Project Report

Global Economic and Social Indicators: Analyzing Key Drivers of Growth and Development

Mehrad Hajati

Ahnaf Iqbal

Bright Adams

Data Analysis STAT 4620/5620 Winter 24-25

https://github.com/MHajatiDAL/DataAnalysisProject

Keywords: Mixed Effects Models, ANOVA, Generalized Linear Models, Linear Models, AIC, Global Development, Economic Growth, Unemployment,

Abstract

This project investigates the impact of two government expenditures. First the impact of government expenditure on education and how it effects female unemployment. Then government expenditure on research and development (R&D) and its effects on high-technology exports. Using global data from the World Bank (2000–2021) and a series of linear and mixed-effects models, we analyze how these investments influence development outcomes. Findings reveal that education spending does not uniformly reduce female unemployment, its effectiveness depends on factors like broadband access, tax revenue, and labor force participation. Also, while R&D expenditure is positively associated with high-tech exports, the strength of this relationship varies significantly across countries. These results show the importance of country specific strategies, showing that policies must be supported by broader infrastructure and economic conditions to have meaningful impact.

Introduction

In today's world, development is driven by a complex combination of economic, social, and technological factors. Among the various indicators used to assess a country's level of development, economic growth, innovation, and employment remain some of the most important factors. Policymakers and researchers are interested in understanding how specific investments and policies such as education spending, taxation, and research and development (R&D) influence these development outcomes. Insights gained from these relationships can guide more effective policy design and resource allocation at both national and international levels.

This study aims to explore two fundamental questions related to economic and social development across countries. The first question investigates, "The conditions under which government expenditure on education impacts unemployment rates, with a particular focus on female unemployment". Education is widely regarded as a long-term investment in human capital; however, its direct and indirect effects on employment outcomes could be influenced by other factors.

The second question examines, "The relationship between research and development expenditure and high-technology exports varies across countries". As countries look to enhance their positions in the global innovation landscape, understanding whether R&D investments translate into tangible outputs like tech exports is important. Yet, this relationship may not be uniform, as it depends on other factors.

Using data sourced from the World Bank covering a wide range of global indicators, this study uses statistical analysis to uncover not only general trends but also country-specific variations. By analyzing these two research questions independently, we aim to provide a good understanding of how policy-driven inputs influence important development outcomes.

The insights generated can inform more targeted, evidence-based policymaking for sustainable economic and social progress.

Data Description

This project uses a dataset originally gathered to support research on the United Nations Sustainable Development Goals (UN SDGs). The dataset was compiled from the World Bank's publicly available indicators. Initially launched in 2019 with a Eurostat focus, the dataset was expanded in 2020 to include a broader selection of World Bank indicators. Data was downloaded directly from the World Bank website and minimally processed to retain the integrity of the original data. Preparation steps and the timeframe/scope of data is included below:

- Extracting country and indicator metadata (e.g., codes, descriptions)
- Consolidating 34 separate indicator CSVs into a single file
- Keeping only essential columns (codes and values) for compactness and clarity
- Geographic coverage: Global (includes EU27 and non-EU countries)
- Time coverage: 2000 to 2021 (some 1999 data may appear as part of 2000)
- Number of indicators: 34 key development indicators
- Number of countries: Varies by indicator, generally over 180 countries

The full list of features, including their code and description can be found in Appendix A.

Data Structure and Preprocessing

The dataset is structured as a tidy CSV with the following key components:

1. Country Code and Name
2. Indicator Code and Name
3. Year (2000–2021)
4. Value: Numeric representation of the indicator (e.g., percentages, index scores, GDP shares)

Initially, the raw data was structured such that each indicator appeared as a row with corresponding years as column headers. This wide format was not suitable for statistical modeling. The indicator names were transformed into individual columns, and the years were turned into rows. Resulting in a long-format dataset where each row represented a single country-year combination with all selected indicators as variables. This structure allowed for easier modelling and visualization.

Research Question 1: Under what conditions and how does government expenditure on education affect unemployment rates?

Unemployment is one of the most persistent challenges affecting countries globally. One proposed solution often suggested by policymakers and economists is increased investment in education. The hypothesis is that by improving access to and quality of education, a nation can ensure its citizens have better skills, which in turn makes them more employable and drives economic productivity.

But does the data support this assumption? More specifically, to what extent does government expenditure on education (as a percentage of GDP) help reduce unemployment rates, particularly for females? This research question explores the potential link between a country's investment in education and its resulting unemployment outcomes, using international data collected over two decades.

Exploratory Data Analysis

To visually explore the relationship between education expenditure and unemployment, we created a scatter plot with government education spending as a % of GDP and unemployment rate as shown in the figure below.

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(ggplot2)
library(lme4)  # For linear mixed models
```

```
Loading required package: Matrix

Attaching package: 'Matrix'
```

```
The following objects are masked from 'package:tidyr':

    expand, pack, unpack
```

```
facttable<-read_csv("./facttable.csv")
```

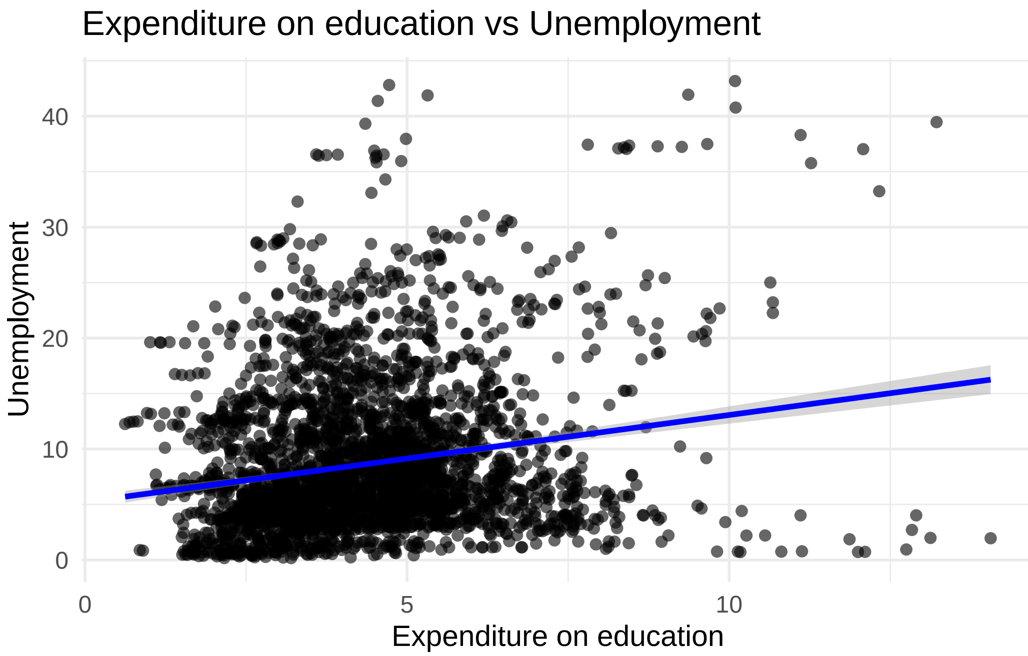
```
Rows: 8778 Columns: 24
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): Country Code, Indicator
dbl (22): 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
facttable_wide <- facttable %>%
  pivot_longer(cols = 3:24, names_to = "year", values_to = "value") %>%
  pivot_wider(names_from = Indicator, values_from = value)
```

```
data <- facttable_wide %>%
  select("Country Code", year, SE.XPD.TOTL.GD.ZS, SL.UEM.TOTL.FE.ZS) %>%
  na.omit()
```

```
ggplot(data, aes(x = SE.XPD.TOTL.GD.ZS, y = SL.UEM.TOTL.FE.ZS)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Expenditure on education vs Unemployment",
       x = "Expenditure on education",
       y = "Unemployment") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Each point on the plot represented a specific country-year observation. A linear trend line was also added using a smoothing function. The plot suggested a positive association contrary to expectations, higher education spending seemed to correlate with higher unemployment rates, though the data also exhibited wide variability. This early finding hinted at possible complexities in the relationship between the two variables.

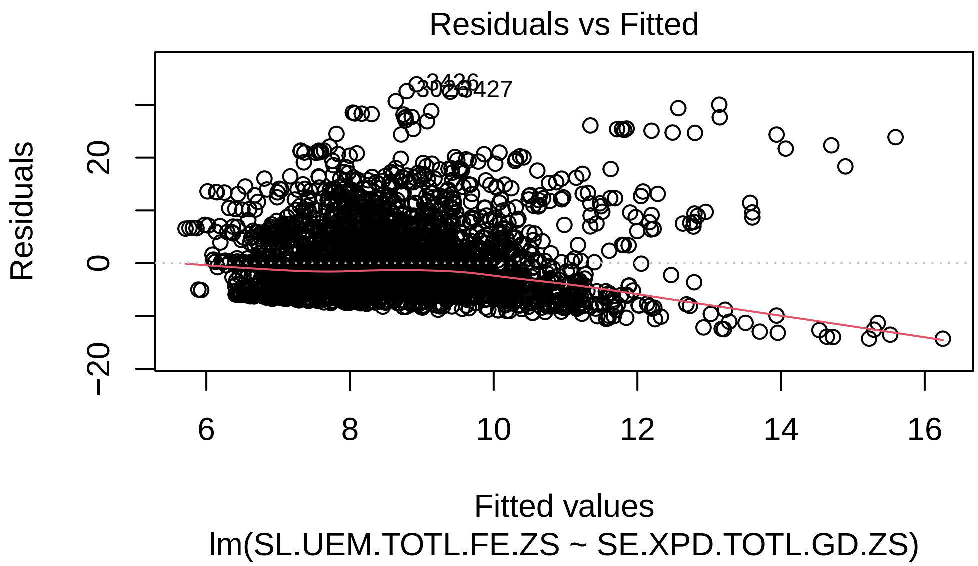
Linear Regression

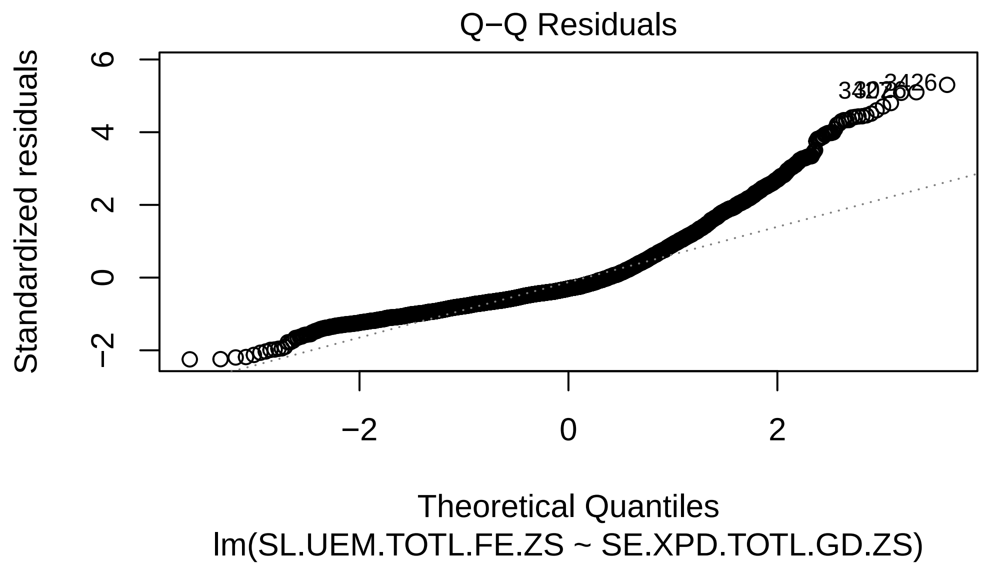
```r
lm_model <- lm(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS, data = data)
summary(lm_model)
```

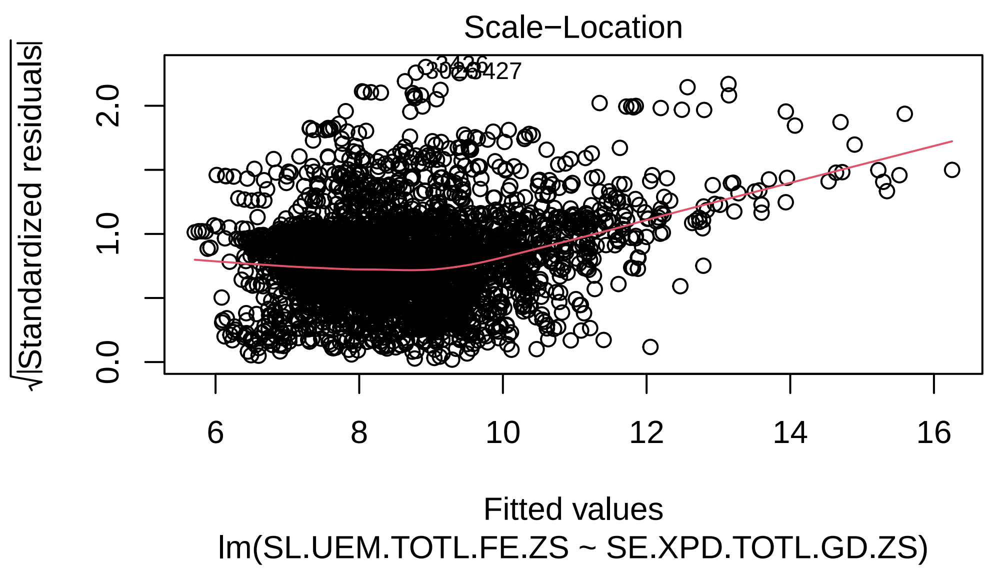
```
Call:
lm(formula = SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.293  -4.090  -1.928   2.474  33.884 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
```

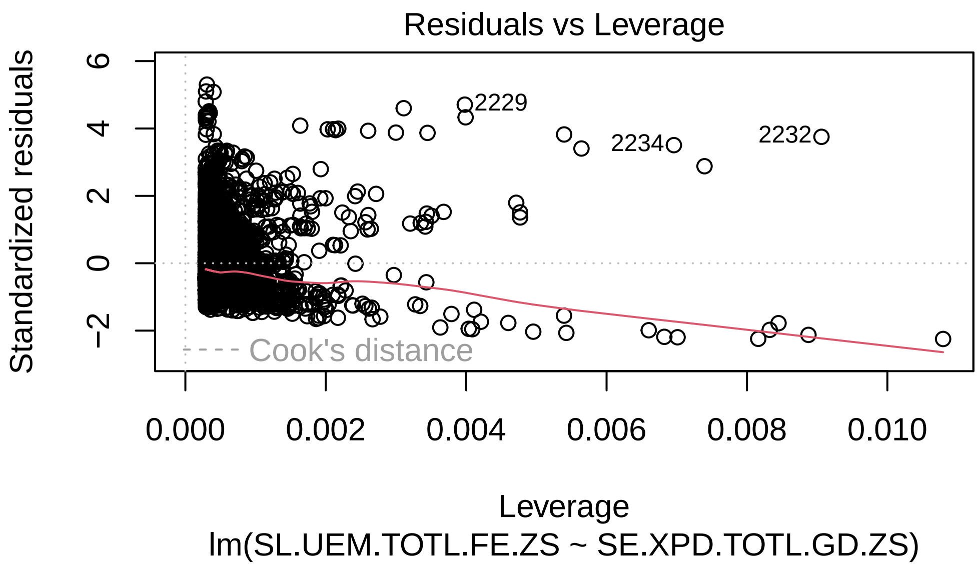
```
(Intercept)        5.22263    0.30991   16.85   <2e-16 ***
SE.XPD.TOTL.GD.ZS  0.78455    0.06723   11.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.391 on 3453 degrees of freedom
Multiple R-squared:  0.03794,   Adjusted R-squared:  0.03766
F-statistic: 136.2 on 1 and 3453 DF,  p-value: < 2.2e-16
```

```r
plot(lm_model)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS)


Scale−Location
√|Standardized residuals|
Fitted values
lm(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS)



Standard diagnostic plots above were examined to assess the linear model's fit, revealing a residuals vs fitted plot with a wide, pattern-less spread suggestive of unequal variance, while the normal Q-Q plot displayed deviations from the diagonal line, particularly at the tails, indicating non-perfectly normally distributed residuals; these diagnostic findings confirmed that the linear model offers a rough estimate but its assumptions are not fully satisfied.

Generalized Linear Model (Gamma distribution)

```
source("./histogram.R")
plot_histogram(data, "SL.UEM.TOTL.FE.ZS","Histogram of Unemployment")
```

```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

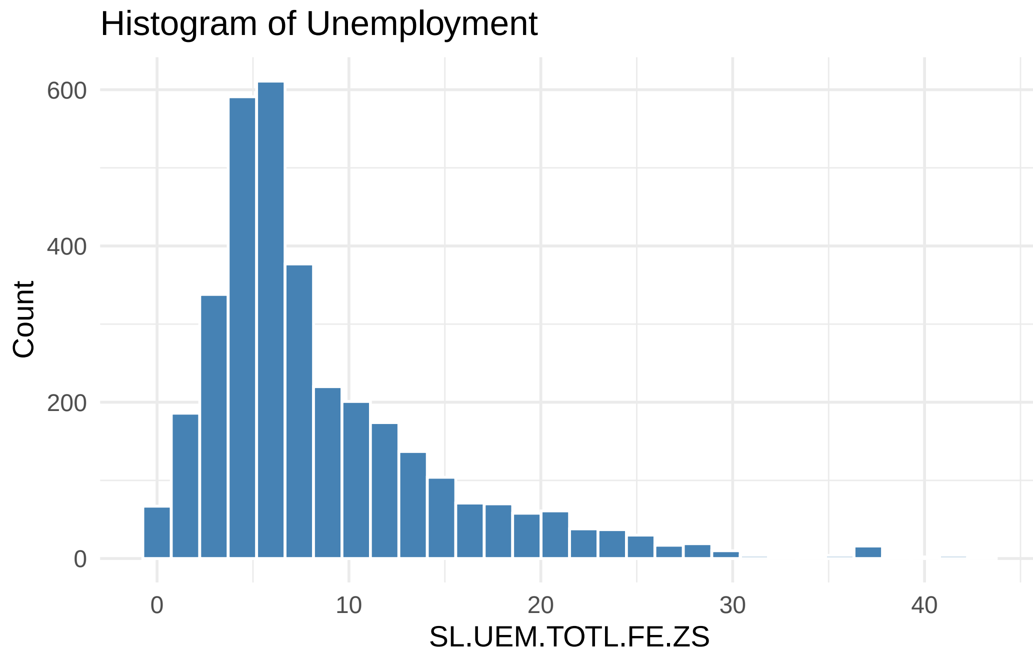


Figure reveals that the response variable, female unemployment rate (SL.UEM.TOTL.FE.ZS), consists exclusively of continuous and strictly positive values. This characteristic violates one of the key assumptions of ordinary linear regression, the assumption of normally distributed residuals. Recognizing this, we explored an alternative modeling approach better suited for positively skewed continuous data. We applied a Generalized Linear Model (GLM) using the Gamma distribution family, which is particularly appropriate for modeling positive continuous outcomes with heteroscedasticity. A Generalized Linear Model (GLM) is an extension of the classical linear regression framework that allows for response variables to have error distribution models other than a normal distribution. GLMs are useful when modeling response variables that are not continuous and normally distributed such as binary, count, or strictly positive data. GLMs consist of three components: a random component specifying the distribution of the response variable, a linear predictor composed of fixed effects, and a link function that connects the mean of the response variable to the linear predictor (Nelder & Wedderburn, 1972)

```
gamma_model <- glm(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS,
                   family = Gamma(link = "inverse"), data = data)
summary(gamma_model)
```

```
Call:
glm(formula = SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS, family = Gamma(link = "inverse"),
    data = data)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.1549004  0.0036053   42.97   <2e-16 ***
SE.XPD.TOTL.GD.ZS -0.0085063  0.0006784  -12.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.5298314)

    Null deviance: 1882.5  on 3454  degrees of freedom
Residual deviance: 1817.0  on 3453  degrees of freedom
AIC: 20884

Number of Fisher Scoring iterations: 6
```

Comparing the Linear Model to the GLM model

```r
AIC(lm_model, gamma_model)
```

```
            df      AIC
lm_model     3 22626.07
gamma_model  3 20884.26
```

Upon comparing model performance using the Akaike Information Criterion (AIC), we observed a substantial improvement: the linear model yielded an AIC of 22,626.07, while the Gamma GLM achieved a significantly lower AIC of 20,884.26. This reduction in AIC suggests that the Gamma model provides a better fit for the data, capturing the underlying distribution of the response variable more effectively than the standard linear regression model.

Generalized Linear Mixed Model (GLMM) – Random Intercept

To better account for the variability in unemployment rates across countries and to capture potential country-specific differences in the relationship between education expenditure and unemployment, a Generalized Linear Mixed Model (GLMM) was used. A GLMM is an extension of the Generalized Linear Model (GLM) that includes both fixed effects and random effects, making it especially useful for analyzing non-normally distributed data with grouped or hierarchical structures (McCulloch, 1996). This model included Country as a random effect, while keeping government expenditure on education (% of GDP) as a fixed effect predictor. The use of a GLMM enables us to model both global trends and local (country-specific) variations, which can lead to more accurate and generalizable insights.

Mixed models, or mixed-effects models, are an extension of traditional regression models that incorporate both fixed effects and random effects.

- Fixed effects are variables whose effects are assumed to be constant across all observations—in this case, government expenditure on education.
- Random effects, on the other hand, capture variability associated with groupings in the data. In this case, the country each observation belongs to. They allow the intercepts (or slopes) to vary for different levels of a grouping factor, recognizing that different countries might inherently have different baseline unemployment levels due to historical, political, or economic contexts.

The decision to use a GLMM over a standard GLM or linear regression was motivated by two major considerations. First, as seen in earlier diagnostic plots, the response variable (female unemployment rate) is strictly positive and continuous, making the Gamma distribution with a log link an appropriate choice for modeling. Second, the data include repeated measures across different years for multiple countries, making observations from the same country likely to be correlated. Ignoring this structure would violate the assumption of independence and could lead to biased estimates and underestimated standard errors.

```
colnames(data)[colnames(data) == "Country Code"] <- "Country"
lmm_model <- glmer(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (1 | Country), family = Gamma(lin
summary(lmm_model)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: Gamma  ( log )
Formula: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (1 | Country)
   Data: data

     AIC      BIC   logLik deviance df.resid
 14152.9  14177.5  -7072.4  14144.9     3451

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.0561 -0.4068 -0.0397  0.3542  6.7350

Random effects:
 Groups   Name        Variance Std.Dev.
```

```
 Country  (Intercept) 0.11623  0.3409
 Residual             0.08337  0.2887
Number of obs: 3455, groups:  Country, 222

Fixed effects:
                   Estimate Std. Error t value Pr(>|z|)
(Intercept)        1.960292   0.066106  29.654   <2e-16 ***
SE.XPD.TOTL.GD.ZS -0.004539   0.005881  -0.772     0.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
SE.XPD.TOTL -0.383
```

In the fitted GLMM, the fixed effect estimate for government expenditure on education was -0.0045, with a standard error of 0.00588 and a non-significant p-value of 0.44. This indicates that, after accounting for country-level random effects, the relationship between education spending and female unemployment is weak and not statistically significant. The intercept estimate was 1.96, and the random intercept variance for country was 0.116, with a standard deviation of 0.34. This highlights considerable variation in baseline unemployment levels across countries, justifying the inclusion of country as a random effect.

From the random effects output, it is evident that countries differ in their levels of unemployment, independent of education expenditure. The inclusion of this random effect structure is statistically supported by a comparison of model performance using the Akaike Information Criterion (AIC).

Comparing the LM, GLM and GLMM

```r
AIC(lm_model, gamma_model, lmm_model)
```

```
            df      AIC
lm_model     3 22626.07
gamma_model  3 20884.26
lmm_model    4 14152.87
```

When comparing the AIC values of the simple linear model and the GLMM, the contrast is striking. The linear model had an AIC of 22,626.07, while the GLMM yielded a significantly lower AIC of 14,152.87. This large drop in AIC (a difference of over 8,400 points) suggests that the GLMM provides a substantially better fit to the data, even after accounting for the extra complexity introduced by the random effect.

This comparison clearly shows that failing to account for the hierarchical or nested structure of the data (i.e., repeated measures within countries) would have led to a suboptimal model. The GLMM's superior performance indicates that country-level differences do meaningfully influence unemployment rates, and modeling these differences explicitly improves our understanding of how education expenditure relates to employment outcomes globally.

The GLMM analysis shows that although the direct, global relationship between government expenditure on education and female unemployment is weak and not statistically significant after controlling for country-level variability, accounting for random effects significantly improves model fit. The presence of substantial variation across countries shows the importance of localized factors in shaping employment outcomes, factors that are not captured solely by spending on education. Therefore, while increased education expenditure alone may not universally reduce unemployment rates, especially among women, its effects may be more influenced by a host of country-specific social and economic variables.

Generalized Linear Mixed Model (GLMM) – Random Intercept and Slope

Following the results from the random intercept model, we extended the analysis by fitting a Generalized Linear Mixed Model (GLMM) that included both a random intercept and a random slope for government expenditure on education across countries. In this model, not only was each country allowed to have its own baseline level of female unemployment (random intercept), but the effect of education expenditure on unemployment (slope) was also allowed to vary across countries. This is important in real-world contexts, where different countries may not only start at different levels of unemployment but also respond differently to increases in education spending due to various socio-economic, cultural, and policy-driven factors.

```r
# Random Intercept Model
model_random_intercept <- glmer(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (1 | Country),family
summary(model_random_intercept)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: Gamma  ( log )
Formula: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (1 | Country)
   Data: data

     AIC      BIC   logLik deviance df.resid
 14152.9  14177.5  -7072.4  14144.9     3451
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.0561 -0.4068 -0.0397  0.3542  6.7350 

Random effects:
 Groups   Name        Variance Std.Dev.
 Country  (Intercept) 0.11623  0.3409  
 Residual             0.08337  0.2887  
Number of obs: 3455, groups:  Country, 222

Fixed effects:
                   Estimate Std. Error t value Pr(>|z|)    
(Intercept)        1.960292   0.066106  29.654   <2e-16 ***
SE.XPD.TOTL.GD.ZS -0.004539   0.005881  -0.772     0.44    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
SE.XPD.TOTL -0.383
```

```r
#Random Intercept and Random Slope Model

model_random_slope <- glmer(SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (SE.XPD.TOTL.GD.ZS | Cou
summary(model_random_slope)
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: Gamma  ( log )
Formula: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (SE.XPD.TOTL.GD.ZS |  
    Country)
   Data: data

     AIC      BIC   logLik deviance df.resid 
 13623.8  13660.7  -6805.9  13611.8     3449 

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.5258 -0.3837 -0.0354  0.3647  7.5228 

Random effects:
 Groups   Name              Variance Std.Dev. Corr
```

```
 Country  (Intercept)       0.62344  0.7896
          SE.XPD.TOTL.GD.ZS 0.02997  0.1731   -0.89
 Residual                   0.06237  0.2497
Number of obs: 3455, groups:  Country, 222

Fixed effects:
                  Estimate Std. Error t value Pr(>|z|)
(Intercept)        1.85762    0.12145  15.296   <2e-16 ***
SE.XPD.TOTL.GD.ZS  0.01966    0.02526   0.779    0.436
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
SE.XPD.TOTL -0.828
```

```r
anova(model_random_intercept, model_random_slope)
```

```
Data: data
Models:
model_random_intercept: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (1 | Country)
model_random_slope: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS + (SE.XPD.TOTL.GD.ZS | Country)
                       npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
model_random_intercept    4 14153 14178 -7072.4    14145
model_random_slope        6 13624 13661 -6805.9    13612 533.1  2  < 2.2e-16

model_random_intercept
model_random_slope     ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from the random slope model show a higher variance in the random intercept (0.623) and a non-negligible variance for the slope (0.030), along with a strong negative correlation (-0.89) between them. This suggests that countries with higher baseline unemployment tend to experience weaker (or even negative) impacts from increased education spending. Despite the fixed effect of education expenditure still being statistically insignificant (estimate = 0.01966, p = 0.436), the model's AIC significantly dropped to 13,623.8 from 14,152.9 in the random intercept model. The deviance also decreased, indicating better model fit. Furthermore, the likelihood ratio test comparing the two models yielded a chi-square statistic of 533.1 with 2 degrees of freedom and a p-value of < 2.2e-16, strongly favoring the more complex random slope model.

The random slope model significantly improves upon the random intercept model by capturing the heterogeneity in how countries respond to education spending in relation to female unemployment. This shows that global policy recommendations should not assume uniform effects of educational investment. Instead, policies must be contextualized, acknowledging that some countries may benefit more from increased education spending in terms of reducing unemployment, while others may not see the same effect. The inclusion of both random intercepts and slopes thus provides a more realistic representation.

Analyzing Interactions using ANOVA

To further investigate the factors influencing female unemployment as we realized the slope varied for each country, we extended our analysis by introducing interaction effects between government expenditure on education and several macroeconomic variables. These included tax revenue, broadband internet access, female labor force participation, inflation, market capitalization, urban population, interest rates, private sector credit, and technology exports. The extended model maintained the random slope structure, allowing for country-level variations in the effect of education expenditure while exploring how this effect might be moderated by other variables.

```r
colnames(facttable_wide)[colnames(facttable_wide) == "Country Code"] <- "Country"
lmm_extended <- lmer(SL.UEM.TOTL.FE.ZS ~
                     SE.XPD.TOTL.GD.ZS * (GC.TAX.TOTL.GD.ZS +
                                          IT.NET.BBND.P2 + SL.TLF.CACT.FE.ZS +
                                          FP.CPI.TOTL.ZG + CM.MKT.LCAP.GD.ZS + EN.URB.LCTY.UI
                     (SE.XPD.TOTL.GD.ZS | Country),
                     data = facttable_wide, REML = FALSE)
summary(lmm_extended)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: SL.UEM.TOTL.FE.ZS ~ SE.XPD.TOTL.GD.ZS * (GC.TAX.TOTL.GD.ZS +
    IT.NET.BBND.P2 + SL.TLF.CACT.FE.ZS + FP.CPI.TOTL.ZG + CM.MKT.LCAP.GD.ZS +
    EN.URB.LCTY.UR.ZS + FR.INR.LNDP + FS.AST.PRVT.GD.ZS + TX.VAL.TECH.MF.ZS) +
    (SE.XPD.TOTL.GD.ZS | Country)
   Data: facttable_wide

     AIC      BIC   logLik deviance df.resid
  2349.4   2453.2  -1150.7   2301.4      535

Scaled residuals:
    Min      1Q  Median      3Q     Max
-6.5500 -0.2857 -0.0505  0.2490  6.8868
```

```
Random effects:
 Groups   Name              Variance Std.Dev. Corr
 Country  (Intercept)       60.444   7.775
          SE.XPD.TOTL.GD.ZS  1.770   1.330    -0.77
 Residual                    2.014   1.419
Number of obs: 559, groups:  Country, 63

Fixed effects:
                                     Estimate Std. Error t value
(Intercept)                         24.313881   5.389402   4.511
SE.XPD.TOTL.GD.ZS                   -2.100910   1.166803  -1.801
GC.TAX.TOTL.GD.ZS                   -0.216868   0.215857  -1.005
IT.NET.BBND.P2                      -0.357473   0.100516  -3.556
SL.TLF.CACT.FE.ZS                   -0.245991   0.092217  -2.668
FP.CPI.TOTL.ZG                      -0.106770   0.095274  -1.121
CM.MKT.LCAP.GD.ZS                   -0.024905   0.014730  -1.691
EN.URB.LCTY.UR.ZS                   -0.004688   0.085046  -0.055
FR.INR.LNDP                          0.044044   0.191860   0.230
FS.AST.PRVT.GD.ZS                    0.041274   0.030879   1.337
TX.VAL.TECH.MF.ZS                    0.031067   0.021882   1.420
SE.XPD.TOTL.GD.ZS:GC.TAX.TOTL.GD.ZS  0.096294   0.044845   2.147
SE.XPD.TOTL.GD.ZS:IT.NET.BBND.P2     0.081655   0.022749   3.589
SE.XPD.TOTL.GD.ZS:SL.TLF.CACT.FE.ZS -0.001452   0.020388  -0.071
SE.XPD.TOTL.GD.ZS:FP.CPI.TOTL.ZG     0.011771   0.019393   0.607
SE.XPD.TOTL.GD.ZS:CM.MKT.LCAP.GD.ZS  0.006268   0.003312   1.893
SE.XPD.TOTL.GD.ZS:EN.URB.LCTY.UR.ZS  0.010126   0.020642   0.491
SE.XPD.TOTL.GD.ZS:FR.INR.LNDP        0.008056   0.033730   0.239
SE.XPD.TOTL.GD.ZS:FS.AST.PRVT.GD.ZS -0.007027   0.007092  -0.991
SE.XPD.TOTL.GD.ZS:TX.VAL.TECH.MF.ZS -0.010576   0.007578  -1.396


Correlation matrix not shown by default, as p = 20 > 12.
Use print(x, correlation=TRUE)  or
    vcov(x)        if you need it
```

```r
library(car)
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following object is masked from 'package:dplyr':

    recode


The following object is masked from 'package:purrr':

    some
```

```r
Anova(lmm_extended, type = "III")
```

```
Analysis of Deviance Table (Type III Wald chisquare tests)

Response: SL.UEM.TOTL.FE.ZS
                                      Chisq Df Pr(>Chisq)    
(Intercept)                         20.3530  1  6.439e-06 ***
SE.XPD.TOTL.GD.ZS                    3.2420  1  0.0717709 .  
GC.TAX.TOTL.GD.ZS                    1.0094  1  0.3150499    
IT.NET.BBND.P2                      12.6478  1  0.0003760 ***
SL.TLF.CACT.FE.ZS                    7.1157  1  0.0076411 ** 
FP.CPI.TOTL.ZG                       1.2559  1  0.2624320    
CM.MKT.LCAP.GD.ZS                    2.8586  1  0.0908838 .  
EN.URB.LCTY.UR.ZS                    0.0030  1  0.9560429    
FR.INR.LNDP                          0.0527  1  0.8184300    
FS.AST.PRVT.GD.ZS                    1.7866  1  0.1813419    
TX.VAL.TECH.MF.ZS                    2.0156  1  0.1556859    
SE.XPD.TOTL.GD.ZS:GC.TAX.TOTL.GD.ZS  4.6107  1  0.0317727 *  
SE.XPD.TOTL.GD.ZS:IT.NET.BBND.P2    12.8833  1  0.0003315 ***
SE.XPD.TOTL.GD.ZS:SL.TLF.CACT.FE.ZS  0.0051  1  0.9432279    
SE.XPD.TOTL.GD.ZS:FP.CPI.TOTL.ZG     0.3684  1  0.5438640    
SE.XPD.TOTL.GD.ZS:CM.MKT.LCAP.GD.ZS  3.5822  1  0.0584008 .  
SE.XPD.TOTL.GD.ZS:EN.URB.LCTY.UR.ZS  0.2407  1  0.6237260    
SE.XPD.TOTL.GD.ZS:FR.INR.LNDP        0.0570  1  0.8112273    
SE.XPD.TOTL.GD.ZS:FS.AST.PRVT.GD.ZS  0.9818  1  0.3217470    
SE.XPD.TOTL.GD.ZS:TX.VAL.TECH.MF.ZS  1.9481  1  0.1627887    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from the Anova table reveal that a few key variables and their interactions with education expenditure have a statistically significant relationship with female unemployment. Specifically, broadband internet access (IT.NET.BBND.P2) and female labor force participation (SL.TLF.CACT.FE.ZS) had significant main effects, suggesting that greater connectivity and workforce participation are associated with changes in unemployment rates. Also, the interactions between education expenditure and tax revenue (GC.TAX.TOTL.GD.ZS) as well as broadband access were significant. Indicating that the effect of education expenditure is amplified or diminished depending on these moderating factors.

Interpretation of Significant Variables

- Broadband Access Strengthens the Effect of Education Spending: The interaction between government education expenditure and broadband internet access (SE.XPD.TOTL.GD.ZS:IT.NET.BBND.P2, $p < 0.001$) was highly significant. Indicating that in countries with better broadband penetration, spending on education is more effective in reducing female unemployment. Fast and reliable internet enables individuals to gain access to digital education resources, acquire tech skills, and find remote jobs. As a result, these countries may see a stronger return on education investment in terms of employability and job creation.
- Tax Revenue Plays a Crucial Role: The interaction between education expenditure and tax revenue (SE.XPD.TOTL.GD.ZS:GC.TAX.TOTL.GD.ZS, $p = 0.031$) was also statistically significant. This suggests that the effectiveness of government education spending on reducing female unemployment is amplified in countries with higher tax revenue. A well-funded government is better positioned to support the education system through scholarships, vocational and technical training, and job placement services. These two investments help bridge the gap between education and job opportunities.
- Female Labor Force Participation is Key: The main effect of female labor force participation (SL.TLF.CACT.FE.ZS, $p = 0.008$) was significant. Countries where more women are already active in the workforce tend to experience lower unemployment rates among females. Suggesting a reinforcing cycle, where higher participation provides more role models, encourages gender-inclusive policies, and creates a cultural norm that supports female employment. Therefore, encouraging female workforce engagement through policies such as childcare support, anti-discrimination laws, and maternity leave can further reduce unemployment.
- Broadband Access is a Standalone Predictor of Employment Outcomes: Aside from its interaction effect, broadband access on its own (IT.NET.BBND.P2, $p < 0.001$) was also a significant predictor. Even without considering education spending, greater access to broadband seems to correlate with reduced female unemployment. This reinforces the importance of digital infrastructure in modern job markets, where access to online platforms, remote work, and digital skill development plays a central role in employability.

To conclude, the inclusion of interaction terms provided deeper insights into the complex dynamics at play. While education expenditure alone did not significantly impact female unemployment, its effectiveness is conditioned by other variables. These findings show that investing in education must be followed by improvements in technology access, public finance systems, and labor market participation to have results. This understanding will guide more targeted policy recommendations for reducing female unemployment.

Research Question 2: How does R&D expenditure relate to high-tech exports across countries?

Innovation-driven economies often rely on investments in Research and Development (R&D) to boost their technological advancement. This analysis explores whether increased R&D expenditure by governments is associated with higher levels of high-tech exports across various countries.

Exploratory Data Analysis

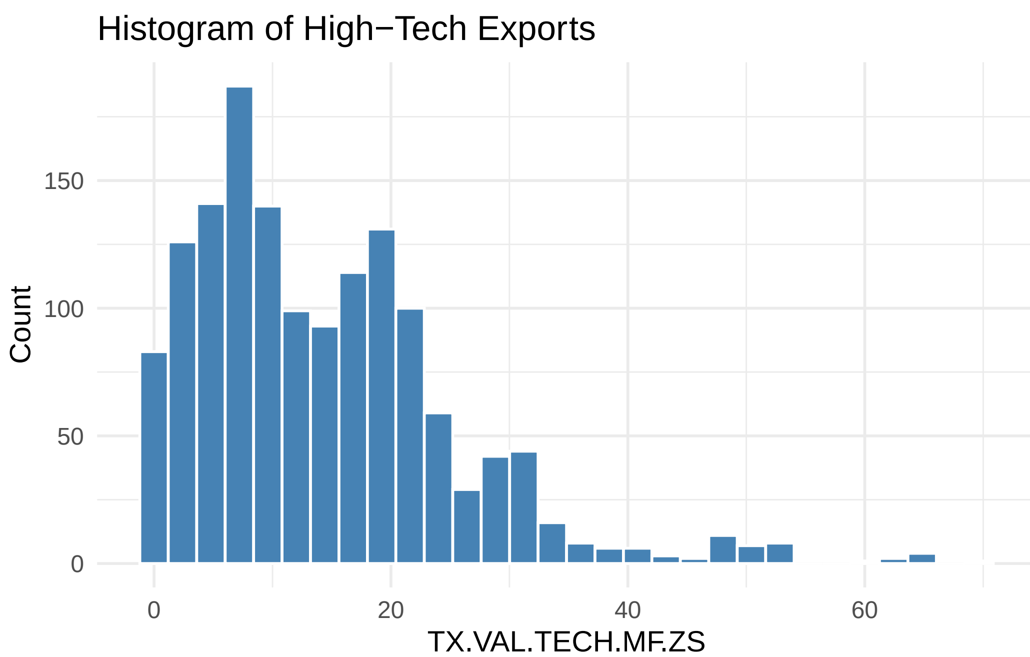
Initial visualizations and summaries of the response variable (Research and Development) showed that it was mostly positive and continuous, which suggested that a Gamma distribution could be appropriate for modeling. However, a few data points were negative, violating the assumptions of the Gamma distribution. Due to this, we proceeded with either a standard linear model or a generalized linear model (GLM) with a Gaussian distribution, both of which can accommodate the full range of observed values, including negatives.

```
facttable<-read_csv("./facttable.csv")
```

```
Rows: 8778 Columns: 24
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): Country Code, Indicator
dbl (22): 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
facttable_wide <- facttable %>%
  pivot_longer(cols = 3:24, names_to = "year", values_to = "value") %>%
  pivot_wider(names_from = Indicator, values_from = value)
data <- facttable_wide %>%
  select("Country Code", year, GB.XPD.RSDV.GD.ZS, TX.VAL.TECH.MF.ZS) %>%
  na.omit()
```

```
source("./histogram.R")
data %>%
  filter(TX.VAL.TECH.MF.ZS < 200) %>%
  drop_na() %>%
  plot_histogram("TX.VAL.TECH.MF.ZS","Histogram of High-Tech Exports")
```

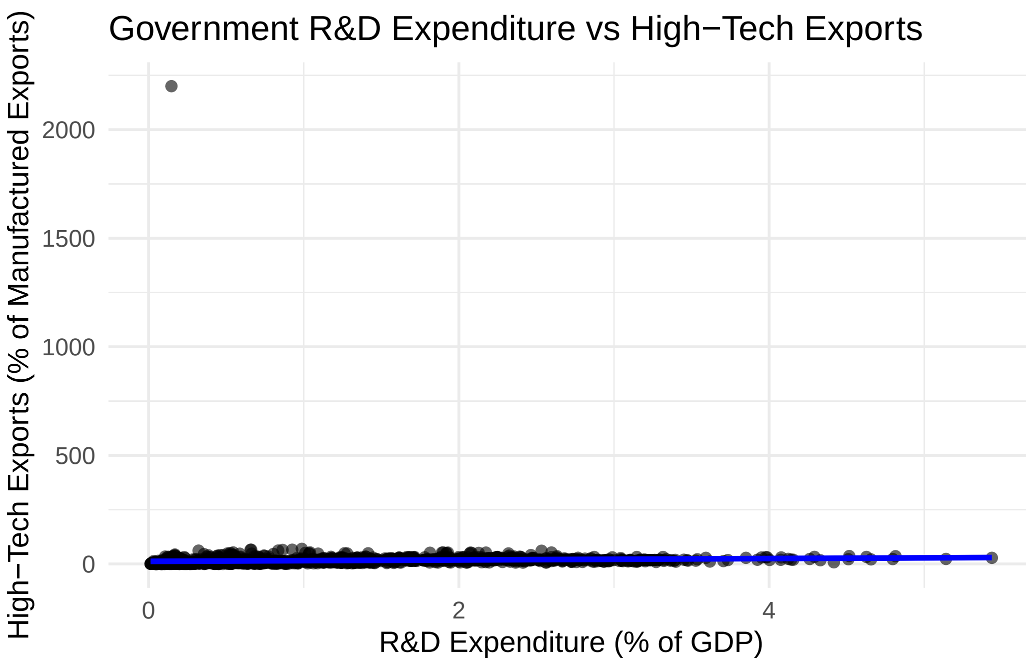


Linear Regression

To explore how Research and Development (R&D) expenditure relates to high-technology exports across countries, we fit a simple linear regression model where the response variable is the percentage of high-technology exports (TX.VAL.TECH.MF.ZS), and the predictor is R&D expenditure as a percentage of GDP.

```
ggplot(data, aes(x = GB.XPD.RSDV.GD.ZS, y = TX.VAL.TECH.MF.ZS)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Government R&D Expenditure vs High-Tech Exports",
       x = "R&D Expenditure (% of GDP)",
       y = "High-Tech Exports (% of Manufactured Exports)") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

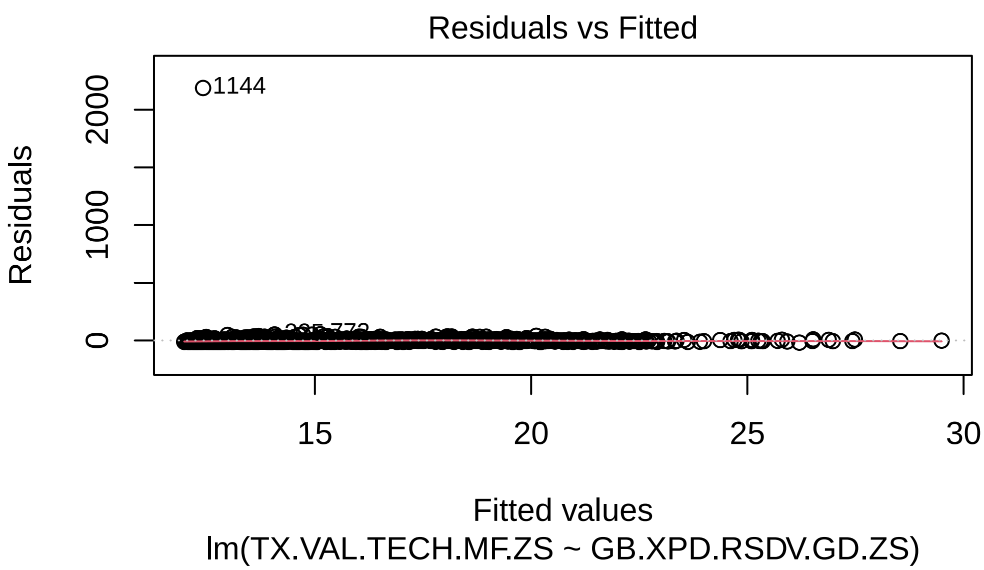


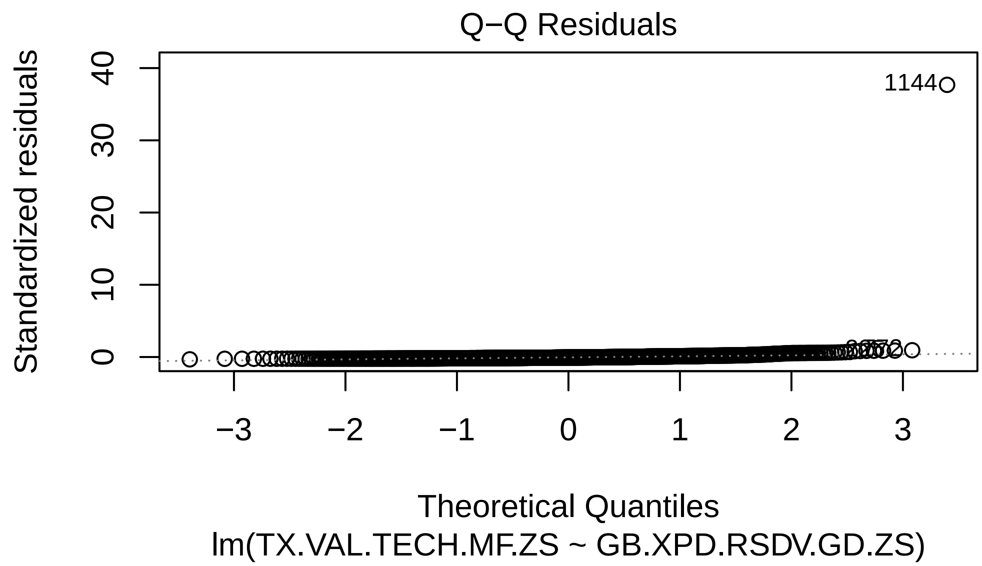

```r
lm_model <- lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS, data = data)
summary(lm_model)
```

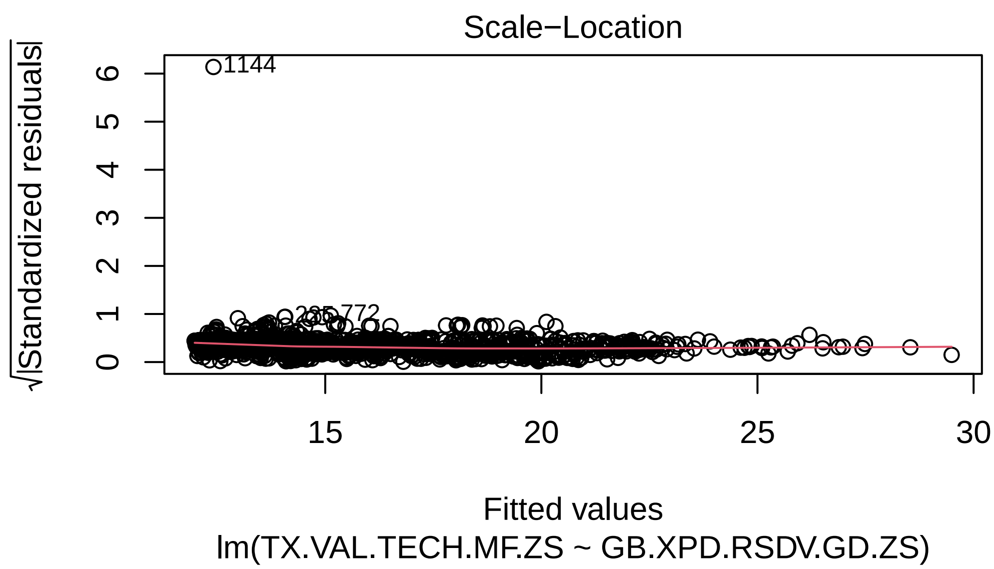
```
Call:
lm(formula = TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
 -18.58   -8.06   -3.98    2.84 2187.83 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         11.939      2.399   4.976 7.26e-07 ***
GB.XPD.RSDV.GD.ZS    3.229      1.550   2.083   0.0374 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.1 on 1462 degrees of freedom
Multiple R-squared:  0.00296,   Adjusted R-squared:  0.002278 
F-statistic:  4.34 on 1 and 1462 DF,  p-value: 0.0374
```

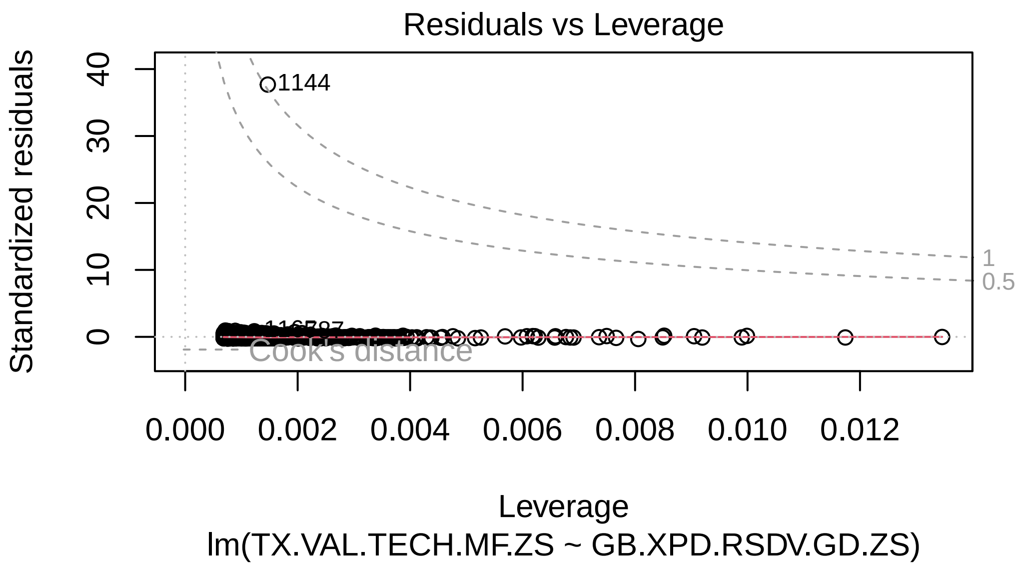
The estimated coefficient for R&D expenditure is 3.23, with a p-value of 0.0374, suggesting that the relationship is statistically significant at the 5% level. This implies that, on average, a 1% increase in R&D expenditure (as a share of GDP) is associated with an approximate 3.23 percentage point increase in high-technology exports, holding other factors constant. Despite the statistical significance, the R-squared value is very low ($R^2 = 0.00296$), indicating that R&D expenditure alone explains less than 0.3% of the variation in high-tech exports across countries. This suggests other factors may be influencing high-tech export levels, and a more complex model may be required to improve explanatory power.

```r
plot(lm_model)
```




Q−Q Residuals
Standardized residuals
0
10
20
30
40
1144
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS)


Scale−Location
√|Standardized residuals|
0
1
2
3
4
5
6
1144
772
15
20
25
30
Fitted values
lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS)



The figure above presents the model diagnostic plots. From both the Residuals vs Fitted and Q-Q plots, it is evident that point 1144 is an outlier. This observation is further supported by the Residuals vs Leverage plot, where point 1144 also appears within the Cook's distance boundary, indicating high influence on the model.

Linear Regression without Outliers

```r
facttable<-read_csv("./facttable.csv")
```

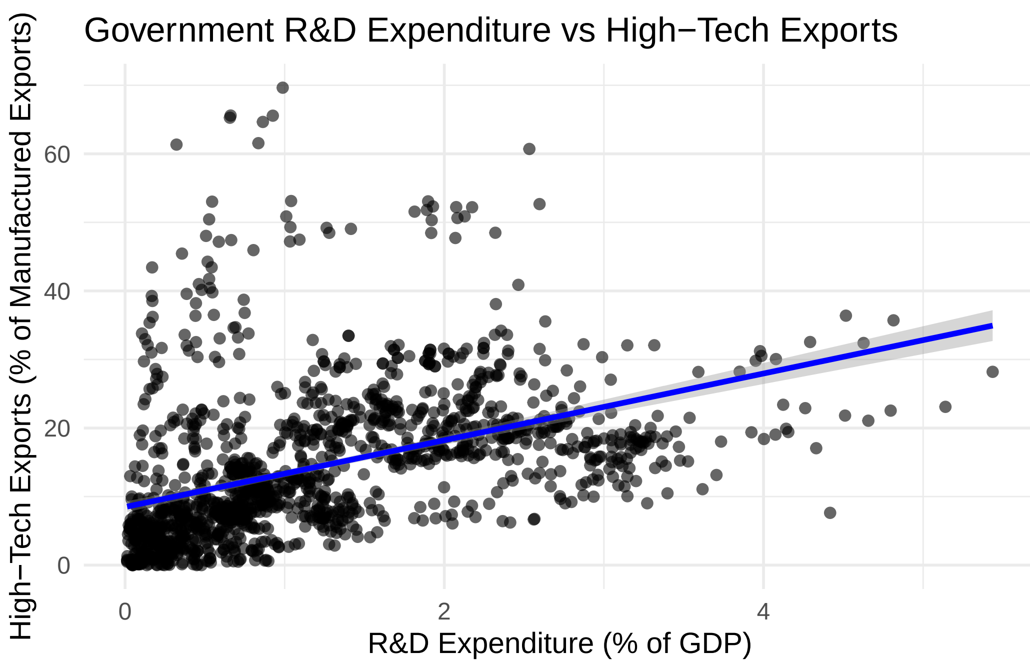
```
Rows: 8778 Columns: 24
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): Country Code, Indicator
dbl (22): 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
facttable_wide <- facttable %>%
  pivot_longer(cols = 3:24, names_to = "year", values_to = "value") %>%
  pivot_wider(names_from = Indicator, values_from = value)
```

```r
# taking out outliers
data <- facttable_wide %>%
  select("Country Code", year, GB.XPD.RSDV.GD.ZS, TX.VAL.TECH.MF.ZS) %>%
  filter(TX.VAL.TECH.MF.ZS < 500) %>%
  na.omit()
```

```r
ggplot(data, aes(x = GB.XPD.RSDV.GD.ZS, y = TX.VAL.TECH.MF.ZS)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "Government R&D Expenditure vs High-Tech Exports",
       x = "R&D Expenditure (% of GDP)",
       y = "High-Tech Exports (% of Manufactured Exports)") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

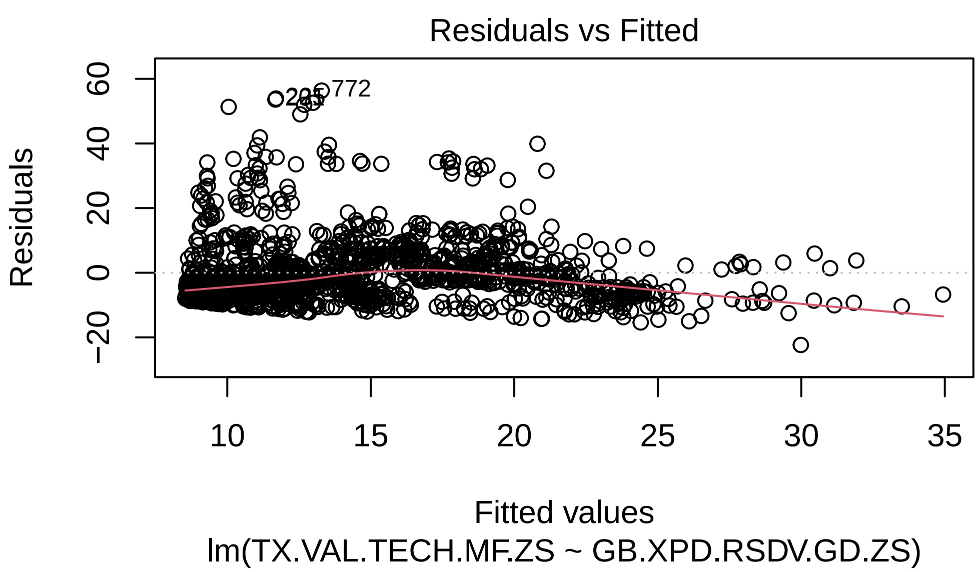


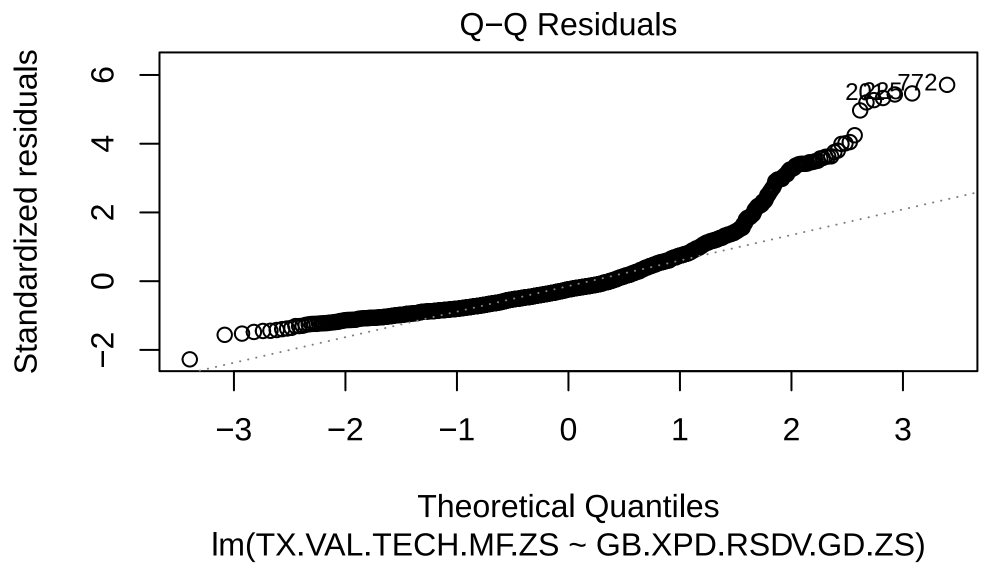

```r
lm_model <- lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS, data = data)
summary(lm_model)
```

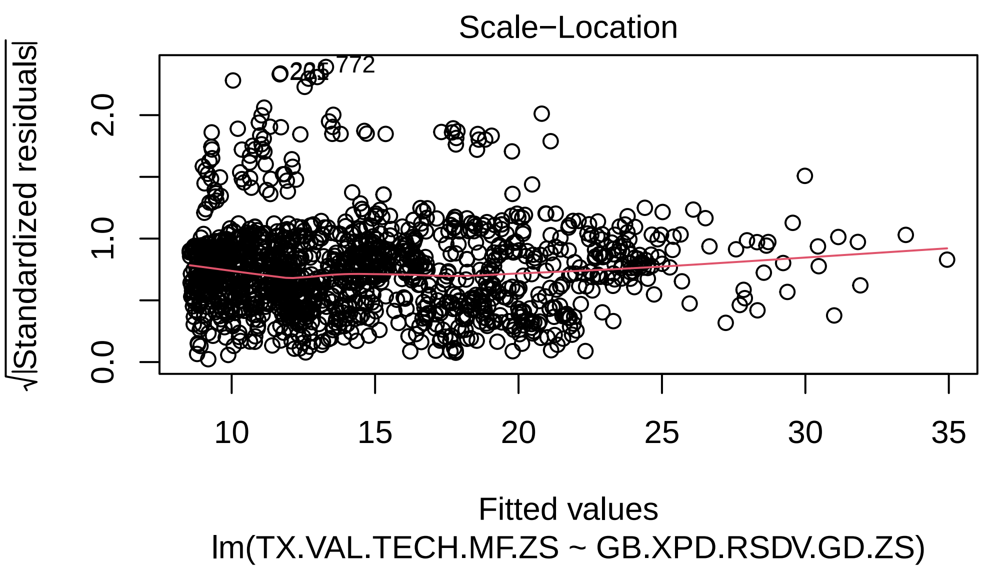
```
Call:
```

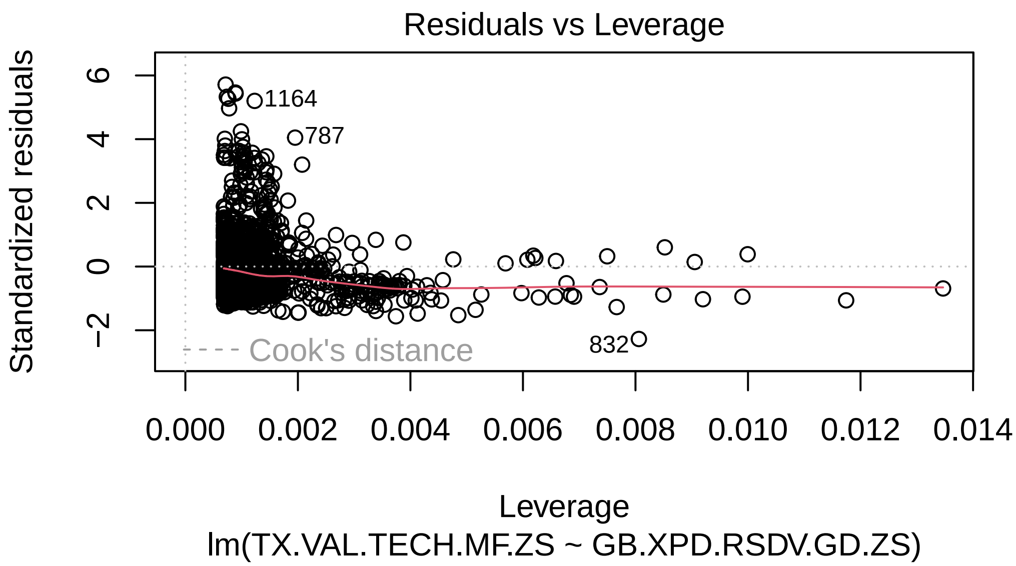
```
lm(formula = TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.357  -6.350  -2.342   3.540  56.360 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         8.4783     0.4078   20.79   <2e-16 ***
GB.XPD.RSDV.GD.ZS   4.8683     0.2634   18.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.868 on 1461 degrees of freedom
Multiple R-squared:  0.1895,	Adjusted R-squared:  0.189 
F-statistic: 341.7 on 1 and 1461 DF,  p-value: < 2.2e-16
```

```r
plot(lm_model)
```




Q−Q Residuals
Standardized residuals
−2
0
2
4
6
2025
2025
772
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS)


Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
2025
772
10
15
20
25
30
35
Fitted values
lm(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS)



Refitting the model without the outlier results in a substantial improvement. The Adjusted R-squared increases from 0.002278 to 0.189, indicating a better fit. Additionally, the p-value for the independent variable becomes significantly smaller, suggesting stronger statistical evidence of its effect. All subsequent analyses were conducted without the outlier.

Generalized Linear Mixed Model (GLMM) – Random Intercept

We fitted a Linear Mixed Model (LMM) with random intercepts for each country. This approach allows us to capture unobserved heterogeneity among countries that may influence export performance independently of R&D spending.

```r
colnames(data)[colnames(data) == "Country Code"] <- "Country"
lmm_model <- glmer(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country),  data = data)
```

```
Warning in glmer(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country), :
calling glmer() with family=gaussian (identity link) as a shortcut to lmer() is
deprecated; please call lmer() directly
```

```r
summary(lmm_model)
```

```
Linear mixed model fit by REML ['lmerMod']
```

```
Formula: TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country)
   Data: data

REML criterion at convergence: 9077.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.8034 -0.2789 -0.0505  0.1913 10.6700

Random effects:
 Groups   Name        Variance Std.Dev.
 Country  (Intercept) 84.99    9.219
 Residual             20.04    4.477
Number of obs: 1463, groups:  Country, 160

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         8.3618     0.8970   9.322
GB.XPD.RSDV.GD.ZS   3.8028     0.5161   7.369

Correlation of Fixed Effects:
            (Intr)
GB.XPD.RSDV -0.550
```

The model output showed that the fixed effect estimate for government R&D expenditure was 3.80 with a standard error of 0.5161, resulting in a t-value of approximately 7.369. While this coefficient is slightly larger than that from the simple linear model, it comes with greater uncertainty. The random intercept for countries had an estimated variance of 84.99, corresponding to a standard deviation of 9.219, so baseline levels of high-technology exports vary considerably from one country to another. The residual variance was 20.04, with a standard deviation of 4.477, like what was observed in the linear model.

Comparing the Linear Model to the LMM

```r
AIC(lm_model, lmm_model)
```

```
          df       AIC
lm_model   3 10854.382
lmm_model  4  9085.301
```

To determine whether the mixed-effects model offers a better fit than the simple linear regression, we compared their AIC values. The linear model had an AIC of 10854.38, while the mixed-effects model achieved a slightly lower AIC of 9085.30. This comparison supports the inclusion of random effects for country. Including these effects allows the model to better account for differences in national contexts, policies, and infrastructure that may influence the effectiveness of R&D investment on export outcomes.

The linear mixed model not only aligns with the linear model's finding that R&D expenditure positively affects high-tech exports but also provides a better representation of the data by accounting for country-level variability. This suggests that national characteristics play a meaningful role in shaping the impact of R&D investment.

Generalized Linear Mixed Model (GLMM) – Random Intercept and Random Slope

To further account for the nested structure of the data, we implemented Generalized Linear Mixed Models (GLMMs). This model helps us to gain a better understanding of the relationship between R&D expenditure and high-tech exports by using fixed and random effects.

```r
model_random_intercept <- lmer(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country), data =
summary(model_random_intercept)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country)
   Data: data

REML criterion at convergence: 9077.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.8034 -0.2789 -0.0505  0.1913 10.6700

Random effects:
 Groups   Name        Variance Std.Dev.
 Country  (Intercept) 84.99    9.219
 Residual             20.04    4.477
Number of obs: 1463, groups:  Country, 160

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         8.3618     0.8970   9.322
GB.XPD.RSDV.GD.ZS   3.8028     0.5161   7.369
```

```
Correation of Fixed Effects:
            (Intr)
GB.XPD.RSDV -0.550
```

```r
model_random_slope <- lmer(TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (GB.XPD.RSDV.GD.ZS | Coun
summary(model_random_slope)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (GB.XPD.RSDV.GD.ZS |
    Country)
   Data: data

REML criterion at convergence: 9017.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.0565 -0.2620 -0.0506  0.2039 11.5948

Random effects:
 Groups   Name              Variance Std.Dev. Corr
 Country  (Intercept)       232.22   15.239
          GB.XPD.RSDV.GD.ZS 333.68   18.267   -0.76
 Residual                    15.72    3.965
Number of obs: 1463, groups:  Country, 160

Fixed effects:
                  Estimate Std. Error t value
(Intercept)          8.138      1.387   5.866
GB.XPD.RSDV.GD.ZS    4.847      1.843   2.630

Correlation of Fixed Effects:
            (Intr)
GB.XPD.RSDV -0.749
```

```r
anova(model_random_intercept, model_random_slope)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: data
Models:
```

```
model_random_intercept: TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (1 | Country)
model_random_slope: TX.VAL.TECH.MF.ZS ~ GB.XPD.RSDV.GD.ZS + (GB.XPD.RSDV.GD.ZS | Country)
                       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model_random_intercept    4 9087.1 9108.2 -4539.5   9079.1
model_random_slope        6 9034.5 9066.3 -4511.3   9022.5 56.523  2  5.322e-13

model_random_intercept
model_random_slope     ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We began with a random intercept model, which assumes that each country has a different baseline level of high-tech exports, but the effect of government R&D expenditure on exports is the same across all countries. The model revealed that R&D expenditure (GB.XPD.RSDV.GD.ZS) had a positive fixed effect on high-tech exports (TX.VAL.TECH.MF.ZS), with an estimate of 3.80 ($t = 7.37$). The variance of the random intercept was 84.99, and the residual variance was 20.04, indicating large variation in baseline high-tech exports across countries. However, this model assumes a uniform slope for all countries.

To address this, we extended the model to a random intercept and random slope model. This model acknowledges that not only do countries start at different levels of high-tech exports, but the effectiveness of R&D expenditure may be different. The results showed improved model fit, with the AIC decreasing from 9087.1 to 9034.5 and the BIC decreasing from 9108.2 to 9066.3. The log-likelihood increased from -4539.5 to -4511.3, and the residual variance slightly decreased. The fixed effect of R&D expenditure remained positive (estimate = 4.847), and importantly, the variance of the random slope was 91.6. This suggests substantial country-level variability in the impact of R&D expenditure on exports. The negative correlation (-0.76) between the random intercept and slope also suggests that countries with lower baseline export levels may benefit more from increased R&D investment, and vice versa. This indicates that allowing the effect of R&D expenditure to vary across countries leads to a significantly better representation of the data. Therefore, the random slope model is preferred.

To conclude, these findings highlight the heterogeneity in how R&D expenditure translates into high-tech export outcomes across countries. While investment in R&D is generally beneficial, its effectiveness is influenced by other country specific factors. Future analysis could explore these moderating factors to better understand why the impact of R&D spending differs and to inform more tailored national strategies for promoting high-tech exports.

Conclusion

This study explored two key questions: how government education spending affects female unemployment, and how R&D expenditure influences high-tech exports across countries. Using the global World Bank data and modeling techniques learnt this semester, we found that these relationships are complex and highly context dependent.

For education spending, simple models suggested no significant link to lower female unemployment. However, when accounting for country level differences using mixed models, we found that the impact of education spending varies widely. Factors like broadband access and tax revenue significantly influence outcomes, showing that education investment alone is insufficient without supporting infrastructure and policies.

R&D spending showed a positive but modest relationship with high-tech exports. Mixed models revealed that countries benefit differently. Those with lower baseline exports often gain more from R&D investments, again showing the importance of national context.

The findings highlight that development strategies must be tailored to individual countries. Good policies require a supportive ecosystem that includes technology, fiscal strength, and labor market readiness for investments to be effective.

Appendix

Appendix A

Code, Indicator Description

Indicator Code	Indicator Description
CM.MKT.LCAP.GD.ZS	Market capitalization of listed domestic companies (% of GDP)
EG.IMP.CONS.ZS	Energy imports, net (% of energy use)
EG.USE.ELEC.KH.PC	Electric power consumption (kWh per capita)
EN.URB.LCTY.UR.ZS	Population in the largest city (% of urban population)
FP.CPI.TOTL.ZG	Inflation, consumer prices (annual %)
FR.INR.LNDP	Interest rate spread (lending rate minus deposit rate, %)
FS.AST.PRVT.GD.ZS	Domestic credit to private sector (% of GDP)
GB.XPD.RSDV.GD.ZS	Research and development expenditure (% of GDP)
GC.NLD.TOTL.GD.ZS	Net lending (+) / net borrowing (-) (% of GDP)
GC.TAX.TOTL.GD.ZS	Tax revenue (% of GDP)
GC.XPN.TOTL.GD.ZS	Expense (% of GDP)
IC.BUS.DISC.XQ	Business extent of disclosure index (0 = less disclosure to 10 = more disclosure)
IC.CRD.INFO.XQ	Depth of credit information index (0 = low to 8 = high)
IC.GOV.DURS.ZS	Time spent dealing with the requirements of government regulations (% of senior management time)
IC.LGL.CRED.XQ	Strength of legal rights index (0 = weak to 12 = strong)
IC.REG.DURS.FE	Time required to start a business, female (days)
IC.REG.DURS.MA	Time required to start a business, male (days)
IC.TAX.TOTL.CP.ZS	Total tax and contribution rate (% of profit)
IT.CEL.SETS.P2	Mobile cellular subscriptions (per 100 people)
IT.MLT.MAIN.P2	Fixed telephone subscriptions (per 100 people)
IT.NET.BBND.P2	Fixed broadband subscriptions (per 100 people)
LP.LPI.OVRL.XQ	Logistics performance index: Overall (1 = low to 5 = high)
MS.MIL.XPND.GD.ZS	Military expenditure (% of GDP)
NY.GDP.MKTP.KD.ZG	GDP growth (annual %)
SE.PRM.TCAQ.ZS	Trained teachers in primary education (% of total teachers)
SE.XPD.TOTL.GD.ZS	Government expenditure on education, total (% of GDP)
SI.POV.GINI	GINI index (World Bank estimate)
SI.POV.URHC	Urban poverty headcount ratio at national poverty lines (% of urban population)
SL.TLF.CACT.FE.ZS	Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)
SL.TLF.CACT.MA.ZS	Labor force participation rate, male (% of male population ages 15+) (modeled ILO estimate)

Indicator Code	Indicator Description
SL.UEM.TOTL.FE.ZS	Unemployment, female (% of female labor force) (modeled ILO estimate)
SL.UEM.TOTL.MA.ZS	Unemployment, male (% of male labor force) (modeled ILO estimate)
SP.URB.TOTL.IN.ZS	Urban population (% of total population)
TX.VAL.TECH.MF.ZS	High-technology exports (% of manufactured exports)

References

McCulloch, C. E. (1996). AN INTRODUCTION TO GENERALIZED LINEAR MIXED MODELS. Conference on Applied Statistics in Agriculture. https://doi.org/10.4148/2475-7772.1314

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370. https://doi.org/10.2307/2344614