In this project, I replicated Table 2 ("OLS Regression") from Agniva Ghosh's replication paper of Acemoglu, Johnson, and Robinson's (2001) study. Using the publicly available AJR dataset, I estimated eight OLS models, where the dependent variable is the logarithm of per capita GDP in 1995 or the logarithm of per capita output in 1988, and the main explanatory variable is institutional quality (average expropriation risk). The replicated coefficients are very close to those reported in the replication paper; the minor differences may stem from updates to the publicly available data and the handling of missing values.

```
# Final Project - Econometrics Replication

# Replicating Table 2: OLS Regressions from AJR (Institutions and Growth)

# Course: AS.440.606.57. FA25 Econometrics

# Author: QIAN HAOLEI

# Date: maketable2.dta


# Load required packages

library(tidyverse)

library(haven)

library(fixest)

library(modelsummary)


# Load the AJR dataset

ajr_dta <- read_dta("data/maketable2.dta")


# Construct the baseline sample (baseco == 1)

base_sample <- ajr_dta %>% filter(baseco == 1)


# Define all eight OLS models for Table 2

model_list <- list(

  "(1)" = feols(loggdp95 ~ avexpr, data = ajr_dta, se = "hetero"),

  "(2)" = feols(loggdp95 ~ avexpr, data = base_sample, se = "hetero"),

  "(3)" = feols(loggdp95 ~ avexpr + lat_abst, data = ajr_dta, se = "hetero"),
```

```
  "(4)" = feols(loggdp95 ~ avexpr + lat_abst + africa + asia + other, data = ajr_dta, se = "hetero"),
  "(5)" = feols(loggdp95 ~ avexpr + lat_abst, data = base_sample, se = "hetero"),
  "(6)" = feols(loggdp95 ~ avexpr + lat_abst + africa + asia + other, data = base_sample, se = "hetero"),
  "(7)" = feols(loghjypl ~ avexpr, data = ajr_dta, se = "hetero"),
  "(8)" = feols(loghjypl ~ avexpr, data = base_sample, se = "hetero")
)

# Define goodness-of-fit statistics to display
gof_map <- list(
  list("raw" = "nobs", "clean" = "Num. Obs.", "fmt" = 0),
  list("raw" = "r.squared", "clean" = "R-squared", "fmt" = 3)
)

# Define coefficient labels for the table
coef_map <- c(
  "avexpr"    = "Average Expropriation Risk",
  "lat_abst" = "Distance from Equator",
  "africa"    = "Africa",
  "asia"      = "Asia",
  "other"     = "Other continents"
)

# Produce Table 2 using modelsummary()
table2_ols <- modelsummary(
  model_list,
  output = "gt",
  title = "Table 2: OLS Regressions",
  coef_map = coef_map,
```

```
  gof_map = gof_map,
  stars = c('*' = .1, '**' = .05, '***' = .01),
  add_rows = tribble(
    ~term, ~`(1)`, ~`(2)`, ~`(3)`, ~`(4)`, ~`(5)`, ~`(6)`, ~`(7)`, ~`(8)`,
    "Base Sample", "No", "Yes", "No", "No", "Yes", "Yes", "No", "Yes",
    "Continent Dummies", "No", "No", "No", "Yes", "No", "Yes", "No", "No"
  ),
  notes = "* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors are in parentheses."
) %>%
  tab_spanner(
    label = "Dependent variable: Log GDP per capita, 1995",
    columns = 2:7
  ) %>%
  tab_spanner(
    label = "Dependent variable: Log output per worker, 1988",
    columns = 8:9
  )

table2_ols
```

Replication code documentation

Section: Replicating Table 2 – OLS Regressions

In this project I replicate Table 2 (“OLS Regressions”) from Agniva Ghosh’s replication of Acemoglu, Johnson, and Robinson (2001). Table 2 reports eight OLS specifications that relate economic performance to institutional quality.

- The dependent variable in columns (1)–(6) is log GDP per capita in 1995 (loggdp95).

- The dependent variable in columns (7)–(8) is log output per worker in 1988 (loghjypl).
- The main explanatory variable is Average Expropriation Risk (avexpr), which measures institutional quality.
- Additional controls include distance from the equator (lat_abst) and continent dummy variables (africa, asia, other).
- Some specifications are estimated on the full sample, while others restrict the sample to the base sample where baseco == 1.

The replication code is organized as follows:

1: Load packages

The script first loads the packages used for data handling, estimation, and table creation:

```
library(tidyverse)

library(haven)

library(fixest)

library(modelsummary)
```

2：Load the AJR dataset

The AJR data are stored in a Stata file maketable2.dta.
The following line imports the dataset into R:

```
ajr_dta <- read_dta("data/maketable2.dta")
```

3: Define the base sample

Table 2 distinguishes between the full sample and a “base sample”. The base sample is created by keeping only observations with baseco == 1:

```
base_sample <- ajr_dta %>% filter (baseco == 1)
```

4: Estimate the eight OLS models

I then estimate eight OLS regressions using feols() from the fixest package.

All models use heteroskedasticity-robust standard errors (se = "hetero").

Columns (1) and (2) regress loggdp95 on avexpr, using the full sample and the base sample respectively.

Columns (3) and (5) add distance from the equator as a control.

Columns (4) and (6) additionally include continent dummies (africa, asia, other).

Columns (7) and (8) repeat the simple regression of institutional quality on log output per worker instead of log GDP per capita.

These models are stored in a list model_list, which is passed to modelsummary().

5: Construct Table 2

I define a goodness-of-fit map to display the number of observations and R-squared in the bottom rows of the table, and a coefficient map to show readable variable names.
Then I call modelsummary() to produce the OLS table with:

robust standard errors in parentheses,

significance stars for the p-values,

two additional rows indicating whether the base sample is used and whether continent dummies are included in each specification.

Finally, I use tab_spanner() to group columns (1)–(6) under the heading
“Dependent variable: Log GDP per capita, 1995”
and columns (7)–(8) under
“Dependent variable: Log output per worker, 1988”.

The resulting output matches the structure and content of Table 2: OLS Regressions in the replication paper.

As an extension of Table 2, I focus on the relationship between institutional quality and log output per worker in 1988(loghjypl). Columns (7) and (8) of Table 2 estimate simple bivariate OLS regressions of loghjypl on the institutional quality

measure avexpr, using the full sample and the base sample respectively.

To check the robustness of this relationship, I estimate two additional specifications:

I first add distance from the equator (lat_abst) as a control variable:

loghjypli=β0+β1avexpri+β2lat_absti+ui.loghjypli=β0+β1avexpri+β2lat_absti+ui.

I then include continent dummy variables (africa, asia, other) in addition to lat_abst:

loghjypli=β0+β1avexpri+β2lat_absti+β3africai+β4asiai+β5otheri+ui.loghjypli=β0+β1avexpri+β2lat_absti+β3africai+β4asiai+β5otheri+ui.

All models are estimated using heteroskedasticity-robust standard errors.

The new results show that the coefficient on institutional quality (avexpr) remains positive and statistically significantafter controlling for geography and continent fixed effects. This suggests that the strong association between institutions and output per worker is not driven solely by geographic factors and is broadly consistent with the findings reported for log GDP per capita in columns (3)–(6) of Table 2.

Extension: OLS for log output per worker with controls

Robustness check for the relationship between institutions

and log output per worker (1988).

Baseline models for comparison (already estimated in model_list):

(7): loghjypl ~ avexpr (full sample)

(8): loghjypl ~ avexpr (base sample)

```
ext_models <- list(
  # Baseline specifications from Table 2 (for comparison)
  "(7)"    = model_list[["(7)"]],
  "(8)"    = model_list[["(8)"]],

  # New extension models with geographic controls
  "(9)"    = feols(loghjypl ~ avexpr + lat_abst,
                   data = ajr_dta, se = "hetero"),

  "(10)" = feols(loghjypl ~ avexpr + lat_abst + africa + asia + other,
                   data = ajr_dta, se = "hetero")

# Produce an extension table
table_ext <- modelsummary(
  ext_models,
  output     = "gt",
  title      = "Extension: OLS regressions for log output per worker",
  coef_map = coef_map,
  gof_map    = gof_map,
  stars      = c("*" = .1, "**" = .05, "***" = .01),
  notes      = "* p < 0.1, ** p < 0.05, *** p < 0.01. Robust standard errors are in
parentheses."
)

table_ext
```

Data Cleaning

This replication uses the publicly available AJR dataset maketable2.dta, which contains a series of indicators of institutional quality, geographical location, and economic performance for a set of countries. Before estimation, two data cleaning steps are necessary:

1: Handling Missing Values

The estimation function automatically removes any observations with missing values in the dependent variable or the explanatory variables used in the given model.

This is why the sample sizes differ across the eight regression models in Table 2.

For example, the R output will display a note similar to the following:

“52 observations deleted due to NA values (left: 15, right: 42).”

This indicates that the combination of regression variables used in the model leads to a different set of available observations.

Therefore, all regression analyses in this replication are based on a complete data sample, consistent with the original paper.

2: Construction of the Base Sample

This paper defines a "base sample" that excludes certain countries due to data quality issues or the presence of outliers.

In the dataset, this is represented by the dummy variable `baseco`.

The base sample is constructed as follows:

base_sample <- ajr_dta %>% filter (baseco == 1)

Specifications (2), (5), (6), and (8) in Table 2 restrict the estimation to this subsample, which typically results in a smaller sample size.

Overall, the data cleaning steps replicate the structure used in the original AJR analysis and ensure that the regression estimates are based on a consistent and comparable sample.

Econometric Assumptions

The Ordinary Least Squares (OLS) estimates in Table 2 are based on the following standard linear regression assumptions:

1: Linear Relationship

It is assumed that the relationship between institutional quality (avexpr) and economic outcomes (loggpg95, loghjypl) is linear.

This is a reasonable approximation for cross-country regression analysis.

2: Exogeneity

The key assumption is:

$$E[\varepsilon i \mid avexpri, Xi]=0.$$

However, in this situation, the exogeneity assumption is likely to be violated for the following reasons:

Countries with higher incomes may have better institutions (reverse causality),

Unobserved historical or cultural factors may simultaneously affect both institutions and economic development (omitted variable bias).

Therefore, the coefficient of the avexpr variable in the OLS regression is likely to be biased.

3: Heteroscedasticity

Cross-country data often exhibit heteroscedastic errors.

Therefore, this study uses heteroscedasticity-robust standard errors (set as se = "hetero" in the feols() function).

No perfect multicollinearity

Including dummy variables for Africa, Asia, and other regions requires omitting a reference group.

The data structure has been designed to avoid the dummy variable trap.

Independence

Each country is treated as an independent observation.

This is standard practice in cross-sectional macroeconomic research.

Because exogeneity is unlikely to hold, the authors of the original paper rely on instrumental variable (IV) estimation in later tables to address bias in the OLS results.

OLS Results Interpretation

In all eight models in Table 2, the coefficient for institutional quality (avexpr) is positive and statistically significant.

This means that:

Countries with better protection against expropriation tend to have higher per capita income and productivity levels.

The coefficient of avexpr decreases slightly as more control variables are added, but remains significant.

This pattern suggests two conclusions:

The relationship is robust.

Adding geographical location (lat_abst) or continent dummy variables does not eliminate the association between institutions and economic performance.

OLS may overestimate the true causal effect.

The decrease in the coefficient after adding control variables means that part of the original OLS correlation reflects omitted factors that are correlated with both institutions and income.

As stated in the replication study:

“These OLS results may be biased. The direction of the bias is uncertain: omitted variables could bias the coefficient upwards or downwards, while reverse causality could bias it upwards.”

Therefore, the authors proceed to use instrumental variable methods (2SLS) to obtain more reliable causal estimates. The OLS results are mainly used as a benchmark.