Culture and Institutions: Data Task 1

Jonas Wallstein & Caroline Belka

2023-05-11

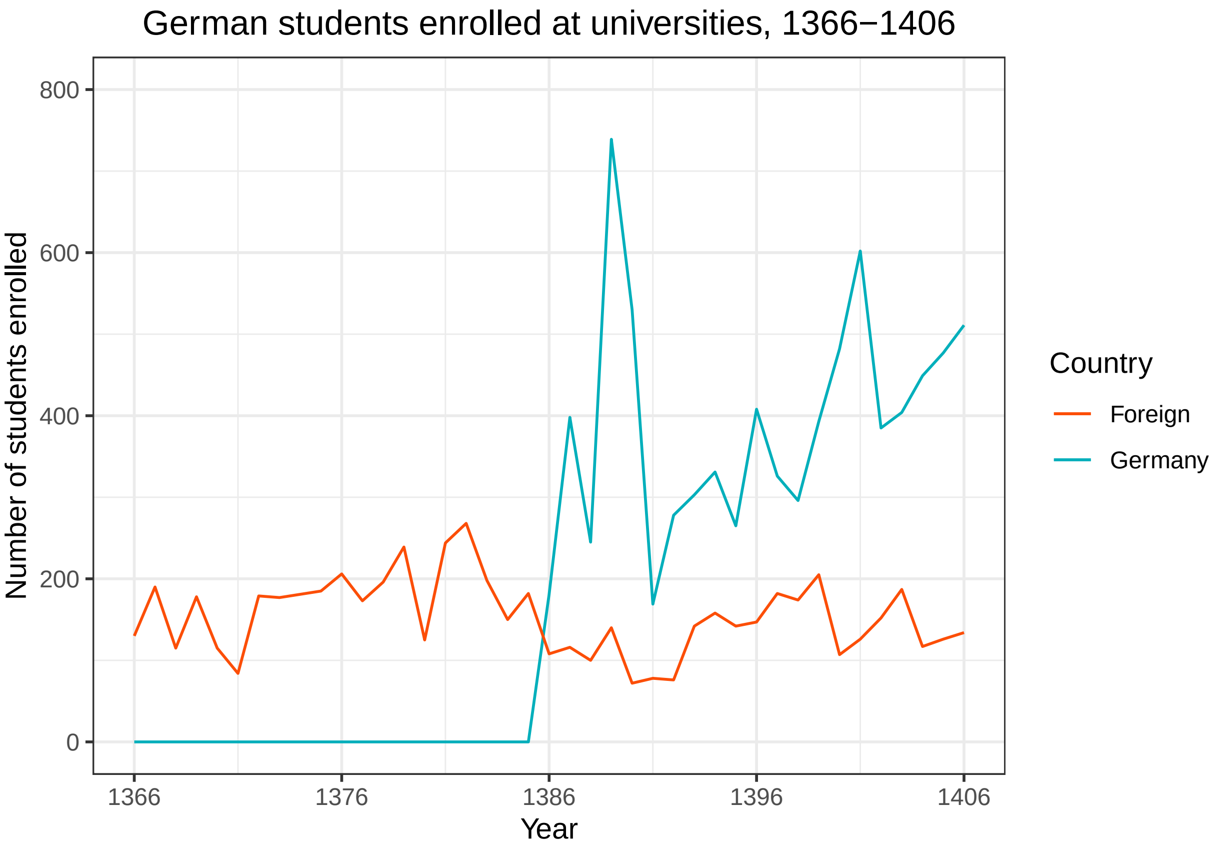
0. Setup

```
rm(list = ls())
library(tidyverse)
library(stargazer)
library(ggplot2)
library(estimatr)

stu_ger <- read_csv("./cieh_2023_datatask1_data/students_germany.csv")
analysis_ger <- read_csv("./cieh_2023_datatask1_data/analysis_germany.csv")
analysis_eng_wales <- read_csv("./cieh_2023_datatask1_data/analysis_england_wales.csv")
analysis_italy <- read_csv("./cieh_2023_datatask1_data/analysis_italy.csv")
```

1. Figure IV

```
theme_set(theme_bw())
ggplot(stu_ger, aes(x=year)) +
  geom_line(aes(y=germany, colour="Germany")) +
  geom_line(aes(y=foreign, colour="Foreign")) +
  scale_color_manual(name = "Country", values = c("Germany" = "#00AFBB",
                                                  "Foreign" = "#FC4E07")) +
  scale_x_continuous(name="Year", breaks=seq(1366, 1406, by=10)) +
  scale_y_continuous(name = "Number of students enrolled", limits = c(0,800)) +
  ggtitle("German students enrolled at universities, 1366-1406") +
  theme(plot.title = element_text(hjust = 0.5))
```



2. Panel

```
# Write function to aggregate over years
# Create "newmarkets": total number of newly established markets
# in Germany per 1000 cities in the given year
aggregate_by_year = function(input_df, output_df, number_cities, delete1386){
  df <- input_df %>%
     group_by(year) %>%
     summarise(newmarkets = sum(markets) * 1000 / number_cities)
  df$post1386 = ifelse(df$year >1386, 1, 0) # create post1368 dummy
  ifelse(delete1386 == TRUE, df <- filter(df, year != 1386), "") # option to delete 1386
  df <- mutate(df, year = year - 1386) # centering the year around 1386 as in the paper
  assign(output_df, df, envir = .GlobalEnv)
  head(df, 5)
}

aggregate_by_year(input_df = analysis_ger, output_df = "by_year_ger",
                  number_cities = 2256, delete1386 = F)
```

```
## # A tibble: 5 x 3
##    year newmarkets post1386
##   <dbl>      <dbl>    <dbl>
## 1   -20      0.887        0
## 2   -19      1.33         0
## 3   -18      3.99         0
## 4   -17      0.443        0
## 5   -16      1.33         0
```

3. Regression

```
by_year_pre1386 <- filter(by_year_ger, year < 0)

rlm1 <- lm_robust(newmarkets ~ year, by_year_pre1386)
summary(rlm1)
```

3a)

```
##
## Call:
## lm_robust(formula = newmarkets ~ year, data = by_year_pre1386)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  1.22480    0.61660   1.986  0.06243 -0.07063  2.52023 18
## year        -0.06066    0.05351  -1.134  0.27183 -0.17307  0.05176 18
##
## Multiple R-squared:  0.04696 ,   Adjusted R-squared:  -0.005991
## F-statistic: 1.285 on 1 and 18 DF,  p-value: 0.2718
```

3b) An additional year (before 1386) goes along with 0.06 fewer newly established markets per 1000 cities. The effect is negative and not significant.

```
predictions <- filter(by_year_ger, year == 0)

predictions$rlm1_prediction <- predict(rlm1, predictions)
print(predictions)
```

3c)

```
## # A tibble: 1 x 4
##    year newmarkets post1386 rlm1_prediction
##   <dbl>      <dbl>    <dbl>           <dbl>
## 1     0          0        0            1.22
```

The predicted value for 1386 is 1.22 new markets, according to the regression model considering the years before 1386.

```
by_year_post1386 <- filter(by_year_ger, year > 0)

rlm2 <- lm_robust(newmarkets ~ year, by_year_post1386)
summary(rlm2)
```

3d)

```
##
## Call:
## lm_robust(formula = newmarkets ~ year, data = by_year_post1386)
##
## Standard error type:  HC2
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)  1.30879    0.52653   2.486  0.02298  0.20260   2.4150 18
## year         0.09065    0.05165   1.755  0.09626 -0.01787   0.1992 18
##
## Multiple R-squared:  0.1254 ,    Adjusted R-squared:  0.07681
## F-statistic:  3.08 on 1 and 18 DF,  p-value: 0.09626
```

3e) An additional year (after 1385) goes along with 0.09 more newly established markets per 1000 cities. The effect is positive and but insignificant. The declining trend in new markets seems to turn towards an increasing trend after 1386.

```
predictions$rlm2_prediction <- predict(rlm2, predictions)
print(predictions)
```

3f)

```
## # A tibble: 1 x 5
##    year newmarkets post1386 rlm1_prediction rlm2_prediction
##   <dbl>      <dbl>    <dbl>           <dbl>           <dbl>
## 1     0          0        0            1.22            1.31
```

The predicted value for 1386 is 1.31 new markets, according to the regression model considering the years after 1386.

```
coef(rlm1)[2] - coef(rlm2)[2]
```

3g)

```
##       year
## -0.1513091
```

The difference between the $\hat{\beta}_1$ from a) and d) is 0.15

```
predictions$delta = predictions$rlm2_prediction - predictions$rlm1_prediction
print(predictions)
```

3h)

```
## # A tibble: 1 x 6
##    year newmarkets post1386 rlm1_prediction rlm2_prediction  delta
##   <dbl>      <dbl>    <dbl>           <dbl>           <dbl>  <dbl>
## 1     0          0        0            1.22            1.31 0.0840
```

The predictions from the two regression models differ by 0.08. The actual number of newly established markets in 1386 is zero.

4. Interaction year × post

```
by_year_ger <- filter(by_year_ger, year != 0)

f = newmarkets ~ year * post1386
```

```
rlm3 <- lm_robust(formula = f, by_year_ger)
summary(rlm3)
```

```
##
## Call:
## lm_robust(formula = f, data = by_year_ger)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   CI Lower CI Upper DF
## (Intercept)    1.22480    0.61660  1.9864  0.05464 -0.0257208  2.47533 36
## year          -0.06066    0.05351 -1.1336  0.26445 -0.1691754  0.04786 36
## post1386       0.08399    0.81082  0.1036  0.91808 -1.5604280  1.72840 36
## year:post1386  0.15131    0.07437  2.0345  0.04932  0.0004776  0.30214 36
##
## Multiple R-squared:  0.09769 ,   Adjusted R-squared:  0.0225
## F-statistic: 1.566 on 3 and 36 DF,  p-value: 0.2144
```

- The estimate for $\hat{\beta}_1$ is -0.061 meaning that with every additional year, the number of newly established markets decreases on average by 0.06 ceteris paribus. The effect is however not significant.
- The estimate for $\hat{\beta}_2$ is 0.084 meaning that the number of expected markets jumps after 1386 by 0.084. The effect is not significant at the 5% level.
- The estimate for $\hat{\beta}_3$ is 0.151 meaning that after 1386 the effect of an additional year on newly established markets is 0.09 (0.151 - 0.061). The effect is significant at the 5% level.

5. Sample split

```
# Calculate Median
median_distdiff = median(analysis_ger$distdiff)

# Subsample where distdiff > median -> 2256/2 = 1128 cities remain
by_year_above = filter(analysis_ger, distdiff >= median_distdiff)
aggregate_by_year(input_df = by_year_above, output_df = "by_year_above",
                  number_cities = 1128, TRUE)
```

```
## # A tibble: 5 x 3
##    year newmarkets post1386
##   <dbl>      <dbl>    <dbl>
## 1   -20      1.77         0
## 2   -19      1.77         0
## 3   -18      3.55         0
## 4   -17      0.887        0
## 5   -16      0            0
```

```
# Regression on subsample where distdiff > median
rlm4 <- lm_robust(formula = f, by_year_above)
summary(rlm4)
```

```
##
## Call:
## lm_robust(formula = f, data = by_year_above)
##
## Standard error type:  HC2
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)    0.97984    0.78154  1.2537 0.218020  -0.6052  2.56488 36
## year          -0.09665    0.06076 -1.5907 0.120422  -0.2199  0.02658 36
## post1386      -0.76521    0.97719 -0.7831 0.438704  -2.7470  1.21661 36
## year:post1386  0.28729    0.08519  3.3723 0.001794   0.1145  0.46006 36
##
## Multiple R-squared:  0.1743 ,    Adjusted R-squared:  0.1055
## F-statistic: 4.273 on 3 and 36 DF,  p-value: 0.01114
```

```
# Subsample where distdiff < median -> 2256/2 = 1128 cities remain
by_year_below = filter(analysis_ger, distdiff < median_distdiff)
aggregate_by_year(input_df = by_year_below, output_df = "by_year_below",
                  number_cities = 1128, TRUE)
```

```
## # A tibble: 5 x 3
##    year newmarkets post1386
##   <dbl>      <dbl>    <dbl>
## 1   -20      0            0
## 2   -19      0.887        0
## 3   -18      4.43         0
## 4   -17      0            0
## 5   -16      2.66         0
```

```
# Regression on subsample where distdiff < median
rlm5 <- lm_robust(formula = f, by_year_below)
summary(rlm5)
```

```
##
## Call:
## lm_robust(formula = f, data = by_year_below)
##
## Standard error type:  HC2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)    1.46976    0.68323  2.1512  0.03824  0.08412  2.85541 36
## year          -0.02466    0.06001 -0.4110  0.68351 -0.14636  0.09704 36
## post1386       0.93318    1.15864  0.8054  0.42587 -1.41664  3.28301 36
## year:post1386  0.01533    0.09891  0.1550  0.87769 -0.18527  0.21593 36
##
## Multiple R-squared:  0.02561 ,   Adjusted R-squared:  -0.05559
## F-statistic: 0.3776 on 3 and 36 DF,  p-value: 0.7697
```

6. Placebo Analysis

Italy

```
aggregate_by_year(analysis_italy, "by_year_italy", 190, T)
```

```
## # A tibble: 5 x 3
##    year newmarkets post1386
##   <dbl>      <dbl>    <dbl>
## 1   -20       0           0
## 2   -19       0           0
## 3   -18       5.26        0
## 4   -17       0            0
## 5   -16       0            0
```

```
rlm6 <- lm_robust(formula = f, by_year_italy)
summary(rlm6)
```

```
##
## Call:
## lm_robust(formula = f, data = by_year_italy)
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error  t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)    2.520776    1.27769  1.97292  0.05622 -0.07050    5.1120 36
## year           0.114761    0.09862  1.16367  0.25221 -0.08525    0.3148 36
## post1386       0.304709    2.60160  0.11712  0.90741 -4.97158    5.5810 36
## year:post1386 -0.007915    0.23918 -0.03309  0.97379 -0.49300    0.4772 36
##
## Multiple R-squared:  0.08584 ,   Adjusted R-squared:  0.00966
## F-statistic: 1.505 on 3 and 36 DF,  p-value: 0.2297
```

England & Wales

```
aggregate_by_year(analysis_eng_wales, "by_year_eng", 2254, T)
```

```
## # A tibble: 5 x 3
##    year newmarkets post1386
##   <dbl>      <dbl>    <dbl>
## 1   -20       2.66        0
## 2   -19       1.77        0
## 3   -18       2.22        0
## 4   -17       0           0
## 5   -16       0           0
```

```
rlm7 <- lm_robust(formula = f, by_year_eng)
summary(rlm7)
```

```
##
## Call:
## lm_robust(formula = f, data = by_year_eng)
##
## Standard error type:  HC2
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
## (Intercept)    3.08224    1.04058   2.962 0.005385  0.97185  5.19263 36
## year           0.10975    0.07757   1.415 0.165717 -0.04757  0.26707 36
## post1386      -2.03381    1.19954  -1.695 0.098615 -4.46658  0.39896 36
## year:post1386 -0.09974    0.09188  -1.085 0.284925 -0.28609  0.08661 36
##
## Multiple R-squared:  0.1468 ,    Adjusted R-squared:  0.07575
## F-statistic: 1.148 on 3 and 36 DF,  p-value: 0.3428
```

7. Regression Table

```r
# Stargazer does not work with lm_robust
# Robust standard errors manually added with starprep argument
lm3 <- lm(formula = f, by_year_ger)
lm4 <- lm(formula = f, by_year_below)
lm5 <- lm(formula = f, by_year_above)
lm6 <- lm(formula = f, by_year_italy)
lm7 <- lm(formula = f, by_year_eng)

stargazer(lm3, lm4, lm5, lm6, lm7, se = starprep(lm3, lm4, lm5, lm6, lm7),
          dep.var.labels = "New markets",
          column.labels = c("Base", "< Median",">= Median", "Italy", "England and Wales"))
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, May 11, 2023 - 13:08:15

Table 1:

	Dependent variable:				
	New markets				
	Base	< Median	>= Median	Italy	England and Wales
	(1)	(2)	(3)	(4)	(5)
year	−0.061	−0.025	−0.097	0.115	0.110
	(0.054)	(0.060)	(0.061)	(0.099)	(0.078)
post1386	0.084	0.933	−0.765	0.305	−2.034*
	(0.811)	(1.159)	(0.977)	(2.602)	(1.200)
year:post1386	0.151**	0.015	0.287***	−0.008	−0.100
	(0.074)	(0.099)	(0.085)	(0.239)	(0.092)
Constant	1.225**	1.470**	0.980	2.521**	3.082***
	(0.617)	(0.683)	(0.782)	(1.278)	(1.041)
Observations	40	40	40	40	40
R^2	0.098	0.026	0.174	0.086	0.147
Adjusted R^2	0.022	−0.056	0.105	0.010	0.076
Residual Std. Error (df = 36)	1.561	2.000	2.016	5.032	1.509
F Statistic (df = 3; 36)	1.299	0.315	2.533*	1.127	2.065
Note:					*p<0.1; **p<0.05; ***p<0.01