The Impact of Economic Development on the Labor Market

By

Sejal Mogalgiddi,

Tony Tran,

Ryan Jones,

and Gabriel Alvarado

Abstract

In this study, we analyze the unemployment rate (%) of 19 countries from 2010 to 2023 based on macroeconomic factors such as GDP, inflation, and economic growth. Our goal is to understand better the extent to which these economic indicators influence unemployment. The dataset from Kaggle provided key economic variables that allowed us to explore relationships between these factors. Initial exploratory data analysis revealed violations of regression assumptions largely due to outliers from Turkey and Saudi Arabia, which were removed. Despite some improvement, further transformations were necessary. A Box-Cox test suggested a log transformation, which we compared against square root and inverse square root transformations to find the best fit. Ultimately, the log transformation was selected for its superior improvement in model performance. Our final model included GDP, inflation, economic growth, and an interaction term between GDP and inflation. The model's R-squared was low, indicating that macroeconomic indicators alone cannot fully explain unemployment trends. These findings highlight the complexity of labor market dynamics and suggest the need to incorporate additional variables. This study provides a foundation for further exploration of the factors relating to unemployment and their implications for policy-making.

Motivation & Data Description

In this project, we aim to analyze a country's unemployment rate (%) based on economic factors such as gross domestic product in billion USD (GDP), inflation, and economic growth. The unemployment rate measures the percentage of a country's population that is actively seeking employment but not being employed. High unemployment rates can cause a decrease in consumer spending and tax revenues while leading to an increase in government spending on

welfare programs (Ganong and Noel, 2016). Accurately understanding what leads to unemployment can help governments, policymakers, and businesses make better decisions regarding policy creation or workforce development.

By analyzing GDP, inflation, and economic growth, we aim to better understand how these factors not only affect a country locally but also other countries on a global scale. These variables are often used to determine the health, direction, and influence of a country's macroeconomic environment. Our research looks to contribute to a greater understanding of macroeconomic relationships by analyzing how these variables interact with one another. By examining how the relationships between GDP, inflation, and unemployment vary depending on a country's economic development, governments can adjust economic policies based on their country's current stage of development.

Our dataset is from Machine Learning Engineer Adil Shamim on Kaggle. This dataset provides key economic indicators, such as GDP, Inflation Rate, Unemployment Rate, and Economic Growth, for 19 different countries around the world from 2010 to 2023. The variables that we will explore in our model are:

- **GDP (in billion USD):** The total market value of all goods and services produced by the country in a given year, measured in billions of USD. GDP reflects the size and health of an economy.
- **Inflation Rate (%):** The percentage change in the general price level of goods and services in the economy over the year. Inflation impacts purchasing power and economic stability.

- **Economic Growth (%)**: The annual percentage increase (or decrease) in GDP, representing whether the economy is expanding or contracting. Positive growth indicates expansion, while negative growth indicates contraction.

Additionally, economic growth will be the variable we use to determine if a country is developing or already developed, given that "Typically, higher levels of GDP per capita are associated with higher standards of living" (Ravikumar, Chinagorom-Abiakalam, Smaldone, 2016).

Overall, our model will be used to analyze unemployment rates based on these economic indicators, improving our understanding of how macroeconomic factors influence employment.

Exploratory Data Analysis

The relationship between GDP, inflation, economic growth, and unemployment will be examined over time in countries with different levels of economic development. GDP helps us establish if there is a trend between economic health and growth. It's also often associated with other variables; for example, a higher GDP is often associated with lower unemployment and stable inflation. It is also important that our dataset includes the GDP of both developed and developing countries to provide a more robust analysis. As for inflation, Depersio (2024) notes that the Phillips curve illustrates the inverse relationship between inflation and unemployment, showing these variables are often linked, and examining them together can shed light on how they influence each other or the relationships between other variables in the economic system. The final variable to be assessed is economic growth, which is defined as the percentage change in the value of all goods and services produced within a nation over a specific period (Chen,

2023). This metric helps capture how economic activity changes over time and can be affected by numerous factors.

In addition to our primary variables, we're also interested in exploring the presence of patterns or trends within the other variables included in our analysis. To achieve this, we will incorporate the following interaction terms into our full model:

- GDP and Economic Growth.
- GDP and Inflation
- Inflation and Economic Growth

These interaction terms will allow us to examine the combined effects of different variables and how they may influence unemployment rates.

Model Diagnostics & Selection

When familiarizing the dataset, we checked the five fundamental assumptions: existence, independence, linearity, homoscedasticity, and normal distribution. The data does exist, and we can analyze it, so existence is not violated.

	Unemployment.Rate....	GDP..in.billion.USD.	Inflation.Rate....	Economic.Growth....
Unemployment.Rate....	1.0000000	-0.1141088	0.2430853	0.1206309
GDP..in.billion.USD.	-0.1141088	1.0000000	-0.1751804	-0.1334499
Inflation.Rate....	0.2430853	-0.1751804	1.0000000	0.3724251
Economic.Growth....	0.1206309	-0.1334499	0.3724251	1.0000000

>

Figure 1. The correlation matrix used to test independence

Looking at the correlation matrix, we see no signs of collinearity as all the values are below 0.8. As a result, this shows that independence is not violated as well.

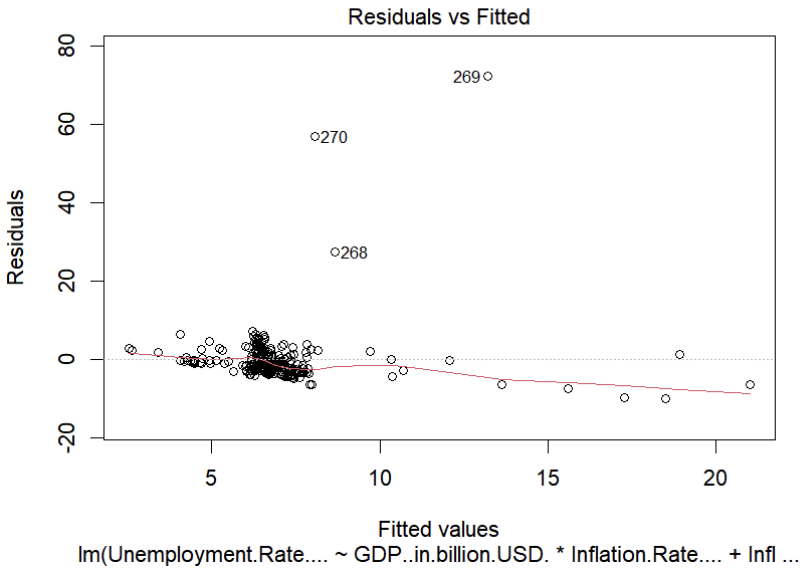


Figure 2. The Residual vs Fitted plot is used to test linearity and homoscedasticity

On the other hand, we can see in Figure 2 that linearity is slightly violated because the red trend line curves down. We also see in Figure 2 that homoscedasticity is slightly violated as the data points spread out in the middle of the plot, whereas the left side and right side are more tightly bunched up. This violation appears to predominantly come from outliers 268, 269, and 280 in the dataset.

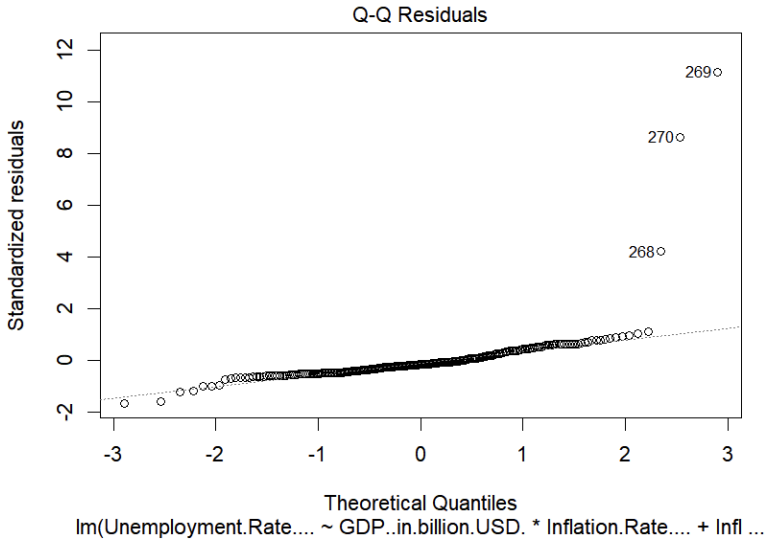


Figure 3. The Normal Q-Q Plot is used to test normal distribution

Looking at the residuals plot in Figure 3, the normal distribution is slightly violated. The same data points that were an issue for homoscedasticity are again an issue here, as they fall far away from the trend line. Curvature away from the line is an indicator that normality is violated and that we will need to fix this as we develop our model.

The violation of linearity suggests that the relationship between independent variables may not be linear and that a more complex model using transformations may be needed to improve accuracy. This is reinforced by the violation of homoscedasticity and normal distribution, which shows a need for transformations, but also indicates that we might need to remove influential outliers.

Continuing our exploratory data analysis revealed issues with linearity, constant variance, and normality, likely due to the presence of outliers in the dataset. To confirm that the issue was indeed outliers, we assessed them with leverage, Cook's distance, and jackknife residuals.

```
> print(high_leverage)
         4          5          6          7          8          9         10         11
0.05668282 0.05430639 0.11233630 0.07016577 0.05742372 0.05866820 0.07077138 0.45283631
        12         13         14         17         28         43         91        182
0.22691552 0.49904791 0.11509602 0.05506535 0.10249895 0.05367417 0.05687120 0.12935622
       219        244        245        253        254        260        261        262
0.05396083 0.12177619 0.26964663 0.10339892 0.20601771 0.07436808 0.12344463 0.15700208
       263        265        266        267        269
0.17985346 0.19423368 0.05588176 0.37733056 0.05264444
```

Figure 4. Leverage output

First, we looked at leverage in Figure 4 and compared it to the cut-off point of 0.0225564, which we got from the equation (2 (3 + 1)) / 266. Here we see observations 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 28, 43, 91, 182, 219, 244, 245, 253, 254, 260, 261, 262, 263, 265, 266, 267, and 269 are all listed as greater than the cut off point we calculated.

```
> cooks_d[cooks_d > 1]
named numeric(0)
```

Figure 5. Cook's distance output

Next, we checked whether any Cook's distance values exceeded one, as that would be a sign that the datapoint was a potential outlier. We found no points that had a cook's distance greater than one, as shown in Figure 5.

```
> head(sort(stud_res), 20)
       263        262        267        261        252        260        251         43
-1.6604356 -1.5707116 -1.2193207 -1.1850532 -0.9725301 -0.9690106 -0.9365599 -0.7151508
       277        273        275        276        259        274        249        278
-0.6931035 -0.6688930 -0.6514988 -0.6411419 -0.6408036 -0.6248603 -0.6224619 -0.6196346
       280        247         41        281
-0.6157146 -0.5900836 -0.5848040 -0.5765363
> tail(sort(stud_res), 20)
       189        190        149        217        211        216          1        199
 0.6236972  0.6236972  0.6256858  0.6396538  0.6691961  0.7097027  0.7192381  0.7882651
       215        202        214        201        212        200        213        182
 0.7975106  0.8018883  0.8298390  0.8622897  0.9007759  0.9268221  0.9660808  1.0334567
       203        268        270        269
 1.1041371  4.3575209 10.2255738 15.4326880
```

Figure 6. Cook's distance output

The final test we looked at was the jackknife residual, where we compared the absolute values of the outputs in Figure 6 to 0.6754309, which we got from the formula we ran in R.

```
print(qt(.25, 266-3-2, lower.tail = FALSE))
```

We can see observations 263, 262, 267, 261, 252, 260, 251, 43, 277, 216, 1, 199, 215, 202, 214, 201, 212, 200, 213, 182, 203, 268, 270, and 269 are all greater than our cut off point we calculated.

Through these tests, we can pinpoint rows 269, 270, 268, 260, 262, 263, 254,259,245,1, and 244 as potential outliers, so we will further assess these rows.

	Country	Year	GDP..in.billion.USD.	Inflation.Rate....	Unemployment.Rate....	Economic.Growth....
43	Japan	2020	6500	0.2	2.8	-4.8
182	Saudi Arabia	2015	646	5.4	10.5	-14.0
260	Turkey	2013	1	592.0	7.4	9.7
261	Turkey	2014	1	767.0	8.2	9.9
262	Turkey	2015	1	857.0	7.7	10.3
263	Turkey	2016	1	853.0	8.5	10.9
267	Turkey	2020	3	717.0	14.6	13.2
269	Turkey	2022	4	500.0	85.5	10.0
268	Turkey	2021	4	0.0	36.1	12.0
270	Turkey	2023	5	0.0	65.0	9.0

> |

Figure 7. Table showing potential outlier observations

This analysis shows that rows 269, 270, 268, and 182 must be further investigated since the unemployment rate is high (85.5%, 65.0%, 36.1%, and 10.5%, respectively). Because these were all from Turkey, we cross-referenced our dataset with what we could find online. We checked both the World Bank Group and Macrotrends to find the rates of unemployment in Turkey between 2010 and 2023.

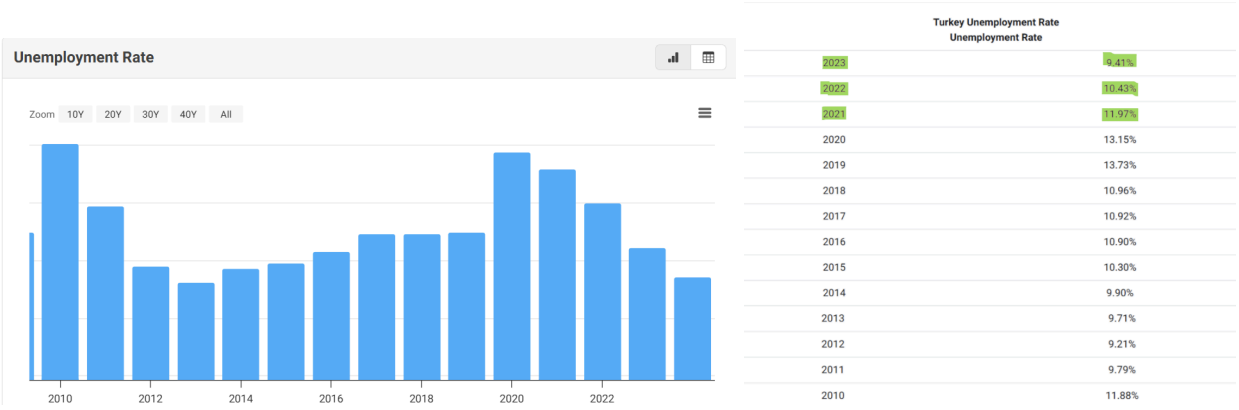
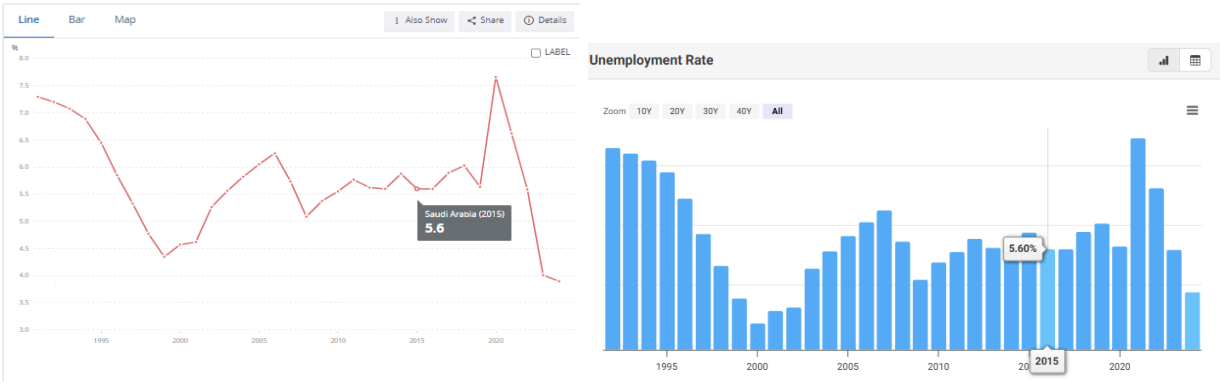


Figure 8. Turkey's unemployment rate from 2010 to 2023 is found on Macrotrends

The recorded unemployment rate we found, shown in Figure 8, does not match our dataset. With this seemingly affecting most, if not all, of Turkey's observations, we decided to completely remove all of the observations from Turkey to reduce outliers and bias.

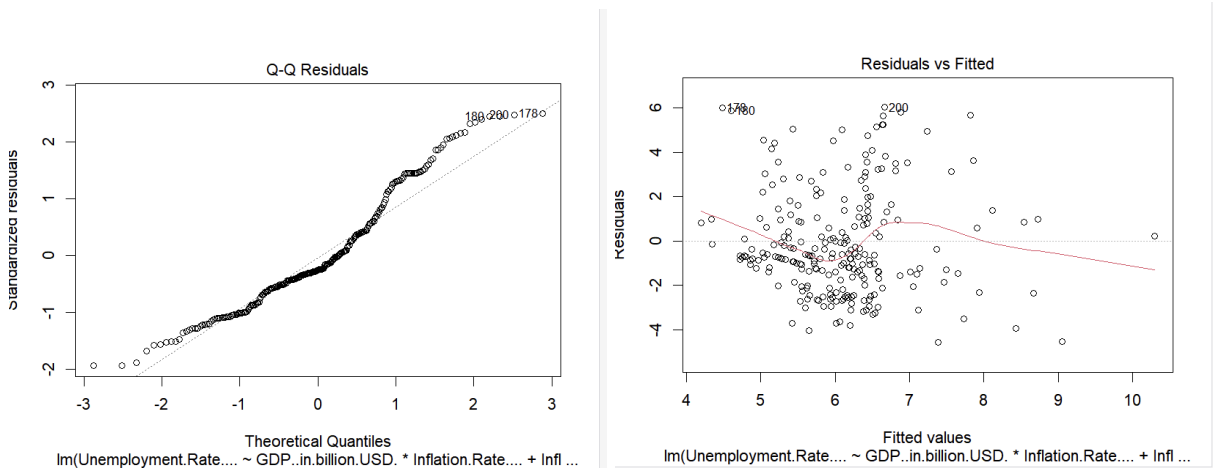


(a) (b)

Figure 9. Saudi Arabia's unemployment rate for 2015 is found on (a) the World Bank Group and (b) Macrotrends

When we looked at Saudi Arabia's unemployment rate in 2015, we found that it was 5.6%, unlike the 10.5% (Figure 9). In this case, we decided to remove the single observation as it did not look like any other observations for Saudi Arabia were an issue. Following the discovery of these inaccuracies in our dataset, we double-checked several other observations to confirm that the rest of the dataset was accurate. This confirmed that the remaining observations were correct and that the outliers we found were isolated.

After removing the outliers, we reevaluated our assumptions to see if any of the violations had been improved or fixed. Both normality and linearity look to be worse in comparison to our original tests, but homoscedasticity did see improvement, as seen in Figure 10 below.



(a) (b)

Figure 10 (a) The Residual vs Fitted plot used to test linearity and homoscedasticity after the removal of outliers, along with (b) the Normal Q-Q Plot used to test normal distribution after the removal of outliers

Following this, we checked again for outliers, which may be skewing the results, but found nothing else that warranted removal. This led us to explore possible transformations to help improve our model, starting with looking at what could be suggested by a Box-Cox test.

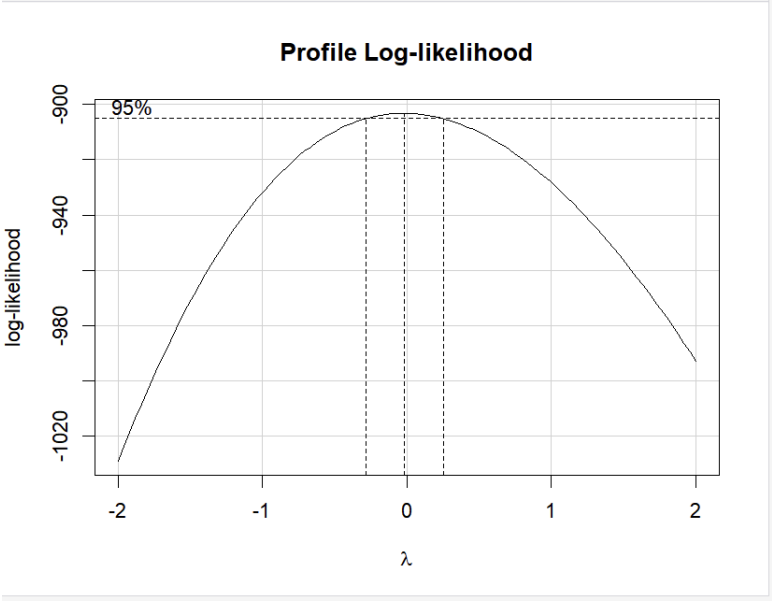
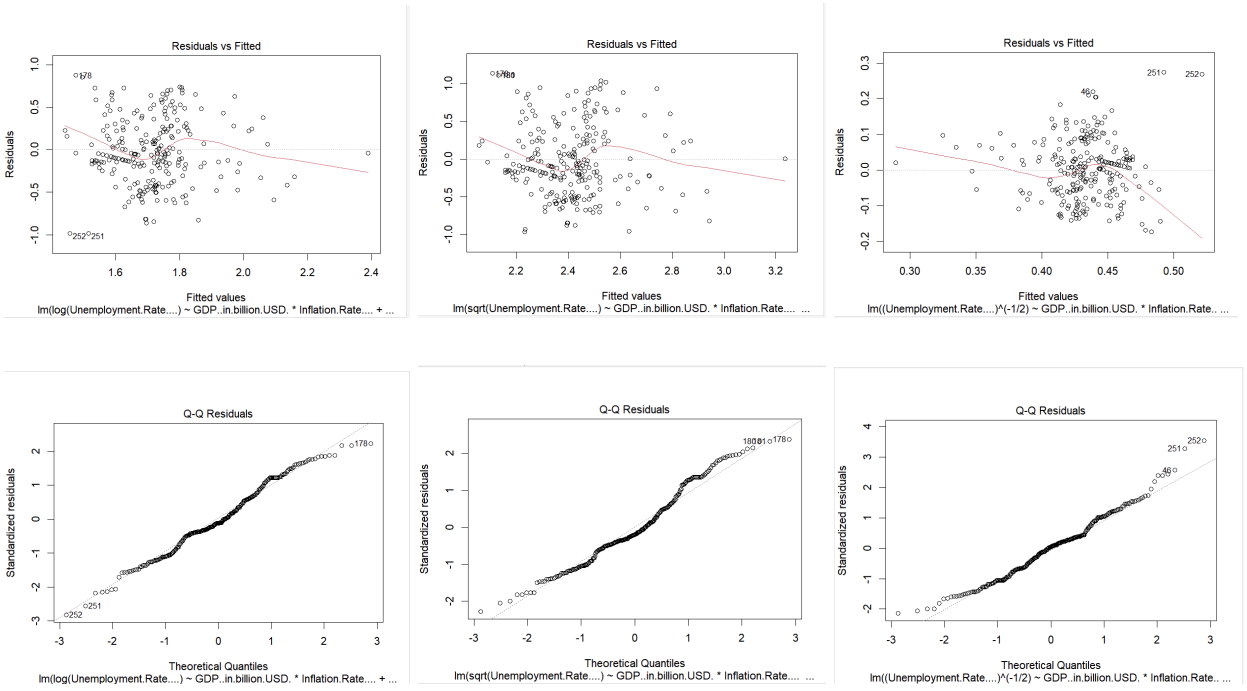


Figure 11: The Box-Cox test used to find which transformation to use

The results of our test, shown in Figure 11, strongly suggest a log transformation. However, Box-Cox tests primarily focus on improving normality, so we decided to attempt multiple

transformations by conducting a square root and inverse square root transformation as well.



LOG TRANSFORMATION SQUARE ROOT TRANSFORMATION INVERSE SQUARE ROOT

Figure 12: The Residual vs Fitted and Q-Q Plot after performing a log, square root, and inverse square root transformations

From each of the graphs in Figure 12, we see some small and large differences between the results of each transformation. The log transformation shows an improvement in homoscedasticity and normality, while linearity shows a worse result. Next to that, the square root transformation sees the same improvement in homoscedasticity and decline in linearity as the log transformation. However, in the Q-Q plot, we see a worse result for normality, which has a lot more curvature to it. Finally, the inverse square root shows a much worse linearity, homoscedasticity, and normality, ultimately making it the worst transformation out of the three. As a result of these tests, we decided to use the log transformation for our final model, like our Box-Cox test showed, which had the best improvement for normality while not making the other violations worse.

STEPWISE SELECTION

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.820219   0.033315  54.636  < 2e-16 ***
Economic.Growth.... -0.032809   0.007249  -4.526 9.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3992 on 249 degrees of freedom
Multiple R-squared:  0.07602,   Adjusted R-squared:  0.07231
F-statistic: 20.49 on 1 and 249 DF,  p-value: 9.308e-06
```

BACKWARDS SELECTION

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              1.849e+00  3.889e-02  47.554  < 2e-16 ***
GDP..in.billion.USD.                    -1.922e-05  8.186e-06  -2.348   0.0197 *
Economic.Growth....                     -3.495e-02  7.249e-03  -4.822 2.49e-06 ***
GDP..in.billion.USD.:Inflation.Rate....  4.470e-06  2.204e-06   2.028   0.0437 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3962 on 247 degrees of freedom
Multiple R-squared:  0.09724,   Adjusted R-squared:  0.08627
F-statistic: 8.868 on 3 and 247 DF,  p-value: 1.325e-05
```

FORWARDS SELECTION

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              6.981e+00  2.418e-01  28.871  < 2e-16 ***
Economic.Growth....                     -2.345e-01  4.504e-02  -5.207 4.05e-07 ***
GDP..in.billion.USD.                    -1.116e-04  5.067e-05  -2.203   0.0285 *
Inflation.Rate....                      -3.874e-04  3.106e-03  -0.125   0.9009
GDP..in.billion.USD.:Inflation.Rate....  2.038e-05  1.357e-05   1.502   0.1344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.439 on 246 degrees of freedom
Multiple R-squared:  0.1099,    Adjusted R-squared:  0.09543
F-statistic: 7.594 on 4 and 246 DF,  p-value: 8.76e-06
```

Figure 13: The summaries of stepwise selection backwards selection, and forward selection

Next, we looked at what variables to include in the final model. We tried stepwise, backwards, and forward selection with the results of each shown in Figure 13. Stepwise selection eliminated all but economic growth, leaving much to be desired with an R-squared of 0.07602. Backwards Selection was an improvement, only removing inflation and the interaction terms involving economic growth, resulting in an R-squared of 0.09724. The best of the three was forward selection, leaving everything but the interaction terms involving economic growth in the model. This also gave the greatest R-squared of 0.1099, making it the best choice of the three for our final model.

Model Reliability & Summary

After finalizing our model using forward selection, we ended up with GDP, economic growth, inflation rate, and the interaction term between GDP and Inflation. The model has an R-squared value of 0.1099, indicating that the model only explains a small portion of the variance in the dataset. This suggests that other factors not included in the model may play a much more significant role in influencing the unemployment rate. Further analysis revealed that both the inflation rate and its interaction term with GDP are not statistically significant at the $\alpha = 0.05$ level. This is unexpected, as our research showed that inflation and unemployment rates often share an inverse relationship. These discrepancies and limitations may stem from the limited number of observations. After removing the outliers, we were left with fewer than 250 total observations, which may have impacted the model's ability to accurately capture the relationship between these variables.

The low R-squared value and the lack of significance of the inflation rate and its interaction term highlight the limitations of our model and suggest the need for further research. Future studies could explore additional variables such as economic policies, world events, or technological advancements. Using alternative modeling techniques, such as a time series, may better represent the complex dynamics of unemployment rates over time. Additionally, gathering a larger dataset could improve the model's accuracy and predictive power.

Conclusion

In this study, we explored the significance of economic development on unemployment by analyzing variables such as GDP, inflation, and economic growth. Through our exploratory data analysis and model diagnostics, we identified and addressed violations in linearity, homoscedasticity, and normality, ultimately opting for a log transformation to improve model performance. Our final model was found using forward selection and included GDP, economic growth, inflation rate, and the interaction term between GDP and Inflation. Although it had the greatest R-squared, it remained low at 0.1099, suggesting

these economic factors alone are not enough to accurately explain unemployment rates. Despite these limitations, our findings show how macroeconomic variables relate to unemployment and the importance of accurate data collection. Future work could incorporate a larger dataset and apply an alternative method, such as a time series, to enhance our model's accuracy. Our research serves as a foundation for further exploration and may assist in further understanding the relationship between unemployment and various economic factors.

Works cited

Chen, James. "Economic Growth Rate." *Investopedia*, 2019, www.investopedia.com/terms/e/economicgrowthrate.asp.

DePersio, Greg. "What Happens When Inflation and Unemployment Are Positively Correlated?" *Investopedia*, 22 Aug. 2020, www.investopedia.com/ask/answers/040715/what-happens-when-inflation-and-unemployment-are-positively-correlated.asp.

Ganong, Peter, and Pascal Noel. *How Does Unemployment Affect Consumer Spending?* Harvard University, 4 Jan. 2016, https://scholar.harvard.edu/files/ganong/files/ganong_jmp_unemployment_spending.pdf.

Macrotrends. *Turkey Unemployment Rate 1991–2024*. Macrotrends LLC, 2024, https://www.macrotrends.net/global-metrics/countries/tur/turkey/unemployment-rate. Accessed 15 Apr. 2025.

Ravikumar, B, et al. "Convergence or Divergence? A Look at GDP Growth across Richer and Poorer Countries." *Stlouisfed.org*, Federal Reserve Bank of St. Louis, 19 Aug. 2024, www.stlouisfed.org/on-the-economy/2024/aug/convergence-divergence-gdp-growth-richer-poorer-countries.

Shamim, Adil. “Economic Indicators & Inflation.” *Kaggle.com*, 2025, www.kaggle.com/datasets/adilshamim8/economic-indicators-and-inflation, https://doi.org/10738981/f74784dabf1b5ea910d773e50d2eaa48. Accessed 27 Feb. 2025.

The World Bank. *Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate) – Turkey*. The World Bank Group, 2024, https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2024&locations=TR&start=2009&view=chart. Accessed 15 Apr. 2025.

The World Bank. *Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate) – Saudi Arabia*. World Bank, https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?locations=SA. Accessed 20 Apr. 2025.

Macrotrends. *Saudi Arabia Unemployment Rate 1991-2024*. Macrotrends LLC, https://www.macrotrends.net/global-metrics/countries/SAU/saudi-arabia/unemployment-rate. Accessed 20 Apr. 2025.

We, the project teams members, certify that below is an accurate account of the percentage of effort contributed by each team member in the project and report.

Project Team Member	Percentage of Total Effort
Sejal Mogalgiddi	25%
Tony Tran	25%
Ryan Jones	25%
Gabriel Alvarado	25%