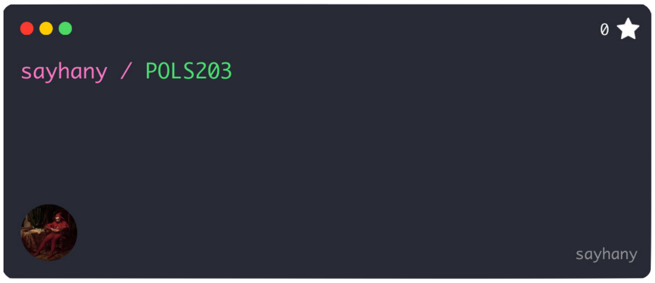
Group L Second Project

Emirhan Yücel 2020302264

Mübin Salih Sarıçiçek 2020302243

Sayhan Yalvaçer 2019202063



Dataset

All data is retrieved from "Our World in Data".

Research question:

What is the relationship between the EU membership and GDP per capita growth during the period 2004 - 2014?

Strategy:

1. We first subsetted the former Eastern Bloc countries and divided them into two groups according to whether they joined the EU in 2004, as many of them did.
2. We omitted some countries either because they joined the union later or underwent some major economic crises during the specified period.
3. Then we proceeded to compare their mean values of growth and applied a t-test in order to test the hypothesis and calculate a confidence interval.
4. In the second part of the project, we extendedour sample to all European countries and added many new independent variables. We built several models with differing numbers of IVs, then, we described their Model metrics and compared them according to *Akaike information criterion* to ultimately decide which model we are going to choose.
5. The most important question to answer was whether the best model will include EU membership as an IV.

The Code

Load the required packages:

```r
if (!require(tidyverse)) install.packages('tidyverse')
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidyverse)

if (!require(lmtest)) install.packages('lmtest')
```

```
## Loading required package: lmtest
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(lmtest) # For Breusch-Pagan test

if (!require(ggthemes)) install.packages('ggthemes')
```

```
## Loading required package: ggthemes
```

```r
library(ggthemes) # Themes for ggplot2

if (!require(broom)) install.packages('broom')
```

```
## Loading required package: broom
```

```r
library(broom) # Extracting model metrics

if (!require(ggfortify)) install.packages('ggfortify')
```

```
## Loading required package: ggfortify
```

```r
library(ggfortify) # Visualizing model metrics

if (!require(MASS)) install.packages('MASS')
```

```
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(MASS) # For stepAIC

if (!require(simputation)) install.packages('simputation')
```

```
## Loading required package: simputation
```

```r
library(simputation) # Simple linear imputation

if (!require(missForest)) install.packages('missForest')
```

```
## Loading required package: missForest
```

```r
library(missForest) # Imputation by using random forest algorithm

if (!require(car)) install.packages('car')
```

```
## Loading required package: 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
library(car) # For vif

if (!require(corrplot)) install.packages('corrplot')
```

```
## Loading required package: corrplot
## corrplot 0.92 loaded
```

```r
library(corrplot) # For plotting correlation matrices
```

Disable scientific notation

```r
options(scipen = 999)
```

Set seed for the reproducibility of imputation results

```r
set.seed(203)
```

Read the dataset using readr

```r
pols_203_final_merged <- read_csv("pols_203_final_merged.csv")
```

```
## Rows: 121078 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): id, Entity, Code
## dbl (14): Year, Total dependency ratio - Sex: all - Age: none - Variant: est...
##
## 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.
```

Explore the structure of the dataset

Our dataset contains 121078 rows and 17 columns

```r
dim(pols_203_final_merged)
```

```
## [1] 121078     17
```

Show the names of the columns

```r
colnames(pols_203_final_merged)
```

```
##  [1] "id"
##  [2] "Entity"
##  [3] "Code"
##  [4] "Year"
##  [5] "Total dependency ratio - Sex: all - Age: none - Variant: estimates"
##  [6] "output_quantity"
##  [7] "Government expenditure on tertiary education as % of GDP (%)"
##  [8] "International tourism, number of arrivals"
##  [9] "Top marginal income tax rate (Reynolds (2008))"
## [10] "Oil production per capita (kWh)"
## [11] "particip_vdem_owid"
## [12] "particip_vdem_high_owid"
## [13] "particip_vdem_low_owid"
## [14] "Per capita electricity (kWh)"
## [15] "GDP per capita (output, multiple price benchmarks)"
## [16] "Time required to start a business (days)"
## [17] "Population"
```

Classes of the columns

```
class(pols_203_final_merged$id)
```

```
## [1] "character"
```

```
class(pols_203_final_merged$Entity)
```

```
## [1] "character"
```

```
class(pols_203_final_merged$Code)
```

```
## [1] "character"
```

```
class(pols_203_final_merged$`Total dependency ratio - Sex: all - Age: none - Variant: estimates`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$output_quantity)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`Government expenditure on tertiary education as % of GDP (%)`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`International tourism, number of arrivals`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`Top marginal income tax rate (Reynolds (2008))`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`Oil production per capita (kWh)`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$particip_vdem_owid)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`Per capita electricity (kWh)`)
```

```
## [1] "numeric"
```

```
class(pols_203_final_merged$`GDP per capita (output, multiple price benchmarks)`)
```

```
## [1] "numeric"
```

```r
class(pols_203_final_merged$`Time required to start a business (days)`)
```

```
## [1] "numeric"
```

```r
class(pols_203_final_merged$Population)
```

```
## [1] "numeric"
```

Mean values of the columns

Age dependency

```r
mean(pols_203_final_merged$`Total dependency ratio - Sex: all - Age: none - Variant: estimates`,
     na.rm = TRUE)
```

```
## [1] 70.97669
```

Agricultural output

```r
mean(pols_203_final_merged$output_quantity,
     na.rm = TRUE)
```

```
## [1] 65420243991
```

Government expenditure on tertiary education as share of GDP

```r
mean(pols_203_final_merged$`Government expenditure on tertiary education as % of GDP (%)`,
     na.rm = TRUE)
```

```
## [1] 0.967423
```

International tourist arrivals

```r
mean(pols_203_final_merged$`International tourism, number of arrivals`,
     na.rm = TRUE)
```

```
## [1] 40150409
```

Top marginal income tax rate

```r
mean(pols_203_final_merged$`Top marginal income tax rate (Reynolds (2008))`,
     na.rm = TRUE)
```

```
## [1] 49.35
```

Oil production per capita

```r
mean(pols_203_final_merged$`Oil production per capita (kWh)`,
     na.rm = TRUE)
```

```
## [1] 25668.78
```

Participatory democratic institutions

```r
mean(pols_203_final_merged$particip_vdem_owid,
     na.rm = TRUE)
```

```
## [1] 0.2603911
```

Per capita electricity generation

```r
mean(pols_203_final_merged$`Per capita electricity (kWh)`,
     na.rm = TRUE)
```

```
## [1] 3834.949
```

GDP per capita

```
mean(pols_203_final_merged$`GDP per capita (output, multiple price benchmarks)`,
     na.rm = TRUE)
```

```
## [1] 14101.82
```

Time required to start a business

```
mean(pols_203_final_merged$`Time required to start a business (days)`,
     na.rm = TRUE)
```

```
## [1] 32.9653
```

Population

```
mean(pols_203_final_merged$Population,
     na.rm = TRUE)
```

```
## [1] 126470437
```

Standard deviations of the numeric columns

Age dependency

```
sd(pols_203_final_merged$`Total dependency ratio - Sex: all - Age: none - Variant: estimates`,
   na.rm = TRUE)
```

```
## [1] 20.20166
```

Agricultural output

```
sd(pols_203_final_merged$output_quantity,
   na.rm = TRUE)
```

```
## [1] 258007206831
```

Government expenditure on tertiary education as share of GDP

```
sd(pols_203_final_merged$`Government expenditure on tertiary education as % of GDP (%)`,
   na.rm = TRUE)
```

```
## [1] 0.5594122
```

International tourist arrivals

```
sd(pols_203_final_merged$`International tourism, number of arrivals`,
   na.rm = TRUE)
```

```
## [1] 174571434
```

Top marginal income tax rate

```
sd(pols_203_final_merged$`Top marginal income tax rate (Reynolds (2008))`,
   na.rm = TRUE)
```

```
## [1] 16.45265
```

Oil production per capita

```
sd(pols_203_final_merged$`Oil production per capita (kWh)`,
   na.rm = TRUE)
```

```
## [1] 167457.5
```

Participatory democratic institutions

```r
sd(pols_203_final_merged$particip_vdem_owid,
   na.rm = TRUE)
```

```
## [1] 0.2098426
```

Per capita electricity generation

```r
sd(pols_203_final_merged$`Per capita electricity (kWh)`,
   na.rm = TRUE)
```

```
## [1] 4952.337
```

GDP per capita

```r
sd(pols_203_final_merged$`GDP per capita (output, multiple price benchmarks)`,
   na.rm = TRUE)
```

```
## [1] 23746.75
```

Time required to start a business

```r
sd(pols_203_final_merged$`Time required to start a business (days)`,
   na.rm = TRUE)
```

```
## [1] 45.96973
```

Population

```r
sd(pols_203_final_merged$Population,
   na.rm = TRUE)
```

```
## [1] 588851231
```

Filter 2004 and 2014

```r
pols_203_final_merged_2004_2014 <- pols_203_final_merged %>%
  filter(Year == 2004 | Year == 2014)
```

Create the vector of countries to be studied

```r
country_vector <- c("Russia", "Germany", "United Kingdom", "France", "Italy",
                    "Spain", "Poland", "Netherlands", "Belgium",
                    "Czech Republic", "Greece", "Portugal", "Sweden", "Hungary",
                    "Belarus", "Austria", "Serbia", "Switzerland", "Denmark",
                    "Finland", "Slovakia", "Norway", "Ireland", "Croatia",
                    "Moldova", "Armenia", "Lithuania", "North Macedonia",
                    "Slovenia", "Latvia", "Estonia", "Montenegro", "Luxemburg",
                    "Malta", "Iceland", "Azerbaijan")
```

Alphabetically rearrange

```r
country_vector <- sort(country_vector) # Sort in ascending order
```

Filter for those countries

```r
pols_203_final_merged_2004_2014 <- pols_203_final_merged_2004_2014 %>%
  filter(Entity %in% country_vector) # Is the country in our country list?
```

Move values to a single column to prepare the pols_203_joined for wrangling

1. **Spot the first appearance of each variable**

```
match(unique(pols_203_final_merged_2004_2014$id), pols_203_final_merged_2004_2014$id)
```

```
##  [1]   1  69 135 166 230 297 365 432 500 565
```

2. **Move the columns**

Data wrangling

```
pols_203_final_merged_2004_2014 <- pols_203_final_merged_2004_2014 %>%
  dplyr::select("id",
                "Entity",
                "Year",
                "value") %>%
  pivot_wider(names_from = "id", # "Tame" the data
              values_from = "value")
```

Split the dataframe into two by year   Filter for 2004

```
pols_203_final_merged_2004 <- pols_203_final_merged_2004_2014 %>%
  filter(Year == 2004) # Year must be equal to 2004
```

Filter for 2014

```
pols_203_final_merged_2014 <- pols_203_final_merged_2004_2014 %>%
  filter(Year == 2014) # Year must be equal to 2014
```

Join the two dataframes

```
pols_203_joined <- inner_join(pols_203_final_merged_2004,
                              pols_203_final_merged_2014,
                              by = c("Entity"),
                              suffix = c("_2004",
                                         "_2014"))
```

Remove the redundant "Year" column

```
pols_203_joined <- pols_203_joined %>%
  dplyr::select(!starts_with("Year"))
```

Rename the "Entities" column

```
names(pols_203_joined)[names(pols_203_joined) == "Entity"] <- "country"
```

Examine

```
pols_203_joined
```

```
## # A tibble: 34 x 21
##    country    total_de~1 agric~2 gov_e~3 touri~4 oil_p~5 democ~6 elect~7 real_~8
##    <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Armenia          51.5 1.19e 9  NA      2.63e5      0    0.393   1924.   5389.
##  2 Austria          46.9 6.24e 9   1.39   1.94e7   1589.   0.665   7579.  39733.
##  3 Azerbaijan       50.3 3.00e 9   0.221  1.28e6  21279.   0.212   2395.   3567.
##  4 Belarus          44.5 9.44e 9   1.41   1.82e6   2385.   0.204   2934.  11319.
##  5 Belgium          52.4 9.02e 9   1.26   6.71e6      0    0.653   8068.  36638.
##  6 Croatia          49.6 2.12e 9   0.692  4.50e7   3420.   0.557   3126.  16701.
##  7 Denmark          50.8 9.74e 9   2.45   2.21e7  41023.   0.743   7464.  39557.
##  8 Estonia          47.0 6.98e 8   0.854 NA            0    0.685   7567.  16688.
##  9 Finland          49.9 3.23e 9   1.99   2.84e6      0    0.677  16373.  38078.
## 10 France           53.7 6.30e10   1.18   1.90e8    281.   0.706   9461.  34926.
## # ... with 24 more rows, 12 more variables:
## #   time_req_to_start_business_2004 <dbl>, population_2004 <dbl>,
## #   total_dependency_ratio_2014 <dbl>, agricultural_output_2014 <dbl>,
## #   gov_exp_tertiary_ed_vs_GDP_2014 <dbl>, tourists_2014 <dbl>,
## #   oil_production_per_cap_2014 <dbl>, democracy_2014 <dbl>,
## #   electricity_per_cap_2014 <dbl>, real_GDP_per_cap_2014 <dbl>,
## #   time_req_to_start_business_2014 <dbl>, population_2014 <dbl>, and ...
```

Create a vector that lists all the EU countries

```
eu <- c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus", "Czech Republic",
        "Denmark", "Estonia", "Finland", "France", "Germany", "Greece",
        "Hungary", "Ireland", "Italy", "Latvia", "Lithuania", "Luxembourg",
        "Malta", "Netherlands", "Poland", "Portugal", "Romania", "Slovakia",
        "Slovenia", "Spain", "Sweden", "United Kingdom")
```

Class them according to whether they are members of the EU

```
pols_203_joined <- pols_203_joined %>%
  mutate(eu = case_when(country %in% eu ~ TRUE,
                        !(country %in% eu) ~ FALSE ))
```

```
summary(pols_203_joined)
```

Summary

```
##    country          total_dependency_ratio_2004 agricultural_output_2004
##  Length:34          Min.   :40.73               Min.   :  116070000
##  Class :character   1st Qu.:45.58               1st Qu.: 2426474000
##  Mode  :character   Median :47.97               Median : 5643441000
##                     Mean   :47.97               Mean   :14178982312
##                     3rd Qu.:50.41               3rd Qu.:13334413750
##                     Max.   :53.68               Max.   :67499301000
##                                                 NA's   :2
##  gov_exp_tertiary_ed_vs_GDP_2004 tourists_2004
##  Min.   :0.2212                  Min.   :    69000
##  1st Qu.:0.8480                  1st Qu.:  1445750
##  Median :1.1012                  Median :  6831500
##  Mean   :1.1786                  Mean   : 21898133
##  3rd Qu.:1.3675                  3rd Qu.: 22061500
##  Max.   :2.4491                  Max.   :190282000
##  NA's   :6                       NA's   :4
##  oil_production_per_cap_2004 democracy_2004   electricity_per_cap_2004
##  Min.   :     0.0            Min.   :0.2040   Min.   : 1433
##  1st Qu.:     0.0            1st Qu.:0.6098   1st Qu.: 3945
##  Median :   172.2            Median :0.6580   Median : 6139
##  Mean   : 15594.7            Mean   :0.6301   Mean   : 7327
##  3rd Qu.:  2137.6            3rd Qu.:0.6887   3rd Qu.: 7579
##  Max.   :379949.8            Max.   :0.8820   Max.   :29450
##  NA's   :1                                    NA's   :1
##  real_GDP_per_cap_2004 time_req_to_start_business_2004 population_2004
##  Min.   : 3558         Min.   :  5.00                  Min.   :   292364
##  1st Qu.:14733         1st Qu.: 16.50                  1st Qu.:  3574935
##  Median :25133         Median : 29.00                  Median :  7622276
##  Mean   :26556         Mean   : 40.52                Mean   : 18915708
##  3rd Qu.:39187         3rd Qu.: 60.75                3rd Qu.: 10961698
##  Max.   :61862         Max.   :137.00                Max.   :144353650
##                        NA's   :3
##  total_dependency_ratio_2014 agricultural_output_2014
##  Min.   :40.31               Min.   :  127950000
##  1st Qu.:47.33               1st Qu.: 2387486000
##  Median :50.45               Median : 6133874000
##  Mean   :49.64               Mean   :14315008824
##  3rd Qu.:52.82               3rd Qu.:11442915750
##  Max.   :58.51               Max.   :87416238000
##
##  gov_exp_tertiary_ed_vs_GDP_2014 tourists_2014
##  Min.   :0.3052                  Min.   :    93900
##  1st Qu.:0.5739                  1st Qu.:  2689000
##  Median :0.8426                  Median :  8985500
##  Mean   :0.8116                  Mean   : 25166291
##  3rd Qu.:1.0647                  3rd Qu.: 31333250
##  Max.   :1.2869                  Max.   :206599008
##  NA's   :31
##  oil_production_per_cap_2014 democracy_2014   electricity_per_cap_2014
##  Min.   :     0.00           Min.   :0.1590   Min.   :  287.6
##  1st Qu.:     0.00           1st Qu.:0.6160   1st Qu.: 3550.8
##  Median :    75.77           Median :0.6565   Median : 5107.4
##  Mean   :  9497.82           Mean   :0.6307   Mean   : 7674.8
##  3rd Qu.:  1311.22           3rd Qu.:0.6990   3rd Qu.: 7482.6
##  Max.   :192163.98           Max.   :0.8820   Max.   :55302.9
##
##  real_GDP_per_cap_2014 time_req_to_start_business_2014 population_2014
##  Min.   : 7270         Min.   : 3.50                   Min.   :   327650
##  1st Qu.:23934         1st Qu.: 6.00                   1st Qu.:  3080780
##  Median :28986         Median :11.50                   Median :  7865878
##  Mean   :34368         Mean   :12.12                   Mean   : 19450566
##  3rd Qu.:45165         3rd Qu.:14.38                   3rd Qu.: 11098288
##  Max.   :84910         Max.   :37.00                   Max.   :144285070
##
##      eu
##  Mode :logical
##  FALSE:11
##  TRUE :23
##
##
##
##
```

Discussion: Some columns have N/As

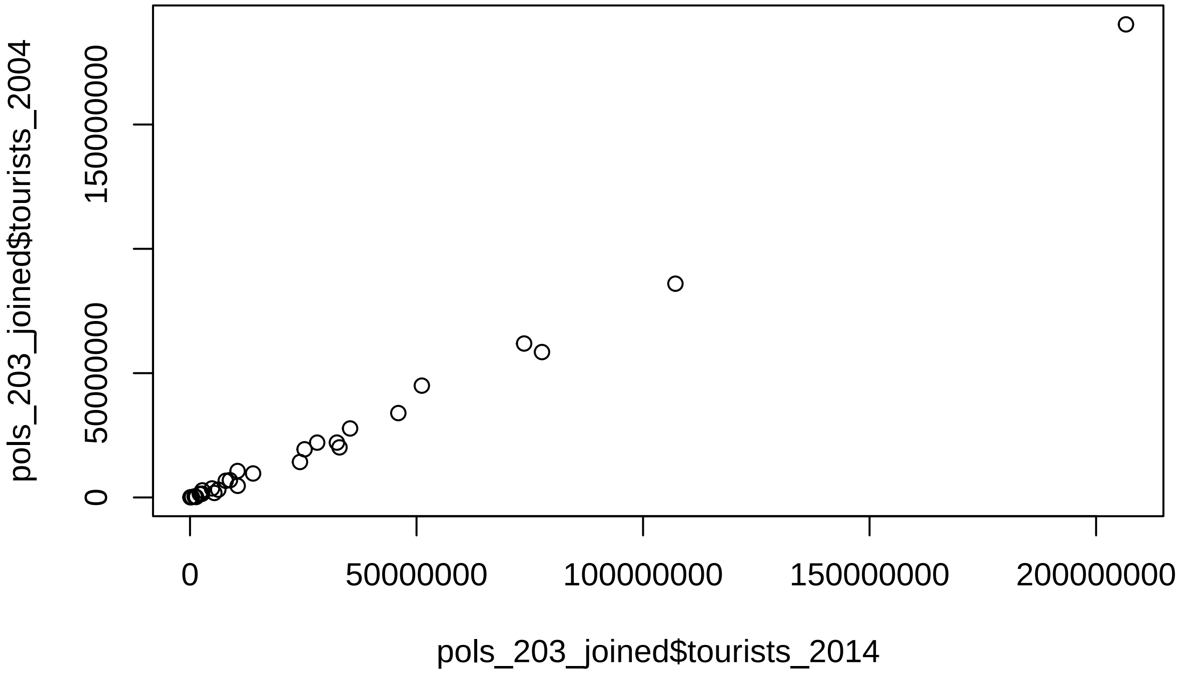
We decided to remove the "gov_exp_tertiary_ed_vs_GDP" columns since they have too many missing values. However, we will fill in the few other values that are missing just for 2004 by using imputation (from Chapter 16 of the textbook, Statistical Methods for the Social Science by Alan Agresti). Because, if we strictly avoid all the data that contains some missing values, we will not be able to consider various IVs and if we delete the country rows that have missing values, our model will be rather biased since the countries we have complete data for are almost always developed countries.

Remove "gov_exp_tertiary_ed_vs_GDP_2004" and "gov_exp_tertiary_ed_vs_GDP_2014"

```
pols_203_joined <- pols_203_joined %>%
  dplyr::select(!c(gov_exp_tertiary_ed_vs_GDP_2004, gov_exp_tertiary_ed_vs_GDP_2014))
```

Imputation by linear regression *tourists_2004*

```
plot(pols_203_joined$tourists_2014, pols_203_joined$tourists_2004)
```



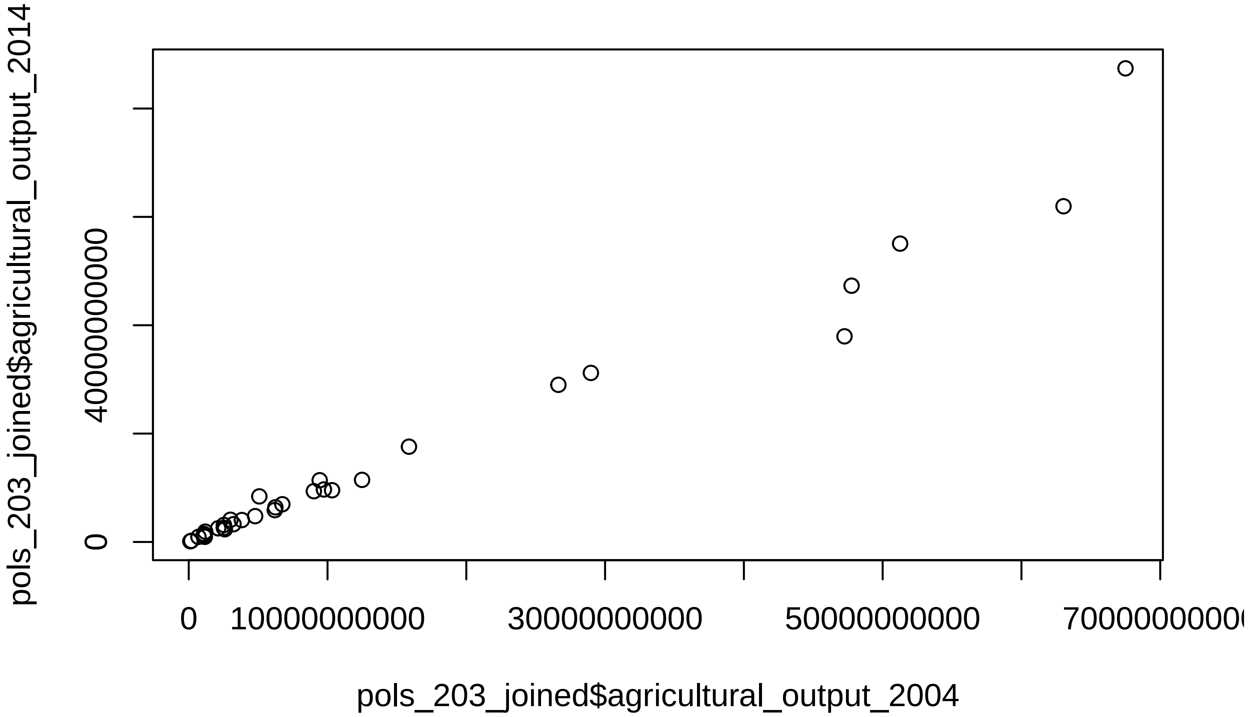
[The relationship is linear].{.underline}

Imputation by linear regression is justified

```
pols_203_joined <- impute_lm(pols_203_joined,
                             tourists_2004 ~ tourists_2014)
```

agricultural_output_2004

```
plot(pols_203_joined$agricultural_output_2004,
     pols_203_joined$agricultural_output_2014)
```

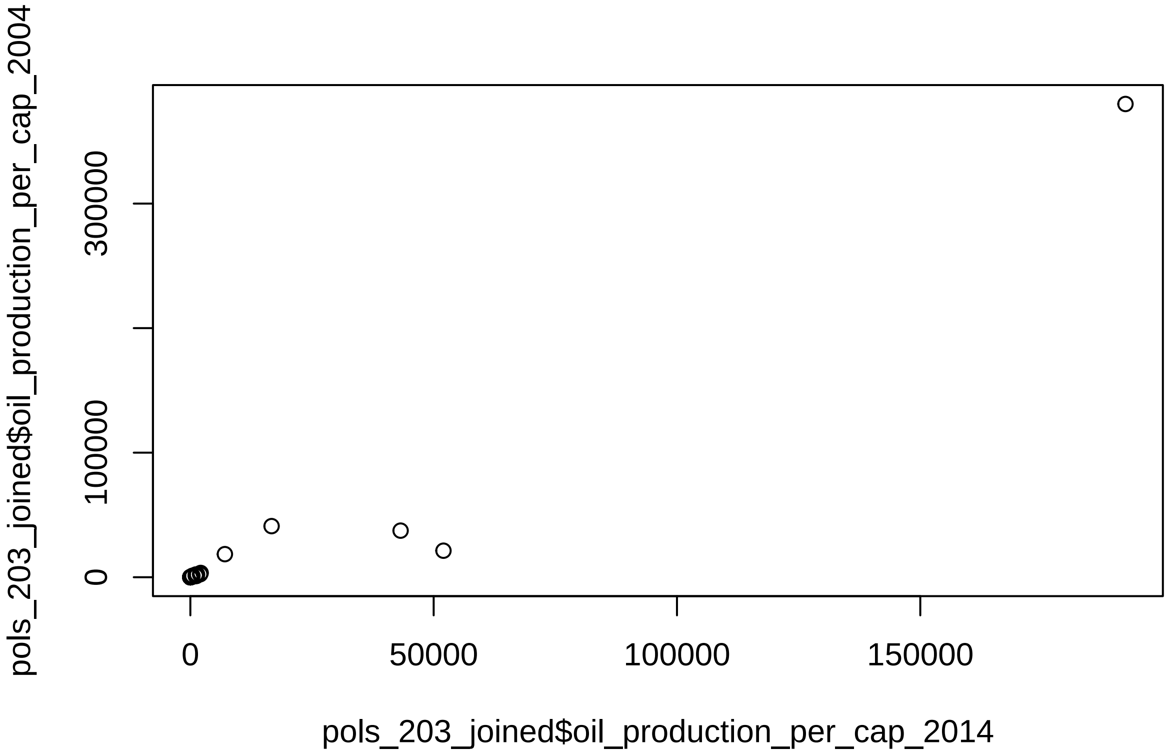


```
# The relationship is non-linear
```

Imputation by linear regression is not justified

oil_production_per_cap_2004

```
plot(pols_203_joined$oil_production_per_cap_2014,
     pols_203_joined$oil_production_per_cap_2004)
```

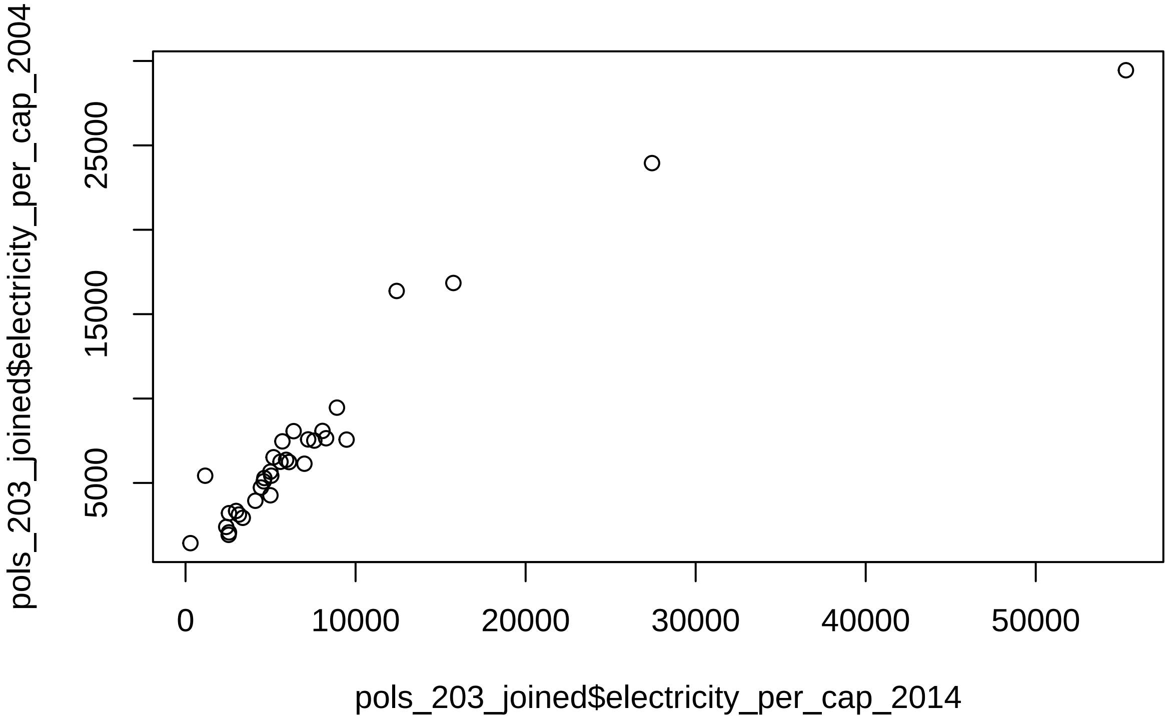


```
# The relationship is non-linear
```

Imputation by linear regression is not justified

electricity__per__cap__2004

```
plot(pols_203_joined$electricity_per_cap_2014,
     pols_203_joined$electricity_per_cap_2004)
```

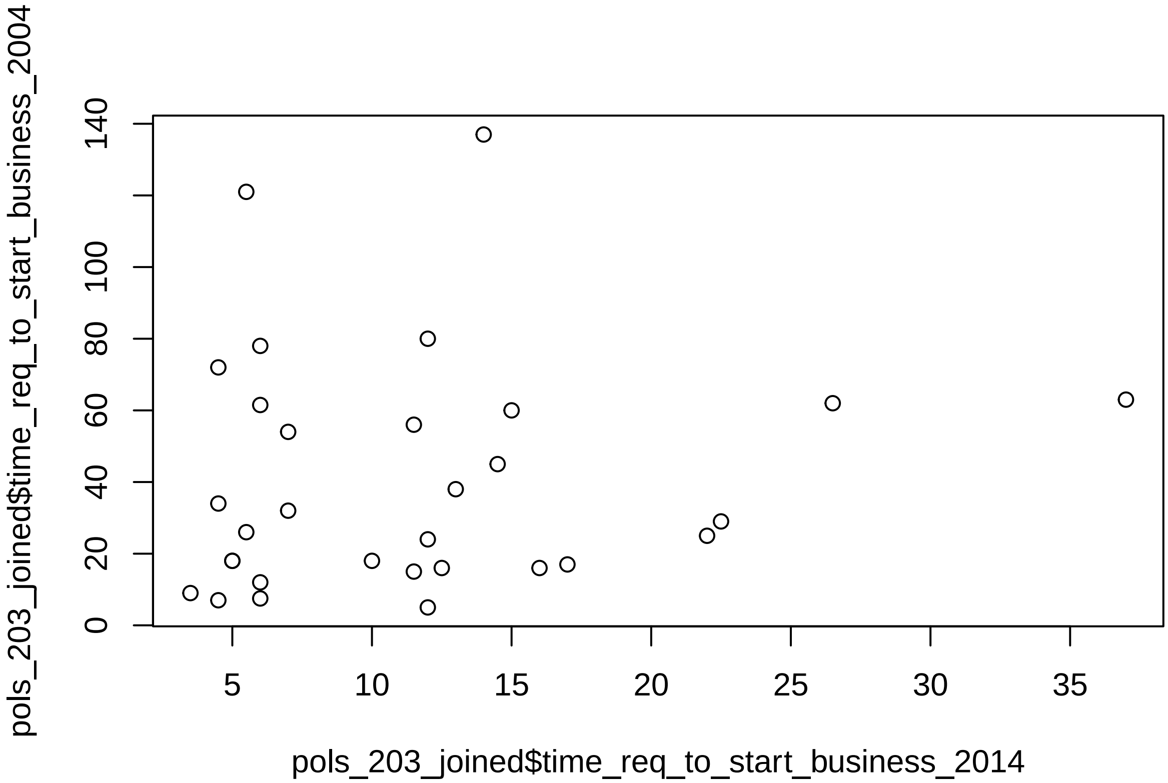


```
# The relationship is non-linear
```

Imputation by linear regression is not justified

time__req__to__start__business__2004

```
plot(pols_203_joined$time_req_to_start_business_2014,
     pols_203_joined$time_req_to_start_business_2004)
```



```
# The is not much relationship
```

Imputation by linear regression is not justified

We decided to fill in the remaining missing values by using a non-parametric

algorithm called "randomForest"

Prepare the dataset for imputation

```
pols_203_joined_4_imp <- dplyr::select(pols_203_joined, -c("country", "eu"))
```

Convert it into an ordinary dataframe

```
pols_203_joined_4_imp <- as.data.frame(pols_203_joined_4_imp)
```

Run the algorithm

```
forest <- missForest(pols_203_joined_4_imp)
```

Convert it back to a tibble

```
forest_tibble <- as_tibble(forest$ximp)
```

Add the "country" and "eu" columns

```
forest_tibble <- bind_cols(pols_203_joined$country, forest_tibble, pols_203_joined$eu)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...20`
```

```
colnames(forest_tibble)[1] <- "country"
colnames(forest_tibble)[20] <- "eu"
```

Repair column names

```
summary(forest_tibble)
```

Summary

```
##    country          total_dependency_ratio_2004 agricultural_output_2004
##  Length:34          Min.   :40.73               Min.   :  116070000
##  Class :character   1st Qu.:45.58               1st Qu.: 2221246000
##  Mode  :character   Median :47.97               Median : 5344481145
##                     Mean   :47.97               Mean   :13556393958
##                     3rd Qu.:50.41               3rd Qu.:11949489500
##                     Max.   :53.68               Max.   :67499301000
##  tourists_2004       oil_production_per_cap_2004 democracy_2004
##  Min.   :    69000   Min.   :     0.0            Min.   :0.2040
##  1st Qu.:  1578250   1st Qu.:     0.0            1st Qu.:0.6098
##  Median :  6184923   Median :   221.7            Median :0.6580
##  Mean   : 19859985   Mean   : 15144.4            Mean   :0.6301
##  3rd Qu.: 21574750   3rd Qu.:  2000.5            3rd Qu.:0.6887
##  Max.   :190282000   Max.   :379949.8            Max.   :0.8820
##  electricity_per_cap_2004 real_GDP_per_cap_2004 time_req_to_start_business_2004
##  Min.   : 1433            Min.   : 3558         Min.   :  5.00
##  1st Qu.: 4026            1st Qu.:14733         1st Qu.: 17.25
##  Median : 5902            Median :25133         Median : 33.00
##  Mean   : 7264            Mean   :26556         Mean   : 42.38
##  3rd Qu.: 7576            3rd Qu.:39187         3rd Qu.: 61.12
##  Max.   :29450            Max.   :61862         Max.   :137.00
##  population_2004     total_dependency_ratio_2014 agricultural_output_2014
##  Min.   :   292364   Min.   :40.31               Min.   :  127950000
##  1st Qu.:  3574935   1st Qu.:47.33               1st Qu.: 2387486000
##  Median :  7622276   Median :50.45               Median : 6133874000
##  Mean   : 18915708   Mean   :49.64               Mean   :14315008824
##  3rd Qu.: 10961698   3rd Qu.:52.82               3rd Qu.:11442915750
##  Max.   :144353650   Max.   :58.51               Max.   :87416238000
##  tourists_2014       oil_production_per_cap_2014 democracy_2014
##  Min.   :    93900   Min.   :     0.00           Min.   :0.1590
##  1st Qu.:  2689000   1st Qu.:     0.00           1st Qu.:0.6160
##  Median :  8985500   Median :    75.77           Median :0.6565
##  Mean   : 25166291   Mean   :  9497.82           Mean   :0.6307
##  3rd Qu.: 31333250   3rd Qu.:  1311.22           3rd Qu.:0.6990
##  Max.   :206599008   Max.   :192163.98           Max.   :0.8820
##  electricity_per_cap_2014 real_GDP_per_cap_2014 time_req_to_start_business_2014
##  Min.   :  287.6          Min.   : 7270         Min.   : 3.50
##  1st Qu.: 3550.8          1st Qu.:23934         1st Qu.: 6.00
##  Median : 5107.4          Median :28986         Median :11.50
##  Mean   : 7674.8          Mean   :34368         Mean   :12.12
##  3rd Qu.: 7482.6          3rd Qu.:45165         3rd Qu.:14.38
##  Max.   :55302.9          Max.   :84910         Max.   :37.00
##  population_2014         eu
##  Min.   :   327650   Mode :logical
##   1st Qu.:  3080780   FALSE:11
##   Median :  7865878   TRUE :23
##   Mean   : 19450566
##   3rd Qu.: 11098288
##   Max.   :144285070
```

There is no N/A left

Calculate the means and per capita values wherever needed

```
forest_tibble <- forest_tibble %>%
  mutate(agricultural_output_per_cap_2004 = agricultural_output_2004 / population_2004,
         agricultural_output_per_cap_2014 = agricultural_output_2014 / population_2014,
         tourists_per_cap_2004 = tourists_2004 / population_2004,
         tourists_per_cap_2014 = tourists_2014 / population_2014,
         total_dependency_ratio_mean = (total_dependency_ratio_2004 +
                                          total_dependency_ratio_2014) / 2,
         oil_production_per_cap_mean = (oil_production_per_cap_2004 +
                                          oil_production_per_cap_2014) / 2,
         democracy_mean = (democracy_2004 +
                             democracy_2014) / 2,
         electricity_per_cap_mean = (electricity_per_cap_2004 +
                                       electricity_per_cap_2014) / 2,
         real_GDP_per_cap_mean = (real_GDP_per_cap_2004 +
                                    real_GDP_per_cap_2014) / 2,
         time_req_to_start_business_mean = (time_req_to_start_business_2004 +
                                              time_req_to_start_business_2014) / 2) %>%
  mutate(tourists_per_cap_mean = (tourists_per_cap_2004 +
                                    tourists_per_cap_2014) / 2) %>%
  mutate(growth = (real_GDP_per_cap_2014 - real_GDP_per_cap_2004) / real_GDP_per_cap_2004) %>%

 dplyr::select(-c("agricultural_output_2004",
                   "agricultural_output_2014",
                   "tourists_2004",
                   "tourists_2014")) %>%
  relocate(eu, .after = growth) # Move eu to the end
```

First part of the project

In our first project, we tried to answer the question of whether former Eastern Bloc countries that joined the EU in 2004 enjoyed higher GDP PPP per capita growth rates.

We will briefly replicate what we did back then as the first part of this project. However, this time we will use the data retrieved from Our World in Data website instead of World Bank DataBank.

Treatment: Joining the EU in 2004 (nominal)

Dependent variable: GDP PPP *per capita* growth (continuous numerical)

Null-hypothesis: Mean GDP PPP *per capita* growth between 2004 - 2014 for the former Eastern Bloc countries that joined the EU is equal to the mean GDP PPP per capita growth for the former Eastern Bloc countries that did not join the union

Alternative hypothesis: Mean GDP PPP *per capita* growth between 2004 - 2014 for the former Eastern Bloc countries that joined the EU is not equal to the mean GDP PPP per capita growth for the former Eastern Bloc countries that did not join the union.

Strategy: Compare the treatment and control groups and then apply a t-test.

Start of the code

Create a vector of the countries that we compared in the first project

```
eastern_bloc_old <- c("Poland", "Czechia", "Estonia", "Hungary", "Latvia",
                      "Lithuania", "Slovakia", "Slovenia", "Belarus", "Russia",
                      "Moldova", "Armenia")
```

Filter them and save as another dataframe

```
eastern_bloc.df <- forest_tibble[forest_tibble$country %in% eastern_bloc_old,]
```

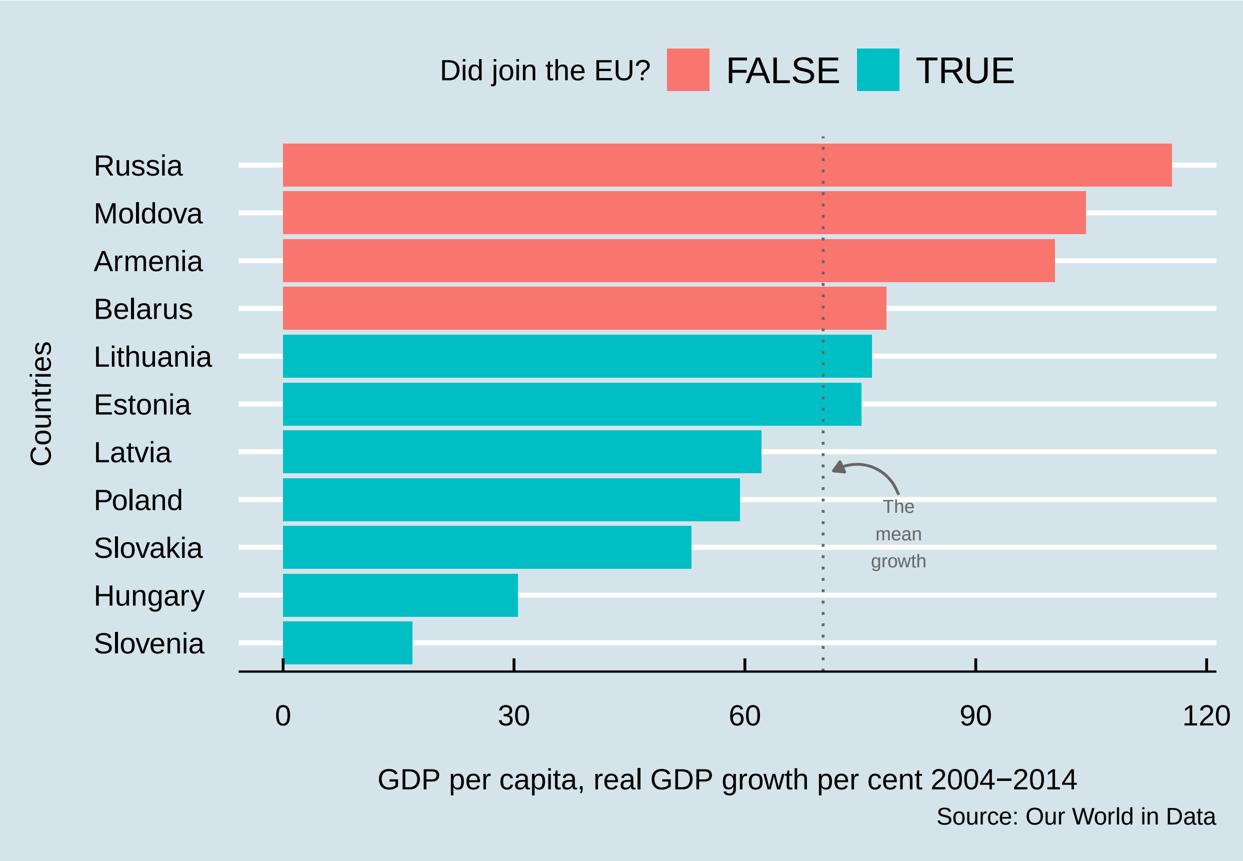
Summarize

```
eastern_bloc.df %>% group_by(eu) %>%
  summarize(mean_growth_per_cent = mean(growth) * 100)
```

```
## # A tibble: 2 x 2
##   eu    mean_growth_per_cent
##   <lgl>                <dbl>
## 1 FALSE                 99.6
## 2 TRUE                  53.3
```

Plot

```
ggplot(eastern_bloc.df,
       aes(x = fct_reorder(country, growth),
           y = growth * 100,
           fill = eu)) +
  geom_col() +
  xlab("Countries") +
  ylab("GDP per capita, real GDP growth per cent 2004-2014") +
  coord_flip() +
  theme_economist() +
  theme(axis.title.x = element_text(margin = margin(t = 15)),
        axis.title.y = element_text(margin = margin(r = 15))) +
  scale_fill_discrete("Did join the EU?") +
  geom_hline(yintercept = mean(eastern_bloc.df$growth * 100),
             color = "grey40",
             linetype = 3) +
  annotate( "text", x = 4, y = 80,
            label = "The\nmean\ngrowth",
            vjust = 1, size = 2.5, color = "grey40") +
  annotate( "curve",
            x = 4.1,
            y = 80,
            xend = 4.6,
            yend = 71.5,
            arrow = arrow(length = unit(0.15, "cm"),
                          type = "closed"), color = "grey40") +
  labs(caption = "Source: Our World in Data")
```



Source: Our World in Data

Significance test

```
t_test <- t.test(eastern_bloc.df$growth[!eastern_bloc.df$eu],
                 eastern_bloc.df$growth[eastern_bloc.df$eu],
                 paired = FALSE,
                 conf.level = 0.95)
```

Is it statistically significant?

```
t_test$p.value < 0.05 # Yes, it is statistically significant (p < 0.05)
```

```
## [1] TRUE
```

t-test

```
t_test
```

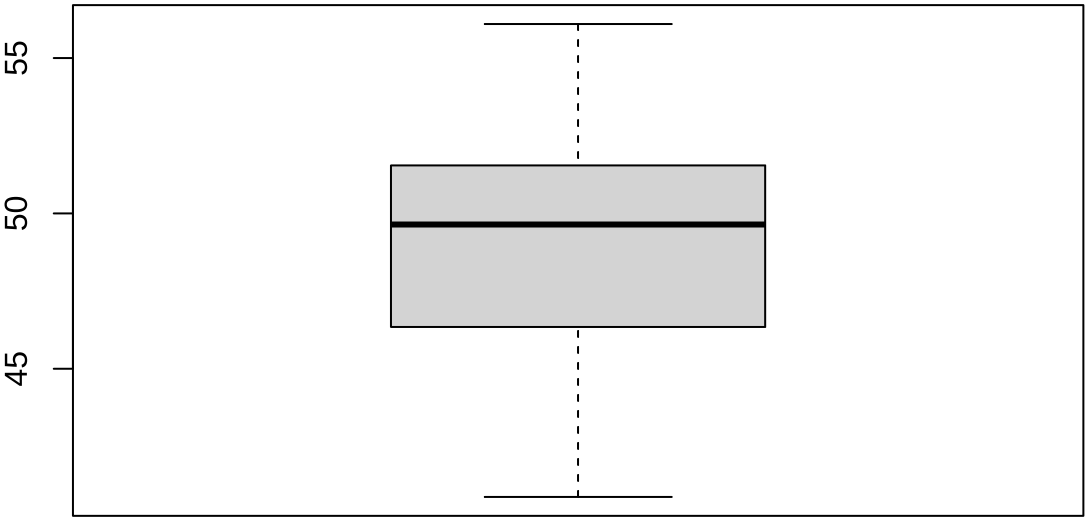
```
##
##  Welch Two Sample t-test
##
## data:  eastern_bloc.df$growth[!eastern_bloc.df$eu] and eastern_bloc.df$growth[eastern_bloc.df$eu]
## t = 4.0324, df = 8.3914, p-value = 0.003421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2001814 0.7249948
## sample estimates:
## mean of x mean of y
## 0.9960751 0.5334871
```

Second part of the project

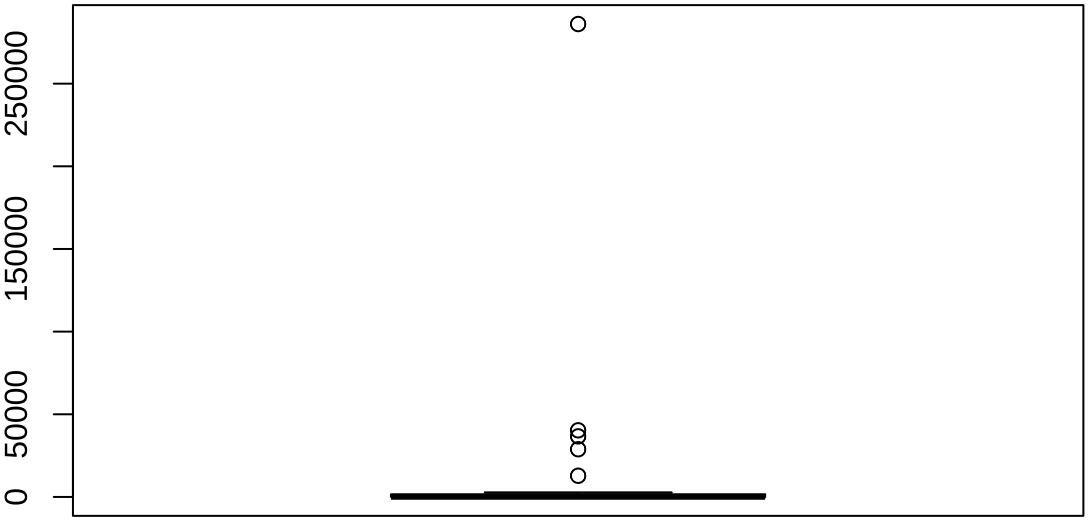
Descriptive statistics

Boxplots

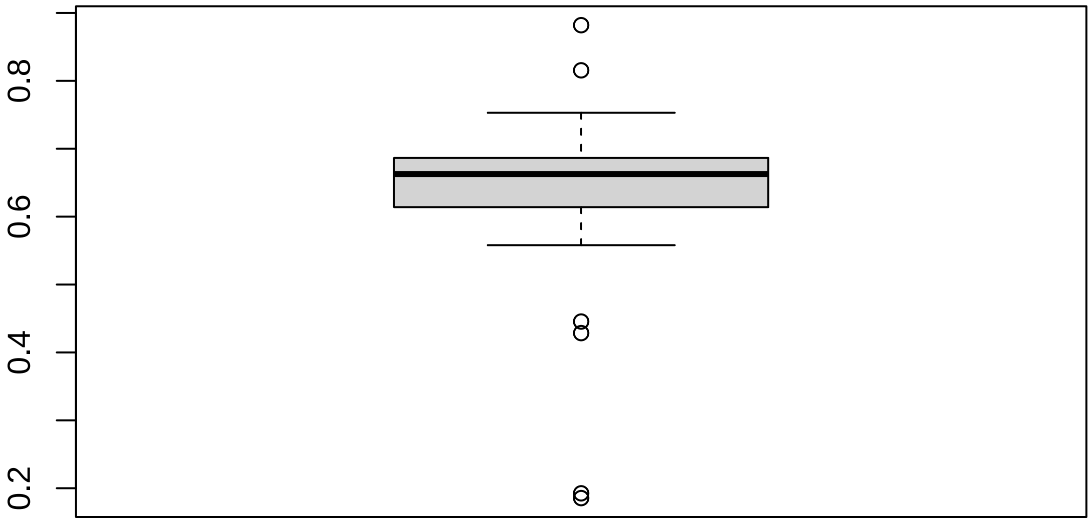
```
boxplot(forest_tibble$total_dependency_ratio_mean)
boxplot(forest_tibble$total_dependency_ratio_mean)
```



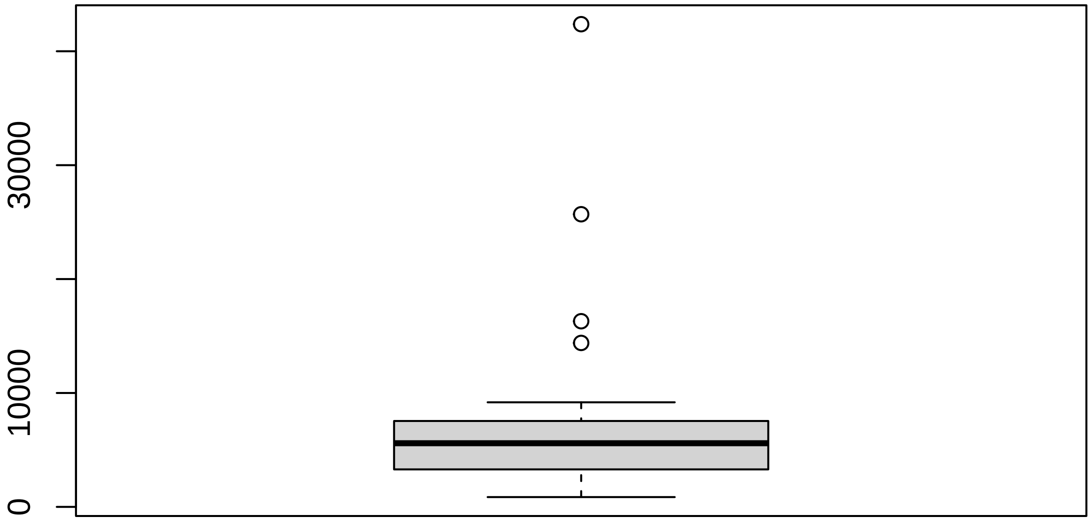
```
boxplot(forest_tibble$oil_production_per_cap_mean)
```



```
boxplot(forest_tibble$democracy_mean)
```

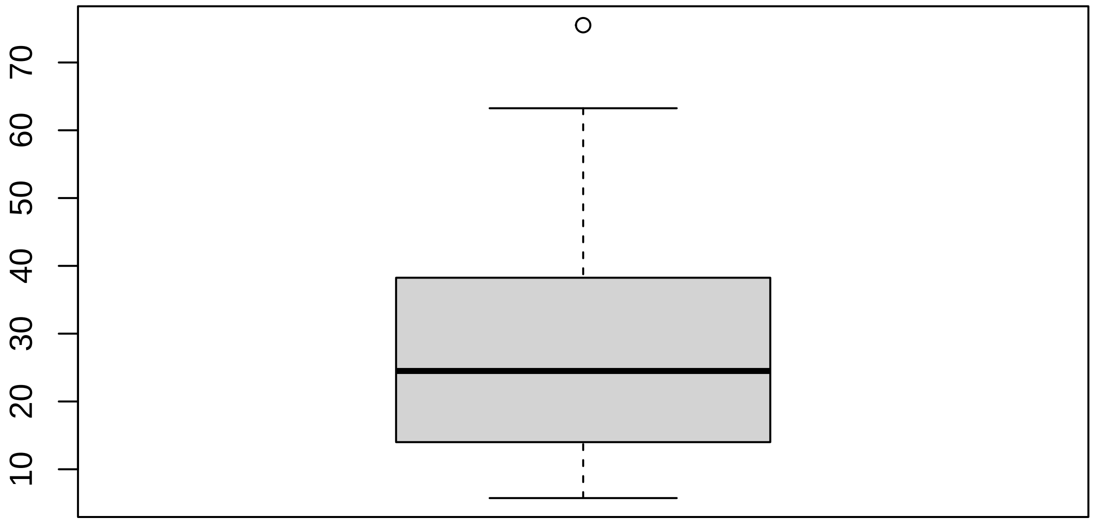


```
boxplot(forest_tibble$electricity_per_cap_mean)$stats
```

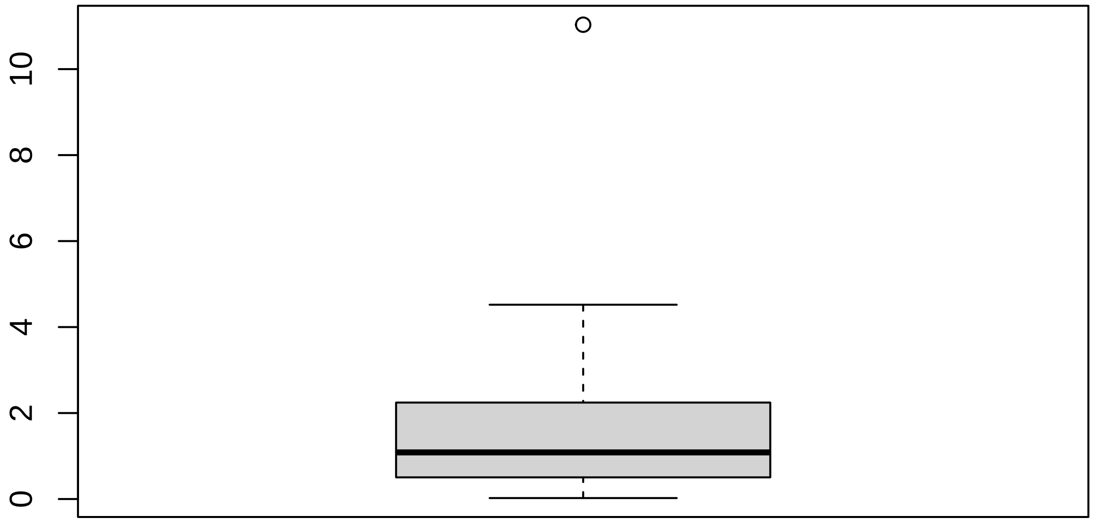


```
##           [,1]
## [1,]  860.3647
## [2,] 3291.9597
## [3,] 5587.8931
## [4,] 7541.5010
## [5,] 9182.6025
```

```
boxplot(forest_tibble$time_req_to_start_business_mean)
```



```
boxplot(forest_tibble$tourists_per_cap_mean)
```



We have several outliers. However, we decided not to remove them yet since we want to avoid overfitting

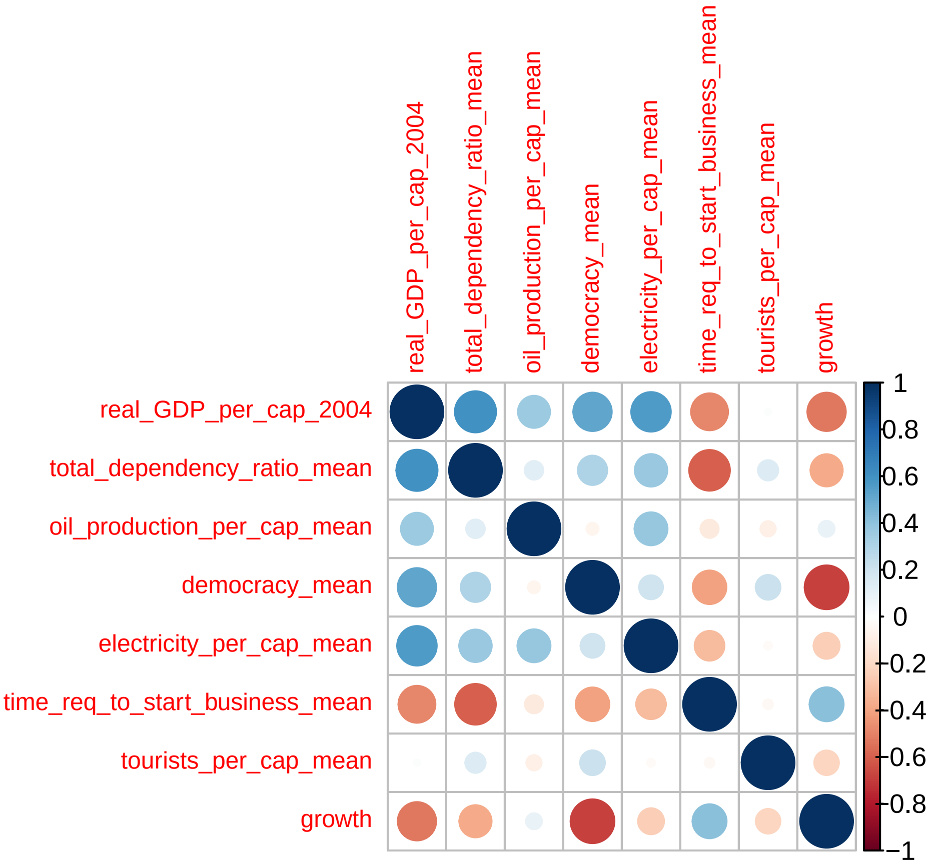
Build a model

Correlation matrix

```
cor_matrix <- cor(forest_tibble[, c(6, 20:23, 25:27)])
```

Plot the correlation matrix

```
corrplot(cor_matrix, tl.cex = 0.75)
```



Findings:

1. Real GDP per capita is **positively** correlated with total dependency ratio, democracy, and electricity generation per capita. Therefore an increase in one of these measures would correspond to an increase in the real GDP per capita.

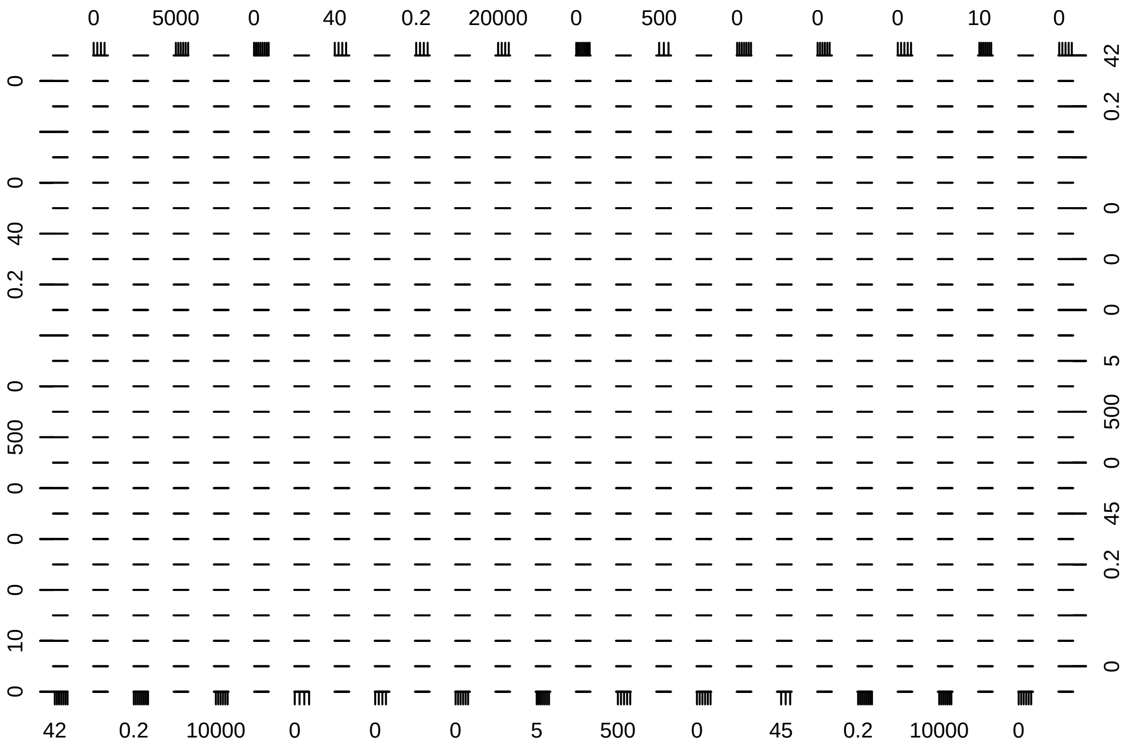
2. Real GDP per capita is **negatively** correlated with growth and the time required to start a business. As the time required to start a business increases Real GDP per capita tend to decrease.

3. The total dependency ratio is **negatively** correlated with the time required to start a business. A longer time required to start a business is related to a lower total age dependency ratio.

4. Democracy score is **strongly** and **negatively** correlated with real GDP per capita growth between 2004 and 2014. We think this could be due to democratic countries already enjoying a high GDP and their growth is affected by the law of marginal benefit. Countries with higher democracy scores have a positive correlation with Real GDP but a strong and negative one with growth.

All possible pairs

```
pairs(forest_tibble[, 2:27])
```



The original model

```
m0 <- lm(growth ~ total_dependency_ratio_mean +
           oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           electricity_per_cap_mean +
           time_req_to_start_business_mean +
           tourists_per_cap_mean +
           real_GDP_per_cap_2004 +
           eu, data = forest_tibble)
```

Model metrics

```
summary(m0) %>% glance()
```

```
## # A tibble: 1 x 8
##   r.squared adj.r.squared sigma statistic p.value    df df.residual  nobs
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>       <int> <dbl>
## 1     0.562         0.422 0.560      4.01 0.00352     8          25    34
```

```
m0_augment <- m0 %>% augment()
```

Cook's distance

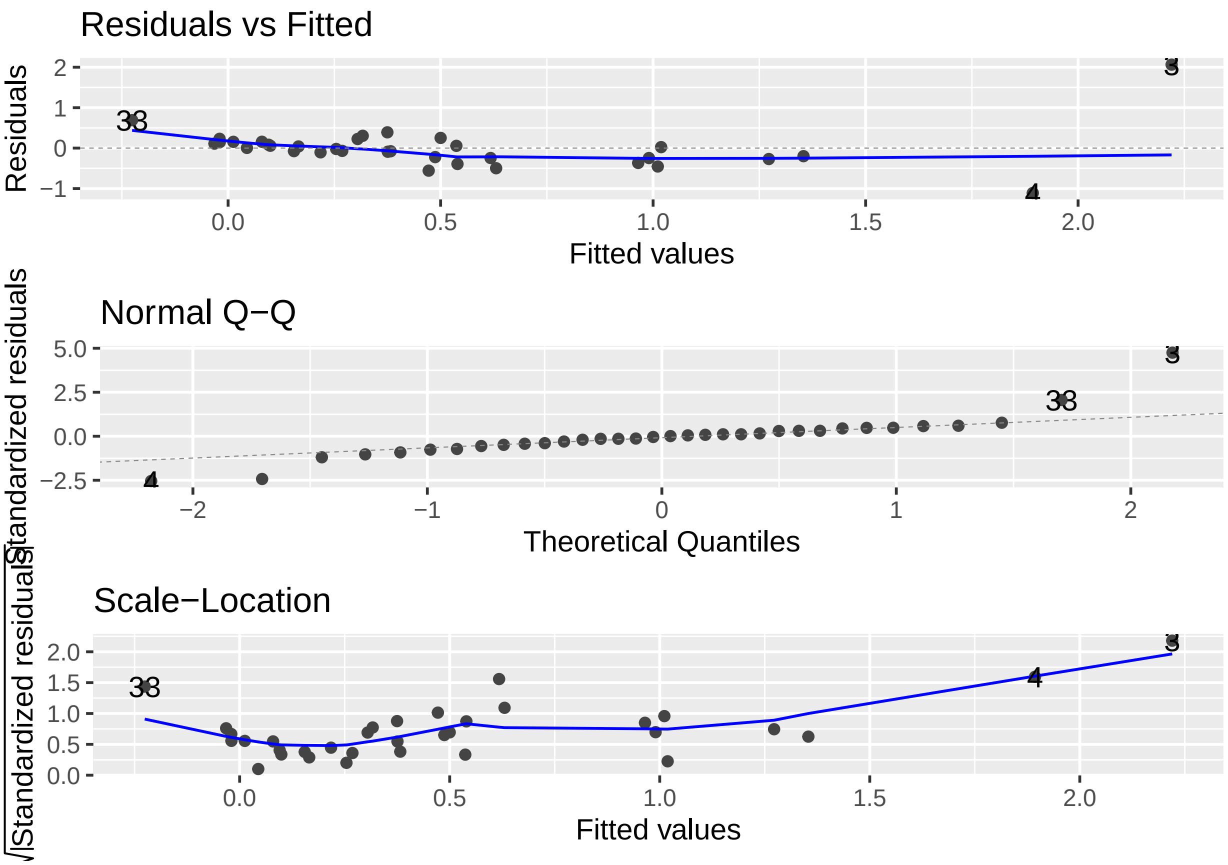
```
cooks.distance(m0)
```

```
##              1              2              3              4              5 
##   0.011275067966 0.001487180474 1.666735795781 0.464163981140 0.000202592021 
##              6              7              8              9             10 
##   0.024604391876 0.003982387063 0.004033832487 0.000331116329 0.000272216982
##               11               12               13               14               15 
##   0.000020940964   0.007955978656   0.000280876402   0.279489651311   0.003815891248 
##               16               17               18               19               20 
##   0.000001368451   0.008466166963   0.014381676817   0.002817004721   0.000102928257 
##               21               22               23               24               25 
##   0.006141316466   0.000237932740   0.012340939177  19.504584461752   0.000319805096 
##               26               27               28               29               30 
##   0.011696353255   0.003604473049   0.025924163082   0.010367879216   0.002574019650 
##               31               32               33               34 
##   0.126949182744   0.000635555812   0.845867971108   0.000604894009 
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 19.19, the highest recorded. T his observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m0, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m0)$p.value < 0.05
```

```
##   BP 
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: Although observations 3, 4, and 36 slightly

distort the curve, it is almost horizontal.

2. Q-Q: Residuals have an S-like distribution.
3. Scale-location: The data looks heteroskedastic since the line is horizontal and shows a steep angle at the right end. The residuals begin to spread wider as it passes 1 on the x-axis.
4. The independent variables explain 0.422 = 42.2% of the variation in the dependent variable

Variance inflation factor

```
vif(m0) # There is moderate (VIF < 5) correlation between the IVs
```

```
##     total_dependency_ratio_mean     oil_production_per_cap_mean
##                        2.541809                        1.556456
##                  democracy_mean        electricity_per_cap_mean
##                        2.109987                        2.001578
## time_req_to_start_business_mean           tourists_per_cap_mean
##                        2.025343                        1.261729
##            real_GDP_per_cap_2004                              eu
##                        3.310268                        2.806703
```

Use “Akaike information criterion” for model selection

```
aic <- stepAIC(m0)
```

```
## Start:  AIC=-31.83
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + oil_production_per_cap_mean + electricity_per_cap_mean +
##     time_req_to_start_business_mean + tourists_per_cap_mean +
##     real_GDP_per_cap_2004 + eu
##
##                                    Df Sum of Sq    RSS     AIC
## - total_dependency_ratio_mean      1   0.03635 7.8888 -33.671
## - tourists_per_cap_mean            1   0.04794 7.9004 -33.621
## - electricity_per_cap_mean         1   0.08387 7.9364 -33.467
## - eu                               1   0.16580 8.0183 -33.118
## - time_req_to_start_business_mean  1   0.20589 8.0584 -32.948
## - oil_production_per_cap_mean      1   0.25891 8.1114 -32.725
## - real_GDP_per_cap_2004            1   0.29482 8.1473 -32.575
## <none>                                         7.8525 -31.828
## - democracy_mean                   1   1.37749 9.2300 -28.333
##
## Step:  AIC=-33.67
## growth ~ oil_production_per_cap_mean + democracy_mean + electricity_per_cap_mean +
##     time_req_to_start_business_mean + tourists_per_cap_mean +
##     real_GDP_per_cap_2004 + eu
##
##                                    Df Sum of Sq    RSS     AIC
## - tourists_per_cap_mean            1   0.04430 7.9331 -35.481
## - electricity_per_cap_mean         1   0.06336 7.9522 -35.399
## - eu                               1   0.13012 8.0190 -35.115
## - time_req_to_start_business_mean  1   0.17334 8.0622 -34.932
## - oil_production_per_cap_mean      1   0.25813 8.1470 -34.576
## - real_GDP_per_cap_2004            1   0.26176 8.1506 -34.561
## <none>                                         7.8888 -33.671
## - democracy_mean                   1   1.72328 9.6121 -28.953
##
## Step:  AIC=-35.48
## growth ~ oil_production_per_cap_mean + democracy_mean + electricity_per_cap_mean + 
##     time_req_to_start_business_mean + real_GDP_per_cap_2004 + 
##     eu
## 
##                                   Df Sum of Sq    RSS     AIC
## - electricity_per_cap_mean         1   0.08403 8.0172 -37.122
## - time_req_to_start_business_mean  1   0.18577 8.1189 -36.694
## - real_GDP_per_cap_2004            1   0.22434 8.1575 -36.533
## - eu                               1   0.23142 8.1646 -36.503
## - oil_production_per_cap_mean      1   0.23533 8.1685 -36.487
## <none>                                         7.9331 -35.481
## - democracy_mean                   1   1.77909 9.7122 -30.601
## 
## Step:  AIC=-37.12
## growth ~ oil_production_per_cap_mean + democracy_mean + time_req_to_start_business_mean + 
##     real_GDP_per_cap_2004 + eu
## 
##                                   Df Sum of Sq    RSS     AIC
## - eu                               1   0.15997 8.1771 -38.451
## - time_req_to_start_business_mean  1   0.18694 8.2041 -38.339
## - oil_production_per_cap_mean      1   0.22297 8.2401 -38.190
## <none>                                         8.0172 -37.122
## - real_GDP_per_cap_2004            1   0.50751 8.5247 -37.035
## - democracy_mean                   1   1.85257 9.8697 -32.054
## 
## Step:  AIC=-38.45
## growth ~ oil_production_per_cap_mean + democracy_mean + time_req_to_start_business_mean + 
##     real_GDP_per_cap_2004
## 
##                                   Df Sum of Sq     RSS     AIC
## - time_req_to_start_business_mean  1   0.12377  8.3009 -39.940
## - oil_production_per_cap_mean      1   0.47514  8.6523 -38.530
## <none>                                          8.1771 -38.451
## - real_GDP_per_cap_2004            1   0.76406  8.9412 -37.414
## - democracy_mean                   1   2.65530 10.8324 -30.890
## 
## Step:  AIC=-39.94
## growth ~ oil_production_per_cap_mean + democracy_mean + real_GDP_per_cap_2004
## 
##                               Df Sum of Sq     RSS     AIC
## - oil_production_per_cap_mean  1   0.48384  8.7847 -40.014
## <none>                                      8.3009 -39.940
## - real_GDP_per_cap_2004        1   1.09154  9.3924 -37.739
## - democracy_mean               1   2.96787 11.2688 -31.547
## 
## Step:  AIC=-40.01
## growth ~ democracy_mean + real_GDP_per_cap_2004
## 
##                         Df Sum of Sq     RSS     AIC
## <none>                                8.7847 -40.014
## - real_GDP_per_cap_2004  1    0.6625  9.4472 -39.542
## - democracy_mean         1    4.1740 12.9587 -28.796
```

```
aic$anova # View the steps
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean + 
##     democracy_mean + oil_production_per_cap_mean + electricity_per_cap_mean + 
##     time_req_to_start_business_mean + tourists_per_cap_mean + 
##     real_GDP_per_cap_2004 + eu
## 
## Final Model:
## growth ~ democracy_mean + real_GDP_per_cap_2004
## 
## 
##                                Step Df   Deviance Resid. Df Resid. Dev
## 1                                                        25   7.852483
## 2     - total_dependency_ratio_mean  1 0.03635162        26   7.888834
## 3           - tourists_per_cap_mean  1 0.04429672        27   7.933131
## 4        - electricity_per_cap_mean  1 0.08403480        28   8.017166
## 5                              - eu  1 0.15996706        29   8.177133
## 6 - time_req_to_start_business_mean  1 0.12376854        30   8.300901
## 7     - oil_production_per_cap_mean  1 0.48383773        31   8.784739
##         AIC
## 1 -31.82805
## 2 -33.67101
## 3 -35.48063
## 4 -37.12237
## 5 -38.45064
## 6 -39.93988
## 7 -40.01371
```

Description of the steps

Remove total_dependency_ratio_mean

```
m1 <- lm(growth ~ oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           electricity_per_cap_mean +
           time_req_to_start_business_mean +
           tourists_per_cap_mean +
           real_GDP_per_cap_2004 +
           eu, data = forest_tibble)
```

```
m1_glance <- m1 %>%
  glance()

m1_glance$r.squared ### Multiple R^2 = 0.5601
```

Model metrics

```
## [1] 0.5601263
```

```
m1_glance$adj.r.squared # Adjusted R^2 = 0.4417
```

```
## [1] 0.4416987
```

```
m1_glance$sigma # RSE 0.5508327
```

```
## [1] 0.5508327
```

```
m1_glance$statistic # F-statistic = 4.729696
```

```
##    value
## 4.729696
```

```
m1_glance$p.value ### p-value = 0.001565 < 0.05
```

```
##       value
## 0.001565157
```

```
m1_glance$AIC # Akaike Information Criterion
```

```
## [1] 64.81681
```

```
m1 %>%
  augment()
```

```
## # A tibble: 34 x 14
##    growth oil_pr~1 democ~2 elect~3 time_~4 touri~5 real_~6 eu    .fitted  .resid
##     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <lgl>   <dbl>   <dbl>
##  1 1.00         0     0.428   2230.   11.5     0.251    5389. FALSE  1.28    -0.278
##  2 0.236     1463.   0.686   7395.   23.5     2.67   39733. TRUE   0.122    0.114
##  3 4.28     36647.   0.186   2394.   63.2     0.193    3567. FALSE  2.19     2.09
##  4 0.784     2194.   0.192   3150.   46       0.368   11319. FALSE  1.92    -1.14
##  5 0.200        0     0.654   7212.   19.2     0.674   36638. TRUE   0.243   -0.0436
##  6 0.286     2771.   0.629   3132.   25.8    11.0     16701. TRUE   0.369   -0.0828
##  7 0.210    28858.   0.73    6580.    6.75    4.52    39557. TRUE  -0.0282  0.238
##  8 0.752        0     0.658   8519.   38.2     3.19    16688. TRUE   0.476    0.276
##  9 0.0804       0     0.663  14395.   17       0.522   38078. TRUE   0.144   -0.0635
## 10 0.157      219.   0.686   9183.    5.75    3.21    34926. TRUE   0.0540   0.103
## # ... with 24 more rows, 4 more variables: .hat <dbl>, .sigma <dbl>,
## #   .cooksd <dbl>, .std.resid <dbl>, and abbreviated variable names
## #   1: oil_production_per_cap_mean, 2: democracy_mean,
## #   3: electricity_per_cap_mean, 4: time_req_to_start_business_mean,
## #   5: tourists_per_cap_mean, 6: real_GDP_per_cap_2004
```

Cook's distance

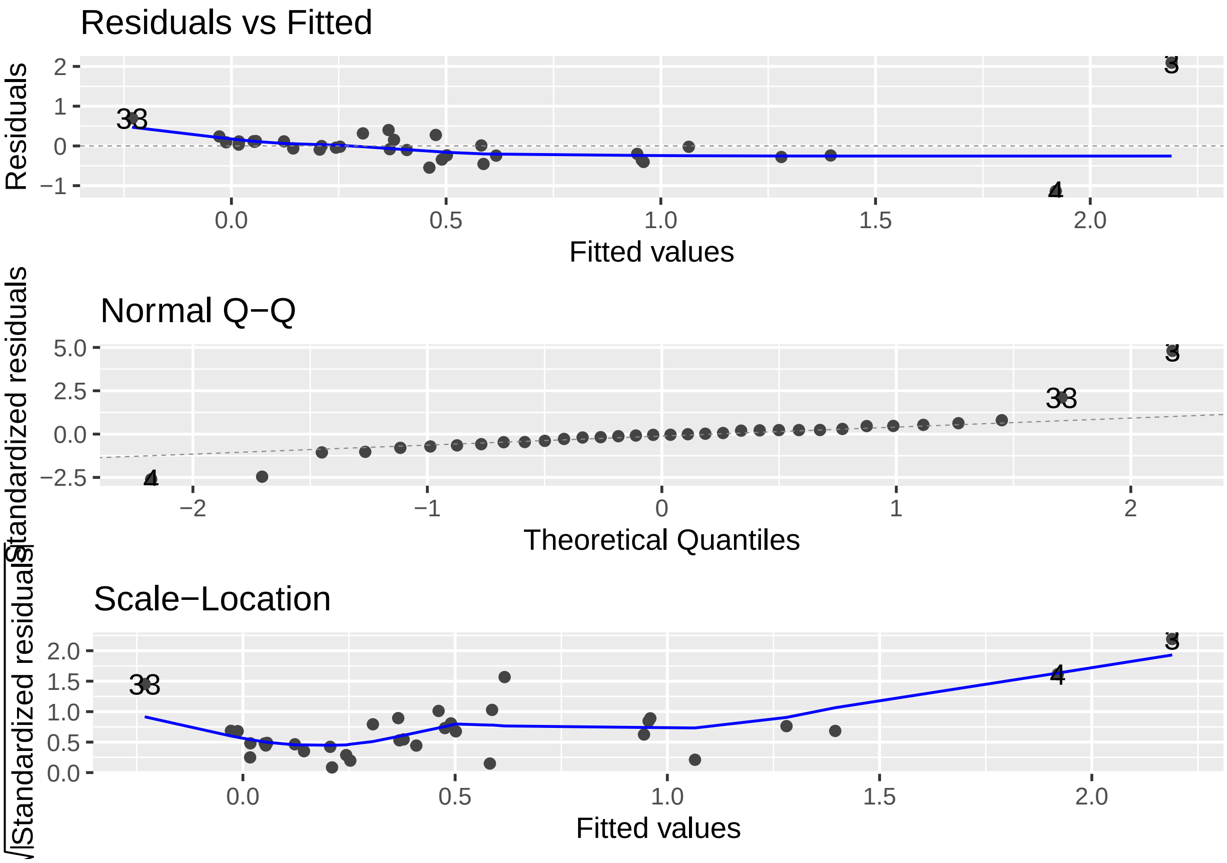
```
cooks.distance(m1)
```

```
##               1              2              3              4              5
##  0.013779717297 0.000501407440 1.720035339904 0.506416624185 0.000075355189
##               6              7              8              9             10
##  0.024134879438 0.004892147962 0.004794189828 0.000270686963 0.000573370596
##              11             12             13             14             15
##  0.000021493950 0.008387172053 0.000462488699 0.181553099831 0.001936031162
##              16             17             18             19             20
##  0.000045593895 0.010293115767 0.017214988324 0.003501582672 0.000063537590
##                21                22                23                24                25
##   0.002925211375   0.000001434352   0.013422185765  22.483376750956   0.000009479385
##                26                27                28                29                30
##   0.004970494391   0.003995279499   0.013319878111   0.001788147270   0.001181670786
##                31                32                33                34
##   0.090711803518   0.001106539492   0.994811357426   0.000531034331
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 22.13, the highest recorded. This observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m1, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m1)$p.value < 0.05 # We can reject the homoskedasticity
```

```
##   BP
## TRUE
```

Findings:

1. Residuals versus fitted: Although observation 3, 4, 36 slightly distort the curve, it is almost horizontal.
2. Q-Q: Residuals have a S-like distribution.
3. Scale-location: Heteroskedasticity is still present but it is smaller in comparison to m0.

4. The independent variables explain 0.4417 = 44.17% of the variation in the dependent variable.

```
vif(m1) # There is moderate (VIF < 5) correlation between the IVs
```

Variance inflation factor

```
##     oil_production_per_cap_mean                  democracy_mean
##                        1.556431                        1.874595
##        electricity_per_cap_mean time_req_to_start_business_mean
##                        1.890089                        1.445817
##           tourists_per_cap_mean           real_GDP_per_cap_2004
##                        1.259100                        3.105230
##                              eu
##                        2.335082
```

Remove tourists__per__cap__mean

```
m2 <- lm(growth ~ oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           electricity_per_cap_mean +
           time_req_to_start_business_mean +
           real_GDP_per_cap_2004 +
           eu,
         data = forest_tibble)
```

```
m2_glance <- m2 %>%
  glance()

m2_glance$r.squared # Multiple R^2 = 0.5577038
```

Model metrics

```
## [1] 0.5576563
```

```
m2_glance$adj.r.squared # Adjusted R^2 = 0.4594
```

```
## [1] 0.4593577
```

```
m2_glance$sigma # RSE = 0.5420223  on 27 DoF
```

```
## [1] 0.5420514
```

```
m2_glance$statistic # F-statistic = 5.673 on 6 and 27 DoF
```

```
##    value
## 5.673086
```

```
m2_glance$p.value # p-value = 0.0006433 < 0.05
```

```
##        value
## 0.0006433372
```

```
m2 %>%
  augment()
```

```
## # A tibble: 34 x 13
##    growth oil_pro~1 democ~2 elect~3 time_~4 real_~5 eu    .fitted  .resid   .hat
##     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <lgl>   <dbl>   <dbl>  <dbl>
##  1 1.00          0    0.428   2230.   11.5    5389. FALSE  1.27   -0.267  0.243
##  2 0.236       1463.   0.686   7395.   23.5   39733. TRUE   0.139   0.0969 0.0727
##  3 4.28       36647.   0.186   2394.   63.2    3567. FALSE  2.19    2.10   0.373
##  4 0.784       2194.   0.192   3150.   46     11319. FALSE  1.93   -1.15   0.372
##  5 0.200          0    0.654   7212.   19.2   36638. TRUE   0.218  -0.0181 0.0667
##  6 0.286       2771.   0.629   3132.   25.8   16701. TRUE   0.535  -0.249  0.0903
##  7 0.210      28858.   0.73    6580.    6.75  39557. TRUE   0.0194  0.191  0.100
##  8 0.752          0    0.658   8519.   38.2   16688. TRUE   0.481   0.271  0.118
##  9 0.0804         0    0.663  14395.   17     38078. TRUE   0.109  -0.0283 0.0968
## 10 0.157        219.   0.686   9183.    5.75  34926. TRUE   0.0720  0.0853 0.0972
## # ... with 24 more rows, 3 more variables: .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, and abbreviated variable names
## #   1: oil_production_per_cap_mean, 2: democracy_mean,
## #   3: electricity_per_cap_mean, 4: time_req_to_start_business_mean,
## #   5: real_GDP_per_cap_2004
```

Cook's distance

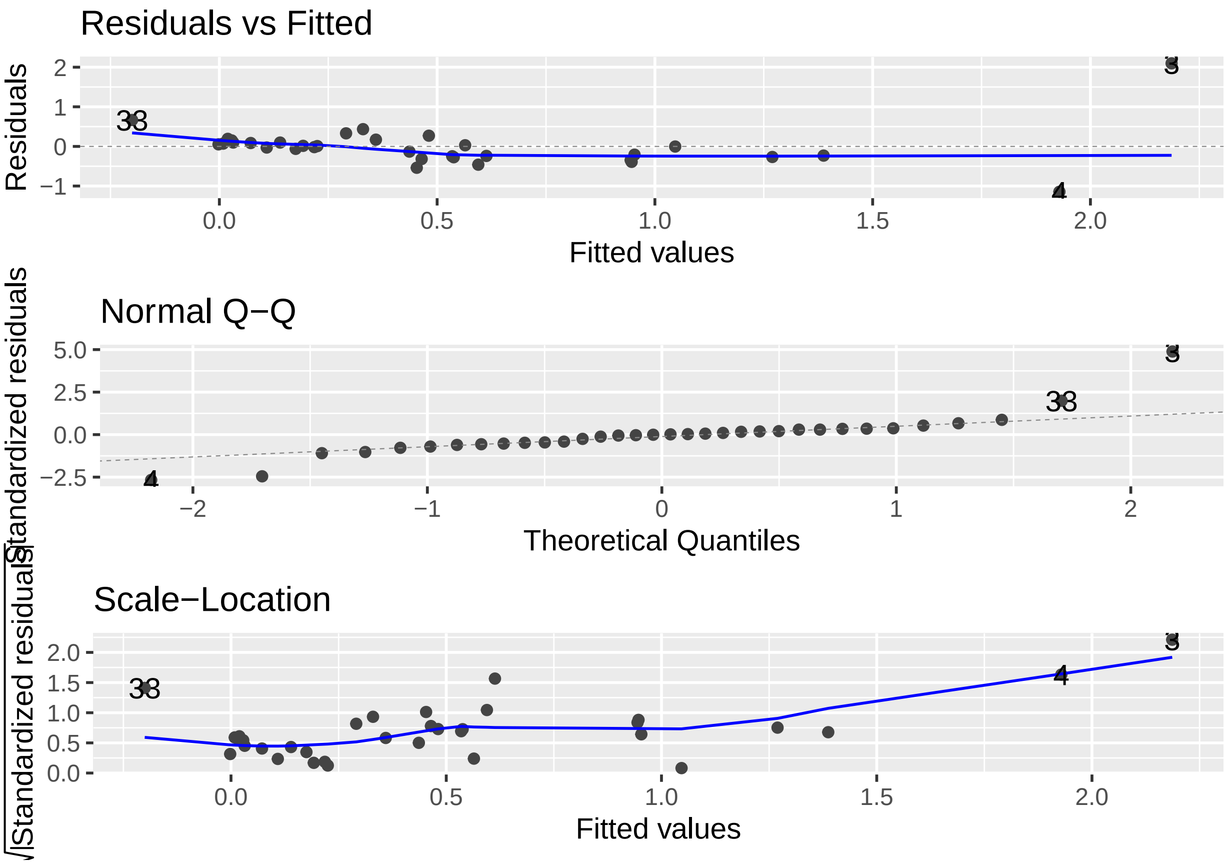
```
cooks.distance(m2)
```

```
##               1               2               3               4               5
##  0.014715381221  0.000386482180  2.033370731204  0.600691084156  0.000012228675
##               6               7               8               9              10
##  0.003282196632  0.002179864742  0.005418193634  0.000046221201  0.000422084989
##              11              12              13              14              15
##  0.000003225755  0.009370256972  0.000677386275  0.108343164932  0.001689895610
##              16              17              18              19              20
##  0.000119574377  0.012784197900  0.018787601800  0.003837212283  0.000001563404
##              21              22              23              24              25
##  0.003685068223  0.000024850599  0.014485371365 25.371081877915  0.000067879105
##              26              27              28              29              30
##  0.004021593695  0.004334475995  0.014165727678  0.002472151652  0.001821139075
##              31              32              33              34
##  0.109783414655  0.001651430642  0.922360049623  0.000221521779
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 25.35, the highest recorded. This observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m2, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m2)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: Although observation 3, 4, 36 slightly distort the curve, it is almost horizontal.
2. Q-Q: Residuals still have a S-like distribution
3. Scale-location: The data is less heteroskedastic than the previous models
4. The independent variables explain 0.4594 = 45.94% of the variation in the dependent variable.

```
vif(m2)
```

Variance inflation factor

```
##     oil_production_per_cap_mean                    democracy_mean
##                        1.530913                          1.862619
##        electricity_per_cap_mean time_req_to_start_business_mean
##                        1.839893                          1.439646
##           real_GDP_per_cap_2004                                eu
##                        2.903884                          1.979227
```

There is moderate (VIF < 5) correlation between the IVs

Remove electricity__per__cap__mean

```
m3 <- lm(growth ~ oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           time_req_to_start_business_mean +
           real_GDP_per_cap_2004 +
           eu, data = forest_tibble)
```

```

m3_glance <- m3 %>%
  glance()

m3_glance$sigma # RSE = 0.535176 on 28 DoF
```

Model metrics

```
## [1] 0.5350956
```

```
m3_glance$r.squared # Multiple R^2 = 0.553
```

```
## [1] 0.5529706
```

```
m3_glance$adj.r.squared # Adjusted m^2 = 0.4731
```

```
## [1] 0.473144
```

```
m3_glance$statistic # F-statistic = 6.927 on 5 and 28 DoF
```

```
##    value
## 6.927141
```

```
m3_glance$p.value # p-value = 0.0002538 < 0.05
```

```
##        value
## 0.0002537722
```

```
m3 %>%
  augment()
```

```
## # A tibble: 34 x 12
##    growth oil_pro~1 democ~2 time_~3 real_~4 eu    .fitted   .resid   .hat .sigma
##     <dbl>     <dbl>   <dbl>   <dbl>   <dbl> <lgl>   <dbl>    <dbl>  <dbl>  <dbl>
##  1 1.00           0   0.428   11.5    5389. FALSE  1.25   -0.249   0.239   0.542
##  2 0.236      1463.   0.686   23.5   39733. TRUE   0.119   0.117   0.0677  0.544
##  3 4.28      36647.   0.186   63.2    3567. FALSE  2.18    2.10    0.373   0.190
##  4 0.784      2194.   0.192   46     11319. FALSE  1.91   -1.13    0.368   0.471
##  5 0.200         0    0.654   19.2   36638. TRUE   0.206  -0.00596 0.0649  0.545
##  6 0.286      2771.   0.629   25.8   16701. TRUE   0.543  -0.257   0.0895  0.542
##  7 0.210     28858.   0.73     6.75  39557. TRUE  -0.0115  0.221   0.0887  0.543
##  8 0.752         0    0.658   38.2   16688. TRUE   0.537   0.215   0.0816  0.543
##  9 0.0804        0    0.663   17     38078. TRUE   0.156  -0.0754  0.0704  0.545
## 10 0.157       219.   0.686    5.75  34926. TRUE   0.0806  0.0767  0.0964  0.545
## # ... with 24 more rows, 2 more variables: .cooksd <dbl>, .std.resid <dbl>, and
## #   abbreviated variable names 1: oil_production_per_cap_mean,
## #   2: democracy_mean, 3: time_req_to_start_business_mean,
## #   4: real_GDP_per_cap_2004
```

Cook's distance

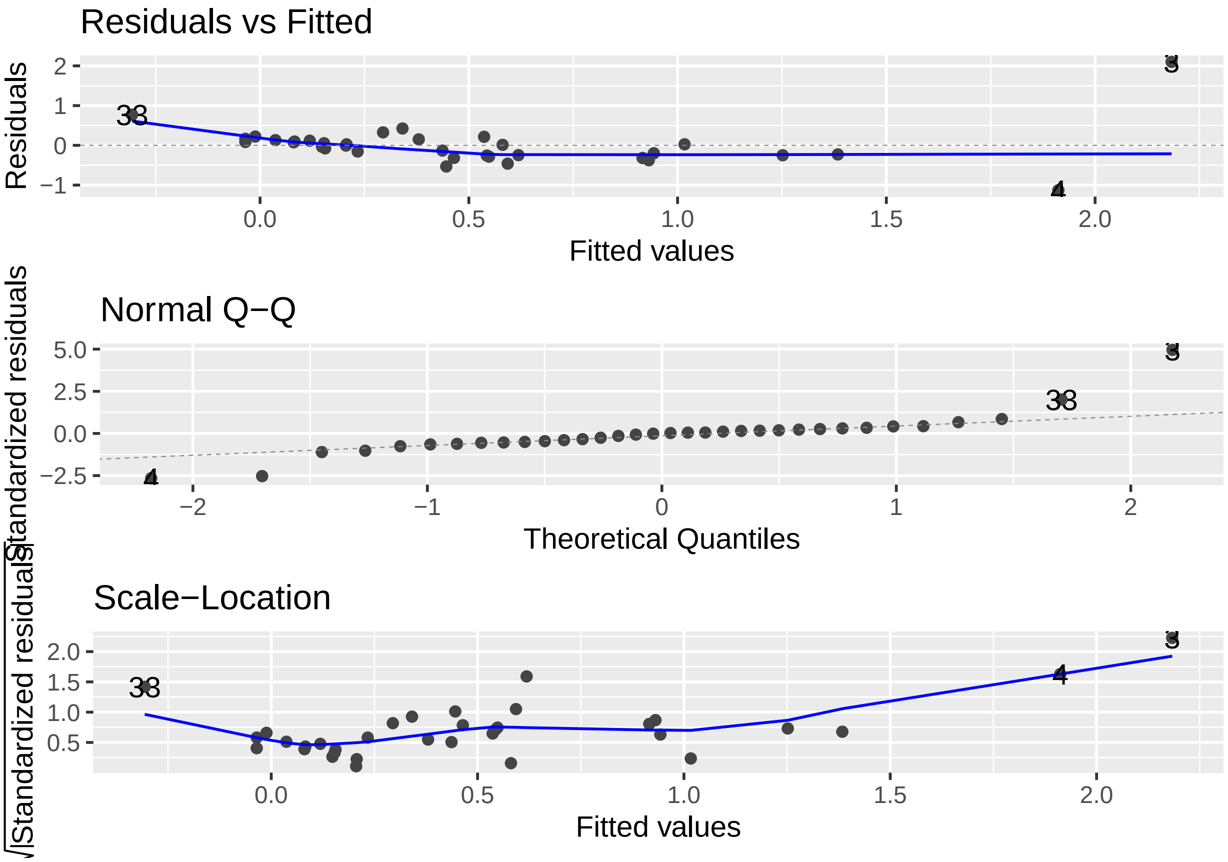
```
cooks.distance(m3)
```

```
##                1                2                3                4                5 
##   0.014945117934   0.000625763186   2.439903837700   0.684823775289   0.000001534721 
##                6                7                8                9               10 
##   0.004155535267   0.003046894163   0.002601510590   0.000269548797   0.000403977400 
##               11               12               13               14               15 
##   0.000037531489   0.010791502899   0.000817554110   0.006580996001   0.003815500192 
##               16               17               18               19               20 
##   0.000312661566   0.014941029477   0.021051471625   0.004863011963   0.000116464987 
##               21               22               23               24               25 
##   0.003947285596   0.000360151791   0.013503721106  31.079223048723   0.000014208858 
##               26               27               28               29               30 
##   0.004812195130   0.005028077304   0.015222767132   0.002172044747   0.001615713477 
##               31               32               33               34 
##   0.130656870029   0.000520528859   0.636232372572   0.000076583669
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 31.07, the highest recorded. This observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m3, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m3)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: The line has a steep angle before 0.0 on the axis. We can say that this model is inferior to the previous models.
2. Q-Q: It has a more prominent S-shape than the previous ones .
3. Scale-location: This is the least homoskedastic model we have examined so far.
4. The independent variables explain 0.4731 = 47.31% of the variation in the dependent variable.

Variance inflation factor

```
vif(m3)
```

```
##     oil_production_per_cap_mean                  democracy_mean
##                        1.527637                        1.850913
## time_req_to_start_business_mean           real_GDP_per_cap_2004
##                        1.439615                        2.184656
##                              eu
##                        1.681792
```

There is moderate (VIF < 5) correlation between the IVs

Remove eu

```
m4 <- lm(growth ~ oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           time_req_to_start_business_mean +
           real_GDP_per_cap_2004, data = forest_tibble)
```

```


m4_glance <- m4 %>%
  glance()

m4_glance$sigma # RSE = 0.531 on 29 DoF
```

Model metrics

```
## [1] 0.5310086
```

```
m4_glance$r.squared # Multiple R^2 = 0.5441
```

```
## [1] 0.544051
```

```
m4_glance$adj.r.squared # Adjusted R^2 = 0.4812
```

```
## [1] 0.4811615
```

```
m4_glance$statistic # F-statistic = 8.651 on 4 and 29 DoF
```

```
##    value
## 8.650902
```

```
m4_glance$p.value # p-value = 0.0001007 < 0.05
```

```
##        value
## 0.0001007223
```

```
m4 %>%
  augment()
```

```
## # A tibble: 34 x 11
##    growth oil_pr~1 democ~2 time_~3 real_~4 .fitted  .resid   .hat .sigma .cooksd
##     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>
##  1 1.00          0   0.428   11.5    5389.  1.23   -0.224  0.235   0.538 1.43e-2
##  2 0.236      1463.   0.686   23.5   39733.  0.136   0.101  0.0659  0.540 5.42e-4
##  3 4.28      36647.   0.186   63.2    3567.  2.19    2.09   0.373   0.207 2.94e+0
##  4 0.784      2194.   0.192   46     11319.  1.90   -1.11   0.367   0.471 8.09e-1
##  5 0.200         0   0.654   19.2   36638.  0.241  -0.0411 0.0572  0.540 7.71e-5
##  6 0.286      2771.   0.629   25.8   16701.  0.620  -0.334  0.0524  0.536 4.63e-3
##  7 0.210     28858.   0.73     6.75  39557.  0.0321  0.178  0.0768  0.539 2.02e-3
##  8 0.752         0   0.658   38.2   16688.  0.591   0.161  0.0633  0.539 1.33e-3
##  9 0.0804        0   0.663   17     38078.  0.188  -0.108  0.0638  0.540 6.00e-4
## 10 0.157       219.   0.686    5.75  34926.  0.126   0.0315 0.0836  0.540 6.98e-5
## # ... with 24 more rows, 1 more variable: .std.resid <dbl>, and abbreviated
## #   variable names 1: oil_production_per_cap_mean, 2: democracy_mean,
## #   3: time_req_to_start_business_mean, 4: real_GDP_per_cap_2004
```

Cook's distance

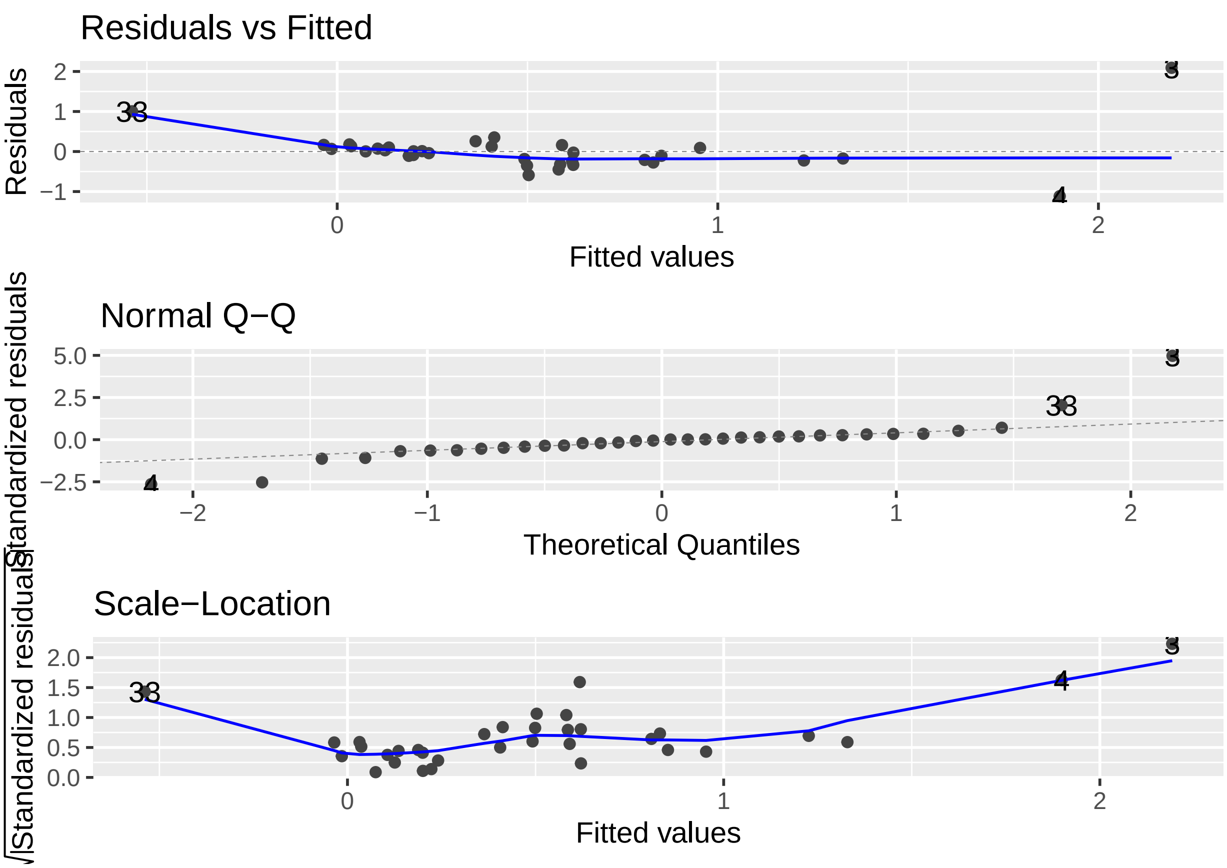
```
cooks.distance(m4)
```

```
##               1               2               3               4               5
##  0.014273309821  0.000541728948  2.938933881889  0.808723241143  0.000077067760
##               6               7               8               9              10
##  0.004625681669  0.002022637242  0.001325340896  0.000599903146  0.000069833468
##              11              12              13              14              15
##  0.000006597230  0.010050438064  0.001403129924  0.000001338224  0.004654853910
##              16              17              18              19              20
##  0.000211501088  0.008667187273  0.012868244871  0.006695596601  0.001323713551
##              21              22              23              24              25
##  0.000751073595  0.000004734636  0.003305385134 37.403971637825  0.000077066784
##              26              27              28              29              30
##  0.006350570716  0.002884472391  0.004677549421  0.001736288762  0.001972248359
##              31              32              33              34
##  0.150752285108  0.000341662552  0.145643409373  0.000440568940
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 37.42, the highest recorded. This observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m4, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m4)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: The distortion before the 0 on the x-axis is increased.
2. Q-Q: The standardized residual still constitute an S-shape. However, the residuals between the 1st and the 2nd quartiles follow a more normal distribution than those of m3.
3. Scale-location: The line is relatively horizontal between 0 and 1 on the x-axis is horizontal. However, the data is still very heteroskedastic.
4. The independent variables explain 0.4812 = 48.12% of the variation in the dependent variable.

Variance inflation factor

```
vif(m4)
```

```
##     oil_production_per_cap_mean                  democracy_mean
##                        1.283962                        1.591261
## time_req_to_start_business_mean           real_GDP_per_cap_2004
##                        1.367563                        1.985210
```

There is moderate (VIF < 5) correlation between the IVs

Apparently, we can explain the growth without using the EU membership data.

This finding refutes what we found in our first project.

Remove time_req_to_start_business_mean

```
m5 <- lm(growth ~ oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           real_GDP_per_cap_2004, data = forest_tibble)
```

Model metrics

```
m5_glance <- m5 %>%
  glance()

m5_glance$sigma # RSE = 0.526 on 30 DoF
```

```
## [1] 0.5260197
```

```
m5_glance$r.squared # Multiple R^2 = 0.5371
```

```
## [1] 0.5371498
```

```
m5_glance$adj.r.squared # Adjusted R^2 = 0.4909
```

```
## [1] 0.4908648
```

```
m5_glance$statistic # F-statistic = 11.61 on 3 and 30 DoF
```

```
##    value
## 11.60526
```

```
m5_glance$p.value # p-value = 0.00003229 < 0.05
```

```
##         value
## 0.00003228842
```

```
m5 %>% augment()
```

```
## # A tibble: 34 x 10
##    growth oil_pr~1 democ~2 real_~3 .fitted  .resid   .hat .sigma .cooksd .std.~4
##     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1 1.00          0    0.428   5389.  1.35   -0.349  0.109    0.531 1.51e-2 -0.702
##  2 0.236     1463.    0.686  39733.  0.121   0.115  0.0643   0.535 8.77e-4  0.226
##  3 4.28     36647.    0.186   3567.  2.13    2.15   0.341    0.208 3.29e+0  5.04
##  4 0.784     2194.    0.192  11319.  1.89   -1.11   0.367    0.468 1.02e+0 -2.65
##  5 0.200         0    0.654  36638.  0.253  -0.0533 0.0560   0.535 1.61e-4 -0.104
##  6 0.286     2771.    0.629  16701.  0.644  -0.358  0.0477   0.531 6.10e-3 -0.698
##  7 0.210    28858.    0.73   39557.  0.0843  0.126  0.0548   0.534 8.76e-4  0.246
##  8 0.752         0    0.658  16688.  0.560   0.192  0.0555   0.534 2.07e-3  0.375
##  9 0.0804        0    0.663  38078.  0.206  -0.126  0.0612   0.534 9.94e-4 -0.247
## 10 0.157      219.    0.686  34926.  0.194  -0.0371 0.0456   0.535 6.23e-5 -0.0722
## # ... with 24 more rows, and abbreviated variable names
## #   1: oil_production_per_cap_mean, 2: democracy_mean,
## #   3: real_GDP_per_cap_2004, 4: .std.resid
```

Cook’s distance

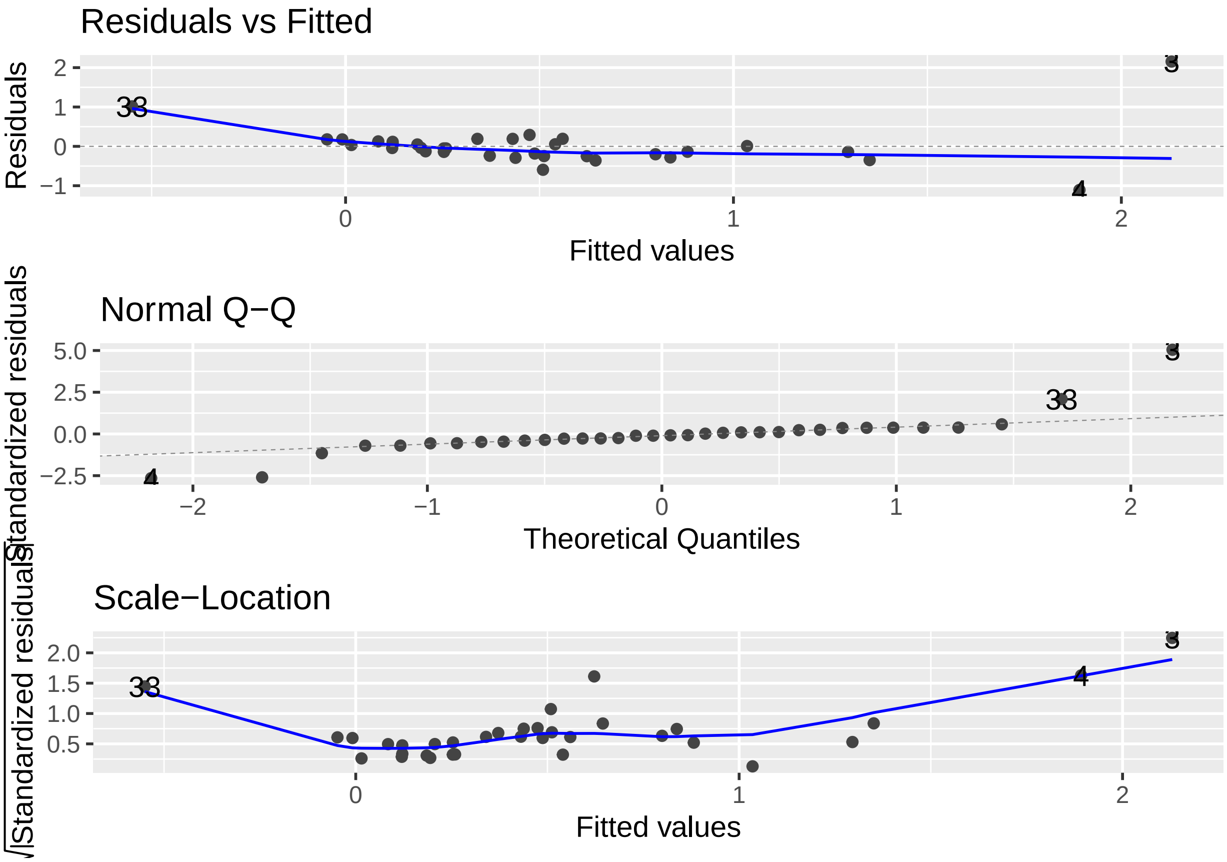
```
cooks.distance(m5)
```

```
##              1              2              3              4              5
##   0.01511340755  0.00087662690  3.29022531880  1.01636318303  0.00016149557
##              6              7              8              9             10
##   0.00610261472  0.00087566526  0.00207229412  0.00099357757  0.00006226195
##              11             12             13             14             15
##   0.00016434756  0.01295573311  0.00169858230  0.00015519790  0.00679303376
##              16             17             18             19             20
##   0.00006744632  0.00408297309  0.00763659046  0.00196513034  0.00000859539
##              21             22             23             24             25
##   0.00139762763  0.00036708089  0.00382764749 48.91649220558  0.00017717889
##              26             27             28             29             30
##   0.00277268782  0.00214993449  0.00622353972  0.00341365158  0.00379368977
##              31             32             33             34
##   0.00191486513  0.00026489942  0.18794894799  0.00093838314
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Norway has a distance of 48.93, the highest recorded. This observation would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m5, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m5)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: The line follows a linear path after the 0.0 point on the x-axis. However, it does not have pattern, which is good for our assumption of normality.
2. Q-Q: The standardized residuals are more normal relative to m5.
3. Scale-location: It is not much different than the m4
4. The independent variables explain 0.4909 = 49.09% of the variation in the dependent variable.

```
vif(m5)
```

Variance inflation factor

```
## oil_production_per_cap_mean              democracy_mean
##                    1.283566                    1.538171
##       real_GDP_per_cap_2004
##                    1.776227
```

There is moderate (VIF < 5) correlation between the IVs

Remove oil_production_per_cap_mean

```
m6 <- lm(growth ~ democracy_mean + real_GDP_per_cap_2004, data = forest_tibble)
```

```
m6_glance <-m6 %>%
  glance()

m6_glance$sigma # RSE = 0.5323 on 31 DoF
```

Model metrics

```
## [1] 0.5323332
```

```
m6_glance$r.squared # Multiple R^2 = 0.5102
```

```
## [1] 0.5101715
```

```
m6_glance$adj.r.squared # Adjusted R^2 = 0.4786
```

```
## [1] 0.4785697
```

```
m6_glance$statistic # F-statistic = 16.14 on 2 and 31 DoF
```

```
##    value
## 16.14373
```

```
m6_glance$p.value # p-value = 0.00001569 < 0.05
```

```
##         value
## 0.00001569216
```

```
m6 %>%
  augment()
```

```
## # A tibble: 34 x 9
##    growth democracy_mean real_GD~1 .fitted  .resid   .hat .sigma .cooksd .std.~2
##     <dbl>          <dbl>     <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1 1.00            0.428     5389.   1.35  -0.346  0.109   0.537 1.94e-2 -0.689
##  2 0.236           0.686    39733.   0.198  0.0385 0.0522  0.541 1.01e-4  0.0743
##  3 4.28            0.186     3567.   2.10   2.19   0.339   0.228 4.36e+0  5.05
##  4 0.784           0.192    11319.   1.99  -1.21   0.346   0.467 1.39e+0 -2.81
##  5 0.200           0.654    36638.  0.329  -0.129  0.0440  0.541 9.50e-4 -0.249
##  6 0.286           0.629    16701.  0.623  -0.337  0.0467  0.537 6.85e-3 -0.648
##  7 0.210           0.73     39557.  0.0692  0.141  0.0544  0.540 1.42e-3  0.272
##  8 0.752           0.658    16688.  0.536   0.216  0.0544  0.540 3.33e-3  0.417
##  9 0.0804          0.663    38078.  0.287  -0.206  0.0479  0.540 2.64e-3 -0.397
## 10 0.157           0.686    34926.  0.251  -0.0932 0.0391  0.541 4.33e-4 -0.179
## # ... with 24 more rows, and abbreviated variable names
## #   1: real_GDP_per_cap_2004, 2: .std.resid
```

Cook's distance

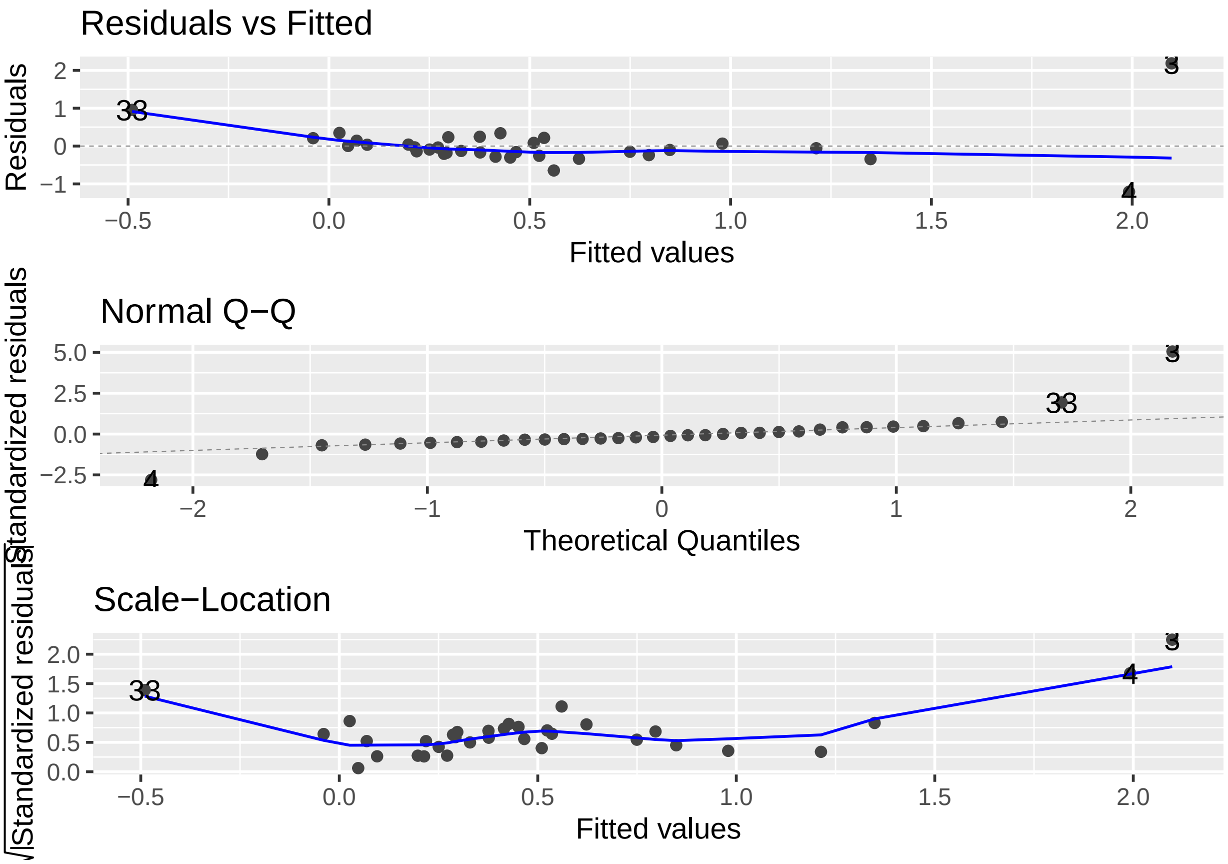
```
cooks.distance(m6)
```

```
##              1              2              3              4              5
## 0.0193799008890 0.0001012836383 4.3560203544552 1.3915876382100 0.0009501640473
##              6              7              8              9             10
## 0.0068482002549 0.0014160545048 0.0033324252223 0.0026390057385 0.0004326824645
##             11             12             13             14             15
## 0.0001058748416 0.0168429199960 0.0016840156922 0.0016836643411 0.0002225714052
##             16             17             18             19             20
## 0.0000002689307 0.0081697313766 0.0125706309321 0.0027860804168 0.0006002448151
##             21             22             23             24             25
## 0.0010124777393 0.0035254583793 0.0026381551528 0.0559377186498 0.0005481461076
##             26             27             28             29             30
## 0.0039012807148 0.0004022577577 0.0056061558811 0.0063172220979 0.0066612304897
##             31             32             33             34
## 0.0030653677449 0.0000969818326 0.2018104939435 0.0018544282407
```

Findings: We used Cook's distance, find outliers that would distort our regression model. Azerbaijan and Belarus have distances of 4.35 and 1.39 respectively, highest recorded. These observations would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m6, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m6)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: There is almost no difference when compared to the previous model.
2. Q-Q: The standardized residuals are distributed more normally than those of m5.
3. Scale-location: The distribution of the standardized residuals still suggest heteroskedasticity which violates our homoskedasticity assumption.
4. The independent variables explain 0.4786 = 47.86% of the variation in the dependent variable.

Variance inflation factor

```
vif(m6)
```

```
##        democracy_mean real_GDP_per_cap_2004 
##              1.387538              1.387538
```

Multicollinearity does not exist (VIF = 1.387538 < 1.5 < 5)

Although the model with real_GDP_per_cap_2004 has more explanatory power, its statistical significance is low. Perhaps, real_GDP_per_cap_2004 and democracy_mean interact. We will try one last model.

```
m7 <- lm(growth ~ democracy_mean * real_GDP_per_cap_2004, data = forest_tibble)
```

```
m7_glance <- m7 %>%
  glance()

m7_glance$sigma # RSE = 0.421 on 30 DoF
```

Model metrics

```
## [1] 0.4210164
```

```
m7_glance$r.squared # Multiple R^2 = 0.7035
```

```
## [1] 0.7034933
```

```
m7_glance$adj.r.squared # Adjusted R^2 = 0.6738
```

```
## [1] 0.6738427
```

```
m7_glance$statistic # F-statistic = 23.73 on 3 and 30 DoF
```

```
##    value
## 23.72605
```

```
m7_glance$p.value # p-value = 0.0000000458 < 0.05
```

```
##            value
## 0.00000004580205
```

```
m7 %>%
  augment()
```

```
## # A tibble: 34 x 9
##    growth democracy_mean real_GD~1 .fitted  .resid   .hat .sigma .cooksd .std.~2
##     <dbl>          <dbl>     <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1 1.00            0.428     5389.   1.59  -0.588  0.126   0.412 8.07e-2  -1.49
##  2 0.236           0.686    39733.   0.183  0.0530 0.0522  0.428 2.30e-4   0.129
##  3 4.28            0.186     3567.   2.88    1.40   0.516   0.208 6.11e+0   4.79
##  4 0.784           0.192    11319.   2.03  -1.25   0.347   0.317 1.79e+0  -3.68
##  5 0.200           0.654    36638.   0.151  0.0491 0.0532  0.428 2.02e-4   0.120
##  6 0.286           0.629    16701.   0.476 -0.190  0.0529  0.427 3.01e-3  -0.464
##  7 0.210           0.73     39557.   0.283 -0.0727 0.0675  0.428 5.79e-4  -0.179
##  8 0.752           0.658    16688.   0.410  0.341  0.0589  0.423 1.09e-2   0.835
##  9 0.0804          0.663    38078.   0.149 -0.0686 0.0533  0.428 3.95e-4  -0.167
## 10 0.157           0.686    34926.   0.217 -0.0600 0.0394  0.428 2.17e-4  -0.145
## # ... with 24 more rows, and abbreviated variable names
## #   1: real_GDP_per_cap_2004, 2: .std.resid
```

Cook's distance

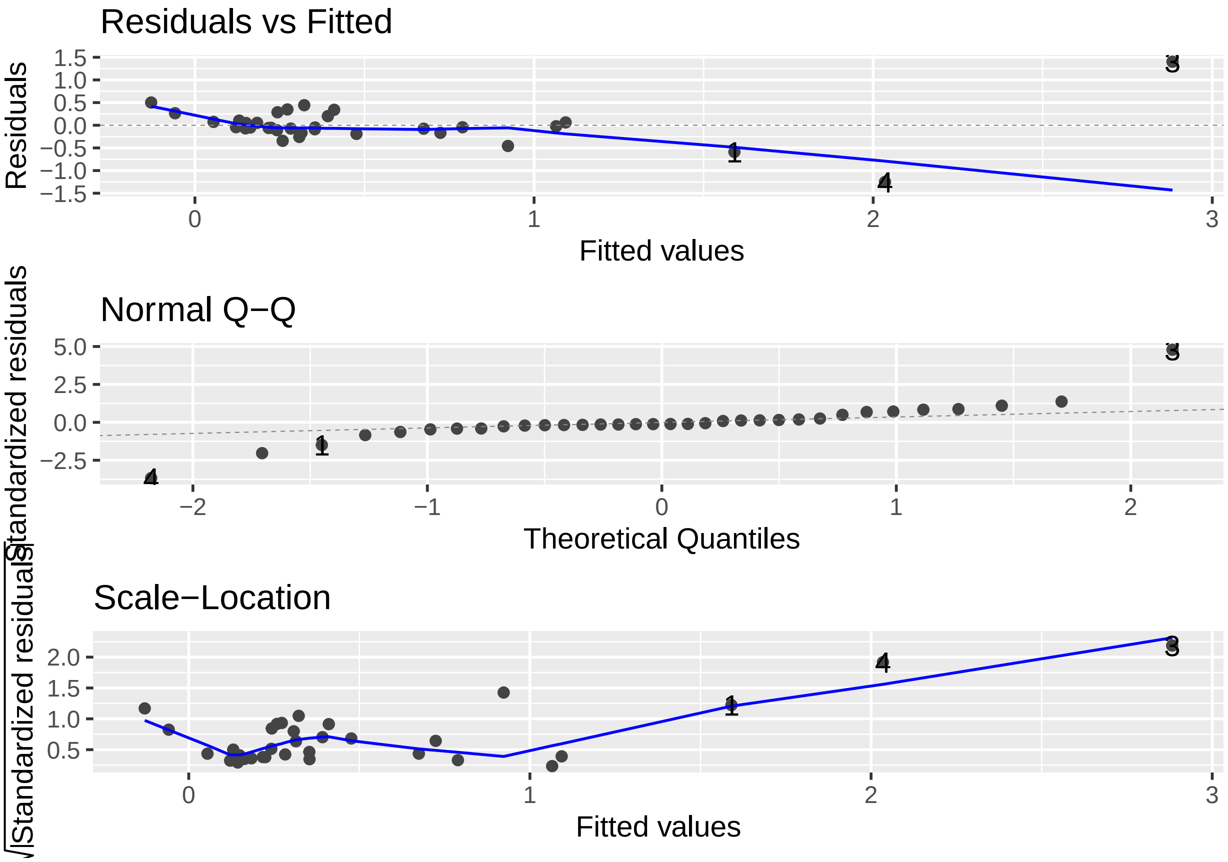
```
cooks.distance(m7)
```

```
##            1            2            3            4            5            6
## 0.0806513669 0.0002300215 6.1137058517 1.7944485378 0.0002017063 0.0030106526
##            7            8            9           10           11           12
## 0.0005791418 0.0109106306 0.0003947103 0.0002165590 0.0009929796 0.0110044891
##           13           14           15           16           17           18
## 0.0002010575 0.0001955208 0.0012753294 0.0079491529 0.0205367374 0.0270623846
##           19           20           21           22           23           24
## 0.0004994801 0.0000864971 0.0002237839 0.0189415581 0.0008249085 0.1475265966
##            25           26           27           28           29           30
## 0.0041460685 0.0016585251 0.0005720242 0.0033492977 0.0116203251 0.0007404913
##            31           32           33           34
## 0.0007163103 0.0001177348 2.5964796022 0.0001940462
```

Findings: We used Cook's distance to find outliers that would distort our regression model. Azerbaijan and Belarus have distances of 6.11 and 1.79 respectively, the highest recorded. These observations would negatively affect our model significantly.

Visualize the model metrics

```
autoplot(m7, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m7)$p.value < 0.05
```

```
##   BP
## TRUE
```

We can reject the homoskedasticity

Findings:

1. Residuals versus fitted: The line is not horizontal at all.
2. Q-Q: The line that represents standardized residuals is extremely distorted.
3. Scale-location: The line behaves in a zig-zag pattern and not horizontal.
4. This model has less explanatory power than all of the previous models.

The independent variables explain 0.6738 = 67.38% of the variation in the dependent variable.

Shapiro-Wilk test

```
shapiro.test(m7$residuals)$p.value < 0.05
```

```
## [1] TRUE
```

Null-hypothesis: distribution is normal **p-value** = 0.0003711 Normality of residuals is rejected

Perhaps we need to remove some variables

```
forest_tibble_2 <- forest_tibble[-c(3, 4, 33),] # Azerbaijan, Belarus, Switzerland are removed
```

Now call the linear model again

```
m8 <- lm(growth ~ democracy_mean * real_GDP_per_cap_2004, data = forest_tibble_2)
```

```


m8_glance <- m8 %>% glance()

m8_glance$sigma # RSE = 0.2017 on 27 DoF
```

Model metrics

```
## [1] 0.2016833
```

```
m8_glance$r.squared # Multiple R^2 = 0.7034933
```

```
## [1] 0.6462473
```

```
m8_glance$adj.r.squared # Adjusted R^2 = 0.6738427
```

```
## [1] 0.6069414
```

```
m8_glance$statistic # F-statistic = 16.44 on 3 and 27 DoF
```

```
##   value
## 16.4415
```

```
m8_glance$p.value # p-value = 0.000002819 < 0.05
```

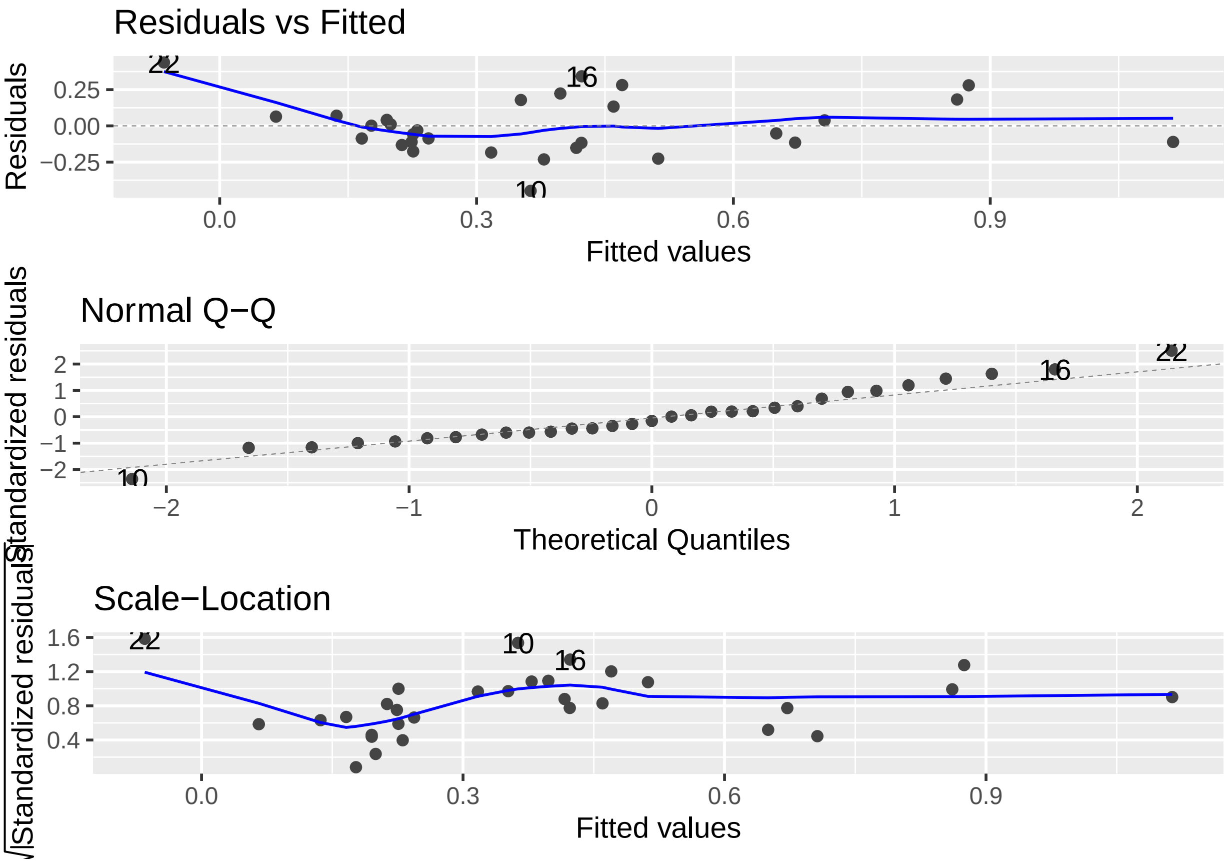
```
##          value
## 0.000002818568
```

```
m8 %>% augment()
```

```
## # A tibble: 31 x 9
##     growth democracy_mean real_G~1 .fitted  .resid   .hat .sigma .cooksd .std.~2
##      <dbl>          <dbl>    <dbl>   <dbl>   <dbl>  <dbl>  <dbl>   <dbl>   <dbl>
##  1  1.00            0.428    5389.   1.11  -0.111  0.543   0.203 1.97e-1 -0.814
##  2  0.236           0.686   39733.   0.195  0.0412 0.0672  0.205 8.06e-4  0.211
##  3  0.200           0.654   36638.   0.231 -0.0309 0.0587  0.205 3.89e-4 -0.158
##  4  0.286           0.629   16701.   0.512 -0.226  0.0633  0.200 2.27e-2 -1.16
##  5  0.210           0.73    39557.   0.200  0.0102 0.200   0.206 1.99e-4  0.0564
##  6  0.752           0.658   16688.   0.470  0.282  0.0684  0.197 3.84e-2  1.45
##  7  0.0804          0.663   38078.   0.213 -0.132  0.0545  0.204 6.57e-3 -0.675
##  8  0.157           0.686   34926.   0.244 -0.0866 0.0472  0.205 2.39e-3 -0.440
##  9  0.233           0.663   39574.   0.195  0.0376 0.0606  0.205 5.98e-4  0.193
## 10 -0.0856          0.607   28452.   0.363 -0.449  0.112   0.183 1.75e-1 -2.36
## # ... with 21 more rows, and abbreviated variable names
## #   1: real_GDP_per_cap_2004, 2: .std.resid
```

Visualize the model metrics

```
autoplot(m8, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m8)$p.value < 0.05
```

```
##    BP
## FALSE
```

We cannot reject the homoskedasticity

Cook's distance

```
cooks.distance(m8)
```

```
##               1               2               3               4               5
## 0.1966334869020 0.0008055953964 0.0003887392504 0.0226693775610 0.0001990438982
##               6               7               8               9              10
## 0.0384190161127 0.0065670088398 0.0023943911374 0.0005981939943 0.1753807348806
##              11              12              13              14              15
## 0.0060335114036 0.0037652044105 0.0050581803838 0.0767462978696 0.0578926199086
##              16              17              18              19              20
## 0.0982300102414 0.0083848642159 0.0464215633216 0.0009351723731 0.0130360878817
##              21              22              23              24              25
## 0.0017590910946 0.5280909174854 0.0095824460393 0.0161716992188 0.2494928588379
##               26              27              28              29              30 
## 0.0079862427435 0.0309188019544 0.0139374273504 0.0108461007718 0.0000008106153 
##               31 
## 0.0042360872993
```

Findings: No observation has a Cook's distance greater than 1. We interpret that as no observation distorts our model significantly.

Findings:

1. Residuals versus fitted: The line is diagonal before the 0.15 point on the x-axis yet it is more horizontal when compared to m7
2. Q-Q: The standardized residuals are distributed normally.
3. Scale-location: Although the line has several curves, it is more horizontal than the previous model.
4. The independent variables explain 0.6738 = 67.38% of the variation in the dependent variable.

Shapiro-Wilk test

```
shapiro.test(m8$residuals)$p.value < 0.05
```

```
## [1] FALSE
```

$p > 0.05$ → We cannot reject the null-hypothesis

We can say that the residuals are normally distributed unlike the previous model.

Although its Rˆ2 is smaller when compared to the previous model, this model has the advantage of normally distributed residuals.

$y = (-2.57155696) \times$ democracy mean $+(-0.00005586) \times$ real GDP per cap 2004 $+(0.00006662) \times$ (democracy mean real GDP per cap 2004) $+2.36268371$ #### Plot the model

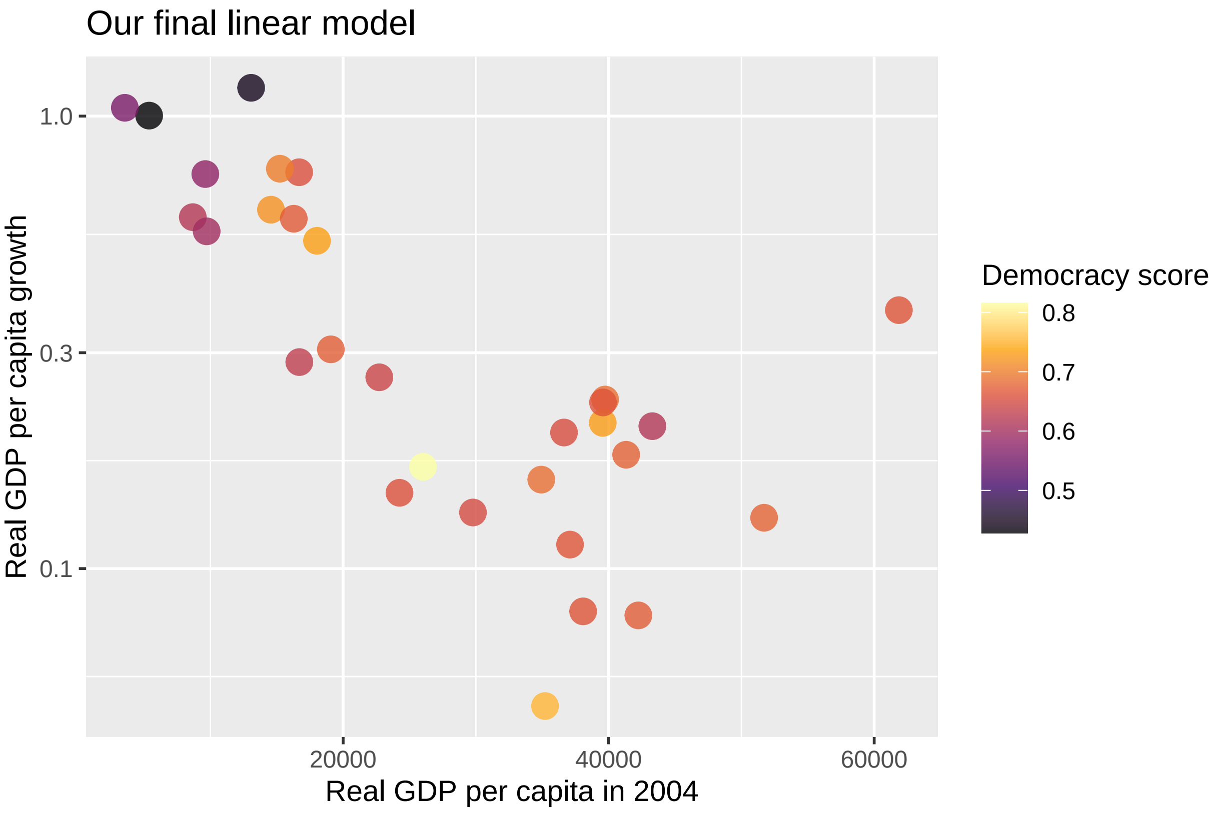
2-dimensional plot

```
ggplot(forest_tibble_2,
       aes(x = real_GDP_per_cap_2004,
           y = growth,
           color = democracy_mean)) +
  scale_y_log10() +
  geom_point(size = 4.2) +
  scale_color_viridis_c(option = "inferno",
                        alpha = 0.8,
                        name = "Democracy score") +
  xlab("Real GDP per capita in 2004") +
  ylab("Real GDP per capita growth") +
  ggtitle("Our final linear model") +
  labs(caption = "Source: Our World in Data")
```

```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Source: Our World in Data

However, in this model real_GDP_per_cap_2004 has a very low significance.

Thus, we will start again with the outliers removed.

```
m9 <- lm(growth ~ total_dependency_ratio_mean +
           oil_production_per_cap_mean +
           democracy_mean +
           oil_production_per_cap_mean +
           electricity_per_cap_mean +
           time_req_to_start_business_mean +
           tourists_per_cap_mean +
           real_GDP_per_cap_2004 +
           eu, data = forest_tibble_2)
```

Call stepAIC again

```
aic2 <- stepAIC(m9)
```

```
## Start:  AIC=-103.89
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + oil_production_per_cap_mean + electricity_per_cap_mean +
##     time_req_to_start_business_mean + tourists_per_cap_mean +
##     real_GDP_per_cap_2004 + eu
##
##                                   Df Sum of Sq     RSS      AIC
## - electricity_per_cap_mean         1   0.01064 0.61846 -105.350
## - eu                               1   0.01780 0.62563 -104.993
## <none>                                         0.60783 -103.888
## - tourists_per_cap_mean            1   0.05519 0.66301 -103.193
## - democracy_mean                   1   0.11842 0.72624 -100.370
## - time_req_to_start_business_mean  1   0.12955 0.73738  -99.898
## - total_dependency_ratio_mean      1   0.13084 0.73867  -99.844
## - oil_production_per_cap_mean      1   0.21602 0.82385  -96.460
## - real_GDP_per_cap_2004            1   0.42491 1.03274  -89.455
##
## Step:  AIC=-105.35
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + time_req_to_start_business_mean + tourists_per_cap_mean +
##     real_GDP_per_cap_2004 + eu
##
##                                   Df Sum of Sq     RSS      AIC
## - eu                               1   0.00747 0.62594 -106.977
## <none>                                         0.61846 -105.350
## - tourists_per_cap_mean            1   0.04642 0.66488 -105.106
## - democracy_mean                   1   0.10873 0.72719 -102.329
## - time_req_to_start_business_mean  1   0.11985 0.73832 -101.859
## - total_dependency_ratio_mean      1   0.12113 0.73959 -101.805
## - oil_production_per_cap_mean      1   0.20557 0.82404  -98.453
## - real_GDP_per_cap_2004            1   0.54885 1.16731  -87.658
##
## Step:  AIC=-106.98
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + time_req_to_start_business_mean + tourists_per_cap_mean +
##     real_GDP_per_cap_2004
##
##                                   Df Sum of Sq     RSS      AIC
## - tourists_per_cap_mean            1   0.03928 0.66522 -107.090
## <none>                                         0.62594 -106.977
## - democracy_mean                   1   0.11143 0.73737 -103.898
## - time_req_to_start_business_mean  1   0.11282 0.73876 -103.840
## - total_dependency_ratio_mean      1   0.11405 0.73998 -103.789
## - oil_production_per_cap_mean      1   0.23067 0.85661  -99.252
## - real_GDP_per_cap_2004            1   0.56721 1.19314  -88.979
##
## Step:  AIC=-107.09
## growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + time_req_to_start_business_mean + real_GDP_per_cap_2004
##
##                                   Df Sum of Sq     RSS      AIC
## <none>                                         0.66522 -107.090
## - time_req_to_start_business_mean  1   0.12749 0.79271 -103.655
## - total_dependency_ratio_mean      1   0.14470 0.80992 -102.989
## - democracy_mean                   1   0.14657 0.81179 -102.918
## - oil_production_per_cap_mean      1   0.22751 0.89273  -99.971
## - real_GDP_per_cap_2004            1   0.53347 1.19869  -90.836
```

```
aic2
```

```
##
## Call:
## lm(formula = growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + time_req_to_start_business_mean + real_GDP_per_cap_2004,
##     data = forest_tibble_2)
##
```

```
## Coefficients:
##                     (Intercept)      total_dependency_ratio_mean
##                     2.830413194                     -0.025978584
##     oil_production_per_cap_mean                   democracy_mean
##                     0.000001985                     -1.043778437
## time_req_to_start_business_mean             real_GDP_per_cap_2004
##                    -0.005068274                     -0.000014340
```

The real final model

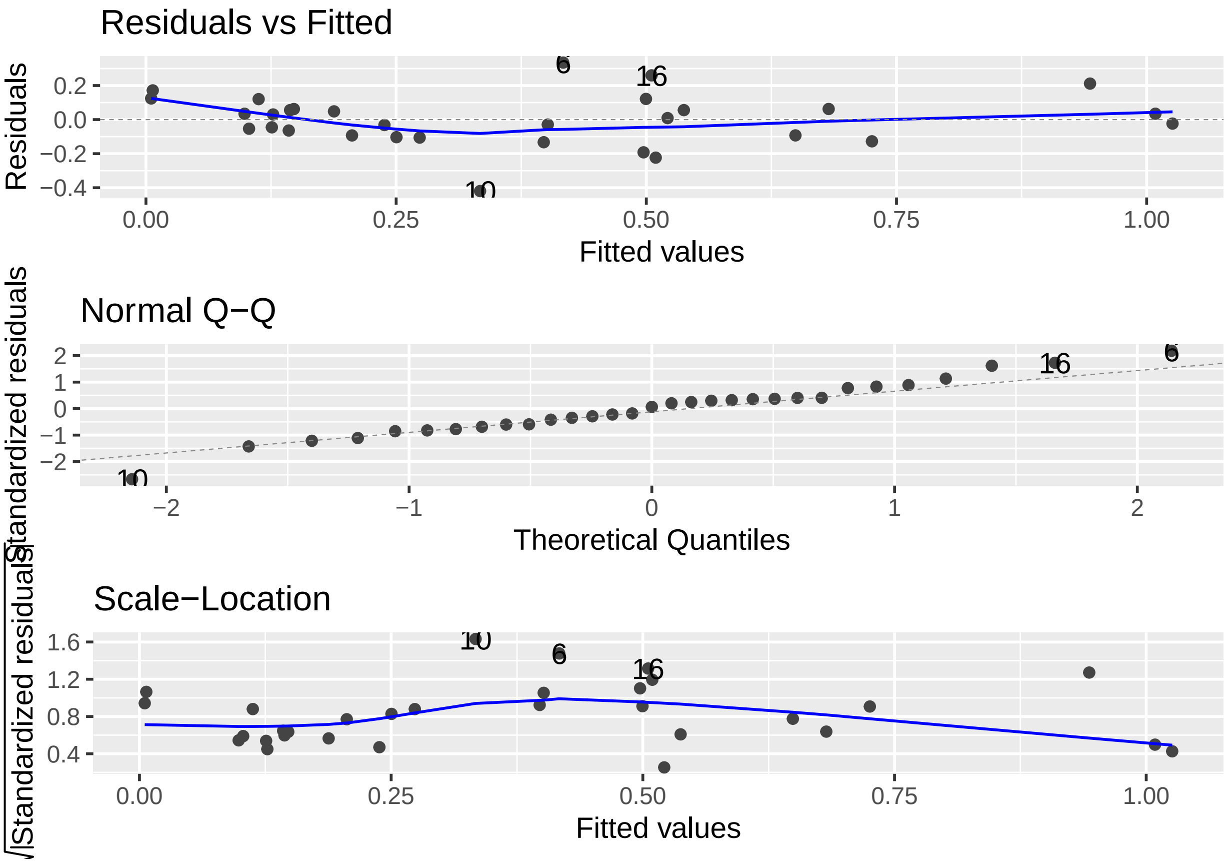
```
m10 <- lm(formula = growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
            democracy_mean + time_req_to_start_business_mean + real_GDP_per_cap_2004,
          data = forest_tibble_2)
```

Model metrics

```
m10_glance <- m10 %>%
  glance()
```

Visualize the model metrics

```
autoplot(m10, which = 1:3, nrow = 3, ncol = 1)
```



Breusch-Pagan test

```
bptest(m10)$p.value < 0.05
```

```
##    BP
## FALSE
```

We cannot reject the homoskedasticity

Summarize

```
summary(m10)
```

```
##
## Call:
## lm(formula = growth ~ total_dependency_ratio_mean + oil_production_per_cap_mean +
##     democracy_mean + time_req_to_start_business_mean + real_GDP_per_cap_2004,
##     data = forest_tibble_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41949 -0.09281  0.00892  0.06228  0.33463
##
## Coefficients:
##                                     Estimate    Std. Error t value  Pr(>|t|)
## (Intercept)                     2.8304131939  0.5910240477   4.789 0.0000644
## total_dependency_ratio_mean    -0.0259785844  0.0111401052  -2.332  0.028054
## oil_production_per_cap_mean     0.0000019852  0.0000006789   2.924  0.007241
## democracy_mean                 -1.0437784373  0.4447271782  -2.347  0.027149
## time_req_to_start_business_mean -0.0050682739  0.0023154021  -2.189  0.038154
## real_GDP_per_cap_2004          -0.0000143399  0.0000032026  -4.478  0.000144
##
## (Intercept)                     ***
## total_dependency_ratio_mean     *
## oil_production_per_cap_mean     **
## democracy_mean                  *
## time_req_to_start_business_mean *
## real_GDP_per_cap_2004           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1631 on 25 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7429
## F-statistic: 18.34 on 5 and 25 DF,  p-value: 0.0000001193
```

Variance inflation factors

```
vif(m10)
```

```
##     total_dependency_ratio_mean     oil_production_per_cap_mean
##                        2.247220                        1.382403
##                  democracy_mean time_req_to_start_business_mean
##                        1.312446                        1.571139
##           real_GDP_per_cap_2004
##                        2.467375
```

We have tested several different independent variables to see their implications on GDP growth of Eastern Bloc countries during the 2004-2014 period.

We decided to use m10 as our final model.

The explanatory variables of our final model are age dependency ratio,

oil production per capita, participatory democracy score, the time required

to start a business, and real GDP per capita in 2004.

Our regression equation for m10 is:

$$
\begin{aligned}
&y = (-0.026034479) \times \text{total-dependency-ratio-mean}+\\
&(0.000001980) \times \text{oil-production-per-cap-mean}+\\
&(-1.019645719) \times \text{democracy-mean}+\\
&(-0.005230104) \times \text{time-req-to-start-a-business-mean}+\\
&(-0.000014419) \times \text{real-GDP-per-cap-2004} + 2.823458000
\end{aligned}
$$

Therefore we can say:

1. 1 unit increase in Total Dependency Ratio is associated with a 0.026034479 decrease in GDP growth.
2. 1 unit increase in Oil Production per Capita is associated with a 0.000001980 increase in GDP growth.
3. 1 unit increase in Democracy is associated with a 1.019645719 decrease in GDP growth.
4. 1 unit increase in Time Required to Start a Business is associated with a 0.005230104 decrease in GDP growth.
5. 1 unit increase in Real GDP per capita in 2004 associated with a 0.000014419 decrease in GDP growth.

By far the most important factor for GDP growth observed here is participatory democracy. Participatory democracy score has a strong and negative effect on GDP per capita growth between 2004 and 2014. We think this could be due to democratic countries already enjoying higher GDP per capita and their growth being limited by the law of marginal benefit.

We want to conclude by reviewing policy implementations countries make.

1. Allowing immigration of young workers to reduce the age dependency.
2. Increasing fertility rates through better healthcare systems to battle aging, and reduce the age dependency ratio.
3. Decreasing the time required to start a business to increase economic growth
4. Boosting the oil production.

There are many effects these policies have on the economy, social welfare, and many other factors when policymakers are taking decisions. We cannot say with certainty that all of these policies may induce expected results in each case since linear regressions do not imply causal relations between variables.

Further research is needed to come up with case-specific policy recommendations.