DATA606 Data Project

Michael Ippolito

DATA606 Data Project

Abstract

It may be self-evident that the amount of money spent on education by a country will have an impact on the country's literacy rate. But is there a statistically significant relationship? And can that relationship be quantified using a predictive model? It is further presumed that literacy rates will be lower in poorer regions of the world and in those having forms of government in which the head of state holds more power and has less accountability to his or her constituents. This project sought to validate those assumptions and to generate a model that will predict the change in literacy rate based on the studied parameters.

The study confirms that literacy rate is impacted positively by educational spending, with a fairly steep relationship: A one-percent increase in educational spending yields an almost a three-percent rise in literacy rate. There is also evidence to suggest a statistically significant relationship between literacy and both the country's geographic subregion and its form of government, although it is noted that not all conditions were sufficiently satisfied for the analysis.

Part 1 - Introduction

It may be self-evident that the amount of money spent on education by a country will have an impact on the country's literacy rate. But is there a statistically significant relationship? And can that relationship be quantified using a predictive model?

It is further presumed that literacy rates will be lower in poorer regions of the world and in those having forms of government in which the head of state holds more power and has less accountability to his or her constituents. This project seeks to validate those assumptions and to generate a model that will predict the change in literacy rate based on the studied parameters.

Part 2 - Data

The data came from several sources:

- *Regional data* is from a public dataset on Github[1].

- *Government spending on education* is posted on worldbank.org and is collected annually from the UNESCO Institute for Statistics; data from the years 1960 to 2020 was available for download[2].

- *Literacy rates* were obtained from an aggregated dataset on wikimedia.org; the data was originally collected by worldbank.org, the CIA World Factbook, and other sources. The dataset includes some historical data as old as 1475 but with a majority of data points between 1960 and 2015[3]. For the purposes of this study, only data from 1960 or newer was used.

- *Forms of government* were taken from Wikipedia[4].

Data preparation included a few tidying steps:

- Removing footnotes and general tidying of the data
- Standardizing country codes across data sets
- Standardizing field names

```r
############################
# Regions
############################

# Load countries by region
# from https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv
regions_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/country

# Standarize field names
regions <- regions_full %>%
  rename(
    "country_numeric_code" = "country.code",
    "country_code" = "alpha.3",
    "subregion" = "sub.region"
  )


############################
# Government spending
############################

# Load govt spending on education data set
# from https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS
edspending_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/gov

# Standardize field names
edspending <- edspending_full %>%
  rename(
    "country" = "Country.Name",
    "country_code" = "Country.Code"
  ) %>%
  select(-`Indicator.Name`, -`Indicator.Code`, -`X`)

# Remove Xs from years in column names
edspending <- rename_with(edspending, function(x) ifelse(substr(x, 1, 1) == "X", substr(x, 2, 5), x))

# Gather columns
edspending <- edspending %>% gather(3:63, key = "year", value = "pct_gdp")

# Convert years to numeric
edspending$year <- as.numeric(edspending$year)


############################
# Literacy rates
############################

# Load literacy rates
```

2

```r
# from https://commons.wikimedia.org/wiki/Data:Cross-country_literacy_rates_-_World_Bank,_CIA_World_Fac
lit_full <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/cross-count

# Standardize field names
lit <- lit_full %>%
    rename(
        "country" = "Entity",
        "country_code" = "Code",
        "year" = "Year",
        "lit_rate" = "Literacy.rates..World.Bank..CIA.World.Factbook..and.other.sources."
    ) %>%
    filter(year >= 1960)


###########################
# Forms of government
###########################

# Load forms of government from CSV if it exists, otherwise parse from Wikipedia
# from https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government
if (url.exists("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/govt_forms.csv"))
    govts <- read.csv("https://raw.githubusercontent.com/mmippolito/cuny/main/data606/project/govt_form
} else {
    # Scrape and parse tables
    govts_html <- read_html("https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government", 
    govts_tables <- html_table(govts_html, fill = TRUE)
    govts <- data.frame(govts_tables[6])
    # Standarize field names
    govts <- govts %>%
        rename(
            country = Name,
            govt = Constitutional.form,
            head_of_state = Head.of.state,
            exec_legitimacy = Basis.of.executive.legitimacy)
    # Remove footnotes
    govts$exec_legitimacy <- str_replace_all(govts$exec_legitimacy, '\\[.+?\\]', '')
    # Change provisional gov't with no head to "no" instead of "n/a"
    govts$head_of_state <- ifelse(govts$head_of_state == "n/a", "no", govts$head_of_state)
    # Lower case
    govts$govt <- tolower(govts$govt)
    govts$head_of_state <- tolower(govts$head_of_state)
    govts$exec_legitimacy <- tolower(govts$exec_legitimacy)
    # Standardize names
    govts <- govts %>%
        mutate(country = case_when(
            country == "Bahamas, The" ~ "Bahamas",
            country == "China, People's Republic of" ~ "China",
            country == "Congo, Democratic Republic of the" ~ "Democratic Republic of Congo",
            country == "Congo, Republic of the" ~ "Congo",
            country == "Côte d'Ivoire" ~ "Cote d'Ivoire",
            country == "Czech Republic" ~ "Czechia",
            country == "East Timor" ~ "Timor",
            country == "Federated States of Micronesia" ~ "Micronesia",
```

```
                country == "Gambia, The" ~ "Gambia",
                country == "Korea, North" ~ "North Korea",
                country == "Korea, South" ~ "South Korea",
                country == "São Tomé and Príncipe" ~ "Sao Tome and Principe",
                country == "Vatican City" ~ "Vatican",
                TRUE ~ country)
        )
    # Add country codes by joining to lit_sum
    govts <-govts %>%
        left_join(
            lit %>% group_by(country, country_code) %>% summarize(n(), .groups = "keep")
            , by = c("country")) %>%
        select(country, govt, head_of_state, exec_legitimacy, country_code)
    # Save to CSV
    # Create short form of exec_legitimacy
    govts <- govts %>%
        mutate(exec_legit_short = case_when(
            exec_legitimacy == "all authority vested in absolute monarch" ~
                "having absolute authority",
            exec_legitimacy == "ministry is subject to parliamentary confidence" ~
                "accountable to legislature",
            exec_legitimacy == "monarch personally exercises power in concert with other institutions" ~
                "sharing power",
            exec_legitimacy == "no constitutionally-defined basis to current regime" ~
                "with no contsitutional legitimacy",
            exec_legitimacy == "power constitutionally linked to a single political movement" ~
                "constitutionally granted power by single party",
            exec_legitimacy == "presidency independent of legislature; ministry is subject to parliament
                "independent of and accountable to legislature",
            exec_legitimacy == "presidency is elected by legislature; ministry may be, or not be, subje
                "elected by legislature",
            exec_legitimacy == "presidency is independent of legislature" ~
                "independent of legislature",
            TRUE ~ '')
        )
    govts$govt_type_short <- paste0(govts$govt, " with ", govts$head_of_state, " head\n", govts$exec_le
}
```

Because the data came from disparate sources, there were gaps that had to be addressed before proceeding with analysis. For example, many countries had sparser data on literacy rates compared to government spending on education:

```
regions %>%
    select(country_code) %>%
    left_join(edspending, by = c("country_code")) %>%
    group_by(country_code) %>%
    summarize(edspending_year_count = n()) %>%
    full_join(lit, by = c("country_code")) %>%
    group_by(country_code, edspending_year_count) %>%
    summarize(litrate_year_count = n(), .groups = "keep") %>%
    full_join(govts, by = c("country_code")) %>%
    group_by(country_code, edspending_year_count, litrate_year_count) %>%
    summarize(govt_count = n(), .groups = "keep") %>%
```

```
    group_by() %>%
    summarize(countries = n(), ed_spending_year_count = sum(edspending_year_count, na.rm = T),
        litrate_year_count = sum(litrate_year_count, na.rm = T), govt_count = sum(govt_count, na.rm = T)
    kable(caption = "<i><font color=#000000><b>Table 1.</b> Summary of observations</font></i>") %>%
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_size = 13)
```

\begin{table}

\caption{Table 1. Summary of observations}

countries	ed_spending_year_count	litrate_year_count	govt_count
253	13149	1268	255

\end{table}

In addition, some countries had no country code listed, while others had no literacy rate or educational spending data:

```
regions %>%
    select(country_code) %>%
    left_join(edspending, by = c("country_code")) %>%
    group_by(country_code) %>%
    summarize(edspending_year_count = n()) %>%
    full_join(lit, by = c("country_code")) %>%
    group_by(country_code, edspending_year_count) %>%
    summarize(litrate_year_count = n(), .groups = "keep") %>%
    full_join(govts, by = c("country_code")) %>%
    group_by(country_code, edspending_year_count, litrate_year_count) %>%
    summarize(govt_count = n(), .groups = "keep") %>%
    arrange(country_code) %>%
    filter(is.na(edspending_year_count) | is.na(litrate_year_count) | is.na(govt_
    kable(caption = "<i><font color=#000000><b>Table 1.5.</b> Data gaps</font></i
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 1.5. Data gaps}

country_code	edspending_year_count	litrate_year_count	govt_count
	NA	109	1
OWID_KOS	NA	1	1
OWID_WRL	NA	4	1
NA	NA	NA	3

\end{table}

For these reasons, the data couldn't be reliably compared on a year-by-year basis. Instead, the data was taken as an aggregate, using mean values for literacy rate and educational spending per country.

```r
###########################
# Data aggregation
###########################

# Calculate mean literacy rate per country
lit_sum <- lit %>%
    group_by(country, country_code) %>%
    summarize(lit_rate = mean(lit_rate, na.rm = T), .groups = "keep")

# Calculate mean education spending per country
ed_sum <- edspending %>% group_by(country, country_code) %>%
    summarize(pct_gdp = mean(pct_gdp, na.rm = T), .groups = "keep")

# Join the summarized tables
j <- lit_sum %>%
    full_join(ed_sum, by = c("country_code")) %>%
    full_join(govts, by = c("country_code")) %>%
    left_join(regions, by = c("country_code"))

# Filter out rows with no country code or subregion; these are regional aggregat
j <- j %>%
    filter(str_length(country_code) > 0 & str_length(region) > 0)

# Filter out NaNs
#j <- j %>%
#  filter(!is.na(pct_gdp)) %>%
#  filter(!is.na(lit_rate))

# Select only relevant fields
j <- j %>%
    select(country_code, country = country.x, region, subregion,
        pct_gdp, lit_rate, govt, head_of_state, exec_legitimacy, exec_legit_short

# Add rankings
j$rank_pct_gdp <- rank(desc(j$pct_gdp))
j$rank_lit_rate <- rank(desc(j$lit_rate))
```

Part 3 - Exploratory data analysis

Data overview:

```r
# Data overview
j %>%
    arrange(country_code) %>%
    kable(caption = "<i><font color=#000000><b>Table 2.</b> Sample data</font></i
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 2. Sample data}

country_code	country	region	subregion
ABW	Aruba	Americas	Latin America and the Caribbea
AFG	Afghanistan	Asia	Southern Asia
AGO	Angola	Africa	Sub-Saharan Africa
AIA	Anguilla	Americas	Latin America and the Caribbea
ALB	Albania	Europe	Southern Europe
AND	Andorra	Europe	Southern Europe
ARE	United Arab Emirates	Asia	Western Asia
ARG	Argentina	Americas	Latin America and the Caribbea
ARM	Armenia	Asia	Western Asia
ASM	American Samoa	Oceania	Polynesia
ATG	Antigua and Barbuda	Americas	Latin America and the Caribbea
AUS	Australia	Oceania	Australia and New Zealand
AUT	Austria	Europe	Western Europe
AZE	Azerbaijan	Asia	Western Asia
BDI	Burundi	Africa	Sub-Saharan Africa
BEL	Belgium	Europe	Western Europe
BEN	Benin	Africa	Sub-Saharan Africa
BFA	Burkina Faso	Africa	Sub-Saharan Africa
BGD	Bangladesh	Asia	Southern Asia
BGR	Bulgaria	Europe	Eastern Europe
BHR	Bahrain	Asia	Western Asia
BHS	Bahamas	Americas	Latin America and the Caribbea
BIH	Bosnia and Herzegovina	Europe	Southern Europe
BLR	Belarus	Europe	Eastern Europe
BLZ	Belize	Americas	Latin America and the Caribbea
BMU	Bermuda	Americas	Northern America
BOL	Bolivia	Americas	Latin America and the Caribbea
BRA	Brazil	Americas	Latin America and the Caribbea
BRB	Barbados	Americas	Latin America and the Caribbea
BRN	Brunei	Asia	South-eastern Asia
BTN	Bhutan	Asia	Southern Asia
BWA	Botswana	Africa	Sub-Saharan Africa
CAF	Central African Republic	Africa	Sub-Saharan Africa
CAN	Canada	Americas	Northern America
CHE	Switzerland	Europe	Western Europe
CHL	Chile	Americas	Latin America and the Caribbea
CHN	China	Asia	Eastern Asia
CIV	Cote d'Ivoire	Africa	Sub-Saharan Africa
CMR	Cameroon	Africa	Sub-Saharan Africa
COD	Democratic Republic of Congo	Africa	Sub-Saharan Africa
COG	Congo	Africa	Sub-Saharan Africa
COK	Cook Islands	Oceania	Polynesia
COL	Colombia	Americas	Latin America and the Caribbea
COM	Comoros	Africa	Sub-Saharan Africa
CPV	Cape Verde	Africa	Sub-Saharan Africa

\end{table}

Literacy rates (dependent variable):

```
# Summary statistics for literacy rate
describe(j$lit_rate) %>%
    kable(caption = "<i><font color=#000000><b>Table 3.</b> Summary stats - liter
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```
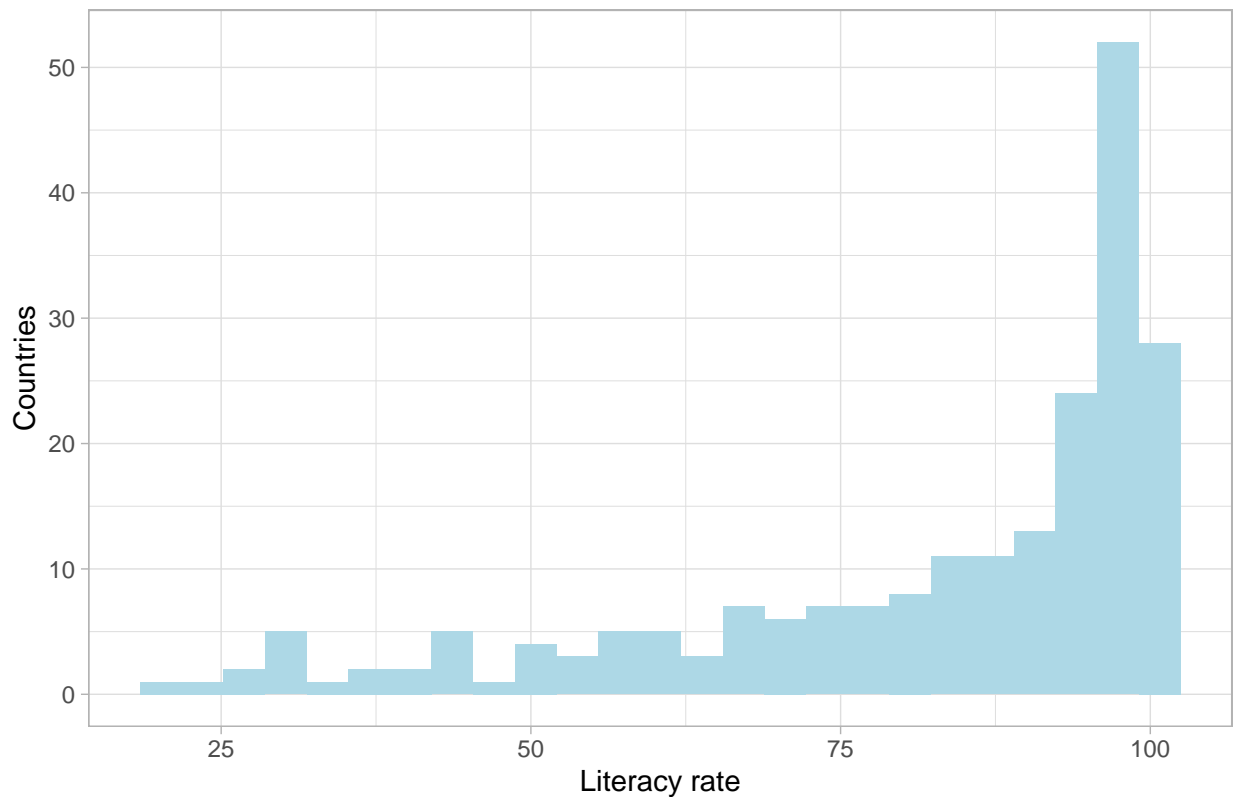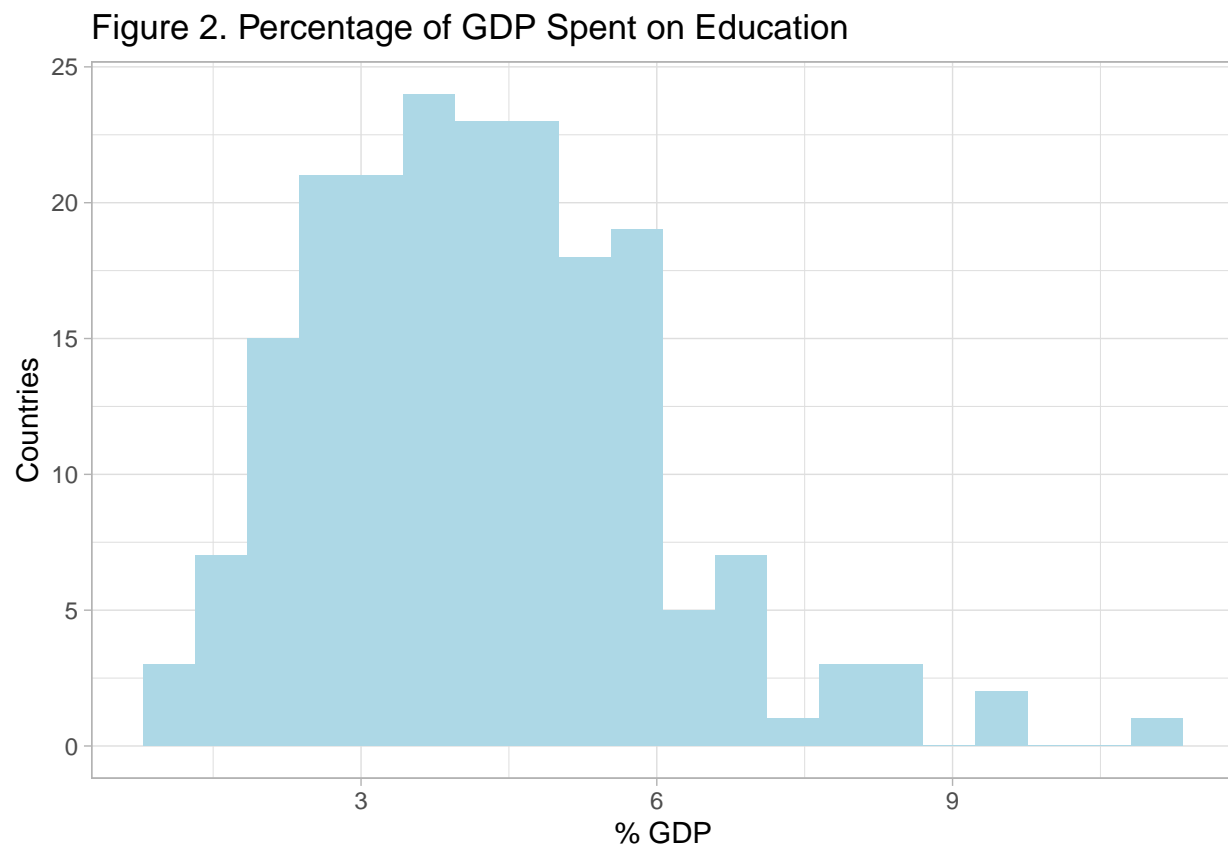
\begin{table}

\caption{Table 3. Summary stats - literacy rate}

	vars	n	mean	sd	median	trimmed	mad	min	max	rang
X1	1	214	82.51319	20.33653	90.89308	86.11195	12.01932	19.40194	100	80.59800

\end{table}

```
# Histogram of literacy rate
j %>%
    drop_na(lit_rate) %>%
    ggplot() +
    geom_histogram(aes(x = lit_rate), bins = 25, fill = 'lightblue') +
    xlab("Literacy rate") + ylab("Countries") +
    ggtitle("Figure 1. Literacy rate") +
    theme_light()
```

Figure 1. Literacy rate

Educational spending (independent, numerical variable):

```
# Summary statistics for percent GDP education spending
describe(j$pct_gdp) %>%
    kable(caption = "<i><font color=#000000><b>Table 4.</b> Summary stats - educa
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 4. Summary stats - educational spending}

	vars	n	mean	sd	median	trimmed	mad	min	max	rang
X1	1	196	4.27805	1.70355	4.071276	4.161513	1.74287	1.230985	11.25354	10.0225

\end{table}

```
# Histogram of GDP education spending
j %>%
    drop_na(pct_gdp) %>%
    ggplot() +
    geom_histogram(aes(x = pct_gdp), bins = 20, fill = 'lightblue') +
    xlab("% GDP") + ylab("Countries") +
```

```
    ggtitle("Figure 2. Percentage of GDP Spent on Education") +
    theme_light()
```

Figure 2. Percentage of GDP Spent on Education

Regions (independent, categorical variable):

```
# Regions table
regions %>%
    group_by(region, subregion) %>%
    summarize(n = n(), .groups = "keep") %>%
    select(-n) %>%
    arrange(region, subregion) %>%
    kable(caption = "<i><font color=#000000><b>Table 4.5.</b> Regions</font></i>"
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 4.5. Regions}

| region	subregion
Africa	Northern Africa
Africa	Sub-Saharan Africa
Americas	Latin America and the Caribbean
Americas	Northern America
Asia	Central Asia
Asia	Eastern Asia
Asia	South-eastern Asia
Asia	Southern Asia
Asia	Western Asia
Europe	Eastern Europe
Europe	Northern Europe
Europe	Southern Europe
Europe	Western Europe
Oceania	Australia and New Zealand
Oceania	Melanesia
Oceania	Micronesia
Oceania	Polynesia

\end{table}

```r
# Bar chart of regions
j %>%
    drop_na(subregion) %>%
    group_by(subregion) %>%
    summarize(n = n()) %>%
    ggplot() +
    geom_bar(aes(x = reorder(subregion, n), y = n), fill = 'lightblue', stat = "i
    xlab("Subregion") + ylab("Countries") +
    ggtitle("Figure 3. Subregions") +
    theme_light() +
    coord_flip()
```

Figure 3. Subregions
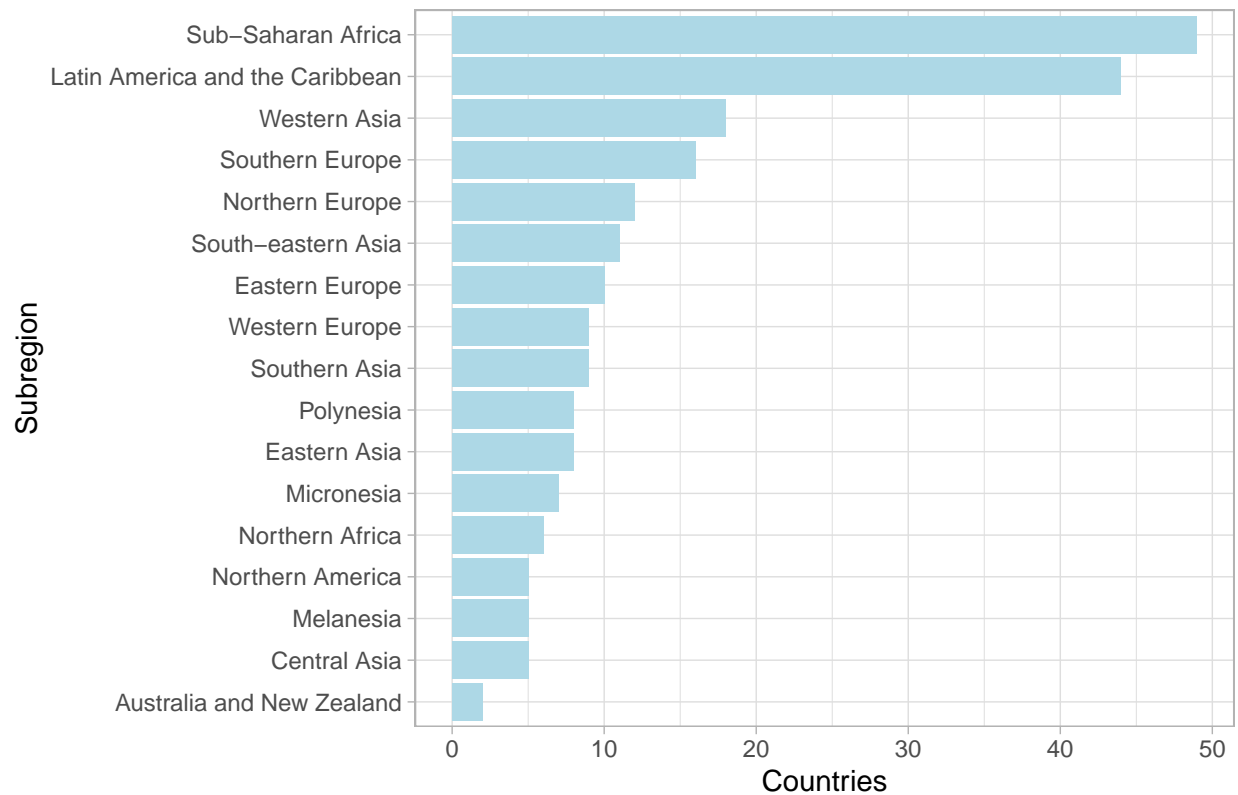
Forms of government (independent, categorical variable):

```r
# Governments table
govts %>%
    group_by(govt, head_of_state, exec_legitimacy) %>%
    summarize(num_countries = n(), .groups = "keep") %>%
    arrange(desc(num_countries)) %>%
    kable(caption = "<i><font color=#000000><b>Table 4.7.</b> Forms of government
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 4.7. Forms of government}

govt	head_of_state	exec_legitimacy
republic	executive	presidency is independent of legislature
republic	ceremonial	ministry is subject to parliamentary confidence
constitutional monarchy	ceremonial	ministry is subject to parliamentary confidence
republic	executive	presidency independent of legislature; ministry is sub
constitutional monarchy	executive	monarch personally exercises power in concert with c
republic	executive	presidency is elected by legislature; ministry may be,
absolute monarchy	executive	all authority vested in absolute monarch
provisional	no	no constitutionally-defined basis to current regime
republic	executive	power constitutionally linked to a single political mov
republic	ceremonial	power constitutionally linked to a single political mov

\end{table}

```
# Bar chart of governments
govts %>%
    group_by(govt_type_short) %>%
    summarize(n = n(), .groups = "keep") %>%
    ggplot() +
    geom_bar(aes(x = reorder(govt_type_short, n), y = n), fill = 'lightblue', sta
    xlab("Form of goverment") + ylab("Countries") +
    ggtitle("Figure 4. Forms of government") +
    theme_light() +
    coord_flip()
```

Figure 4. Forms of government
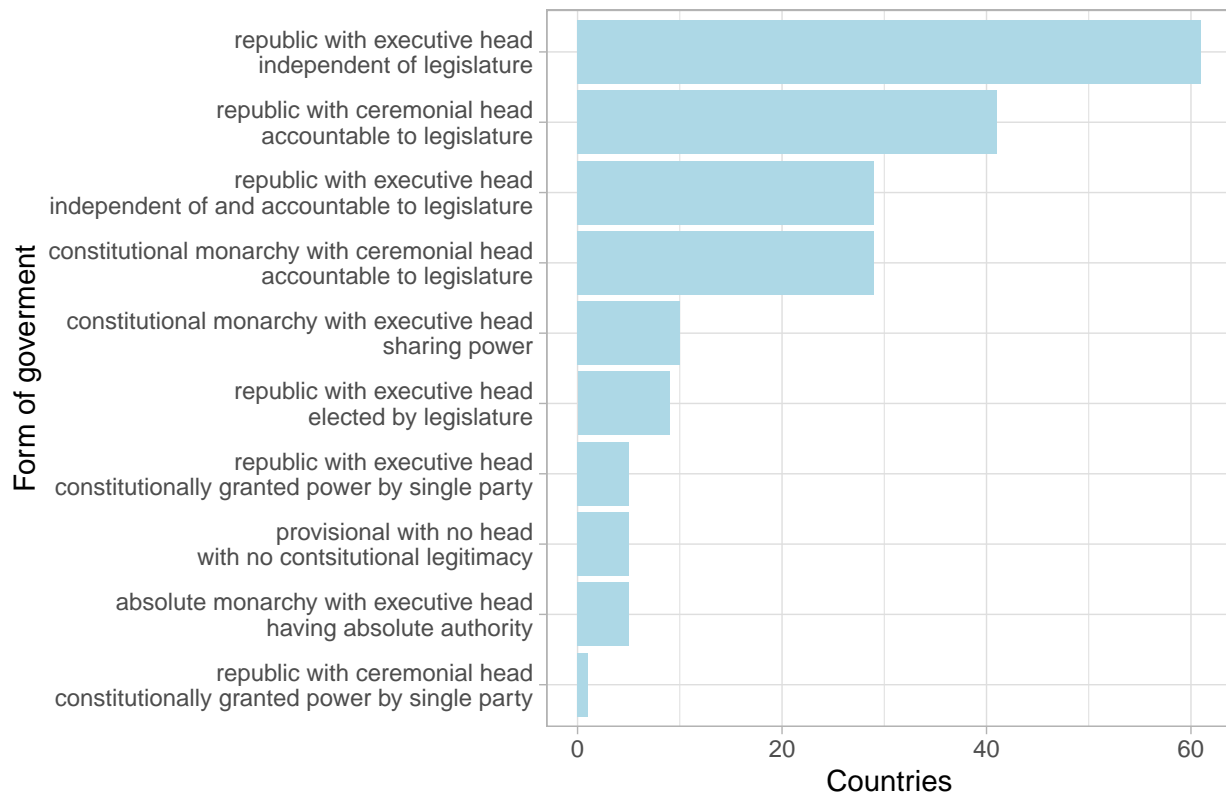
Explore relationships between explanatory and response variables:

```
# Scatter plot of literacy rate vs education spending
j %>%
    drop_na(pct_gdp, lit_rate) %>%
    ggplot() +
    geom_point(aes(x = pct_gdp, y = lit_rate)) +
    theme_light() +
    xlab("Percent GDP") +
    ylab("Literacy rate") +
    ggtitle("Figure 5. Literacy rate vs educational spending")
```
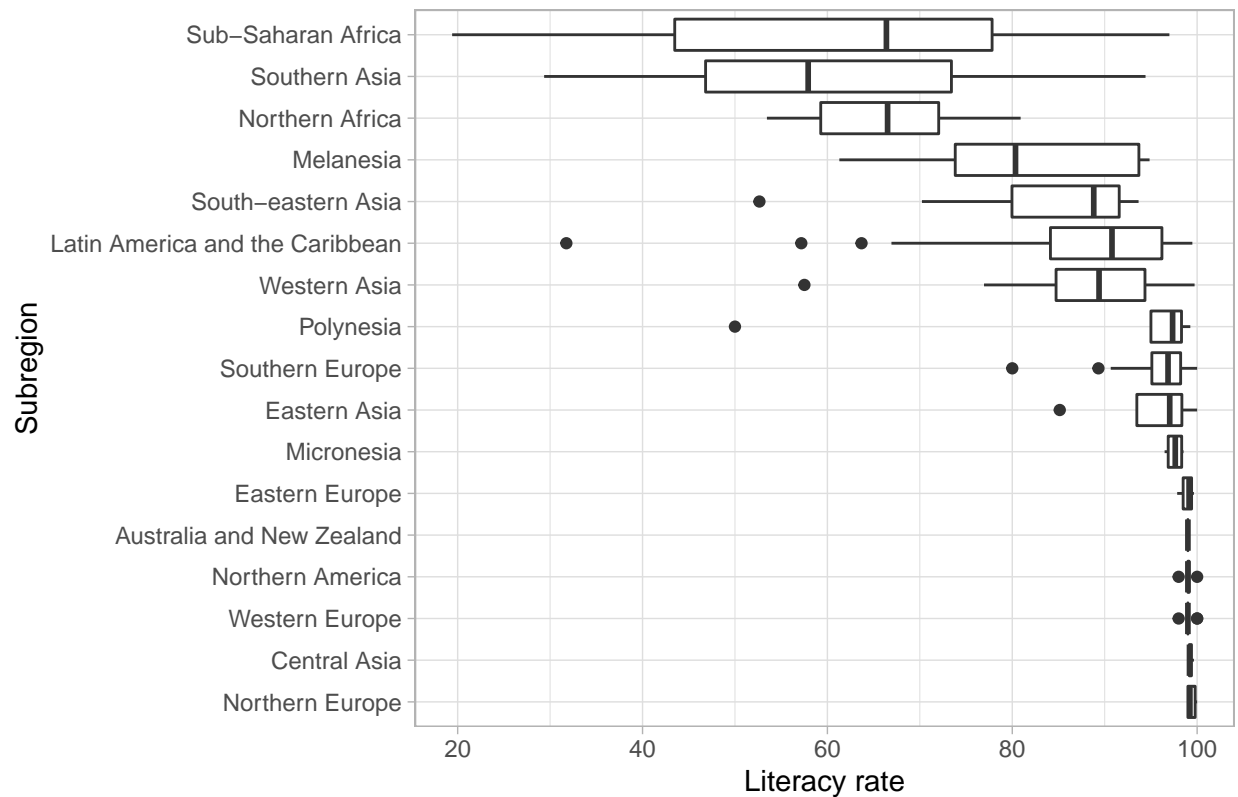
Figure 5. Literacy rate vs educational spending
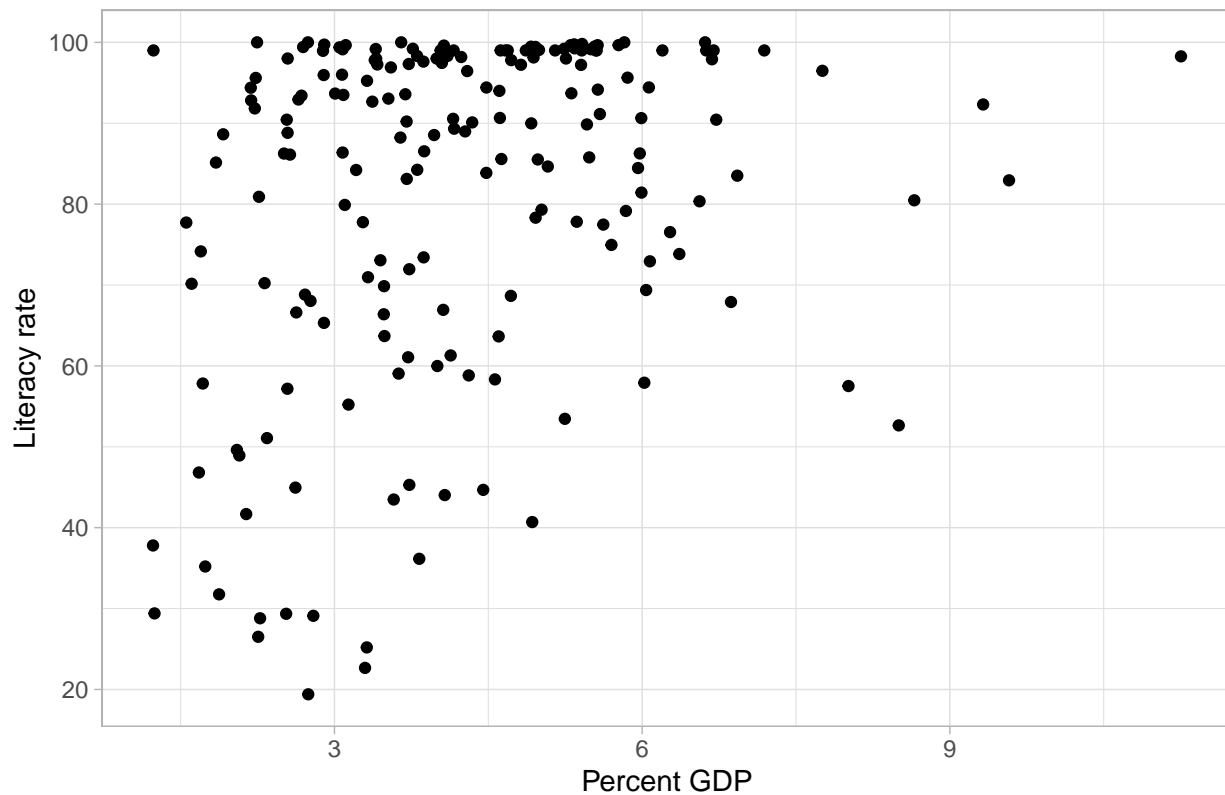

```
# Box plots of literacy rate by region
j %>%
    drop_na(lit_rate) %>%
    ggplot() +
    geom_boxplot(aes(x = reorder(subregion, -lit_rate), y = lit_rate)) +
    coord_flip() +
    theme_light() +
    xlab("Subregion") +
    ylab("Literacy rate") +
    ggtitle("Figure 6. Literacy rate by subregion")
```

Figure 6. Literacy rate by subregion

```r
# Box plots of literacy rate by form of government
j %>%
    drop_na(lit_rate) %>%
    drop_na(govt_type_short) %>%
    ggplot() +
    geom_boxplot(aes(x = reorder(govt_type_short, -lit_rate), y = lit_rate)) +
    xlab("Form of goverment") + ylab("Literacy rate") +
    ggtitle("Figure 7. Literacy rate by form of government") +
    theme_light() +
    coord_flip()
```

Figure 7. Literacy rate by form of government
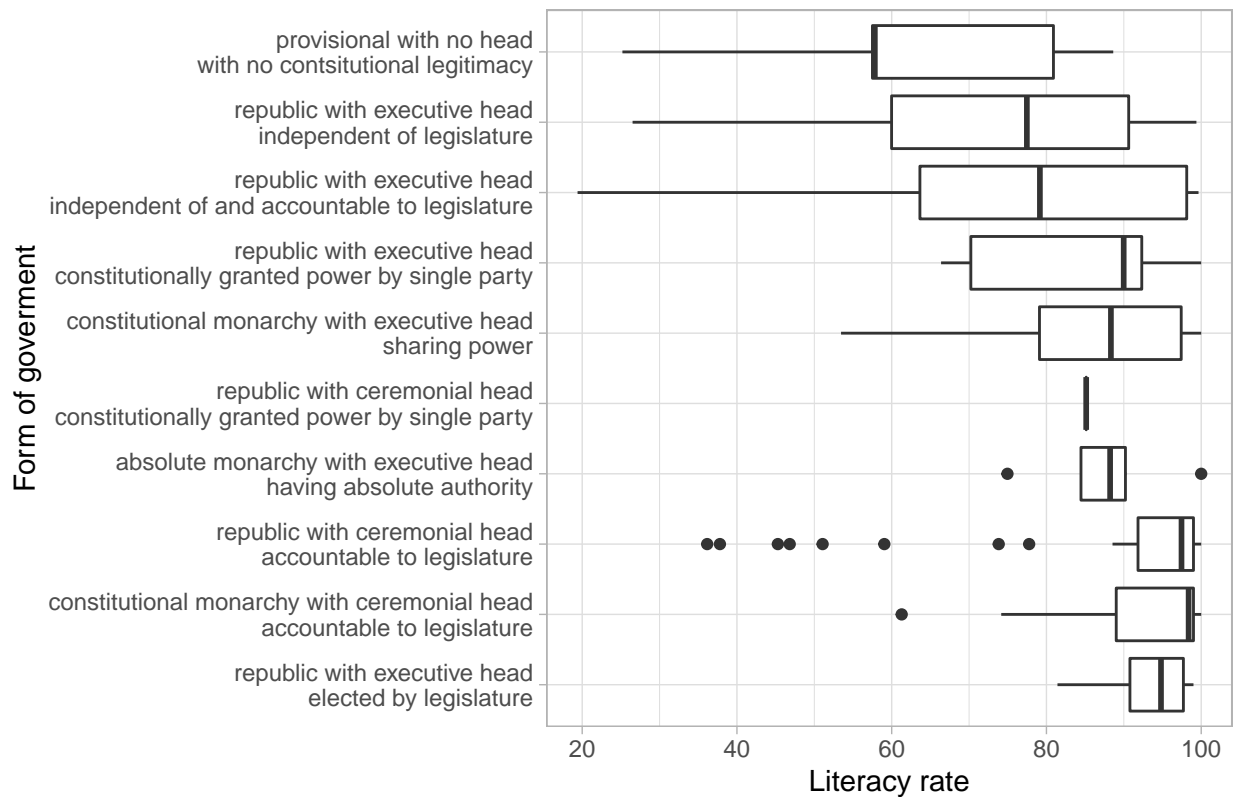

```r
# Set world class for maps
world <- ne_countries(scale = "medium", returnclass = "sf", type = "countries")

# Map of literacy rate
world %>%
    inner_join(j, by = c("iso_a3" = "country_code")) %>%
    drop_na(lit_rate) %>%
    ggplot() +
    geom_sf(aes(fill = lit_rate)) +
    scale_fill_viridis_c(option = "cividis", trans = "sqrt", direction = -1) +
    ggtitle("Figure 8. Literacy rate") +
    labs(fill = "Literacy rate")
```
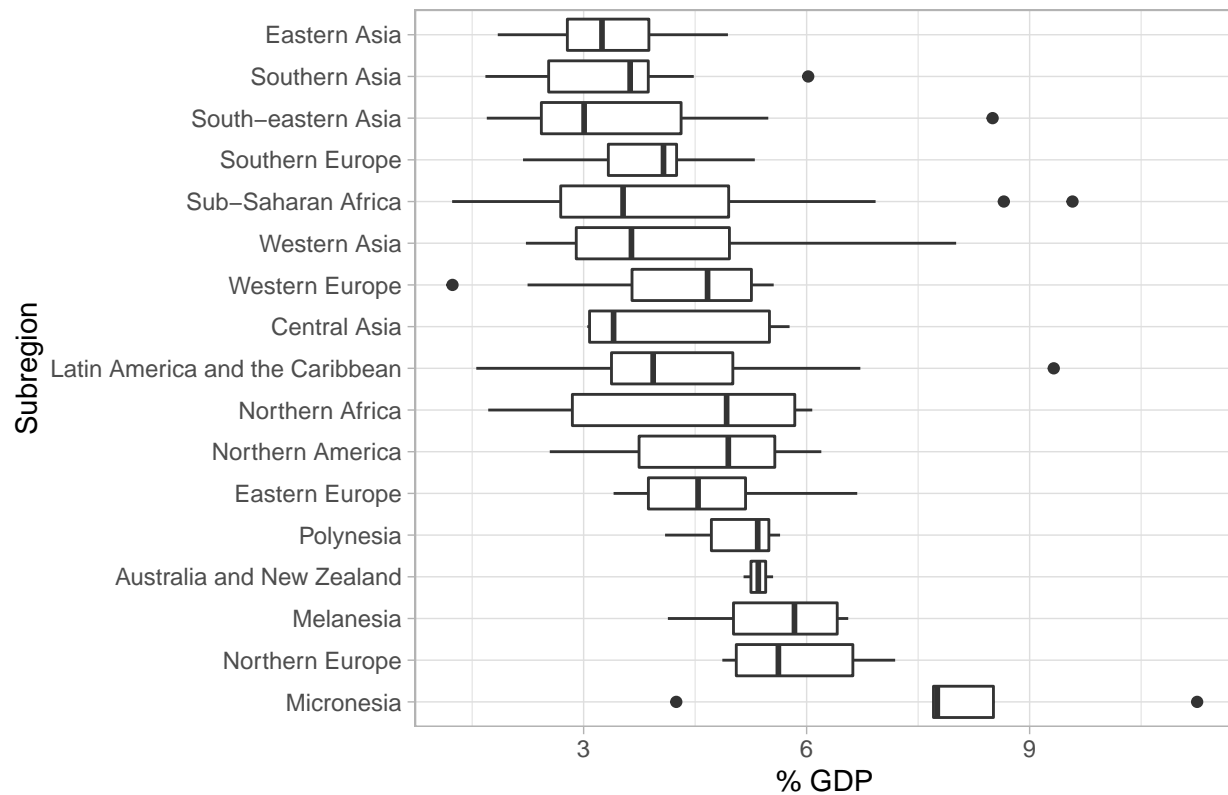
Figure 8. Literacy rate
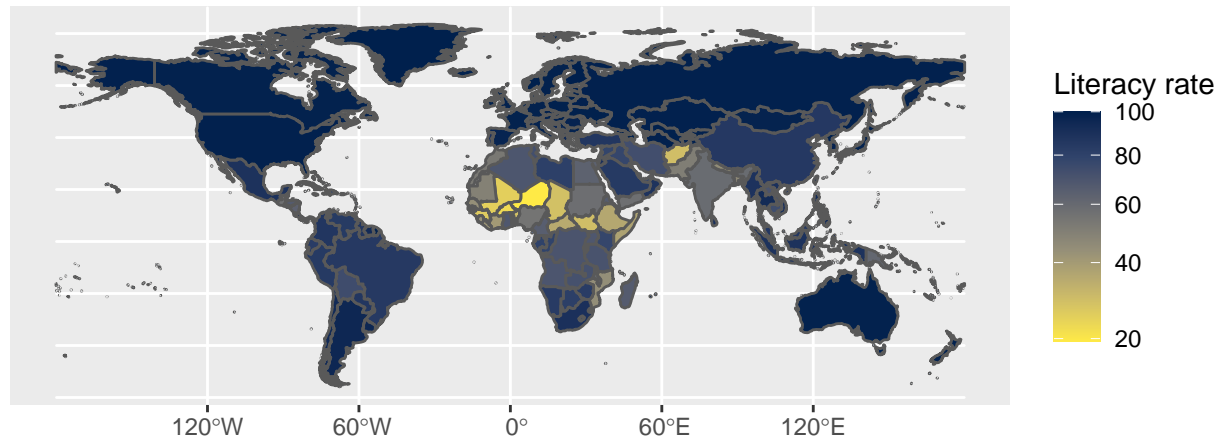
Explore independence of explanatory variables:

```r
# Box plots of ed spending by subregion
j %>%
    drop_na(pct_gdp) %>%
    ggplot() +
    geom_boxplot(aes(x = reorder(subregion, -pct_gdp), y = pct_gdp)) +
    coord_flip() +
    theme_light() +
    xlab("Subregion") +
    ylab("% GDP") +
    ggtitle("Figure 9. Educational spending by subregion")
```

Figure 9. Educational spending by subregion

```
# Box plots of ed spending by form of government
j %>%
    drop_na(pct_gdp, govt_type_short) %>%
    ggplot() +
    geom_boxplot(aes(x = reorder(govt_type_short, -pct_gdp), y = pct_gdp)) +
    coord_flip() +
    theme_light() +
    xlab("Form of government") +
    ylab("% GDP") +
    ggtitle("Figure 10. Educational spending by form of government")
```
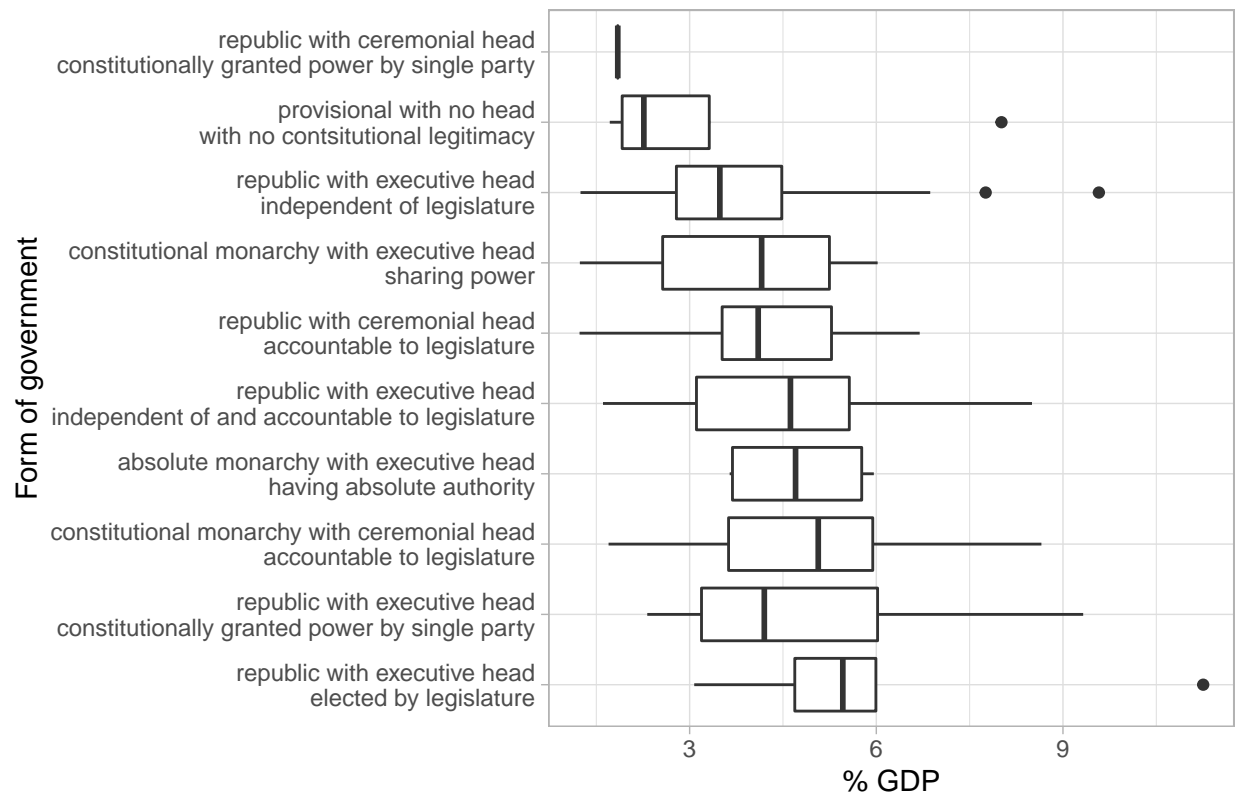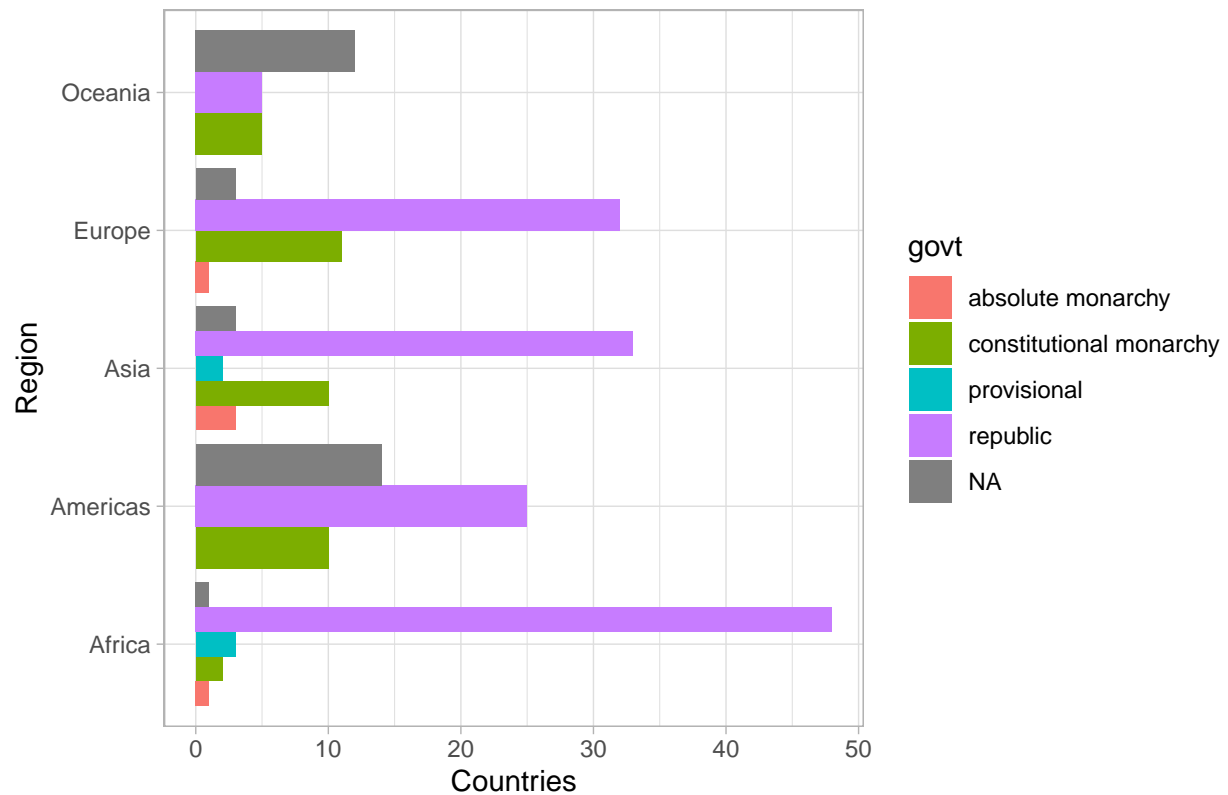
Figure 10. Educational spending by form of go

```
# Bar plot of form of government by region
j %>%
    ggplot() +
    geom_bar(aes(x = region, fill = govt), position = "dodge") +
    coord_flip() +
    theme_light() +
    xlab("Region") +
    ylab("Countries") +
    ggtitle("Figure 11. Form of government by region")
```

Figure 11. Form of government by region

```
# Map of subregions
world %>%
    inner_join(j, by = c("iso_a3" = "country_code"))  %>%
    ggplot() +
    geom_sf(aes(fill = subregion.y)) +
    ggtitle("Figure 12. Subregions of the world") +
    labs(fill = "Subregion")
```
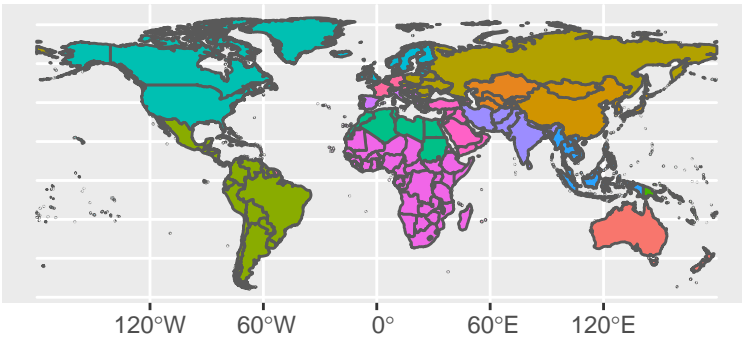
Figure 12. Subregions of the world

Subregion

- Australia and New Zealand
- Central Asia
- Eastern Asia
- Eastern Europe
- Latin America and the Caribbean
- Melanesia
- Micronesia
- Northern Africa
- Northern America
- Northern Europe
- Polynesia
- South–eastern Asia
- Southern Asia
- Southern Europe
- Sub–Saharan Africa
- Western Asia
- Western Europe

```
# Map of forms of government
world %>%
    inner_join(govts, by = c("iso_a3" = "country_code")) %>%
    ggplot() +
    geom_sf(aes(fill = govt)) +
    scale_fill_brewer(palette="PuBu") +
    ggtitle("Figure 13. Forms of government") +
    labs(fill = "Government")
```
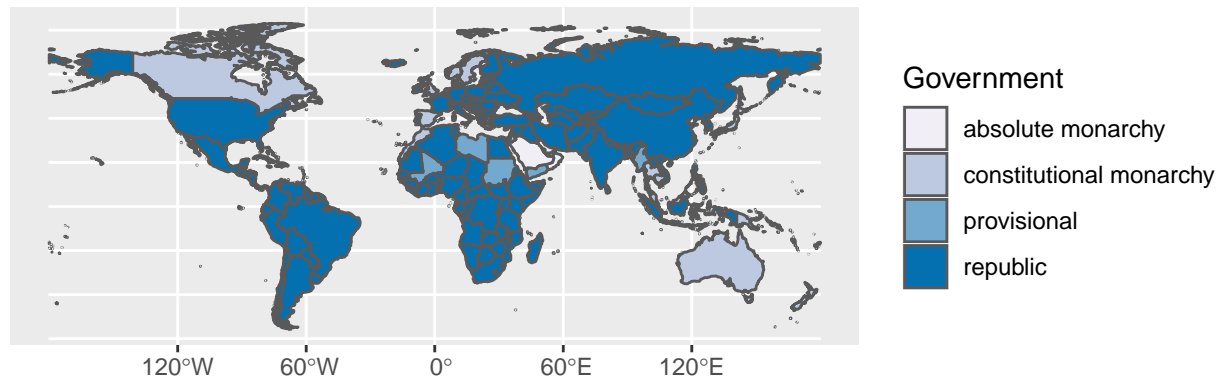
Figure 13. Forms of government

Government

- absolute monarchy
- constitutional monarchy
- provisional
- republic

```r
# Map of educational spending
world %>%
    inner_join(j, by = c("iso_a3" = "country_code")) %>%
    drop_na(pct_gdp) %>%
    ggplot() +
    geom_sf(aes(fill = pct_gdp)) +
    scale_fill_viridis_c(option = "cividis", trans = "sqrt", direction = -1) +
    ggtitle("Figure 14. Educational spending") +
    labs(fill = "% GDP")
```

Figure 14. Educational spending
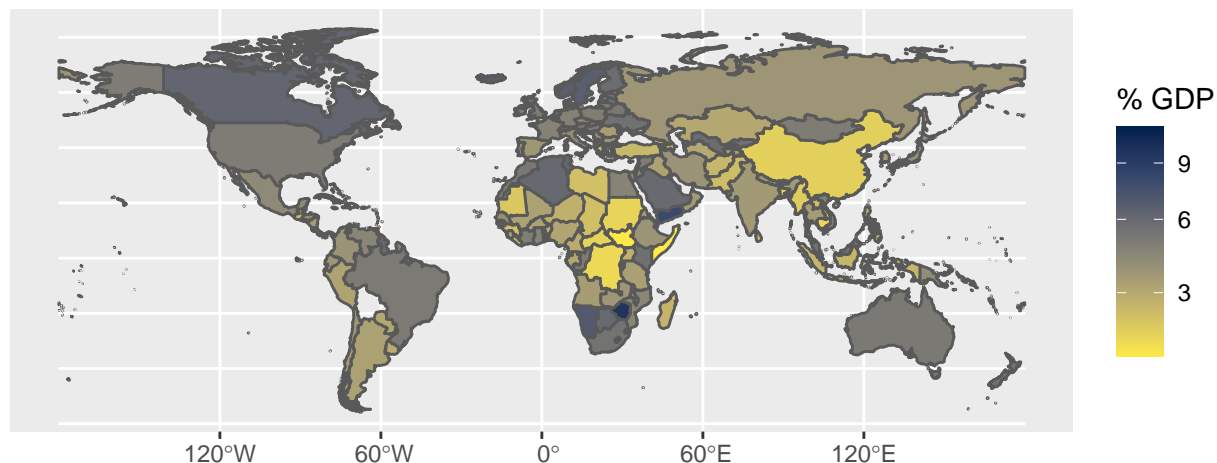
Part 4 - Inference

One goal of the project was to evaluate whether there is a statistically significant difference in literacy rate and educational spending across regions as well as across government types. To evaluate this, a series of ANOVA analyses were performed. The null hypothesis for each analysis was that there was no difference in rate for the variable in question; for example, for the first test, $H_0 =$ there is no difference in educational spending rate among subregions. The alternative hypothesis would be the opposite: $H_A =$ there is a statistically significant difference in educational spending rate among subregions. Similar comparisons were made for the remaining ANOVA tests.

It is noted that not all of the conditions for ANOVA were met. Independence was assumed, but while education spending was fairly normal across subregions and government types, the same could not be said of literacy rate, which was strongly left-skewed. Similarly, the box plots also show that variance was not constant for literacy rates, while it was more constant for educational spending.

```
# Literacy rates by subregion
region_summ <- j %>%
    group_by(region, subregion) %>%
```

```r
    summarize(
        n_pct_gdp = n(),
        sd_pct_gdp = sd(pct_gdp, na.rm = T),
        mean_pct_gdp = mean(pct_gdp, na.rm = T),
        n_lit_rate = n(),
        sd_lit_rate = sd(lit_rate, na.rm = T),
        mean_lit_rate = mean(lit_rate, na.rm = T),
        .groups = "keep"
    )

# Literacy rates by govt
govt_summ <- j %>%
    drop_na(govt_type_short) %>%
    group_by(govt_type_short) %>%
    summarize(
        n_pct_gdp = n(),
        sd_pct_gdp = sd(pct_gdp, na.rm = T),
        mean_pct_gdp = mean(pct_gdp, na.rm = T),
        n_lit_rate = n(),
        sd_lit_rate = sd(lit_rate, na.rm = T),
        mean_lit_rate = mean(lit_rate, na.rm = T),
        .groups = "keep"
    ) %>%
    mutate(sd_pct_gdp = ifelse(is.na(sd_pct_gdp), 0, sd_pct_gdp)) %>%
    mutate(sd_lit_rate = ifelse(is.na(sd_lit_rate), 0, sd_lit_rate))

# Find overall rate - ed spending
mean_pct_gdp = mean(j$pct_gdp, na.rm = T)
sd_pct_gdp = sd(j$pct_gdp, na.rm = T)
print(paste0("Mean spending on education across all countries = ", round(mean_pct
```

```
## [1] "Mean spending on education across all countries = 4.278%, sd = 1.704"
```

```r
# Find overall rate - literacy rate
mean_lit_rate = mean(j$lit_rate, na.rm = T)
sd_lit_rate = sd(j$lit_rate, na.rm = T)
print(paste0("Mean literacy rate across all countries = ", round(mean_lit_rate, 3
```

```
## [1] "Mean literacy rate across all countries = 82.513%, sd = 20.337"
```

```r
print("")
```

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

```
# Histograms - ed spending by subregion
j %>%
    drop_na(pct_gdp) %>%
    ggplot(aes(x = pct_gdp)) +
    geom_histogram(binwidth = 0.5) +
    facet_wrap(~subregion, ncol = 3) +
    ggtitle("Figure 15. Educational spending by subregion") +
    xlab("Educational spending (% GDP)") + ylab("")
```
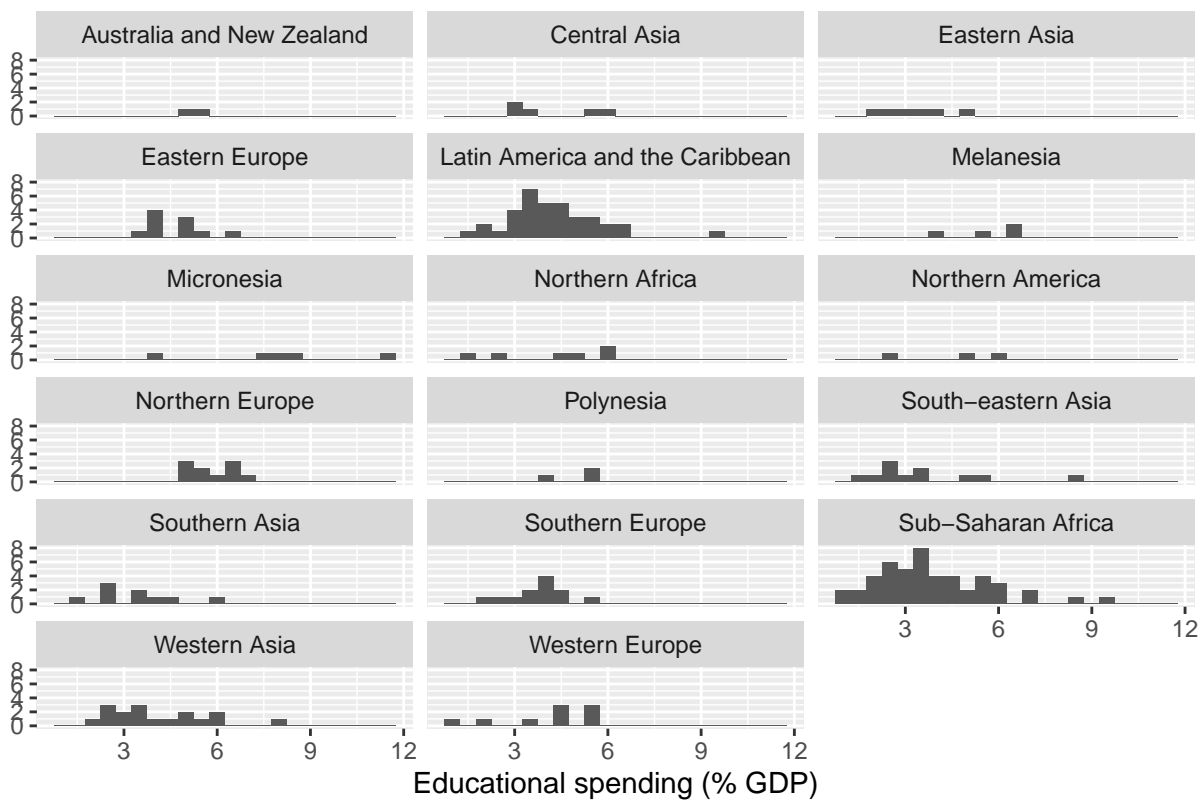
Figure 15. Educational spending by subregion


```
# Histograms - literacy rate by subregion
j %>%
    drop_na(lit_rate) %>%
    ggplot(aes(x = lit_rate)) +
    geom_histogram(binwidth = 7) +
    facet_wrap(~subregion, ncol = 3) +
    ggtitle("Figure 16. Literacy rate by subregion") +
    xlab("Literacy rate (%)") + ylab("")
```
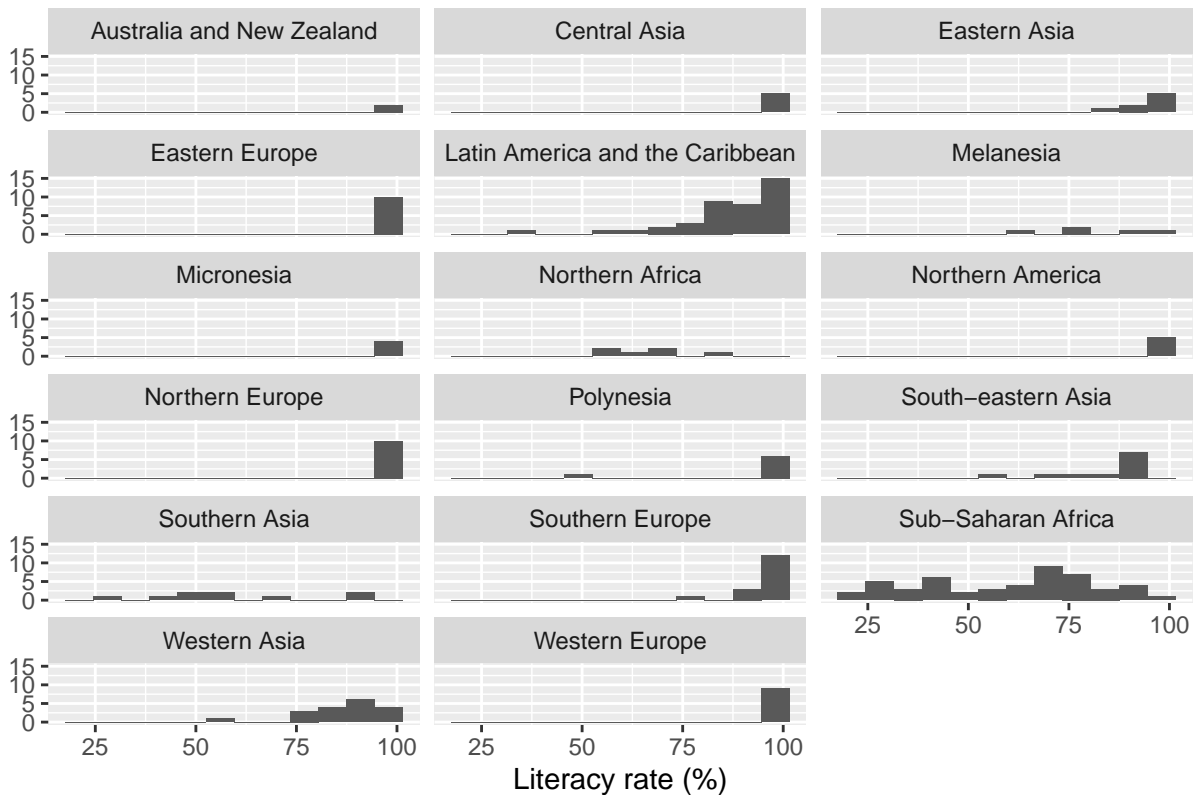
Figure 16. Literacy rate by subregion


```
# Histograms - ed spending by govt
j %>%
    drop_na(pct_gdp) %>%
    ggplot(aes(x = pct_gdp)) +
    geom_histogram(binwidth = 0.5) +
    facet_wrap(~govt_type_short, ncol = 3) +
    ggtitle("Figure 17. Educational spending by government type") +
    xlab("Educational spending (% GDP)") + ylab("")
```

Figure 17. Educational spending by government type

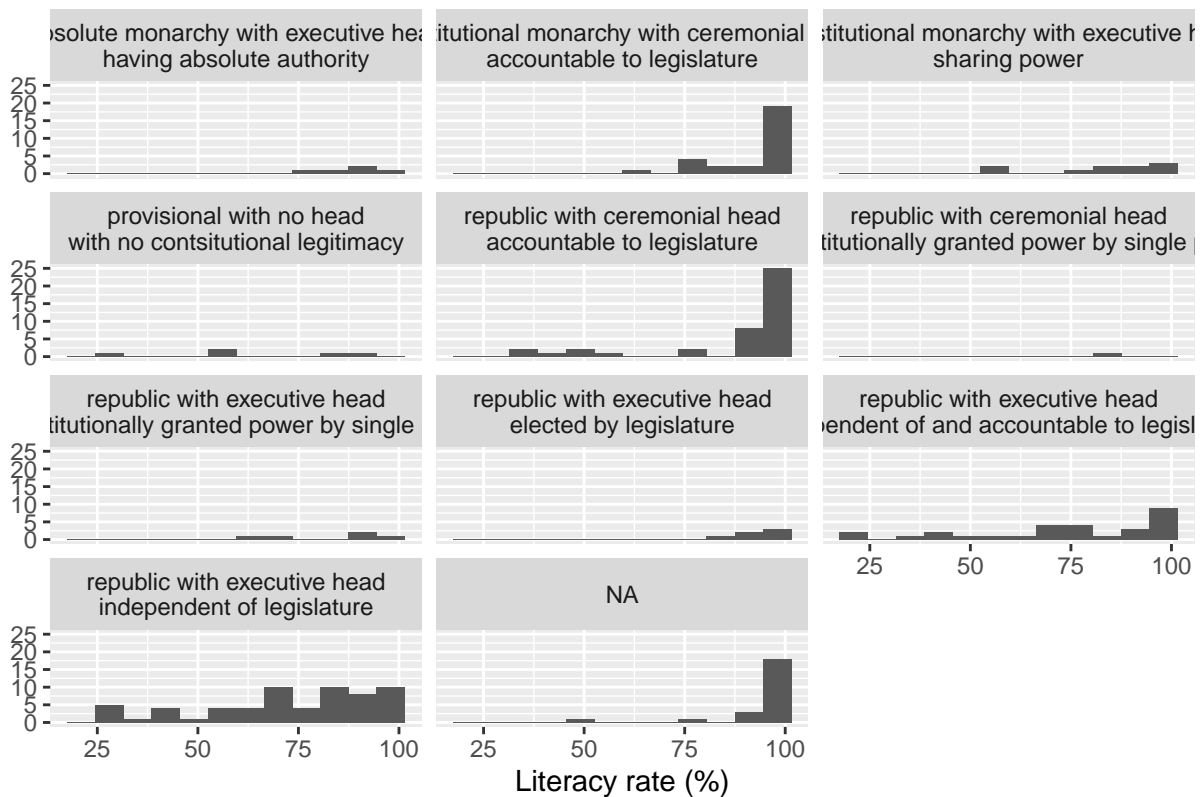Figure 17. Educational spending by government type
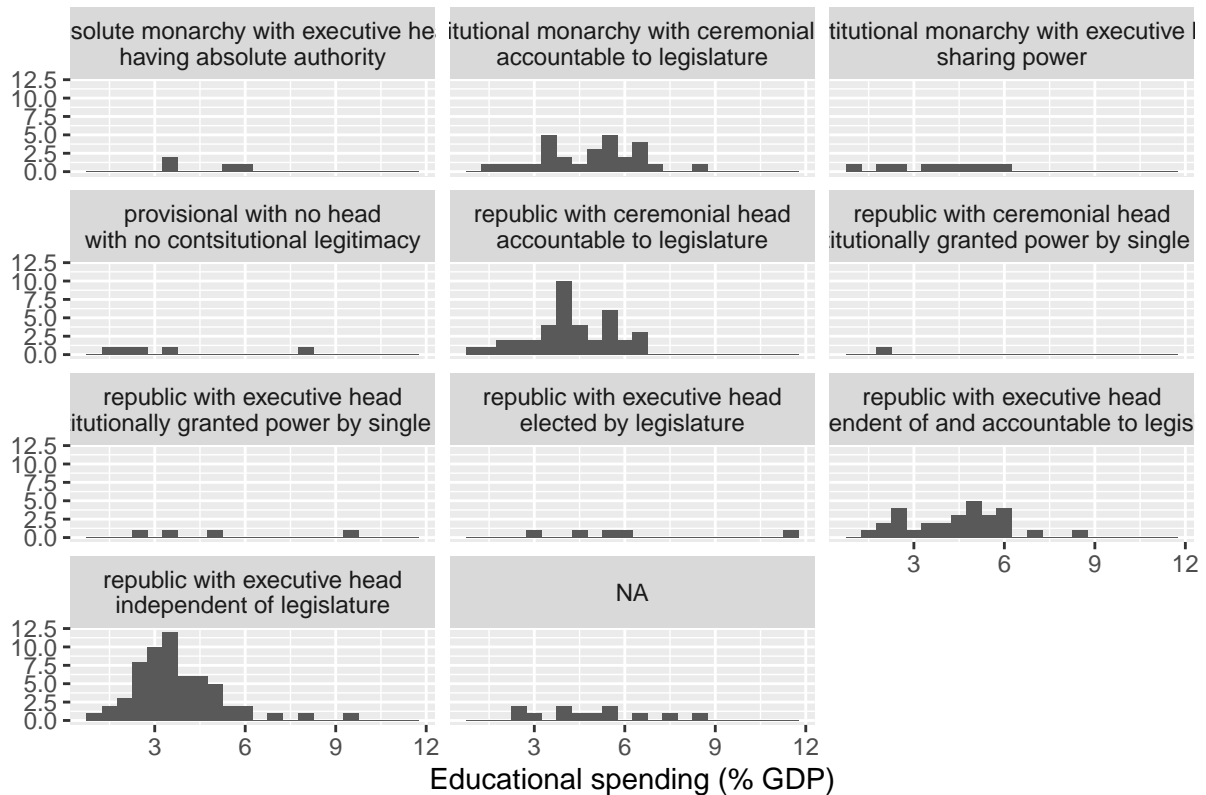
```
# Histograms - literacy rate by govt
j %>%
    drop_na(lit_rate) %>%
    ggplot(aes(x = lit_rate)) +
    geom_histogram(binwidth = 7) +
    facet_wrap(~govt_type_short, ncol = 3) +
    ggtitle("Figure 18. Literacy rate by government type") +
    xlab("Literacy rate (%)") + ylab("")
```

Figure 18. Literacy rate by government type


```r
# ANOVA - ed spending vs subregion
aov_ed_region <- aov(pct_gdp ~ subregion, data = j)
print(paste0("ANOVA results - educational spending vs subregion"))
```

```
## [1] "ANOVA results - educational spending vs subregion"
```

```r
summary(aov_ed_region)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## subregion     16  128.1   8.006   3.273 5.33e-05 ***
## Residuals    179  437.8   2.446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 28 observations deleted due to missingness
```

```r
# ANOVA - ed spending vs govt
aov_ed_govt <- aov(pct_gdp ~ govt_type_short, data = j)
print(paste0("ANOVA results - educational spending vs government type"))
```

```
## [1] "ANOVA results - educational spending vs government type"
```

```
summary(aov_ed_govt)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## govt_type_short   9   56.2   6.244   2.348 0.0159 *
## Residuals       174  462.8   2.660
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 40 observations deleted due to missingness
```

```
# ANOVA - lit rate vs subregion
aov_lit_region <- aov(lit_rate ~ subregion, data = j)
print(paste0("ANOVA results - literacy rate vs subregion"))
```

```
## [1] "ANOVA results - literacy rate vs subregion"
```

```
summary(aov_lit_region)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## subregion   16  48057  3003.5   14.78 <2e-16 ***
## Residuals  197  40035   203.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 10 observations deleted due to missingness
```

```
# ANOVA - lit rate vs govt
aov_lit_govt <- aov(lit_rate ~ govt_type_short, data = j)
print(paste0("ANOVA results - literacy rate vs government type"))
```

```
## [1] "ANOVA results - literacy rate vs government type"
```

```
summary(aov_lit_govt)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## govt_type_short   9  14384    1598    4.25 5.15e-05 ***
## Residuals       181  68065     376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 33 observations deleted due to missingness
```

Next, a linear model was fit to the literacy rate vs ed spending data. It should be noted that not all of the conditions for least squares are *not* met:

1. Linearity: The data appear to be fairly linear, with a positive relationship.
2. Nearly normal residuals: As shown in the histogram and QQ plot, the residuals are significantly skewed to the left.
3. Constant variability: Variability isn't constant, but narrows for higher values.
4. Independent observations: The observations are independent.

A second model was fit that ignored high-leverage points (percent gdp > 7.75%). Contrary to what I expected, the line didn't change much.

```
# Linear model - literacy rate vs ed spending
lit_lm <- lm(lit_rate ~ pct_gdp, data = j, na.action = na.omit)
summary(lit_lm)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp, data = j, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -56.631  -11.855    7.824   15.197   28.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.2369     3.9417  16.804  < 2e-16 ***
## pct_gdp       3.5691     0.8671   4.116 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.03 on 189 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.08227,    Adjusted R-squared:  0.07742
## F-statistic: 16.94 on 1 and 189 DF,  p-value: 5.749e-05
```

```
# Line of best fit
j %>%
    drop_na(pct_gdp, lit_rate) %>%
    ggplot(aes(x = pct_gdp, y = lit_rate)) +
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm", se = T) +
    coord_cartesian(ylim = c(0, 100)) +
    xlab("Government spending (% GDP)") + ylab("Literacy rate (%)") +
    ggtitle("Figure 19. Literacy rate vs educational spending")
```

Figure 19. Literacy rate vs educational spending
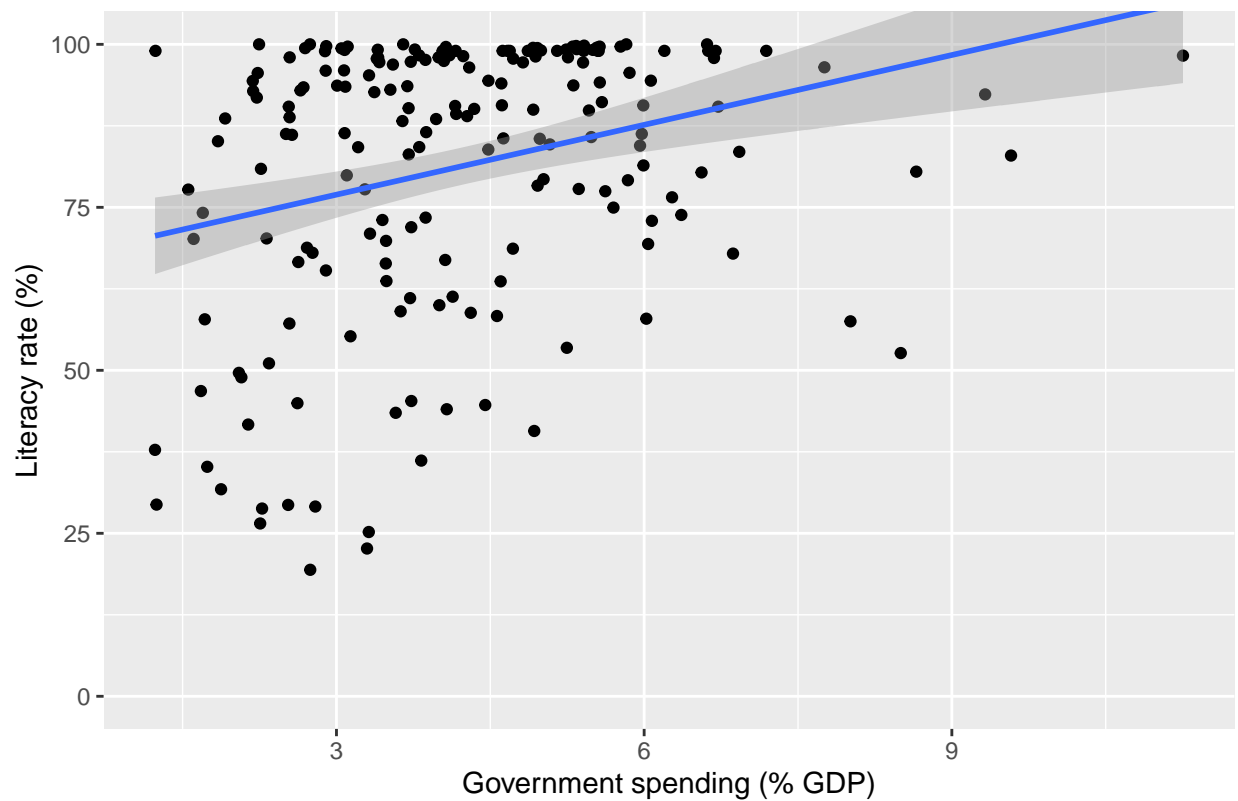
```r
# Residuals plot
lit_lm %>%
    ggplot(aes(x = .fitted, y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    xlab("Fitted values") + ylab("Residuals") +
    ggtitle("Figure 20. Fitted values vs residuals")
```

Figure 20. Fitted values vs residuals
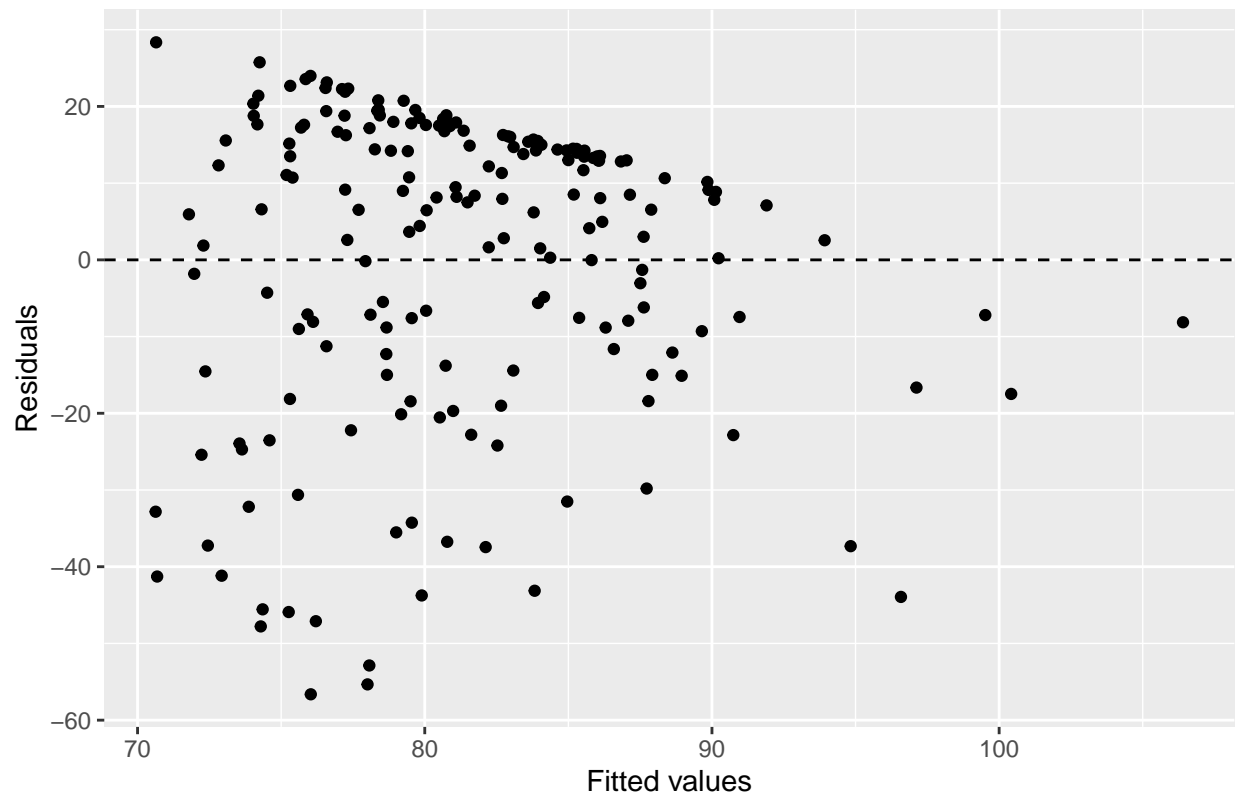
```
# Residuals histogram
lit_lm %>%
    ggplot(aes(x = .resid)) +
    geom_histogram(binwidth = 4) +
    xlab("Residuals") + ylab("Count") +
    ggtitle("Figure 21. Residuals histogram")
```

Figure 21. Residuals histogram
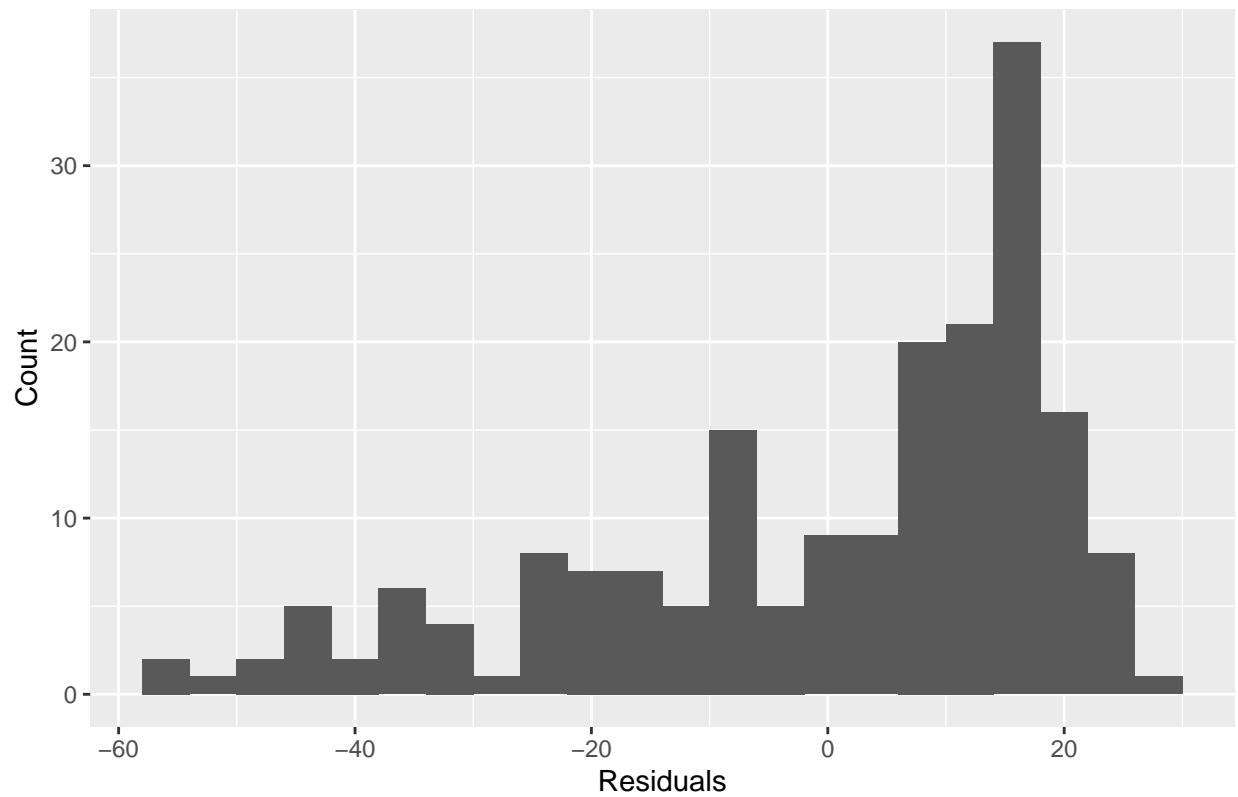

```
# Probability plot
lit_lm %>%
    ggplot(aes(sample = .resid)) +
    stat_qq() +
    xlab("Theoretical values") + ylab("Actual values") +
    ggtitle("Figure 22. Probability plot")
```

Figure 22. Probability plot
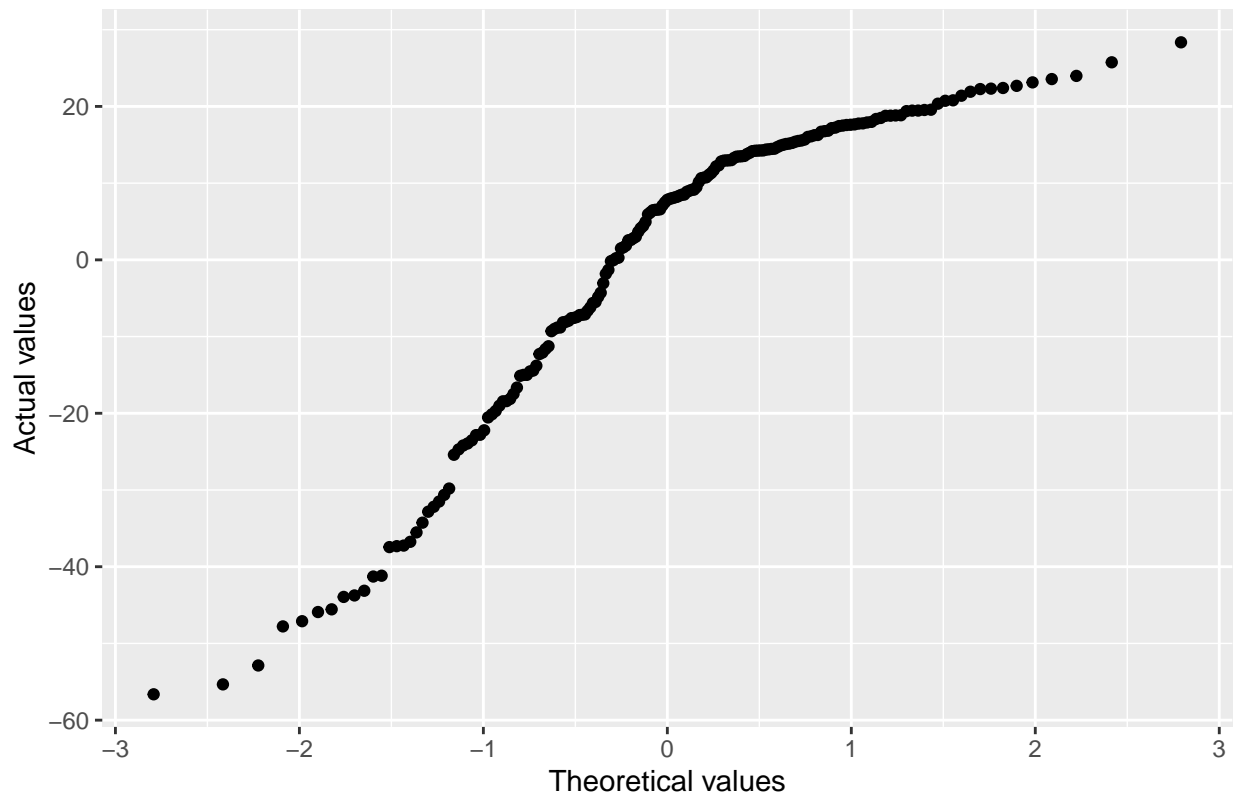

```r
# Removing high-leverage/influential points
lit_lm_2 <- lm(lit_rate ~ pct_gdp, data = filter(j, pct_gdp < 7.75), na.action =
summary(lit_lm_2)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp, data = filter(j, pct_gdp < 7.75),
##     na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.135 -12.064   6.308  14.451  32.390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   60.118      4.452   13.50  < 2e-16 ***
## pct_gdp        5.254      1.040    5.05 1.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 19.7 on 182 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.1229, Adjusted R-squared:  0.1181
## F-statistic:  25.5 on 1 and 182 DF,  p-value: 1.067e-06
```

```
j %>%
    drop_na(pct_gdp, lit_rate) %>%
    filter(pct_gdp < 7.75) %>%
    ggplot(aes(x = pct_gdp, y = lit_rate)) +
    geom_point() +
    geom_smooth(formula = y ~ x, method = "lm", se = T) +
    coord_cartesian(ylim = c(0, 100)) +
    xlab("Fitted values") + ylab("Residuals") +
    ggtitle("Figure 22.5. Literacy rate vs educational spending (high-leverage va
```


Figure 22.5. Literacy rate vs educational spending (high–leverage values re
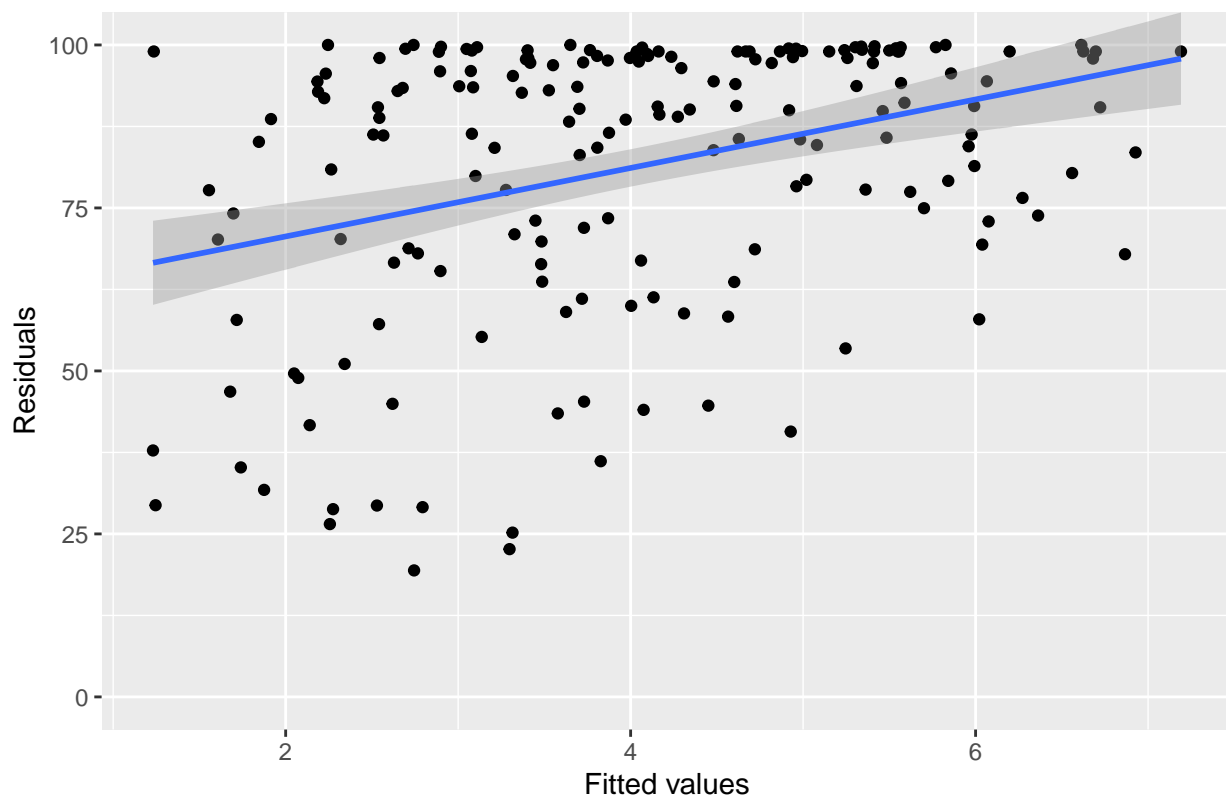
One additional goal of the project was to evaluate whether adding the two categorical variables (subregion and government type) would influence the linear model. First, a backward elimination by p-value strategy was used to eliminate variables. But that resulted in removing the pct_gdp variable and ending up with two categorical variables having p-values greater than 0.05. So a different strategy was employed using forward selection.

```r
# Prepare variables
incl_pct_gdp <- c('include', '', '', 'include', 'include', 'include')
incl_subregion <- c('', 'include', '', 'include', '', 'include')
incl_govt_type <- c('', '', 'include', '', 'include', 'include')
adjrsq <- c()

# Fit linear model - pct_gdp
lm_mult <- lm(lit_rate ~ pct_gdp, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp, data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.631 -11.855   7.824  15.197  28.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  66.2369     3.9417  16.804  < 2e-16 ***
## pct_gdp       3.5691     0.8671   4.116 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.03 on 189 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.08227,   Adjusted R-squared:  0.07742
## F-statistic: 16.94 on 1 and 189 DF,  p-value: 5.749e-05
```

```r
# adj r-squared = 0.07742

# subregion
lm_mult <- lm(lit_rate ~ subregion, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ subregion, data = j)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q     Max
## -55.177  -2.589   0.404   7.421  36.752
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           99.00000   10.08022   9.821  < 2e-16
## subregionCentral Asia                  0.29941   11.92707   0.025 0.979998
## subregionEastern Asia                 -3.59469   11.27003  -0.319 0.750095
## subregionEastern Europe               -0.07527   11.04233  -0.007 0.994568
## subregionLatin America and the Caribbean -12.06678  10.32915  -1.168 0.244127
## subregionMelanesia                   -18.19513   11.92707  -1.526 0.128730
## subregionMicronesia                   -1.43254   12.34569  -0.116 0.907743
## subregionNorthern Africa             -32.64453   11.63963  -2.805 0.005543
## subregionNorthern America             0.02667   11.92707   0.002 0.998218
## subregionNorthern Europe              0.39791   11.04233   0.036 0.971291
## subregionPolynesia                    -8.53966   11.42989  -0.747 0.455873
## subregionSouth-eastern Asia          -15.35649   10.95835  -1.401 0.162683
## subregionSouthern Asia               -38.13625   11.14410  -3.422 0.000756
## subregionSouthern Europe              -3.62906   10.69169  -0.339 0.734649
## subregionSub-Saharan Africa          -38.75206   10.28388  -3.768 0.000217
## subregionWestern Asia                -11.17725   10.62548  -1.052 0.294121
## subregionWestern Europe               0.11111   11.14410   0.010 0.992055
##
## (Intercept)                          ***
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa             **
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia               ***
## subregionSouthern Europe
## subregionSub-Saharan Africa          ***
## subregionWestern Asia
## subregionWestern Europe
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 14.26 on 197 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.5455, Adjusted R-squared:  0.5086
## F-statistic: 14.78 on 16 and 197 DF,  p-value: < 2.2e-16
```

```
# adj r-squared = 0.5086

# govt type
lm_mult <- lm(lit_rate ~ govt_type_short, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ govt_type_short, data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.511  -7.458   5.459  11.145  26.619
##
## Coefficients:
##
## (Intercept)
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
```

```
##
## (Intercept)
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.39 on 181 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.1745, Adjusted R-squared:  0.1334
## F-statistic:  4.25 on 9 and 181 DF,  p-value: 5.15e-05

# adj r-squared = 0.1334
```

```
# Now start with pct_gdp type and add in subregion
lm_mult <- lm(lit_rate ~ pct_gdp + subregion, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp + subregion, data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.675  -3.745   1.107   7.960  37.623
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             85.7399    10.4087   8.237 4.17e-14
## pct_gdp                                  2.4785     0.6754   3.670 0.000323
## subregionCentral Asia                    3.2480    11.5778   0.281 0.779399
## subregionEastern Asia                    0.5249    11.3536   0.046 0.963178
## subregionEastern Europe                  1.6350    10.7033   0.153 0.878771
## subregionLatin America and the Caribbean -9.9564   10.0655  -0.989 0.323966
## subregionMelanesia                     -22.3053    11.9564  -1.866 0.063797
## subregionMicronesia                    -11.9280    14.0872  -0.847 0.398316
## subregionNorthern Africa               -30.1010    11.2929  -2.665 0.008416
## subregionNorthern America                1.6622    12.6132   0.132 0.895308
## subregionNorthern Europe                -0.8357    10.6984  -0.078 0.937825
## subregionPolynesia                       1.5042    13.8114   0.109 0.913402
## subregionSouth-eastern Asia            -11.1689    10.6730  -1.046 0.296810
## subregionSouthern Asia                 -33.3619    10.8698  -3.069 0.002492
## subregionSouthern Europe                 1.0278    10.5937   0.097 0.922821
## subregionSub-Saharan Africa            -35.9655    10.0097  -3.593 0.000426
## subregionWestern Asia                   -7.5503    10.3518  -0.729 0.466760
## subregionWestern Europe                  3.0897    10.8222   0.285 0.775604
##
## (Intercept)                            ***
## pct_gdp                                ***
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia                       .
## subregionMicronesia
## subregionNorthern Africa               **
```

```
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia                    **
## subregionSouthern Europe
## subregionSub-Saharan Africa               ***
## subregionWestern Asia
## subregionWestern Europe
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.8 on 173 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.5616
## F-statistic: 15.32 on 17 and 173 DF,  p-value: < 2.2e-16
```

```
# adj r-squared = 0.5616 (higher than pct_gdp alone)

# pct_gdp type + govt_type
lm_mult <- lm(lit_rate ~ pct_gdp + govt_type_short, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp + govt_type_short, data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.908  -9.014   4.168  11.318  31.083
##
## Coefficients:
##
## (Intercept)
## pct_gdp
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
```

```
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
```

```
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.15 on 173 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.2177, Adjusted R-squared:  0.1724
## F-statistic: 4.813 on 10 and 173 DF,  p-value: 4.179e-06
```

```r
# adj r-squared = 0.1724 (lower than pct_gdp alone)

# pct_gdp type + subregion + govt_type
lm_mult <- lm(lit_rate ~ pct_gdp + subregion + govt_type_short, data = j)
adjrsq <- c(adjrsq, summary(lm_mult)$adj.r.squared)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp + subregion + govt_type_short,
##     data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.598  -4.943   0.617   7.198  36.567
##
## Coefficients:
##
## (Intercept)
## pct_gdp
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa
## subregionNorthern America
```

```
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia
## subregionSouthern Europe
## subregionSub-Saharan Africa
## subregionWestern Asia
## subregionWestern Europe
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia
## subregionSouthern Europe
## subregionSub-Saharan Africa
## subregionWestern Asia
## subregionWestern Europe
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
```

46

```
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia
## subregionSouthern Europe
## subregionSub-Saharan Africa
## subregionWestern Asia
## subregionWestern Europe
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
```

```
## subregionSouth-eastern Asia
## subregionSouthern Asia
## subregionSouthern Europe
## subregionSub-Saharan Africa
## subregionWestern Asia
## subregionWestern Europe
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le,
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
##
## (Intercept)
## pct_gdp
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia
## subregionMicronesia
## subregionNorthern Africa
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia
## subregionSouthern Europe
## subregionSub-Saharan Africa
## subregionWestern Asia
## subregionWestern Europe
## govt_type_shortconstitutional monarchy with ceremonial head\naccountable to le,
## govt_type_shortconstitutional monarchy with executive head\nsharing power
## govt_type_shortprovisional with no head\nwith no contsitutional legitimacy
## govt_type_shortrepublic with ceremonial head\naccountable to legislature
## govt_type_shortrepublic with ceremonial head\nconstitutionally granted power b
## govt_type_shortrepublic with executive head\nconstitutionally granted power by
## govt_type_shortrepublic with executive head\nelected by legislature
## govt_type_shortrepublic with executive head\nindependent of and accountable to
## govt_type_shortrepublic with executive head\nindependent of legislature
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.05 on 157 degrees of freedom
##   (40 observations deleted due to missingness)
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.5546
## F-statistic: 9.765 on 26 and 157 DF,  p-value: < 2.2e-16
```

```
# adj r-squared = 0.5546 (lower than pct_gdp and subregion)

# data frame
df_models <- data.frame(adj_rsquared = adjrsq, pct_gdp = incl_pct_gdp,
    subregion = incl_subregion, govt_type = incl_govt_type)
df_models %>%
    kable(caption = "<i><font color=#000000><b>Table 5.</b> Linear model results<
    kable_styling(bootstrap_options = c("striped", "hover", "condensed"), font_si
```

\begin{table}

\caption{Table 5. Linear model results}

adj_rsquared	pct_gdp	subregion	govt_type
0.0774153	include		
0.5086218		include	
0.1334152			include
0.5616186	include	include	
0.1724391	include		include
0.5546388	include	include	include

\end{table}

```
# Best model:
lm_mult <- lm(lit_rate ~ pct_gdp + subregion, data = j)
summary(lm_mult)
```

```
##
## Call:
## lm(formula = lit_rate ~ pct_gdp + subregion, data = j)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.675  -3.745   1.107   7.960  37.623
##
## Coefficients:
```

```
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            85.7399    10.4087   8.237 4.17e-14
## pct_gdp                                 2.4785     0.6754   3.670 0.000323
## subregionCentral Asia                   3.2480    11.5778   0.281 0.779399
## subregionEastern Asia                   0.5249    11.3536   0.046 0.963178
## subregionEastern Europe                 1.6350    10.7033   0.153 0.878771
## subregionLatin America and the Caribbean -9.9564  10.0655  -0.989 0.323966
## subregionMelanesia                     -22.3053   11.9564  -1.866 0.063797
## subregionMicronesia                    -11.9280   14.0872  -0.847 0.398316
## subregionNorthern Africa               -30.1010   11.2929  -2.665 0.008416
## subregionNorthern America               1.6622    12.6132   0.132 0.895308
## subregionNorthern Europe               -0.8357    10.6984  -0.078 0.937825
## subregionPolynesia                      1.5042    13.8114   0.109 0.913402
## subregionSouth-eastern Asia            -11.1689   10.6730  -1.046 0.296810
## subregionSouthern Asia                 -33.3619   10.8698  -3.069 0.002492
## subregionSouthern Europe                1.0278    10.5937   0.097 0.922821
## subregionSub-Saharan Africa            -35.9655   10.0097  -3.593 0.000426
## subregionWestern Asia                   -7.5503   10.3518  -0.729 0.466760
## subregionWestern Europe                 3.0897    10.8222   0.285 0.775604
##
## (Intercept)                            ***
## pct_gdp                                ***
## subregionCentral Asia
## subregionEastern Asia
## subregionEastern Europe
## subregionLatin America and the Caribbean
## subregionMelanesia                      .
## subregionMicronesia
## subregionNorthern Africa               **
## subregionNorthern America
## subregionNorthern Europe
## subregionPolynesia
## subregionSouth-eastern Asia
## subregionSouthern Asia                 **
## subregionSouthern Europe
## subregionSub-Saharan Africa            ***
## subregionWestern Asia
## subregionWestern Europe
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.8 on 173 degrees of freedom
##   (33 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.5616
## F-statistic: 15.32 on 17 and 173 DF,  p-value: < 2.2e-16
```

Using forward selection, the best model is from using educational spending by itself, without subregion or government type as additional predictors.

Part 5 - Conclusion

Based on the results of this analysis, literacy rate is confirmed to be impacted positively by educational spending. It is noted that the impact is fairly steep, with every percentage point of GDP spent on education having a three-percent rise in literacy rate.

The results of the ANOVA analysis indicate strongly that there is a statistically significant difference between both literacy rate and educational spending among regions and, additionally, among various government types. Not surprisingly, higher literacy rates are enjoyed by republics having an elected head of state who is accountable to his or her constituents, while lower literacy rates were observed in countries with provisional governments. While it was expected that literacy rates in sub-Saharan Africa would be among the lowest, it was somewhat surprising that Central Asia boasted the highest overall literacy rates. These countries include the former Soviet republics of Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan. And while their overall literacy rate is, indeed, very high, according to the UN's definition of literacy, many people in these countries are "functionally" illiterate.[5] (UNESCO defines literacy as "the ability to use reading, writing and numeracy skills for effective functioning and development of the individual and the community. A person is literate who can, with understanding, both read and write a short statement on his or her everyday life."[6] Conversely, "a person is functionally literate who can engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing, and calculation for his own and the community's development."[7]) Another interesting outlier is the amount of money spent on education by Micronesia. While its average 7.9% of GDP spent on education, it should be noted that the U.S. funds around 90% of that amount in exchange for access to the islands for military purposes.[8]

As noted above, not all conditions for ANOVA were met: Literacy rate data were significantly left-skewed, and the variance was not constant. Similarly, not all conditions for fitting a linear model were met: The residuals were skewed to the left, and variability narrowed at higher values.

In conclusion, there was significant statistical evidence to suggest literacy rate is also affected by subregion of the world and by the type of government in place. Perhaps more significantly, there was strong correlation between government

spending on education (as a percentage of GDP) and literacy rate, with a three-fold increase in literacy for every percentage point increase in education spending.

References

[1] https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes/blob/master/all/all.csv [2] https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS [3] https://commons.wikimedia.org/wiki/Data:Cross-country_literacy_rates_-_World_Bank,_CIA_World_Factbook,_and_other_sources_(OWID_2762).tab [4] https://en.wikipedia.org/wiki/List_of_countries_by_system_of_government [5] https://www.dvv-international.de/en/adult-education-and-development/editions/aed-662006/education-for-all-and-literacy/review-of-central-asian-countries [6] https://www.fipfoundation.org/pictograms-support/about-health-literacy/#:~:text=UNESCO%20provides%20a%20definition%20of,his%20or%20her%20everyday%20life.%2 [7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5102880/ [8] https://www.grassrootinstitute.org/2016/09/us-policies-may-hurt-micronesians-more-than-they-help/