We need to find a good title guysss

FEM11149 - Introduction to Data Science

Maya Archer (25 %), Saumnya (25 %),
Rishi Ashok Kumar 560527 (25 %), Nicolas Gonzalez 780037 (25%)

September, 2025

1. Introduction

Economic growth is a central concern for governments worldwide, especially when planning to increase public spending that may require external borrowing. Since borrowed funds must eventually be repaid, it is crucial to ensure that the economy grows at a sustainable pace. A commonly used indicator for this purpose is the **Gross Domestic Product (GDP) growth rate**, which provides a measure of how rapidly an economy is expanding over time. In this project, we work as part of a **data science consulting team** for governmental agencies. One of the agencies has tasked us with investigating whether a country's GDP growth can be predicted using a combination of **economic and demographic variables**.

2. Data

Our analysis uses the **Global Jobs Indicators Database (World Bank)**. From the available variables, we selected a subset grouped as follows:
- **Demographics & human capital**: population growth, population density, and male life expectancy at birth.
- **Labor market**: labor force participation rate, unemployment rate, and employers' share of total employment.
- **External sector**: net trade in goods and services, export value index.
- **Financial development**: private credit bureau coverage.
- **Development level (controls)**: GDP per capita (PPP, constant 2011 international $).

We also created transformed variables: *signed log of net trade*, *population growth squared*. These account for nonlinearities, while keeping the model interpretable.

3. Results.

At first, we wanted to test whether including **net trade in goods and services** would improve the explanatory power of our GDP growth model. When **net trade** was included, its coefficient was essentially zero and statistically insignificant ($p \approx 0.93$). Moreover, the **Adjusted R^2** decreased slightly (0.137 without vs. 0.131 with), suggesting that the variable adds noise rather than explanatory value. The ANOVA test comparing the two models confirmed that net trade does not significantly improve model fit ($p \approx 0.92$).

Therefore, we conclude that **net trade should not be included** in the final model. Instead, the variables that consistently show significant relationships with GDP growth are *Employers, total (% of employment)* (negative and highly significant), *Export value index* (positive and significant), *Life expectancy at birth,*

male (positive and significant), and *GDP per capita, PPP (constant 2011 international $)* (negative and significant). Other controls, such as unemployment, labor force participation, credit coverage, and population dynamics, do not show significant effects in this specification.

Additionally, we added two transformations to test for nonlinearities: a signed log of net trade (`log_net_trade_signed`) to allow for diminishing returns and preserve the deficit/surplus sign, and a quadratic term in population growth (`pop_growth_sq`) to capture potential U- or inverted-U-shaped demographic effects. In the augmented model, neither coefficient is statistically significant (`log_net_trade_signed`: $p \approx 0.25$; `pop_growth_sq`: $p \approx 0.80$), and overall fit does not improve (Adjusted R^2 falls from ≈ 0.131 without the terms to ≈ 0.127 with them). Baseline **net trade** also remains insignificant ($p \approx 0.72$). By contrast, the **export value index** (positive), **employers' share** (negative), **male life expectancy** (positive), and **GDP per capita (PPP)** (negative) remain significant and stable.

```
#
# n1 <- 110
# idx <- sample(1:nrow(dfm1), n1)
#
# dfm_train <- dfm1[idx, ]
# dfm_test <- dfm1[-idx, ]
#
# x <- dfm_train %>%
#   select(!gdp_growth)
#
# y = dfm_train$gdp_growth
#
# fitControl <- trainControl(method =  'repeatedcv', number = 10, repeats = 5 )
#
# lasso <- train(
#   x = as.matrix(x),
#   y = y,
#   method = "glmnet",
#   family,
#   preProcess = c("center", "scale"),
#   tuneGrid = expand.grid(
#     alpha  = 1,
#     lambda = seq(-4, 2, length.out = 1000)
#   ),
#   trControl = fitControl
# )
#
# help('train')
# plot(lasso)
# lasso
#
# library(dplyr)
#
# # See best Lambda
# lam <- lasso$bestTune$lambda
# lam
#
# # Coefficientes for the Lamda.
# b <- coef(lasso$finalModel, s = lam)  # matriz dispersa
# coefs <- as.matrix(b)
#
# df_coef <- data.frame(
```

```
#   variable = rownames(coefs),
#   beta = coefs[, 1],
#   row.names = NULL
# ) %>%
#   filter(variable != "(Intercept)")
#
# eps <- 1e-6
#
# # Actual Variables kept and those killed.
# vars_seleccionadas <- df_coef %>% filter(abs(beta) > eps) %>% arrange(desc(abs(beta)))
# vars_matadas       <- df_coef %>% filter(abs(beta) <= eps) %>% arrange(variable)
#
#
# nrow(vars_seleccionadas)
# nrow(vars_matadas)
```

4. Methods.

5. Conclusion and Discussion.

6. Code.

7. References.