Data Management & Analysis Final Project

Replication and Extention for Acemoglu, Naidu, Restrepo and Robinson (2019)

(Name:) (University ID:) (Name:) (University ID:)

(Submission Due:) 2025/02/06

Contents

0.1 Setup

```
pacman::p_load(
  rmdformats,
  knitr,
  tinytex,
  haven,
  tidyverse,
  kableExtra
)

## Global options
options(max.print="75")
opts_chunk$set(fig.align="center",
               echo=TRUE,
               cache=TRUE,
               prompt=FALSE,
               tidy=FALSE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
```

1 About this Report

1.1 Project Type

1.2 Summary of the Paper

1.2.1 What the problem is

1.2.2 Why it is important

1.2.3 How you solve the problem

1.2.4 What we find

1.3 Data

```
data <- read_dta("data/raw/DDCGdata_final.dta")

# Define the function to summarize the data
summarize_data <- function(data, n = 10) {
  cat("Sample size (number of rows):", nrow(data), "\n")
  cat("Number of variables (columns):", ncol(data), "\n")
  cat("Variable names (first", n, "names):\n")
  print(head(colnames(data), n))
}

summarize_data(data)
```

```
Sample size (number of rows): 9384
Number of variables (columns): 1177
Variable names (first 10 names):
 [1] "country_name"              "wbcode"
 [3] "year"                      "gdppercapitaconstant2000us"
 [5] "lp_bl"                     "ls_bl"
 [7] "lh_bl"                     "taxratio"
 [9] "region"                    "wbcode2"
```

1.4 Empirical Methods

1.4.1 Event Study (Figure.1)

1.4.2 Dynamic Panel Data Model (Table.2)

2 Replication

2.1 Figure.1

2.1.1 Preprocessing

```
# 1. Basic transformations: Rename ID, sort by year, create lagged democracy indicator, and compute lag
data_f1 <- data %>%
  rename(id = "_ID") %>%                      # Rename "_ID" to "id"
  group_by(id) %>%                            # Group by id
  arrange(year) %>%                           # Sort by year within each id
  mutate(
    prev_dem   = dplyr::lag(dem, 1),           # Create lagged democracy indicator using dplyr::lag()
    transition = case_when(                   # Define the transition:
      dem == 1 & prev_dem == 0 ~ 1,            # 1: transition from non-democracy (0) to democracy (1)
      dem == 0 & prev_dem == 0 ~ 0,            # 0: remains non-democracy (0)
      TRUE ~ NA_real_                        # Otherwise, set to NA
    ),
    lag1 = dplyr::lag(y, 1),                   # y lag 1 period
    lag2 = dplyr::lag(y, 2),                   # y lag 2 periods
    lag3 = dplyr::lag(y, 3),                   # y lag 3 periods
    lag4 = dplyr::lag(y, 4)                    # y lag 4 periods
  ) %>%
  filter(                                     # Filter out rows with any missing lag values
    !is.na(lag1) & !is.na(lag2) & !is.na(lag3) & !is.na(lag4)
  ) %>%
  ungroup()

# 2. Create GDP differences for past periods (from t = -15 to -2)
#    New columns are named "gdpDiff_m[abs(t)]", computed as lag(y, abs(t)) - lag1
for (t in -15:-2) {
  col_name <- paste0("gdpDiff_m", abs(t))
  data_f1 <- data_f1 %>%
    group_by(id) %>%
    arrange(year) %>%
    mutate(!!col_name := dplyr::lag(y, abs(t)) - lag1) %>%
```

```r
    ungroup()
}

# 3. Set the immediate past period (m1) difference to 0 and compute the current period difference (0)
data_f1 <- data_f1 %>%
  mutate(
    gdpDiff_m1 = 0,          # Set immediate past difference to 0
    gdpDiff_0  = y - lag1     # Current period difference: y - lag1
  )

# 4. Create GDP differences for future periods (from t = 1 to 30)
#    New columns are named "gdpDiff_p[t]", computed as lead(y, t) - lag1
for (t in 1:30) {
  col_name <- paste0("gdpDiff_p", t)
  data_f1 <- data_f1 %>%
    group_by(id) %>%
    arrange(year) %>%
    mutate(!!col_name := dplyr::lead(y, t) - lag1) %>%
    ungroup()
}

# 5. Keep only rows where the transition value is not missing
data_f1 <- data_f1 %>% filter(!is.na(transition))
```

2.1.2 Estimation

```r
# Define a function to estimate the Average Treatment Effect on the Treated (ATET)
estimateATET <- function(outcome_col) {
  # Filter out rows with missing outcome or transition values
  sub_data <- data_f1 %>% filter(!is.na(.data[[outcome_col]]), !is.na(transition))
  if(nrow(sub_data) == 0) return(NA)

  # Create a factor for 'year' with sorted levels
  year_levels <- sort(unique(sub_data$year))
  sub_data <- sub_data %>% mutate(year_factor = factor(year, levels = year_levels))

  # Split the data into control (transition == 0) and treated (transition == 1) groups
  control_data <- sub_data %>% filter(transition == 0)
  treated_data <- sub_data %>% filter(transition == 1)

  # Return NA if the control group lacks sufficient observations or year variation
  if(nrow(control_data) < 2 || length(unique(control_data$year)) < 2) return(NA)

  # Fit a linear model on the control group with year dummies (without intercept)
  model_formula <- as.formula(paste(outcome_col, "~ year_factor - 1"))
  control_model <- tryCatch(lm(model_formula, data = control_data),
                            error = function(e) NULL)
  if(is.null(control_model)) return(NA)

  # Predict outcomes for the treated group using the control model
  predicted_outcomes <- tryCatch(predict(control_model, newdata = treated_data),
                                 error = function(e) rep(NA, nrow(treated_data)))
```

```
  # Compute treatment effects as the difference between actual and predicted outcomes
  treatment_effects <- treated_data[[outcome_col]] - predicted_outcomes

  # Return the mean treatment effect on the treated (ATET)
  mean(treatment_effects, na.rm = TRUE)
}

# Define relative time periods: pre-treatment (-15 to -1) and post-treatment (0 to 30)
relative_times <- c(seq(-15, -1), seq(0, 30))
atets <- numeric(length(relative_times))

# Loop over each relative time period to estimate ATET for the corresponding outcome column
for(i in seq_along(relative_times)) {
  t_val <- relative_times[i]
  # Set the column name based on whether the period is before (m), during (0),
  # or after (p) treatment
  if(t_val < 0) {
    col_name <- paste0("gdpDiff_m", abs(t_val))
  } else {
    col_name <- if(t_val == 0) "gdpDiff_0" else paste0("gdpDiff_p", t_val)
  }
  atets[i] <- estimateATET(col_name)
}

# Create a data frame with the relative time periods and their corresponding ATET estimates
results_df <- data.frame(RelativeTime = relative_times, ATET = atets)
```

2.1.3 Plot

```
figure_1 <- ggplot(results_df, aes(x = RelativeTime, y = ATET)) +
  geom_line(color = "black") +
  scale_x_continuous(breaks = seq(-15, 30, 5)) +
  labs(x = "Years around Democratization",
       y = "Change in GDP per capita (log points)") +
  theme_bw()

ggsave("output/figure_1.pdf", width = 14, height = 8, units = "cm")
```

2.2 Table.1

```
# Define variable names and display labels
var_info <- tibble(
  var = c("gdppercapitaconstant2000us",  # data column name
          "ginv",
          "tradewb",
          "prienr",
          "secenr",
          "taxratio",
```

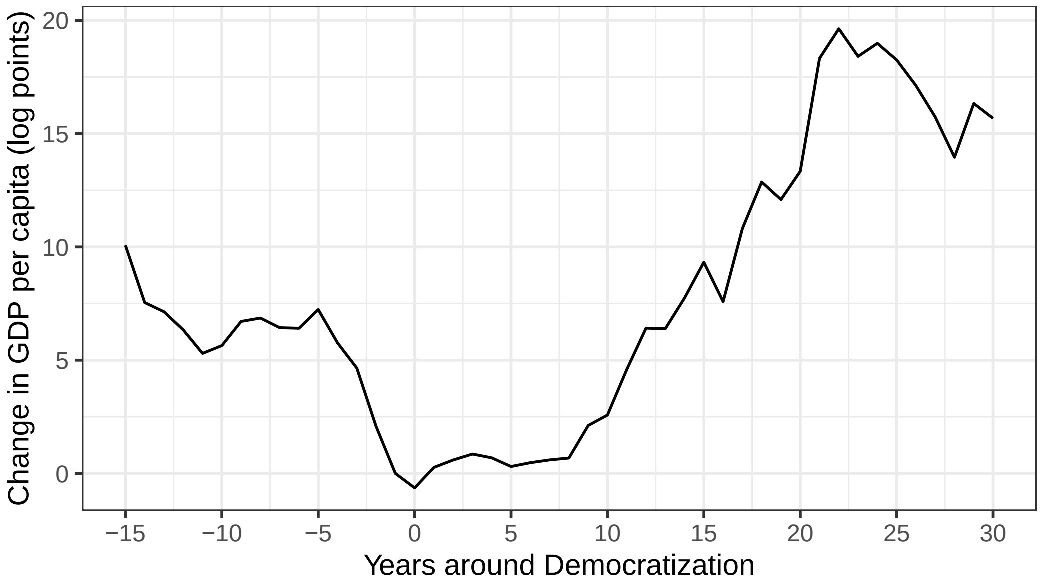

Figure 1: Event Study

```
           "mortnew",
           "unrestn",
           "marketref"),
  label = c("GDP per capita",           # table label
             "Investment share of GDP",
             "Trade share of GDP",
             "Primary-school enrollment rate",
             "Secondary-school enrollment rate",
             "Tax revenue share of GDP",
             "Child mortality per 1,000 births",
             "Unrest rate",
             "Market reforms index (0-100)")
)

# Calculate descriptive statistics
calc_summary <- function(df, var, group_var) {
  df %>%
    filter(!is.na(.data[[var]])) %>%       # remove NAs
    group_by({{ group_var }}) %>%
    summarise(
      Observations = n(),                 # count
      Mean         = mean(.data[[var]], na.rm = TRUE),
      SD           = sd(.data[[var]], na.rm = TRUE),
      .groups      = "drop"
    ) %>%
    mutate(Variable = var)                # record variable name
}

# List of variables to summarize
var_list <- var_info$var
```

```r
# Summarize variables, reshape by democracy status, and add labels
summary_table <- lapply(var_list, function(x) calc_summary(data, x, dem)) %>%
  bind_rows() %>%
  pivot_wider(
    names_from  = dem,
    values_from = c(Observations, Mean, SD),
    names_glue  = "{.value}_dem{dem}"
  ) %>%
  # Rename columns for clarity
  rename(
    Observations_Nondem = Observations_dem0,
    Mean_Nondem         = Mean_dem0,
    SD_Nondem           = SD_dem0,
    Observations_Dem    = Observations_dem1,
    Mean_Dem            = Mean_dem1,
    SD_Dem              = SD_dem1
  ) %>%
  # Join with var_info to add labels
  left_join(var_info, by = c("Variable" = "var")) %>%
  # Rearrange columns and use label as Variable
  select(label,
         Observations_Nondem, Mean_Nondem, SD_Nondem,
         Observations_Dem,    Mean_Dem,    SD_Dem
  ) %>%
  rename(Variable = label)

# Format table for PDF output with a multi-column header
colnames(summary_table) <- c("Variable",
                             "Observations", "Mean", "SD",
                             "Observations", "Mean", "SD")

latex_table <- summary_table %>%
  kbl(
    caption = "Summary Statistics by Democracy Status",
    format = "latex",
    booktabs = TRUE,
    digits = 2
  ) %>%
  add_header_above(c(" " = 1, "Nondemocracies" = 3, "Democracies" = 3)) %>%
  kable_styling(latex_options = c("HOLD_position", "striped"))

# Save the table as a TeX file
save_kable(latex_table, file = "output/table_1.tex")
```

Table 1: Summary Statistics by Democracy Status

Variable	Nondemocracies Observations	Nondemocracies Mean	Nondemocracies SD	Democracies Observations	Democracies Mean	Democracies SD
GDP per capita	3376	2074.46	3838.65	3558	8149.97	9334.83
Investment share of GDP	3225	21.82	10.23	3340	23.28	7.41
Trade share of GDP	3175	71.63	51.06	3485	77.15	41.04
Primary-school enrollment rate	2861	90.29	29.51	2823	101.60	15.86
Secondary-school enrollment rate	2424	45.76	31.77	2538	75.40	29.78
Tax revenue share of GDP	3122	0.16	0.09	2564	0.21	0.10
Child mortality per 1,000 births	4142	77.29	49.64	3615	33.26	32.65
Unrest rate	3739	28.70	45.24	3610	21.91	41.37
Market reforms index (0–100)	3476	21.89	23.26	2829	52.11	24.75

2.3 Table.2

Table 2: Effect of Democracy on (Log) GDP per Capita

	Dependent variable:			
	Log GDP per Capita			
	(1)	(2)	(3)	(4)
Democracy	0.973***	0.651***	0.787***	0.887***
	(0.240)	(0.229)	(0.228)	(0.239)
lag1	0.973***	1.266***	1.238***	1.233***
	(0.003)	(0.012)	(0.013)	(0.013)
lag2		−0.300***	−0.207***	−0.214***
		(0.012)	(0.020)	(0.021)
lag3			−0.026	−0.021
			(0.019)	(0.021)
lag4			−0.043***	−0.039*
			(0.012)	(0.020)
lag5				−0.019
				(0.020)
Persistence:	0.973	0.967	0.963	0.96
Long run effect:	35.587	19.599	21.24	22.008
Effect after 25 years:	17.791	13.8	16.895	17.715
Observations	6,790	6,642	6,336	5,688

Note: *p<0.1; **p<0.05; ***p<0.01

2.4 Figure.2

2.5 Table.3

2.6 Table.4

2.7 Table.5

3 Extention