Relationship Between Unemployment Rates Between Ages 25-34 From 2008 to 2025*

Kei Matsumoto

September 22, 2025

Employment through the years always changes with trends of industries as well as international events/disasters. However, with the increase in possible startup opportunities through new discoveries like AI, has unemployment rates changed over the years? A dataset provided by the California Open Data pool highlights the overall unemployment rates over a monthly period from 2008 to 2025 within specific age groups. This paper will cover the possibility to fit a simple linear regression model to see if there is a relation between unemployment rate and time for people between ages 25 and 34. Overall in this study, it is hard to find if the year correlates to the unemployment rates through linear regression as a different model is most likely better to analyze this data.

1 Introduction

Just by ear, there are people who say that there are better job opportunities with the rise in new advancements in AI and others who say that it is much more of a challenge than with the increase in competitiveness for those positions. Most people will start their careers after graduating from either undergraduate or graduate school around the age of 25-34 which will be the age range that will be analyzed.

This paper will only cover simple linear regression and linear modeling with a lack of further analysis. In this paper, we will conclude that linear regression may not be the optimal model to use to see the relationship between time and unemployment rates as the data highlights a possibility of a cyclic pattern due to sudden spikes of unemployment in specific years. One large spike found was during 2019-2021 which can be due to the breakout of Covid-19 with many jobs having in-person interactions. The linear regression will be analyzed through the

[*] Project repository available at: https://github.com/keimatsumoto1/261A

programming language R using the basic functions within R and major packages such as “lm” and “ggplot” to visualize the data.

The paper will start with the description of the data used, into the methods conducted, and lastly the explanation of the results we have possibly found through the use of the statistical analysis within our knowledge of linear regression.

2 Data

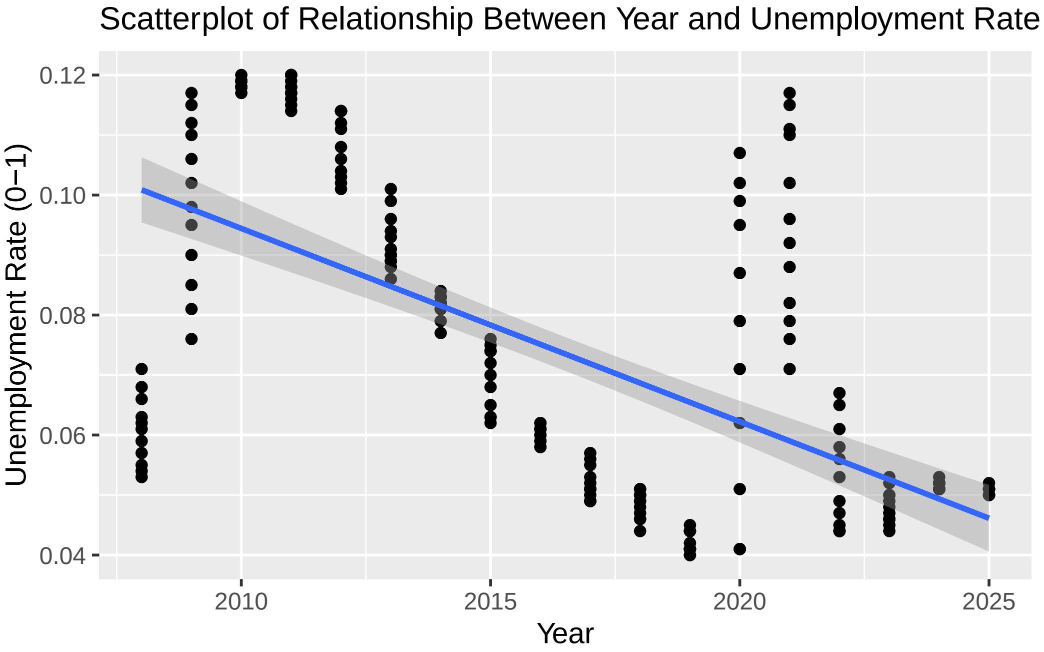

Figure 1: Scatter plot comparing the year of survey study as the x axis, and the rate of unemployment between 0 and 1 as the y axis with the fitted line of regression.

This data was retrieved from the California Open Data Pool titled “Unemployment Rate by Age Groups” where the Current Population Survey (CPS) conducted a non-seasonally adjusted survey between 7 different age groups for each month of 2008 to 2025. The data is retrieved from the Employment Development Department who releases all of the employment statistics such as civilian labor forces, unemployment rates, and industry employment by geography.(CPS, 2025) The EDD collaborates with the U.S. Bureau of Labor Statistics and the exact calculation of the unemployment rate is defined by (unemployed citizens / civilian labor force). Some months are missing in this dataset where January of 2008 and September through December of 2025 are not included. There does not appear to be any missing data within the dataset besides the last months of 2025. Therefore, the data for 2025 may not be as accurate as a quarter of the year is missing compared to other years due to the date of analysis is still being updated through each month.

The data uses measurements of percentages of unemployed people in the designated age group from the survey with other columns being demographic area, date, year, month, and each age group's rate of unemployment.

3 Methods

In this study we will be conducting a simple linear regression to analyze the relationship between the rate of unemployment and time specifically in years ranging from 2008 to 2015.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 will represent the intercept of the model which in this analysis will be the unemployment rate during the first time the study was conducted which is February of 2008.

β_1 will represent the slope of the model which in this analysis will be the decrease in unemployment rate through each increase of 1 year.

This model analysis will be conducted through the programming language R (insert citation) with several packages such as ggplot and readr.

4 Results

$$\beta_0 = 6.56$$

$$\beta_1 = -0.0032195$$

$$R^2 = 0.3847$$

Within our simple linear regression we found that our $\beta_0$is 6.56 and our β_1 is -0.0032195. Our β_0 in this scenario does not apply to our analysis due to the unemployment rate not being able to exceed 1. We can possibly say that our unemployment rate is close to 1 before 2008 which is due to the lack of studies made before 2008 in our experiment. Our β_1 is slightly negative indicating a very minor decrease over the years between 2008 and 2025. To be exact, for every additional year that passes, the rate of unemployment will on average decrease by 0.0032195 or 0.32195%.

In addition, our linear regression analysis includes and R-squared value of 0.3847 indicating that the fit of a linear model may not be ideal for this data.

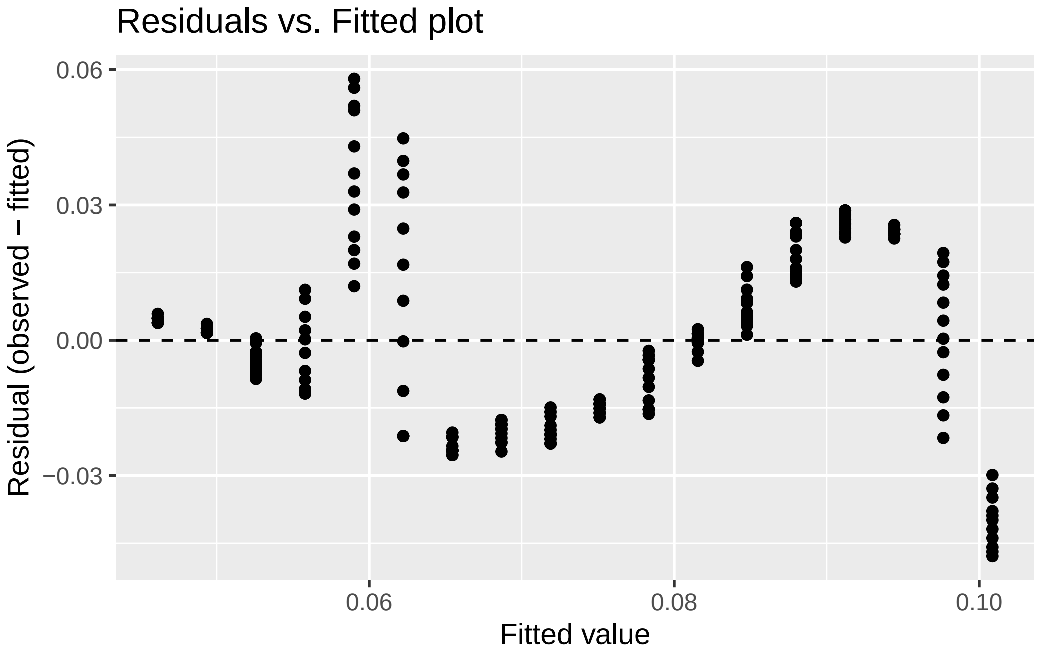

Figure 2: Residual plot of the x axis being the fitted values and the y axis being the residuals for a simple linear regression with year being the respone and unemployment rate being the predictor.

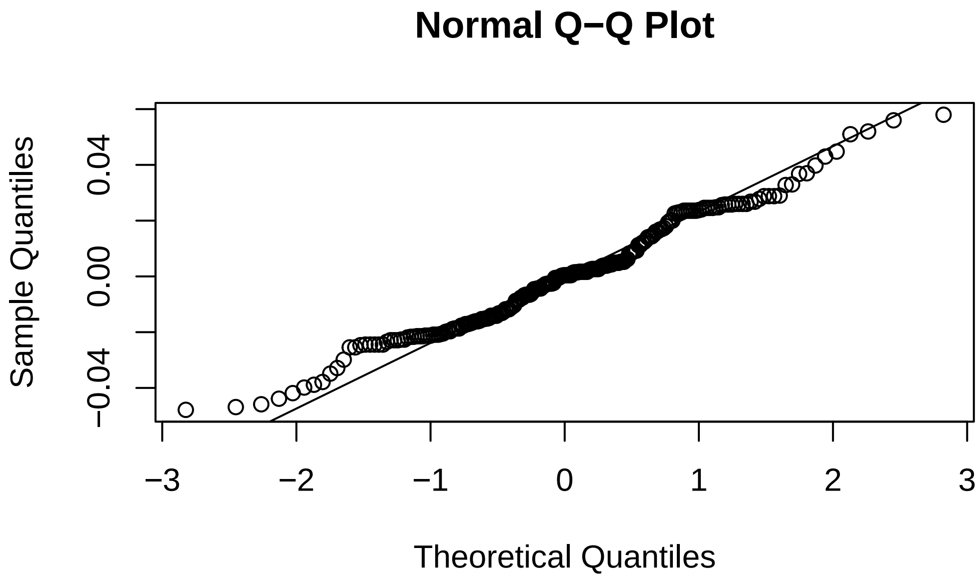

Figure 3: QQ plot between the sample (x) and theoretical (y) quantiles

Our residuals vs fitted plots highlight that our model isn't the best fit for a simple linear regression where the variation of points are not in a form that shows no trends. With our data's sample size being 211 observations, we can possibly infer that our variances are equal.

Our qq plot on the other hand suggests that our model has a possibility of having a good fit to our data. However, our tail and head of the QQ plot highlights the variance from the fit which can add on to our R-squared value being low.

5 Discussion

In our analysis, we tested to see if we can fit a simple linear regression to find if unemployment rates have a relationship over time. We discovered that 2010 and 2020 have large peeks and sharp rises of unemployment rates which can be due to outside influences. One extreme hint for 2020 could be COVID 19 affecting the unemployment rate drastically. A study from US Bureau of Statistics indicates that in 2019, roughly 3/4 of the workers in the United States had to work with in-person human interaction. Combining this statistic with the lock down protocol from COVID-19 can possibly connect to the drastic increase in unemployment during 2020. In addition, 2008 was the time of the Great Recession which can be identified by the large spike in the beginning of our data. By noticing that the decrease then continues after the large spike, we can possibly see that if the unemployment rate would have slowly decreased over time without any interference. Due to our R-squared value being considerably low and multiple violations against our simple linear regression conditions, a better model can be possibly used to fit this data. Possible indications of a cyclical or time-series related model can be found from observing the scatter plot.

6 References

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Unemployment rate by age groups - Unemployment rate by age groups - California Open data. (n.d.).
https://data.ca.gov/dataset/unemployment-rate-by-age-groups/resource/be49fea7-af13-4781-8113-4bb66ba508e9

U.S. Bureau of Labor Statistics. (n.d.). Three-fourths of workers had to interact with the public in 2019; 4.3 percent worked around crowds. U.S. Bureau of Labor Statistics. https://www.bls.gov/opub/ted/2020/three-fourths-of-workers-had-to-interact-with-the-public-in-2019-4-3-percent-worked-around-c rowds

7 Things that I will need think about adding in the final draft

1. Citation with bib file
2. Citations within text
3. Possibly using more than just ages 25-34 and add a better combination graph with deeper analysis (would this be beneficial or too much?)
4. sectioning