Relationship Between Unemployment Rates and Time For People Ages 25-34 From 2008 to 2025*

Kei Matsumoto

September 22, 2025

Employment through the years always changes with new industry trends as well as international events/disasters. Job oppurtunities may have increased with the influence of new advancements such as A.I. but is this guaranteed to be an increase? Has unemployment rates changed over time? A dataset provided by California Open Data 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 difficult to conclude if time correlates to the unemployment rates through linear regression, as indications to fit a different model are evident.

1 Introduction

Just by ear, there are people who say that there are better job opportunities with the rise in new advancements in A.I. On the other hand, some others will say that it is much more of a challenge with the increase in competitiveness for those positions. A recent study from J.P. Morgan Global Research highlights how A.I. has both increased and decreased employment rates depending on the work field. Areas such as the white-collar sectors have received benefit through utilization of A.I. while computer engineering college graduates are seeing difficulties with the job market (J.P. Morgan Global Research, 2025). In addition, a study conducted by the National Center for Education Statistics (NCES) discovered that for 2019, many employment rates increase with the higher degree of education attained (NCES, 2019). To see the age groups of those who have finished at least a Bachelor's Degree and possibly even a Master's Degree, we will research the age group of 25-34 year old people.

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

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 programming language R using the basic functions within R and major packages such as "readr" and "ggplot2" to initialize and visualize the data.

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

2 Data

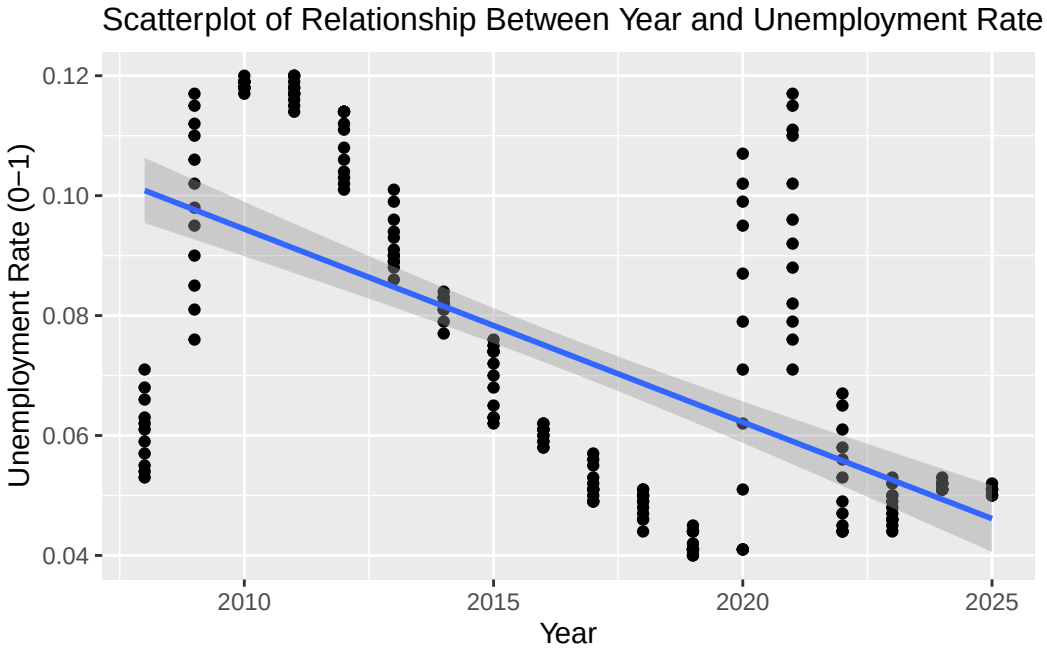

Figure 1: Scatter plot comparing the year of survey study as the x axis, and the rate of unemployment between 0 and 1 as the y axis with the fitted line of regression.

This data was retrieved from California Open Data 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.

In addition, we are only analyzing our data of people within the age of 25 to 34 years old. Our conclusions and analysis in this paper are unable be used to predict the unemployment rates of other age groups as they are completely different sources of data. Additional limitations can be the reduced amount of data points for each month/year in comparison to the entire dataset causing possible over fitting.

By analyzing the scatterplot above, we can notice that for each year there are multiple points for every month. This can affect our variances and violate homoscedacity with our unequal variance even if our number of data points is high.

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 2025.

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

β_0 will represent the intercept of the model which in this analysis will be the expected 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 expected change in unemployment rate through each increase of 1 year.

Y_i will represent the unemployment rate of the i-th Year from 2008 to 2025.

X_i will represent the known constant which is the i-th year from 2008 to 2025.

ϵ_i will represent the random error with a mean of zero and finite variance of σ^2.

This model analysis will be conducted through the programming language R with several packages such as ggplot and readr (R Core Team, 2023).

4 Assumptions

The assumptions we need to validate for our analysis are the following:

1. Validity where the variables are appropriate for the question we are studying.
2. Representativeness where the sample data represents the population of interest.
3. Linearity where following function must be true: $E(Y_i|X_i) = \beta_0 + \beta_1 X$
4. Independence of errors where all errors of our model are assumed to be independent.
5. Equal variance of errors where our errors' variances must be equal throughout the model.
6. Normality of errors where our errors must be normally distributed.

5 Results

By first analyzing the scatterplot made earlier, we can notice that the trend is not in a linear formation with a trend with 2 different peaks being evident. This shape will violate our assumption of linearity for simple linear regression where our relationship between Y_i and X_i may not be linear and suggests a different possible model.

We also see that our data is dependent to time which reflects how our observations are not independent. This will violate one of the key assumptions (independent observations) to linear regression and suggest that our standard errors are biased.

Within our simple linear regression, we found that our $\hat{\beta}_0$ is 6.56 and our $\hat{\beta}_1$ is -0.0032195. Our $\hat{\beta}_0$ 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 $\hat{\beta}_1$ 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%. This indicates that may exist a very small but steady decline in unemployment rates for those between 25 and 34.

In addition, we found that our R-squared value is 0.3847 indicating that our model only explains around 38.47% of the total variance of our data. This value is low emphasizing how our model may not be a significant fit for the data.

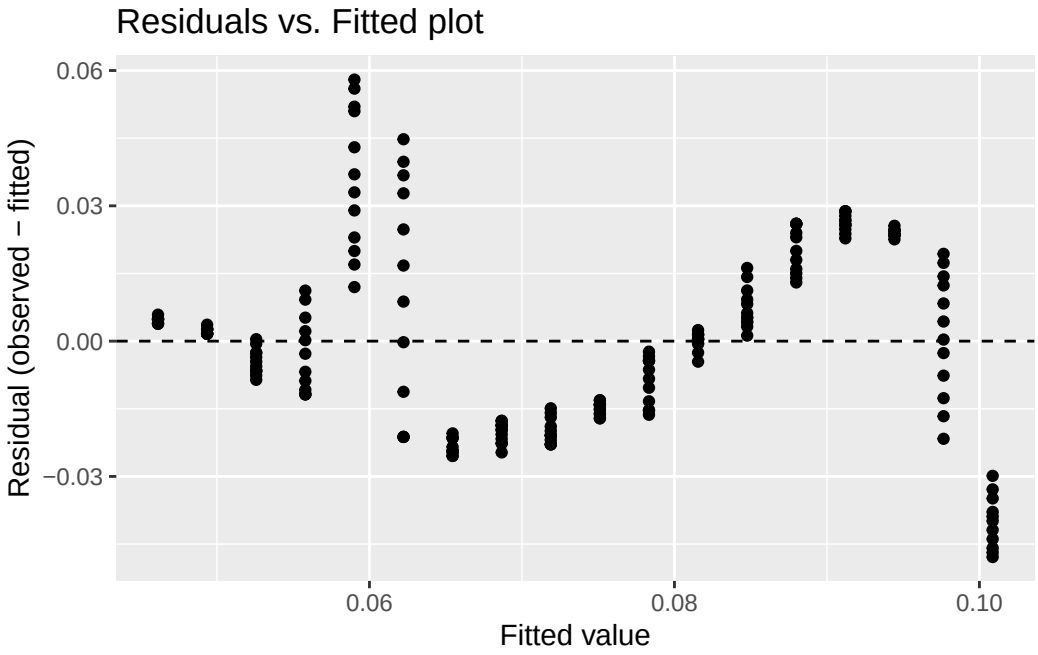

Figure 2: Residual plot of the x axis being the fitted values and the y axis being the residuals for a simple linear regression with year being the response and unemployment rate being the predictor.

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. This will violate multiple assumptions such as error independence, error equal variance, and linearity. Even if our data's sample size of 211 observations is quite large, these assumptions are violated. This suggests possible transformations or usage of a different model being highly recommended for this data.

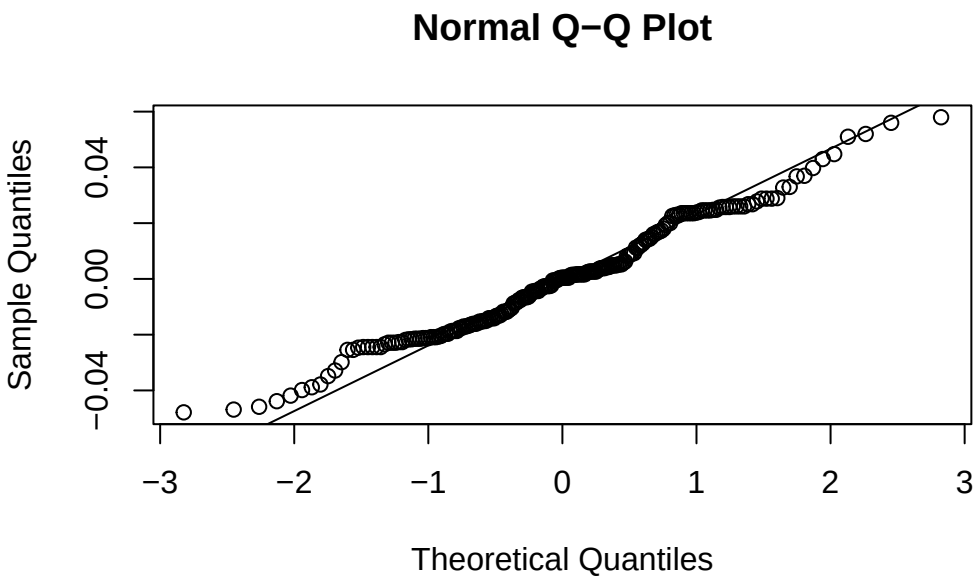

Figure 3: QQ plot between the sample (x) and theoretical (y) quantiles with the line representing the theoretical normal distribution

When observing our QQ plot at a glance, we can see that it might suggest 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 suggesting that our errors are not normal.

6 Discussion

In our analysis, we tested to see if we can fit a simple linear regression to find if unemployment rates have a relationship with time. We discovered that 2010 and 2020 have large peaks and sharp rises of unemployment rates which can be due to outside influences.

2008 was one of our spikes in unemployment rates. The Great Recession which was around the time can be a possible issue identified by the large spike in the beginning of our data. Another spike was identified around 2020 which highlights possibilities of influence by Covid-19 affecting the unemployment rate drastically. A study from US Bureau of Statistics indicates that in 2019 found that roughly 3/4 of the workers in the United States had to work with in-person human interaction (U.S. Bureau of Labor Statistics, 2023). Combining this statistic with the lock down protocol from Covid-19 can possibly connect to the drastic increase in unemployment during 2020. By noticing that the decrease then continues after the two large

spikes, we can possibly see that the unemployment rate possibly would have slowly decreased over time without any interference.

During our analysis, we found that there has been multiple violations against our assumptions with linear regression. These include independent observations, linearity, error normality, error equal variance, and error independence. These violations of assumptions indicate that simple linear regression is not able to be used to analyze the relation between time and unemployment rates.

Our scatterplots and our fitted vs observed residual plots indicate that there are non linear trends. Due to the residual plot not having randomness, our assumptions of error equal variance, normality, and independence are violated. In addition, our R-squared value is low and our QQ-plot also suggests that our errors are not considered normal.

Overall, we can conclude that a different or transformation can be possibly used to fit this data. We discovered that there may exist a small negative correlation between time and unemployment rates but we are unable to justify this through simple linear regression. Our data is found to be dependent to time with indications of a cyclical pattern from observing the scatter plot.

7 References

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

Morgan, J. P. (2025). AI's Impact on Job Growth | J.P. Morgan Global Research. Retrieved from Jpmorgan.com website: https://www.jpmorgan.com/insights/global-research/artificial-intelligence/ai-impact-job-growth

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

Relationship Between Educational Attainment and Labor Underutilization. (n.d.). Retrieved from nces.ed.gov website: https://nces.ed.gov/pubs2019/2019039/index.asp

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-crowds

Wickham H, Hester J, Bryan J (2023). _readr: Read Rectangular Text Data_. R package version 2.1.4, https://CRAN.R-project.org/package=readr.