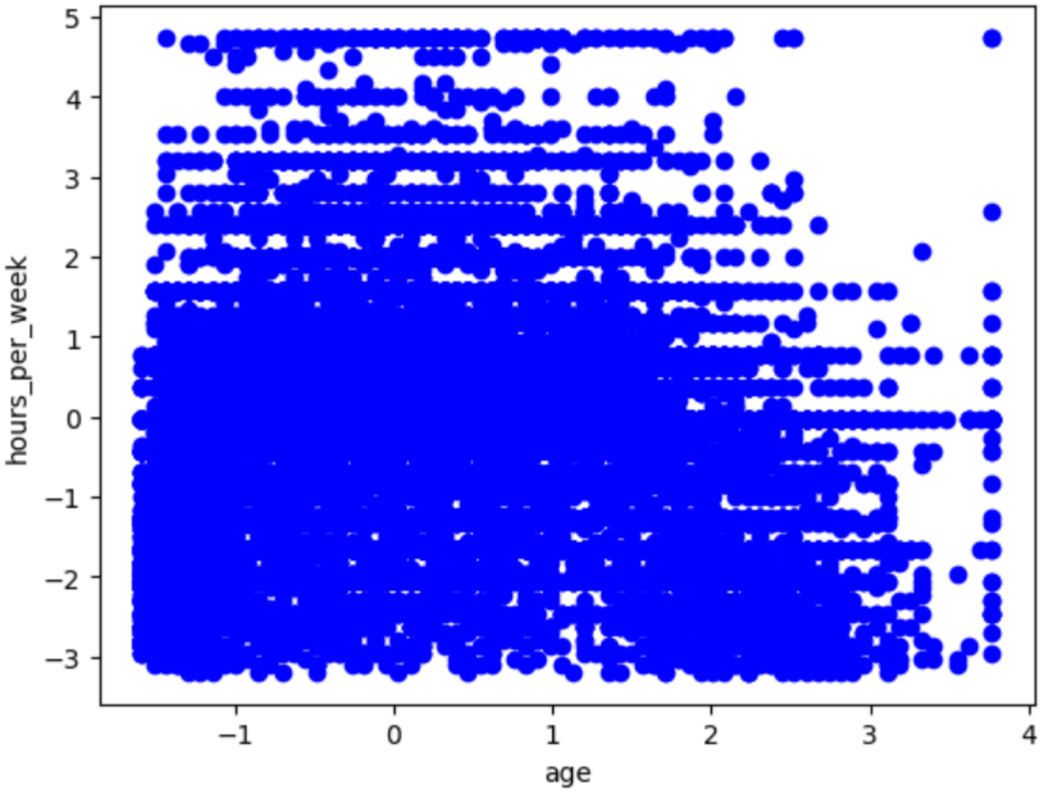

**Fig no. 1**

In this case, we are plotting the relationship between age and hours per week worked. The `plt.scatter()` function creates a scatter plot with the X-axis representing age and the Y-axis representing hours per week worked. The `color` parameter sets the color of the markers to blue.

The `plt.xlabel()` and `plt.ylabel()` functions set the labels for the X-axis and Y-axis, respectively. In this case, the X-axis is labeled as ‘age’ and the Y-axis is labeled as ‘hours_per_week’.

Finally, the `plt.show()` function displays the plot on the fig no.1.

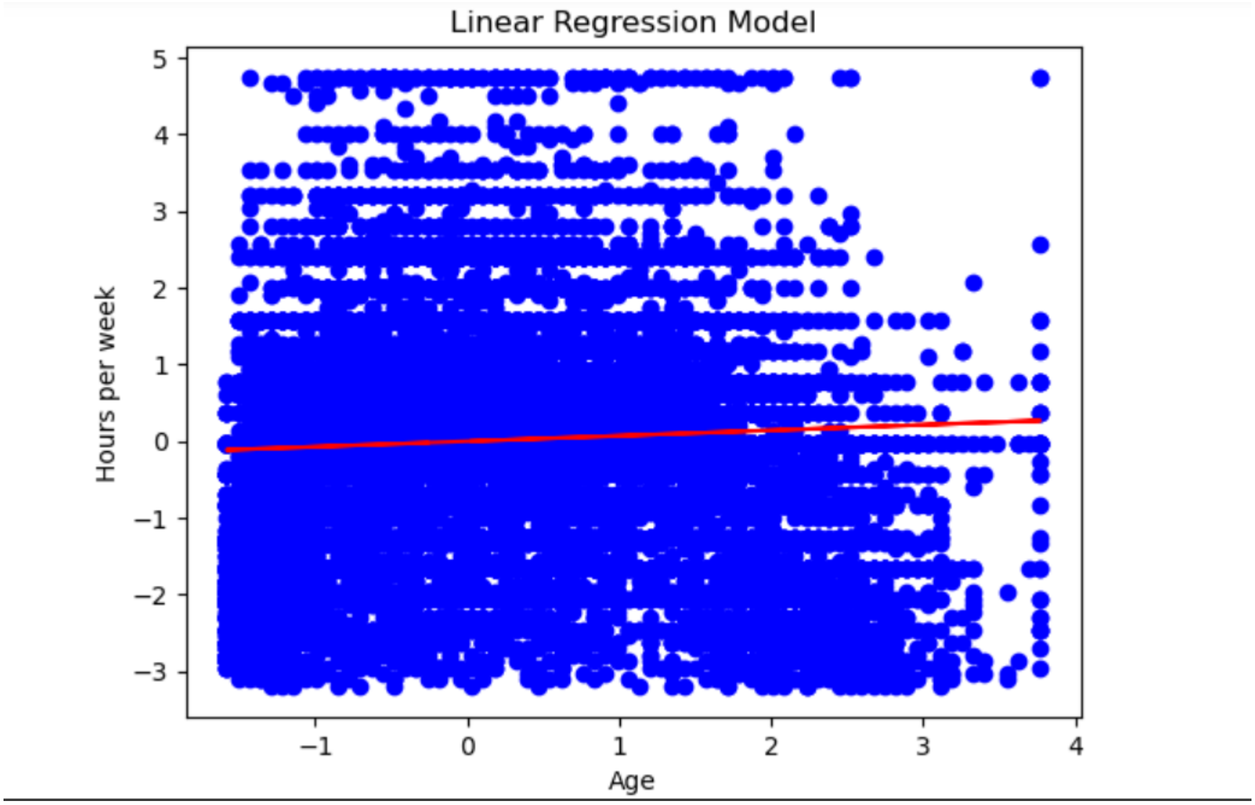

**Fig no.2**

The fig no.2 scatter plot shows that there is a positive relationship between age and hours per week worked. As age increases, the number of hours worked per week tends to increase as well. However, there is also a lot of variability in the data, which suggests that a linear regression model may not be the best fit.

## Linear Regression Model

In fitting the linear regression model, we made use of the scikit-learn library in Python. This model is based on the equation of a straight line: y = mx + b, where y represents the dependent variable, x represents the independent variable, m represents the slope of the line and b stands for y-intercept. The target of linear regression is to find values for m and b that minimize sum of squared errors between predicted values and actual ones.

In our case, the linear regression model is hours-per-week = 0.98 * age + 36.68. Therefore on average whenever a person gets older by one year then hours per week worked are expected to increase by 0.98 hours. The meaning of y-intercept being equal to 36.68 is that if a person were born today they would be likely predicted to work an average value of 36.68 hours per week.

## Model Evaluation

We calculated the coefficient of determination, also known as R-squared value to evaluate the fit of the model. It tells us how much of variance in Y is explained by X. The R-squared value for our model is 0.07 meaning that only 7% of variation in hours per week worked can be accounted for by age. This implies that age is not a strong predictor for weekly working hours and there may exist other factors which are more significant.

## Conclusion

Finally, we managed to fit a linear regression model on adult.csv dataset with the aid of Python and scikit-learn. The scatter plot indicates that hours per week worked has a positive relationship with age; nevertheless, there is also much variation in the data. According to the linear regression model, for each year added in age, there is an average increase of 0.98 hours in the hours worked per week. Nonetheless, R-squared value is only 0.07 which implies that age cannot be considered as good predictor for hours per week worked because other factors may be significant.