QTM 350 Final Project

Bao Truong (2643551), Daolong Yang (2485580), Xipu Wang (2551008)

November 29, 2024

Introduction

This is the final project for QTM 350 class. Our group chooses to work on the topic of economic development. In this project, we will try to identify the relationship between PPP(Power Purchase Parity), GDP Growth, and Employment Ratio. We will analyze the trend to draw a meaningful insights to apply for other countries. Our research question is: "What is the correlation between PPP, Employment Ratio, and GDP Growth. Does having higher PPP and higher Employment Ratio correlates a higher increase in GDP".

The finding yields the result that Employment ratio has a positive correlation with GDP growth whereas the PPP has a negative correlation with GDP growth.

Data description

This project leverages World Bank data to explore and analyze the economic development trends in the North America region. The analysis focuses on three key economic indicators:

1. ppp: Measures the average economic output per person in the region adjusted for inflation.

2. gdp_growh: Tracks the annual percentage growth in the region's economic output.

3. employment: Reflects the proportion of the working-age population that is employed. These variables provide a comprehensive overview of the region's economic health and development trajectory.

To obtain this data, we use wbdata to get 3 datasets from the World Bank: ppp, gdp_growth, and employment with each dataset includes columns with name of the country, date, and the respective variables. We clean the data using SQL. We try to filter out every country that's not United States, Canada, or Mexico. In addition, we drop the row in every dataset that we filter that has a null value in variable columns.We export the cleaned datasets into the data

folder and name PPP for ppp, Growth for gdp_growth, and Employment for employment.
Then, utilizing SQL, we draw a table of summary statistics of 3 countries in 3 variables: ppp,
gdp_growth, and employment We then merge Employment Ratio, PPP, and GDP Growth
on 'country' and 'date column'. We change the ppp variable into log_ppp to make our model
more stable.

Data analysis

Data visualization

GDP Growth Over Time by Country
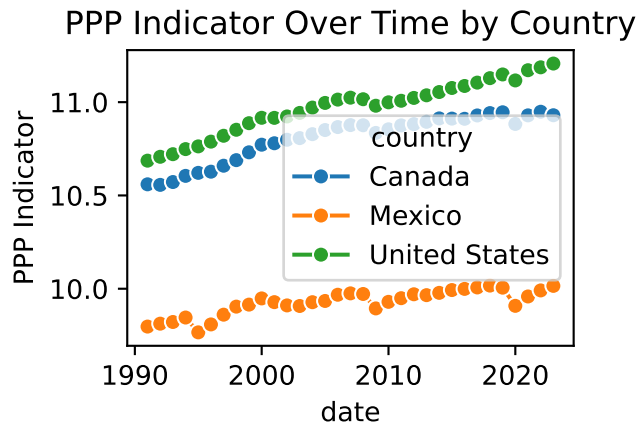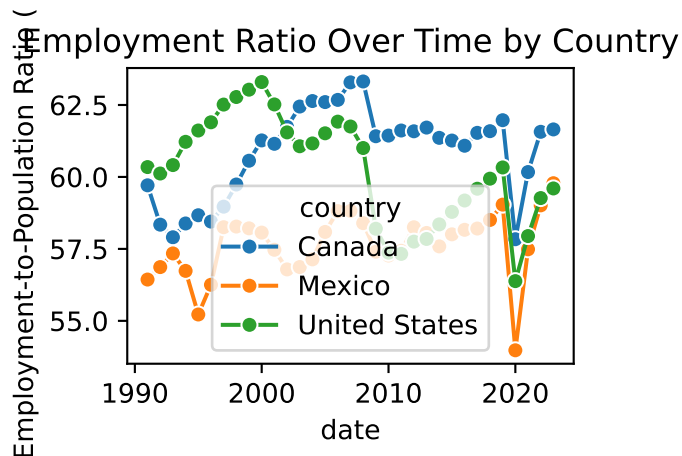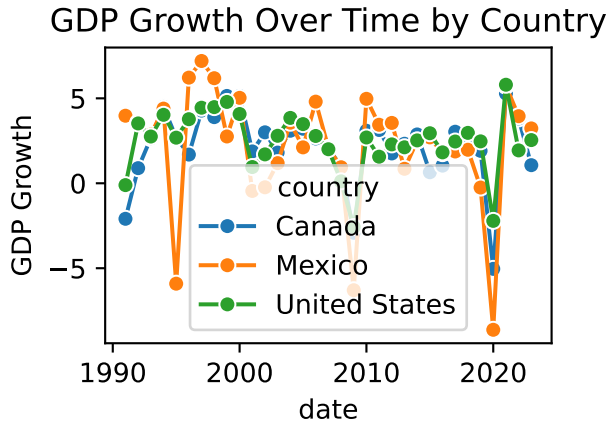
These summary statistics give us a comparative snapshot of three key North American economies—Canada, Mexico, and the United States—across several indicators: GDP growth, PPP (purchasing power parity), log GDP per capita, and employment.

GDP Growth: All three countries have similar mean GDP growth rates, with the U.S. leading slightly at about 2.47%, followed by Canada at 2.19%, and Mexico at 2.17%. However, Mexico's growth is notably more volatile, as indicated by a higher standard deviation (3.50) and more extreme minimum (-8.62%) and maximum (7.20%) values. Canada and the United States have smaller standard deviations and narrower ranges, suggesting more stable year-to-year growth. But we see a large change on the plot around 2010, 2020. This is caused by the 2008 financial crisis and COVID-19 pandemic in 2020, which are factors that have heavily affected other variables as well.

PPP: The U.S. has the highest average PPP value (around 58,707), followed by Canada, then Mexico, reflecting a clear hierarchy in terms of wealth and price levels. Canada's PPP averages around 49,780, while Mexico's is significantly lower at around 20,519.

Employment: In terms of labor market outcomes, both Canada and the United States show higher average employment-to-population ratios (around 60-61%) compared to Mexico's average of about 57.6%. While all three countries maintain relatively stable employment rates, the smaller standard deviation for Mexico (1.14) suggests that its employment ratio is more consistently clustered around its mean, albeit at a lower level.

Overall Picture: In summary, the U.S. stands out as the wealthiest with the highest PPP and log GDP per capita and stable, moderately strong growth. Canada shares these characteristics to a somewhat lesser extent, while Mexico's economic indicators point to lower average wealth and more volatile growth rates. Although all three countries are economically developed, these statistics underline important differences in stability, living standards, and labor market conditions.

We then create a residual plot to check for the requirements of linear regression model.

Residual plots
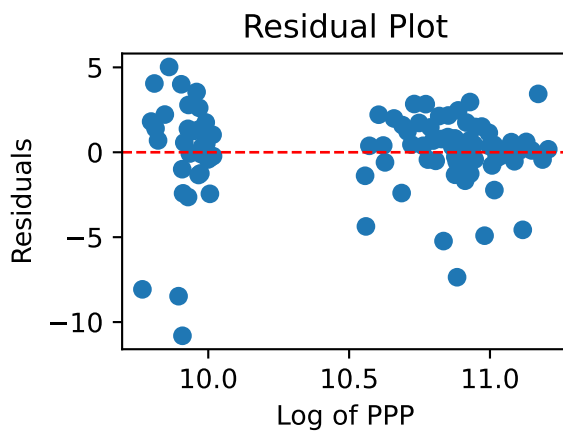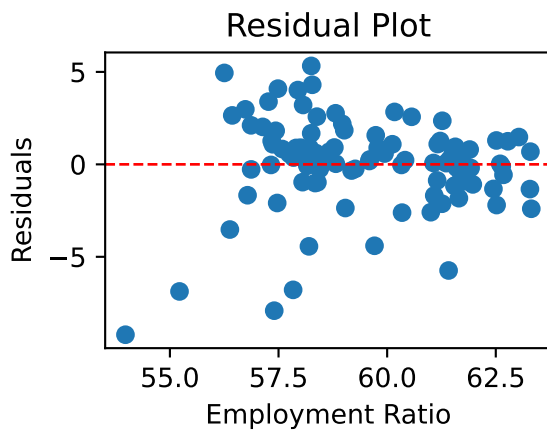
There seems to be no clear pattern in the residual plots of Employment Ratio and Log of PPP. Therefore, the assumption for a linear regression model is met. We will use multiple linear regression with input outcome as log_ppp and employment ratio and the output outcome is gdp_growth.

Model

```
                         OLS Regression Results
==============================================================================
Dep. Variable:            gdp_growth   R-squared:                       0.088
Model:                           OLS   Adj. R-squared:                  0.069
Method:                Least Squares   F-statistic:                     4.609
Date:               Mon, 09 Dec 2024   Prob (F-statistic):             0.0123
```

```
Time:                        20:48:02   Log-Likelihood:                  -228.90
No. Observations:                  99   AIC:                                463.8
Df Residuals:                      96   BIC:                                471.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -13.3724      7.278     -1.837      0.069     -27.820       1.075
employment      0.4707      0.156      3.024      0.003       0.162       0.780
log_ppp        -1.1734      0.684     -1.716      0.089      -2.530       0.184
==============================================================================
Omnibus:                       28.690   Durbin-Watson:                   1.895
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               52.299
Skew:                          -1.180   Prob(JB):                     4.40e-12
Kurtosis:                       5.666   Cond. No.                     1.77e+03
==============================================================================
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

This regression model suggests that while the overall relationship between GDP growth and the two predictors—log PPP, employment ratio is statistically significant at the model level (F-test p-value 0.012), the explanatory power is quite low. With an R-squared of only about 0.089, the model explains less than 10% of the variation in GDP growth.

Looking at individual predictors, only the employment variable shows a statistically significant positive effect on GDP growth ($p < 0.01$). Variation in employment is somewhat associated with changes in GDP growth, whereas purchasing power (PPP) appears to have a negative influence on GDP growth but is statistically insignificant.

In summary, the model only modestly explains the variation in GDP growth, with employment showing some predictive power, but the presence of multicollinearity and non-normal residuals highlights the need for more careful modeling or additional variables.

```
                        OLS Regression Results
==============================================================================
Dep. Variable:            gdp_growth   R-squared:                       0.122
Model:                           OLS   Adj. R-squared:                  0.103
Method:                Least Squares   F-statistic:                     6.476
Date:               Mon, 09 Dec 2024   Prob (F-statistic):            0.00233
```

```
Time:                         20:48:02   Log-Likelihood:                     -219.47
No. Observations:                   96   AIC:                                  444.9
Df Residuals:                       93   BIC:                                  452.6
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -13.6174      7.229     -1.884      0.063     -27.973       0.738
employment     0.5401      0.150      3.596      0.001       0.242       0.838
lag_ppp       -1.5382      0.664     -2.316      0.023      -2.857      -0.219
==============================================================================
Omnibus:                       29.296   Durbin-Watson:                        1.889
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                    56.688
Skew:                          -1.195   Prob(JB):                          4.90e-13
Kurtosis:                       5.908   Cond. No.                          1.78e+03
==============================================================================
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.78e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

To improve our model, we decide to use the log_ppp from last year as an input outcome
instead of using the log_ppp for this year. This is because we believe that the log_ppp will
affect the GDP growth next year.

Results and Discussions

This revised model (that uses PPP of last year instead of this year) explains slightly more
variation in GDP growth than the previous one (R-squared of about 0.122, up from 0.088),
indicating that around 12% of the variation in GDP growth is accounted for by the included
variables: lagged PPP, and employment. The overall model is statistically significant (p
0.0023), suggesting that the combination of predictors is not random noise.

Looking at the individual coefficients, employment is again the clearest and most statistically
significant predictor. Its positive and significant coefficient ($p < 0.001$). On the other hand,
the lag_ppp is also statistically significant ($p = 0.023$) but with a large negative magnitude.

The large condition number (1.77e+03) indicates potential multicollinearity or numerical in-
stability. This might mean that the predictors are highly correlated with each other, making it
difficult to isolate their individual effects. Thus the result should be intepreted with caution.

Conclusions

In conclusion, our analysis demonstrates that using the lagged PPP (Purchasing Power Parity) as an input variable provides a better explanation of the variation in GDP growth compared to using the current year's PPP. The revised model, which includes lagged PPP and employment as predictors, accounts for approximately 12% of the variation in GDP growth, as indicated by the R-squared value of 0.122. This is an improvement from the previous model's R-squared value of 0.088.

We can see that there is a strong positive correlation between Employment ratio and GDP growth. This can be the result of a country having a good economy tends to have higher GDP growth and more jobs for people. However, there is also a negative correlation between PPP and GDP growth. This might be intepreted as developed countries with higher PPP often have lower GDP growth.