Analysis Harvard Library Data

Reka Forgo

2025-11-25

```
# Load necessary libraries
pacman::p_load(ggplot2, dplyr, tidyr, dslabs, car, pastecs, broom, performance, see)
```

```
# Load the Harvard library dataset
data <- read.csv("all_data_02.csv")

#head(data)

data <- read.csv("all_data_02.csv")


#rename language column
colnames(data)[colnames(data) == "languge"] <- "language"


#keep only text resources and drop rows with missing essential fields

data_clean <- data %>%
  filter(
    resource_type == "text",
    !is.na(title),
    !is.na(author),
    !is.na(publication_date),
    !is.na(language),
    !is.na(genre),
    !is.na(creation_date),
    publication_date != "")
```

```
#remove "No linguistic content & Not applicable"

data_clean <- data_clean %>%
  filter(language != "No linguistic content; Not applicable")

#save csv clean
write.csv(data_clean, "data_cleaned.csv", row.names = FALSE)

#unique(data_clean$language)
```

```
#count english books
english_books <- data_clean %>%
  filter(language == "English") %>% nrow()
```

```r
english_books
```

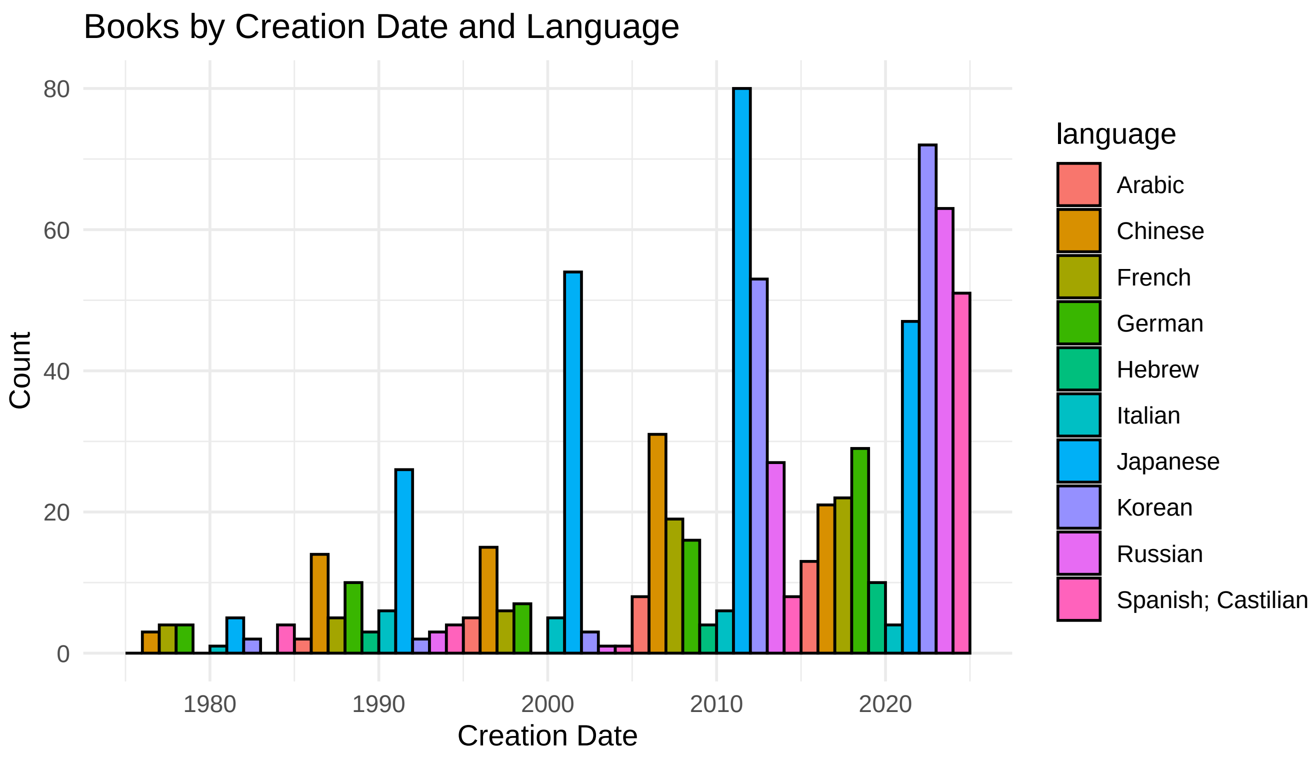
```
[1] 4690
```

```r
#publication date format

data_clean <- data_clean %>%
mutate(
  creation_date = gsub("[^0-9]", "", creation_date),
  creation_date = ifelse(nchar(creation_date) == 4, creation_date, NA),
  creation_date = as.integer(creation_date))


range(data_clean$creation_date, na.rm = TRUE)
```

```
[1] 1969 2027
```

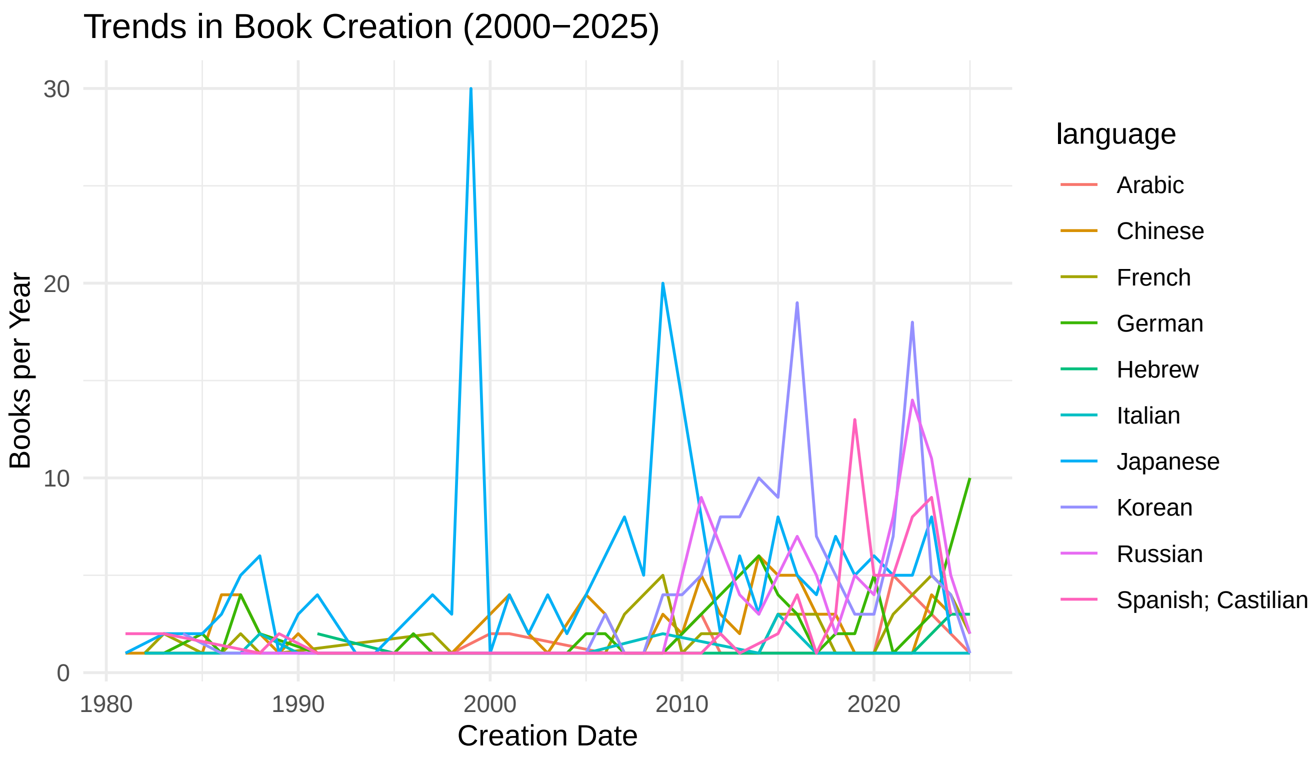
```r
#filter for 1980 - 2025

data_clean <- data_clean %>%
  filter(creation_date >= 1980 & creation_date <= 2025)
```

```r
#plot number of books in selected languages by publication date

top_languages <- data_clean %>%
  count(language, sort = TRUE) %>%
  slice_head(n = 11) %>%
  pull(language)

data_top <- data_clean %>%
  filter(language %in% top_languages, language != "English")


ggplot(data_top, aes(x = creation_date, fill = language)) +
  geom_histogram(binwidth = 10, position = "dodge", color = "black") +
  labs(title = "Books by Creation Date and Language",
       x = "Creation Date",
       y = "Count") +
  theme_minimal()
```

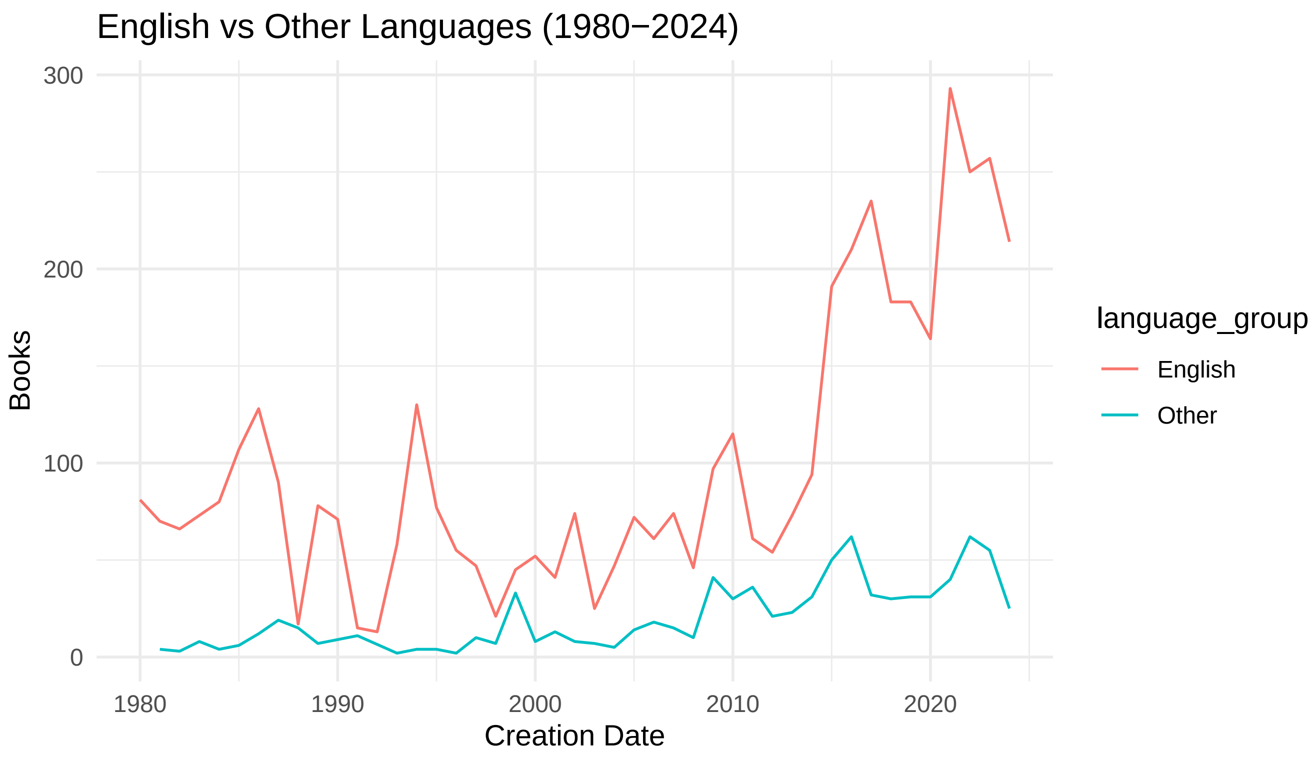


```
data_yearly <- data_clean %>%
  count(creation_date, name = "total_books")

data_language_yearly <- data_clean %>%
  count(creation_date, language, name = "language_books")
```

```
trend_data <- data_language_yearly %>%
  filter(creation_date >= 1980, creation_date <= 2025,
         language %in% top_languages, language != "English")

ggplot(trend_data, aes(x = creation_date, y = language_books, color = language)) +
  geom_line() +
  labs(title = "Trends in Book Creation (2000-2025)",
       x = "Creation Date",
       y = "Books per Year") +
  theme_minimal()
```



```r
data_grouped <- data_clean %>%
  mutate(language_group = ifelse(language == "English", "English", "Other")) %>%
  count(creation_date, language_group) %>%
  filter(between(creation_date, 1980, 2024))

ggplot(data_grouped, aes(x = creation_date, y = n, color = language_group)) +
  geom_line() +
  labs(title = "English vs Other Languages (1980-2024)",
       x = "Creation Date",
       y = "Books") +
  theme_minimal()
```



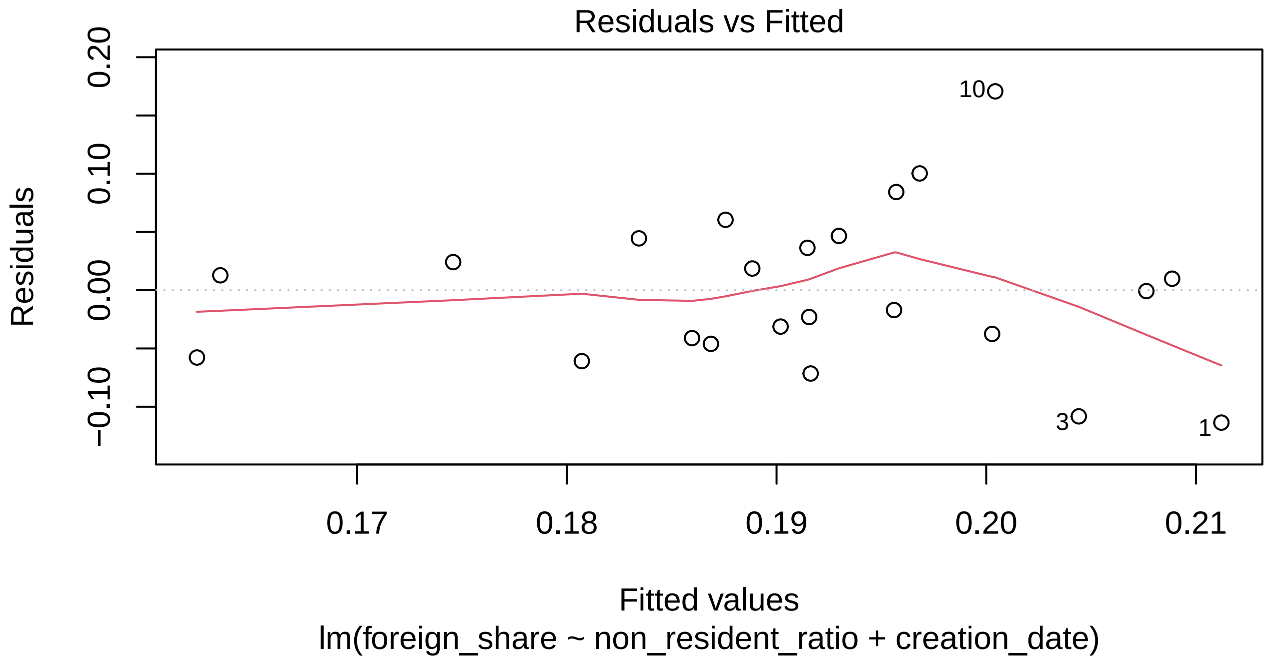
```
demographics <- read.csv("harvard_demographics - corrected.csv") %>%
  mutate(non_resident_ratio = non_resident / total_nr,
         resident_ratio = (total_nr - non_resident) / total_nr)

data_grouped <- data_grouped %>%
  left_join(demographics, by = c("creation_date" = "year"))
```

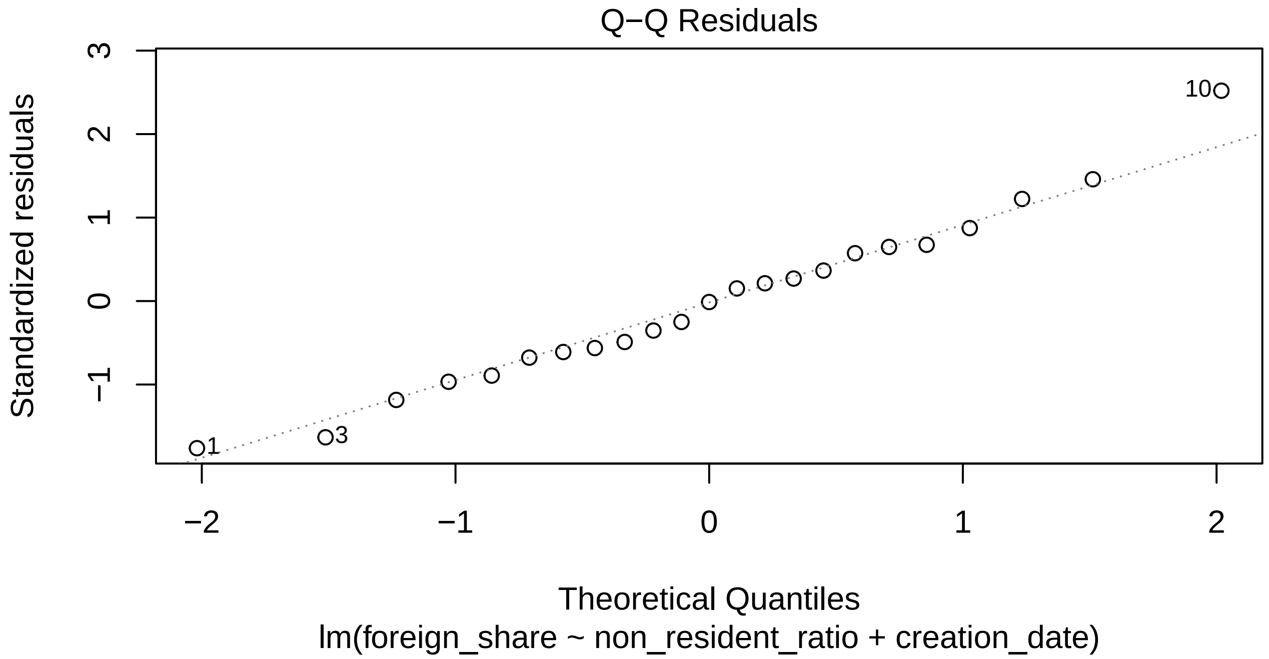
```
data_yearly <- data_clean %>%
  count(creation_date, name = "total_books")

# collapse languages into English vs Foreign
data_pooled <- data_clean %>%
  mutate(language_group = ifelse(language == "English", "English", "Foreign")) %>%
  count(creation_date, language_group, name = "books")

# reshape into wide form: English, Foreign per year
data_pooled <- data_pooled %>%
  tidyr::pivot_wider(
    names_from = language_group,
    values_from = books,
    values_fill = 0
  )

# join totals
data_pooled <- data_pooled %>%
  left_join(data_yearly, by = "creation_date")

# compute shares
```

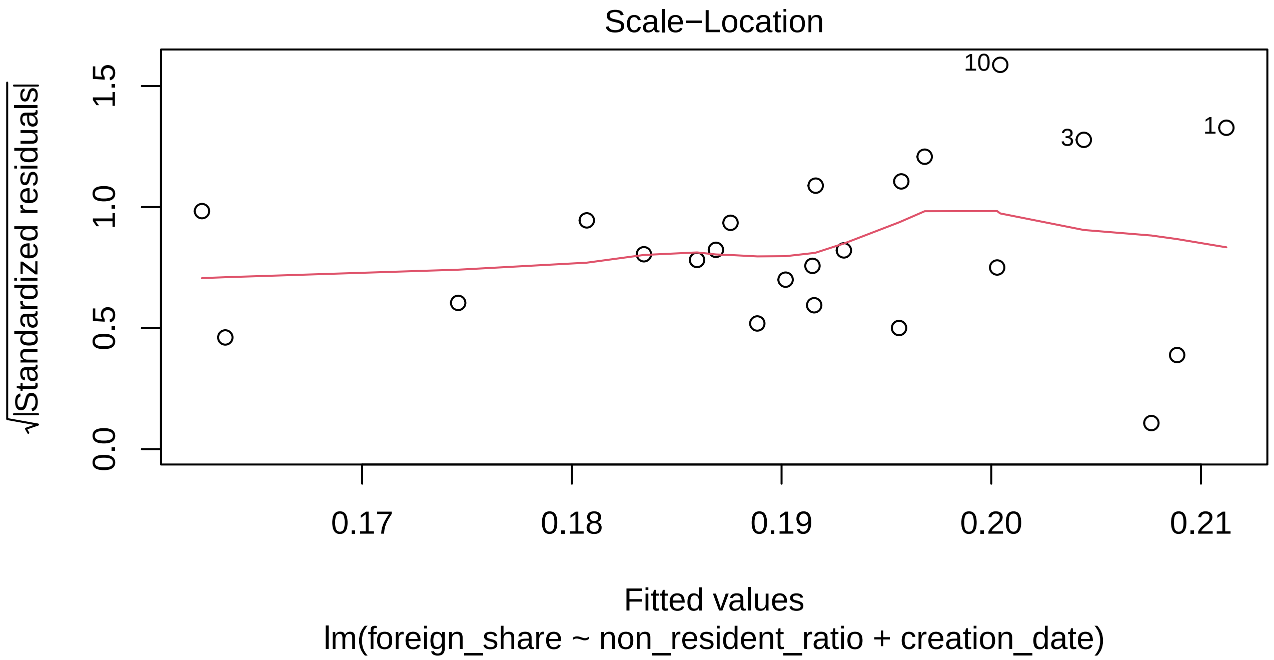
```
data_pooled <- data_pooled %>%
  mutate(
    english_share = English / total_books,
    foreign_share = Foreign / total_books
  )

# join demographics
data_pooled <- data_pooled %>%
  filter(creation_date %in% demographics$year) %>%
  left_join(demographics, by = c("creation_date" = "year"))

model <- lm(foreign_share ~ non_resident_ratio + creation_date, data = data_pooled)

summary(model)
```

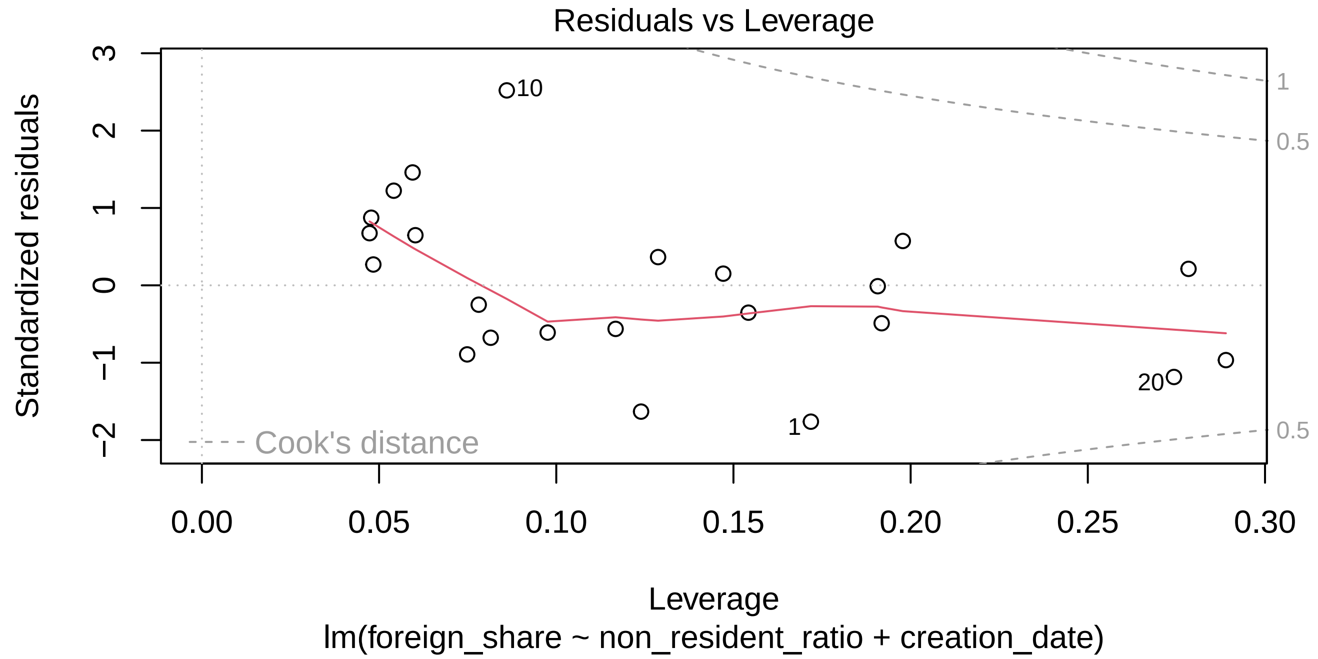
```
Call:
lm(formula = foreign_share ~ non_resident_ratio + creation_date,
    data = data_pooled)

Residuals:
      Min        1Q    Median        3Q       Max
-0.113651 -0.043566 -0.000739  0.040440  0.170705

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.450165  24.201236  -0.308    0.761
non_resident_ratio -1.335717   2.943780  -0.454    0.655
creation_date       0.003942   0.012338   0.319    0.753

Residual standard error: 0.07086 on 20 degrees of freedom
Multiple R-squared:  0.03379,   Adjusted R-squared:  -0.06283
F-statistic: 0.3497 on 2 and 20 DF,  p-value: 0.7091
```

```
#check assumptions 1 by 1

plot(model)
```

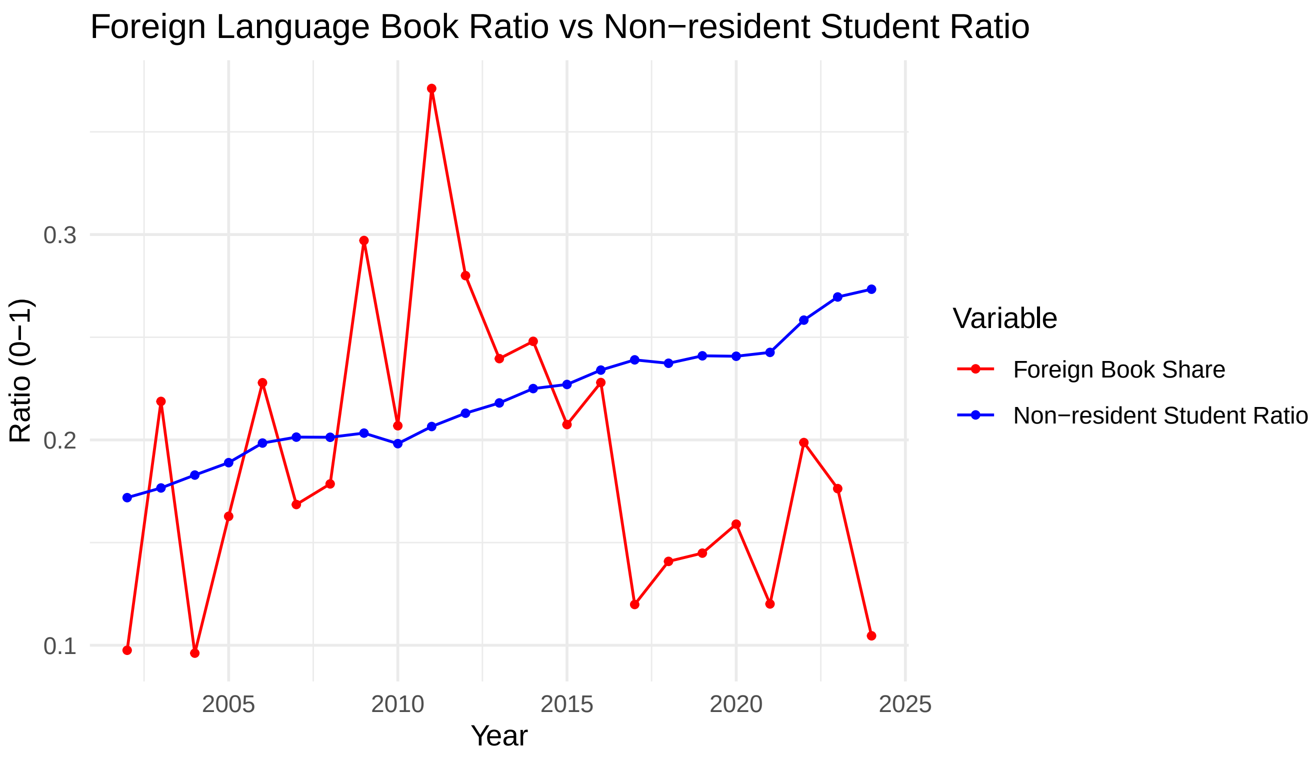

Residuals vs Fitted
Residuals
0.20
0.10
0.00
−0.10
0.17
0.18
0.19
0.20
0.21
10
3
1
Fitted values
lm(foreign_share ~ non_resident_ratio + creation_date)


Q−Q Residuals
Standardized residuals
3
2
1
0
−1
10
1
3
−2
−1
0
1
2
Theoretical Quantiles
lm(foreign_share ~ non_resident_ratio + creation_date)





```r
combined <- data_pooled %>%
  select(year = creation_date, foreign_share) %>%
  left_join(
    demographics %>% select(year, non_resident_ratio),
    by = "year"
  ) %>%
  pivot_longer(
    cols = c(foreign_share, non_resident_ratio),
    names_to = "variable",
    values_to = "value"
  )

# plot both lines together
ggplot(combined, aes(x = year, y = value, color = variable)) +
  geom_line(size = 0.5) +
  geom_point(size = 1) +
  scale_color_manual(
    values = c(
      foreign_share = "red",
      non_resident_ratio = "blue"
    ),
    labels = c(
      foreign_share = "Foreign Book Share",
      non_resident_ratio = "Non-resident Student Ratio"
    )
  ) +
  labs(
    title = "Foreign Language Book Ratio vs Non-resident Student Ratio",
    x = "Year",
    y = "Ratio (0-1)",
    color = "Variable"
  ) +
  theme_minimal()
```


Foreign Language Book Ratio vs Non−resident Student Ratio
Ratio (0−1)
0.1
0.2
0.3
2005
2010
2015
2020
2025
Year
Variable
Foreign Book Share
Non−resident Student Ratio