



The effect of socio-economic indicators on the fertility rates

An empirical analysis of fertility rates among the regions of Europe

Marcell Granát

March 21, 2021

Contents

Introduction	1
Data	1
Total fertility rates	1
Human development	1
A decent standard of living	3
Long and healthy life	4
Knowledge	4
Family benefits	5
Youth unemployment	5
Explore data	5
Model building	5
Framework I: with unemployment	5
Framework II: without unemployment	12
Appendix: R codes	16

Abstract

1. probléma felvetése 2. kutatási kérdés bevett változók miként befolyásolják a TTA-t
3. módszertan megnevezése adattranszformáció (indexek összeállítása, PCA), panel, lasso
4. saját eredmény röviden
- 2 végző output összefoglalása

List of Tables

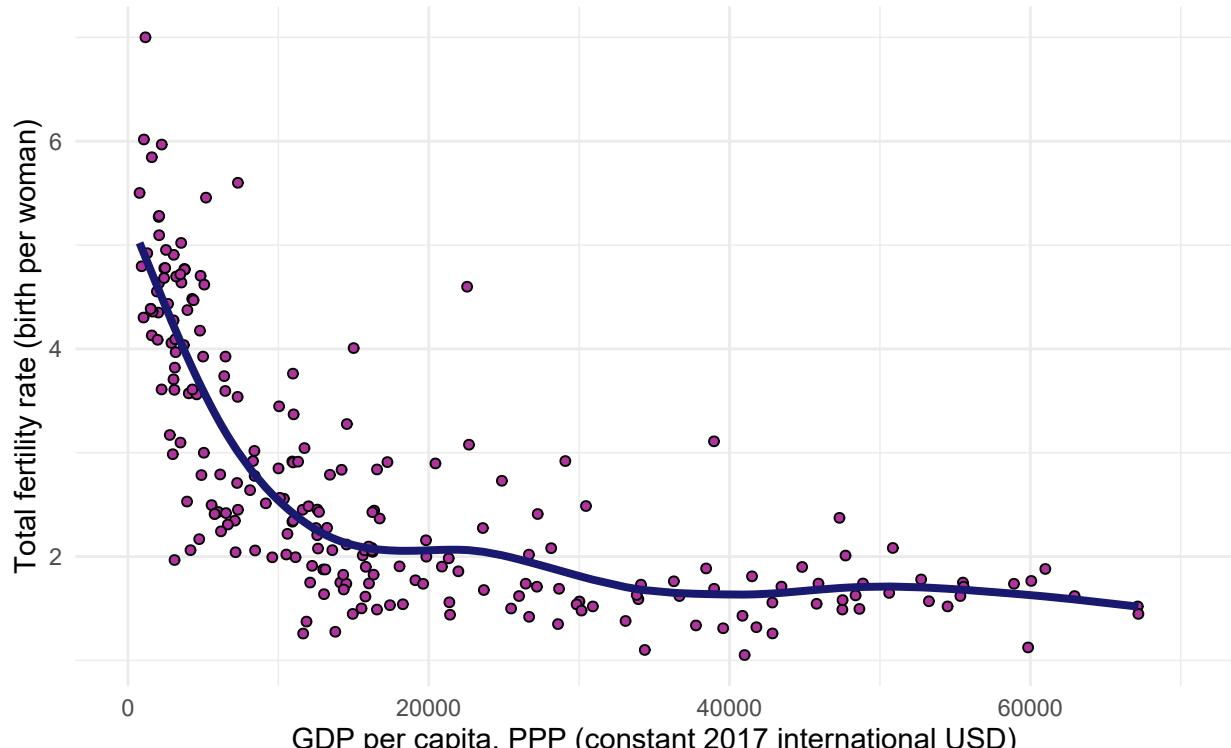
1	Indicators of similarity between the Human Development Indices provided by UNDP and GDL	3
2	Indicators of similarity between the income component of the Human Development Indices provided by GDL and the estimation based on regional GDP	3
3	Indicators of similarity between the health component of the Human Development Indices provided by GDL and the estimation based on regional life expectancy	4
4	Indicators of similarity between the knowledge component of Human Development Indices provided by UNDP and the calculated principal components using educational attainment level	4
5	Models	5
6	T-tests	11
7	T-tests	12

List of Figures

1	Seemingly negative effect of gross domestic product on fertility rates based on nation level observations (2017)	1
2	Total fertility rates through Europe in 2017	2
3	HDI and its components based on the dataset from Global Data Lab (2017)	3
4	PCAs and the explained variance	4
5	Pairwise correlation among TFR and calculation human development indices	5
6	Pairwise correlation among TFR and youth unemployment rates	6
7	Correlation between TFR and family benefits	7
8	Regression tree explaining the TFR using only the calculated HD indices (cp = 0.01)	8
9	Regression tree explaining the TFR using all the mentioned explanatory variables (cp = 0.01)	9
10	Panel models on the total fertility rates	10
11	Number of used observations by countries when the model contains youth unemployment	11
12	Estimated coefficient of the fixed panel model controlling for youth unemployment indicators	13
13	Number of used observations by countries when the model does not contain youth unemployment	14
14	Estimated coefficient of the fixed panel model omitting youth unemployment indicators	15

Introduction

1. motiváció
2. szűk szakirodalom eredményei
3. hipotézis
4. hipotézishez kapcsolódó szakirodalom eredményei
5. tanulmány fő lépései
6. módszertan mögötti indoklás
7. főbb eredmények
8. mit ad hozzá az irodalomhoz



Own editing based on the Figure 5-2. from Kreiszné Hudák (2019).
 The trend is drawn via splines.
 Source of the data: World Bank.

Figure 1: Seemingly negative effect of gross domestic product on fertility rates based on nation level observations (2017)

Data

Total fertility rates

Human development

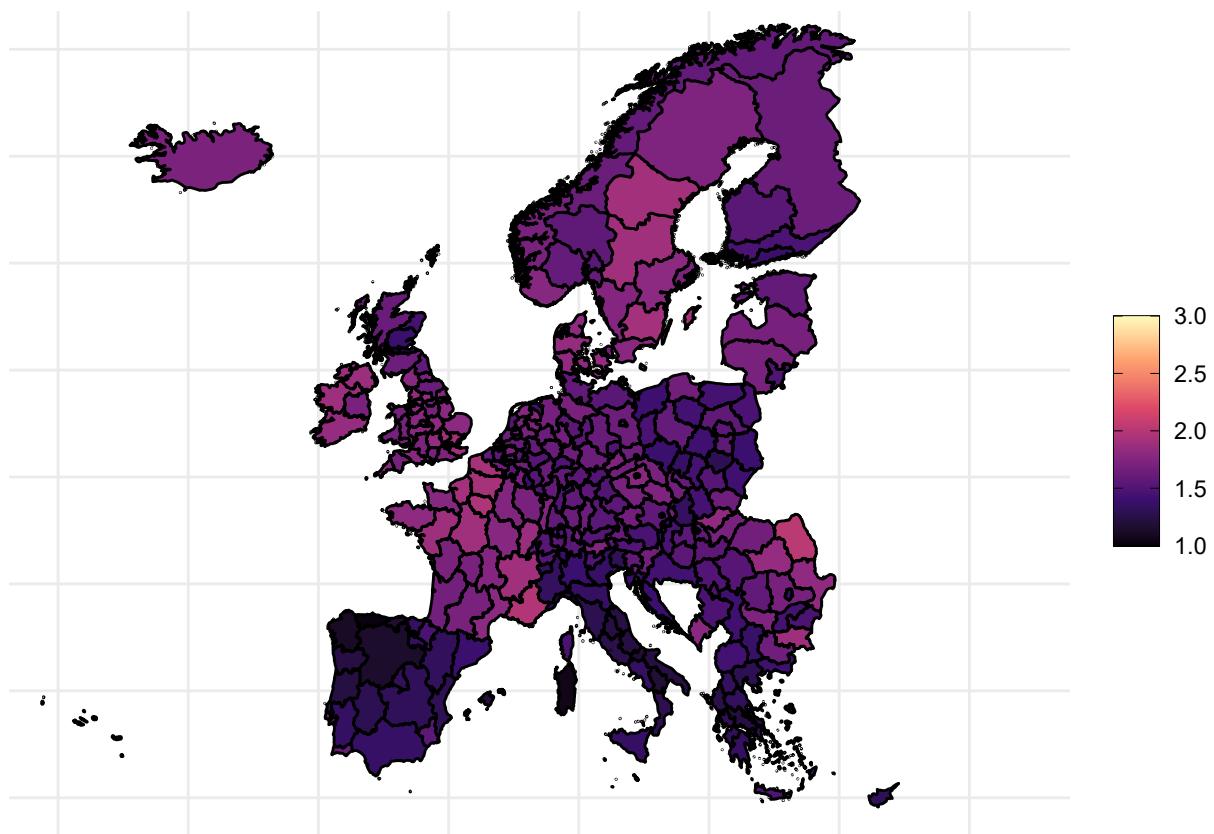


Figure 2: Total fertility rates through Europe in 2017

Table 1: Indicators of similarity between the Human Development Indices provided by UNDP and GDL

Indicator	Value
R^2	99.76%
Spearman R^2	99.71%
Mean absolute deviation	0.007
Mean absolute percentage deviation	1.20%

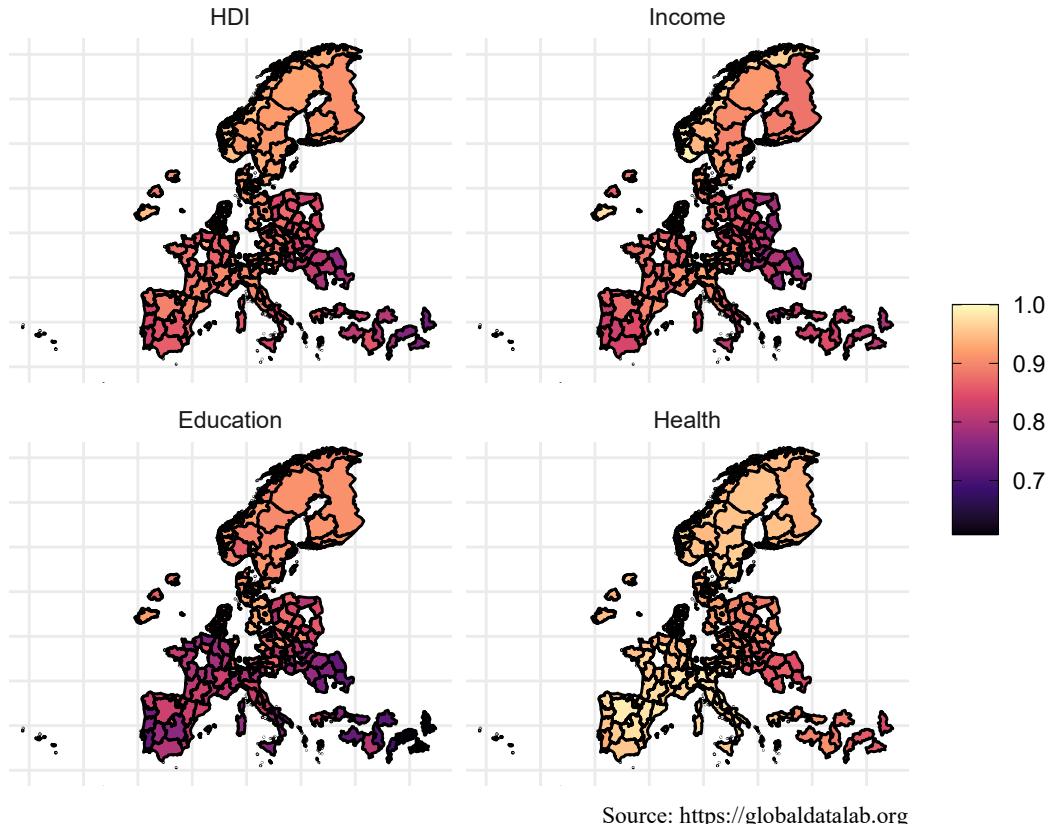


Figure 3: HDI and its components based on the dataset from Global Data Lab (2017)

A decent standard of living

Table 2: Indicators of similarity between the income component of the Human Development Indices provided by GDL and the estimation based on regional GDP

Indicator	Value
R^2	92.20%
Spearman R^2	94.18%
Mean absolute deviation	0.0520
Mean absolute percentage deviation	6.62%

Long and healthy life

Table 3: Indicators of similarity between the health component of the Human Development Indices provided by GDL and the estimation based on regional life expectancy

Indicator	Value
R^2	98.24%
Spearman R^2	98.38%
Mean absolute deviation	0.0041
Mean absolute percentage deviation	0.45%

Knowledge

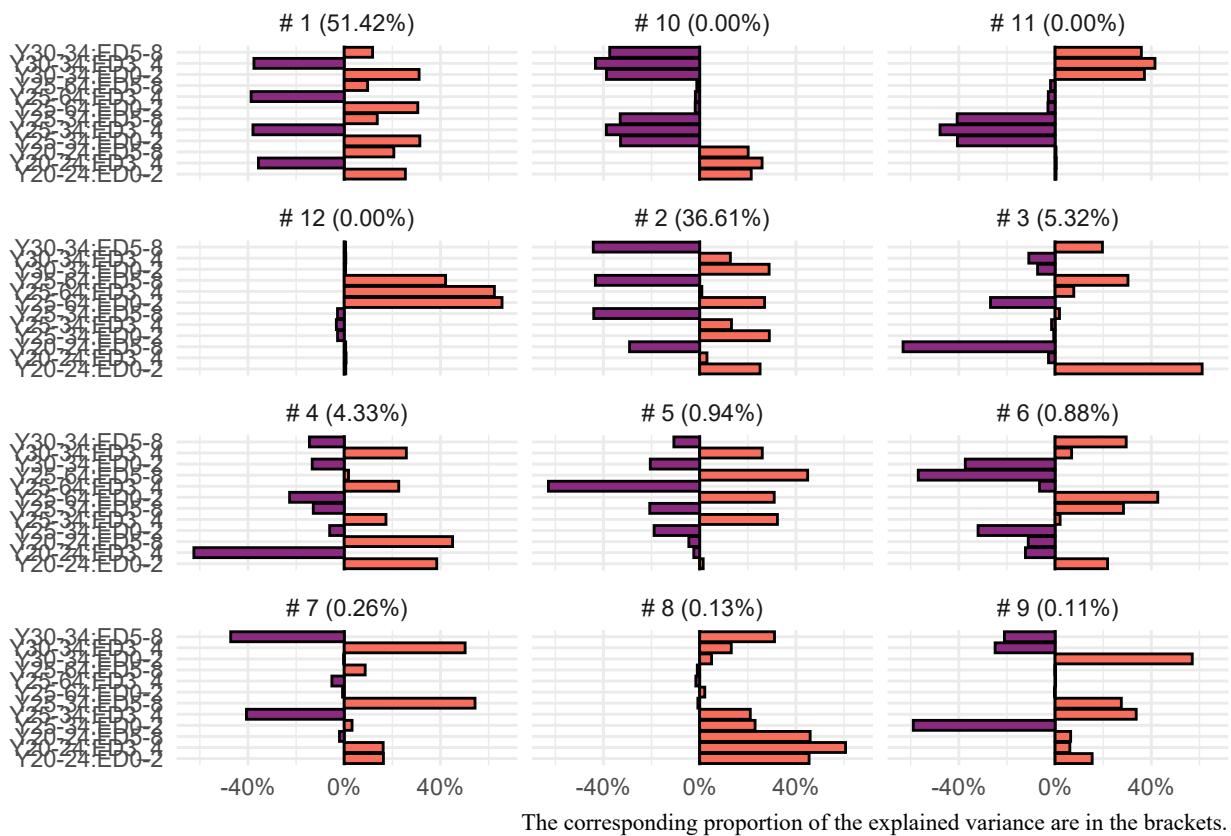


Figure 4: PCAs and the explained variance

Table 4: Indicators of similarity between the knowledge component of Human Development Indices provided by UNDP and the calculated principal components using educational attainment level

Indicator	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7	Comp 8	Comp 9
R^2	9.11%	47.66%	15.03%	0.15%	0.32%	2.40%	0.14%	0.00%	0.05%
Spearman R^2	5.47%	45.55%	14.28%	0.12%	0.38%	2.19%	1.01%	7.87%	0.03%

Family benefits

Youth unemployment

Explore data

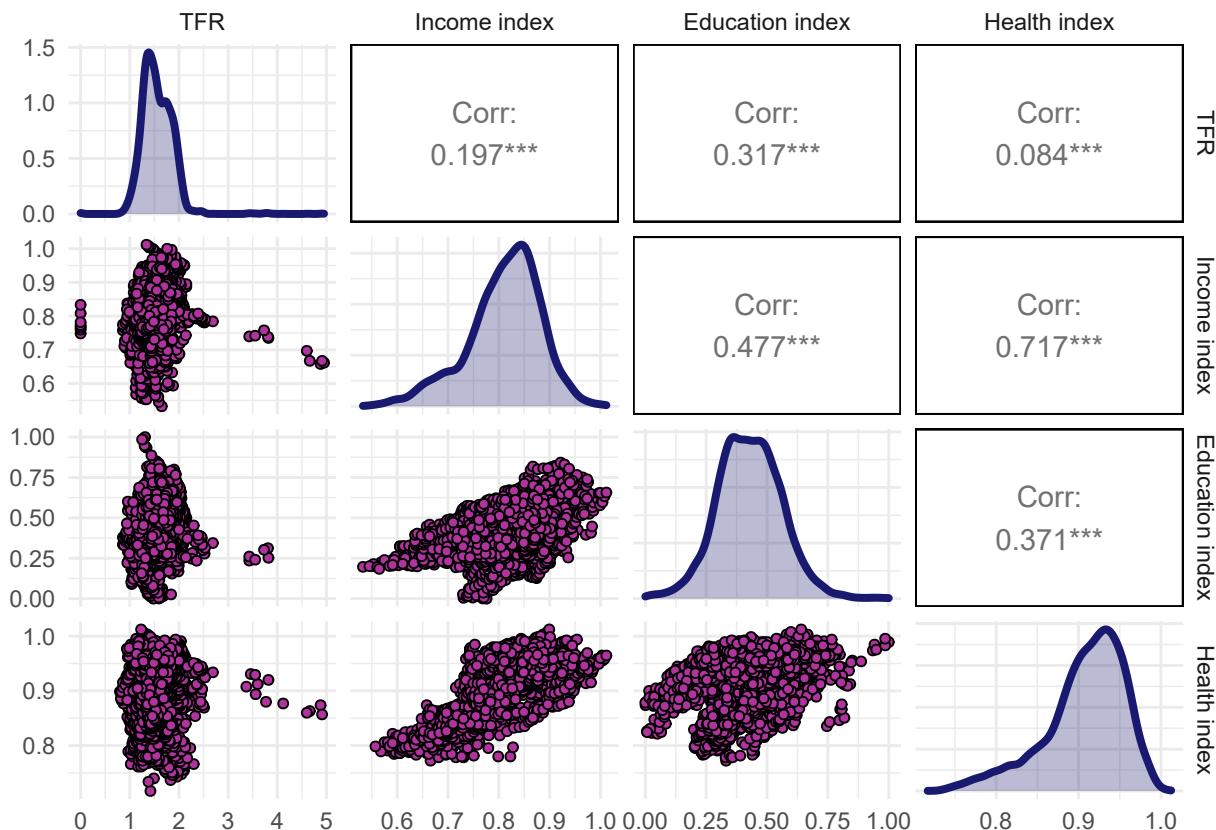


Figure 5: Pairwise correlation among TFR and calculation human development indices

Model building

Table 5: Models

Indicator	Model I.	Model II.
Pooltest	0.00%	0.00%
Phtest	0.00%	0.00%
Adjusted R^2	20.19%	29.05%
Observations	91	246

Framework I: with unemployment

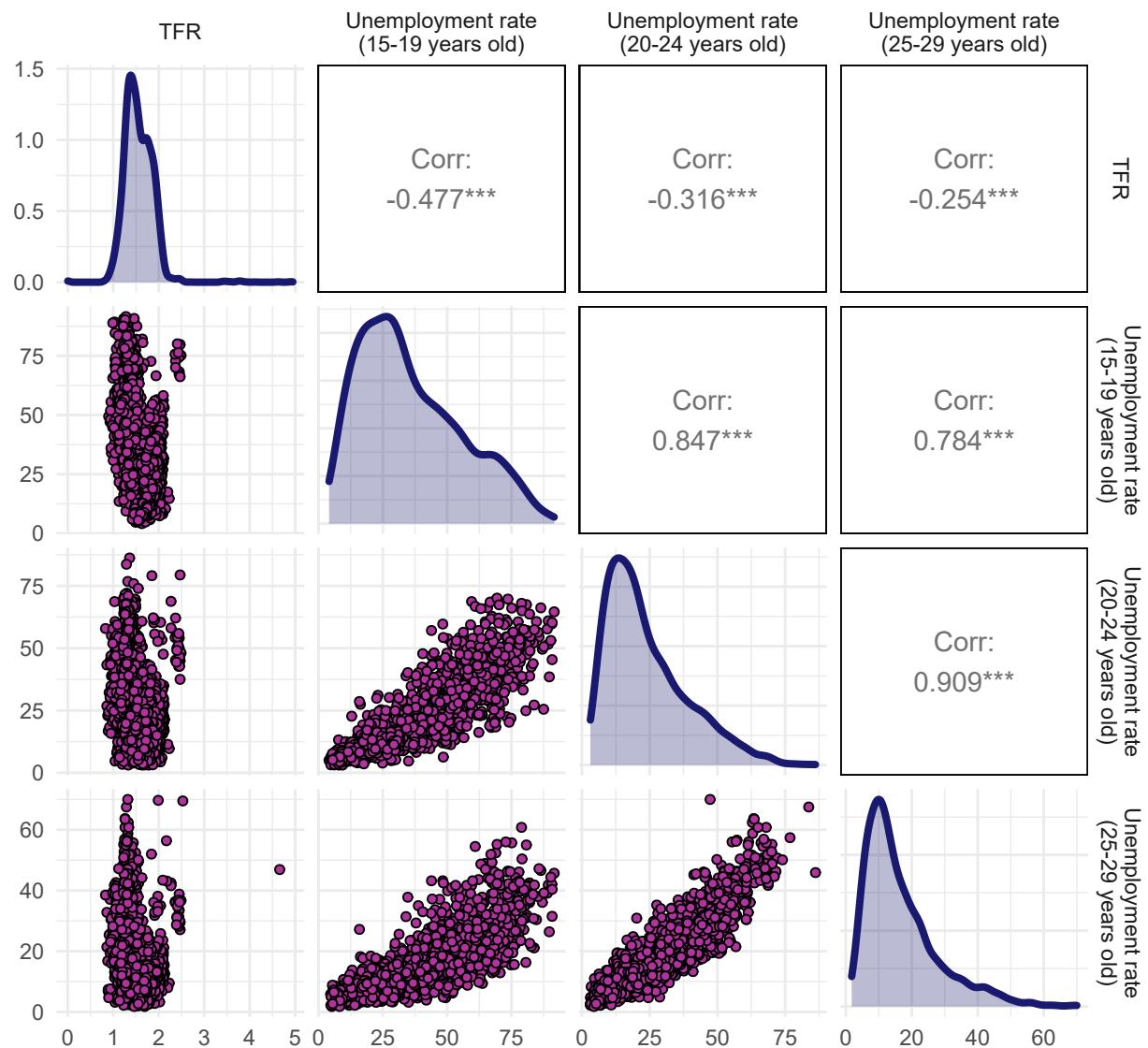


Figure 6: Pairwise correlation among TFR and youth unemployment rates

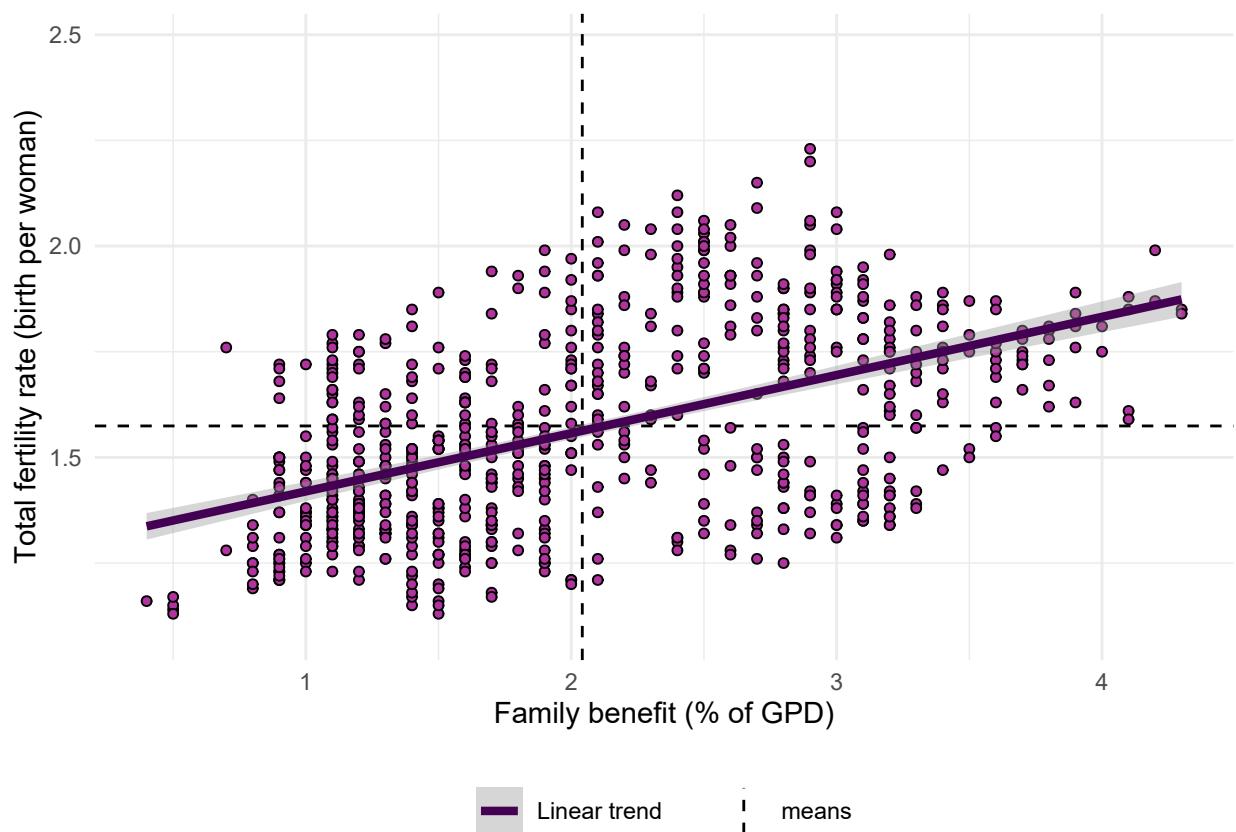


Figure 7: Correlation between TFR and family benefits

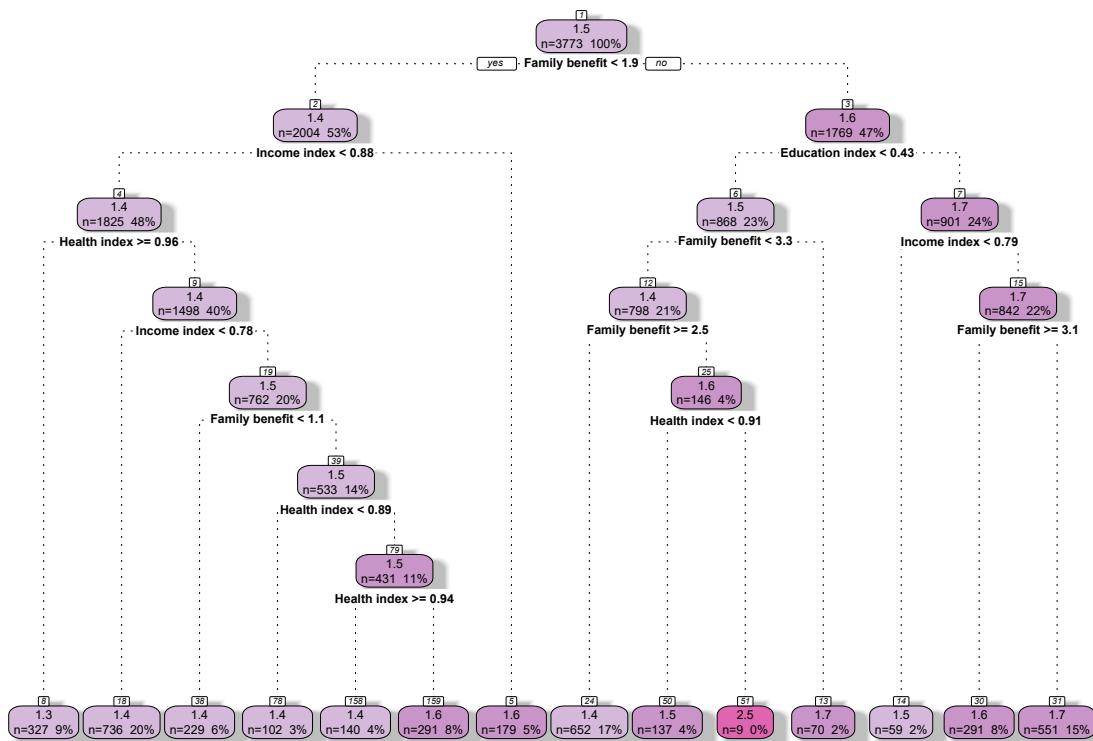


Figure 8: Regression tree explaining the TFR using only the calculated HD indices ($cp = 0.01$)

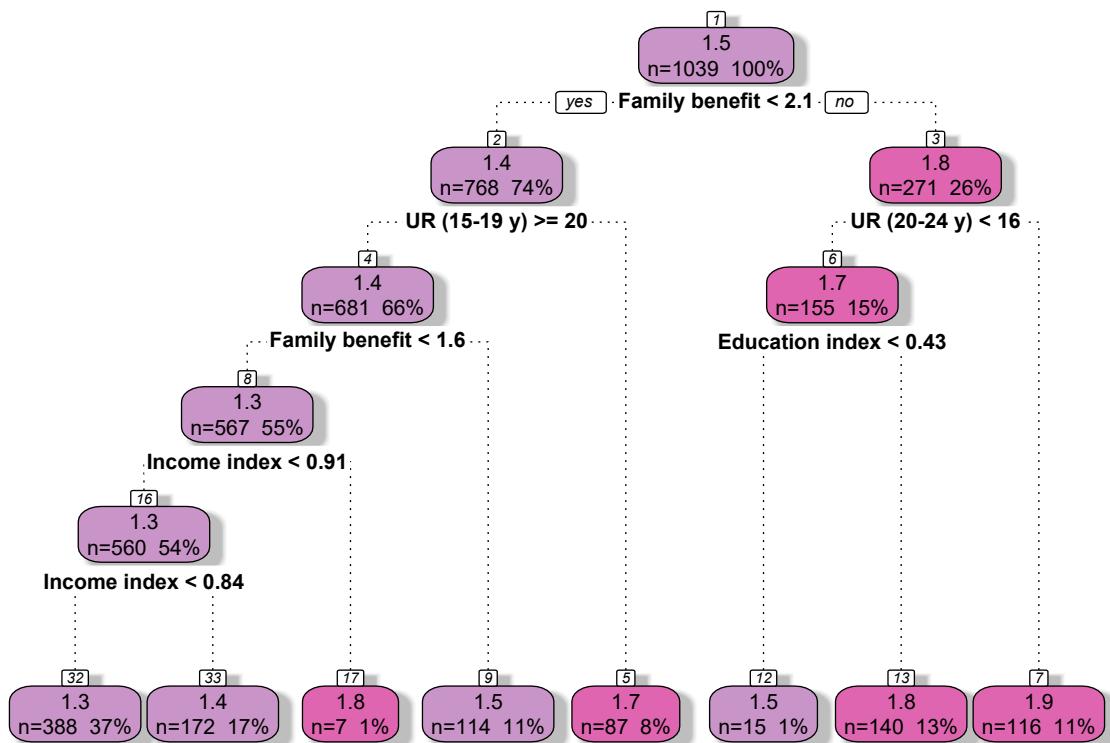


Figure 9: Regression tree explaining the TFR using all the mentioned explanatory variables ($cp = 0.01$)

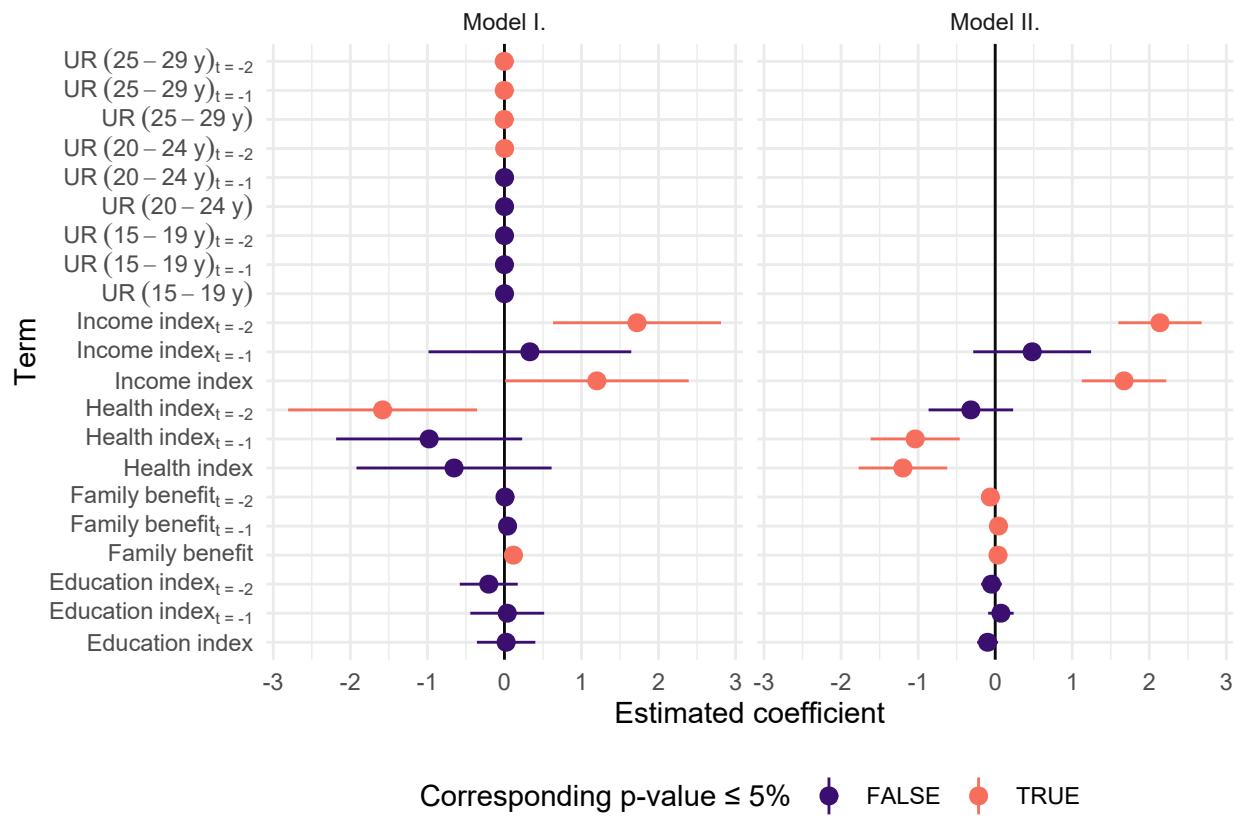


Figure 10: Panel models on the total fertility rates

Table 6: T-tests

Variable	Mean in total sample	Mean in used sample	Number of observations in the total sample	T-statistic	P-value
TFR	1.5575	1.4801	7249	8.0546	0.00%
Education index	0.4246	0.4241	5373	0.0910	92.75%
Health index	0.9089	0.9375	6983	-19.0193	0.00%
Income index	0.8120	0.8298	4291	-7.7143	0.00%
Family benefit	2.0759	1.6302	6155	14.7745	0.00%
UR (15-19 y)	36.6883	44.4313	1713	-8.5191	0.00%
UR (20-24 y)	24.3994	25.5202	2935	-1.7969	7.26%
UR (25-29 y)	17.0107	17.0332	2664	-0.0493	96.07%

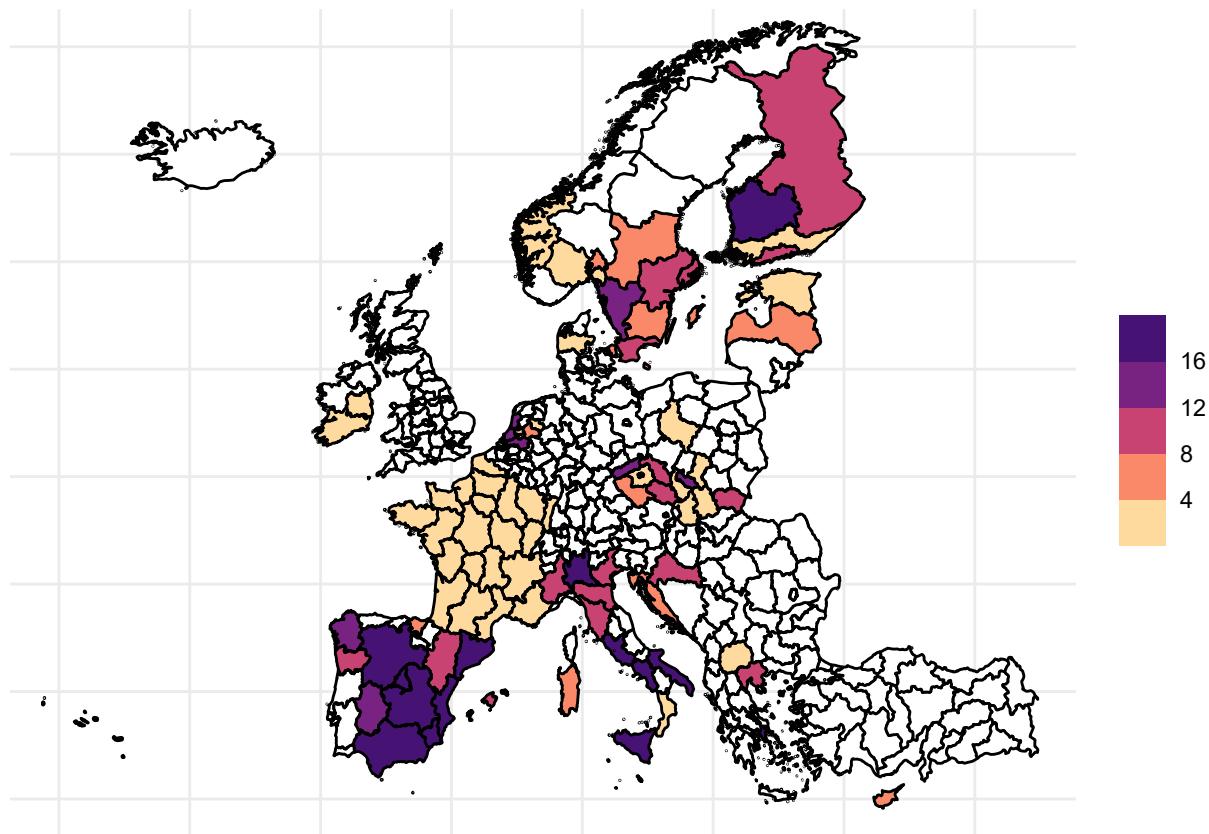
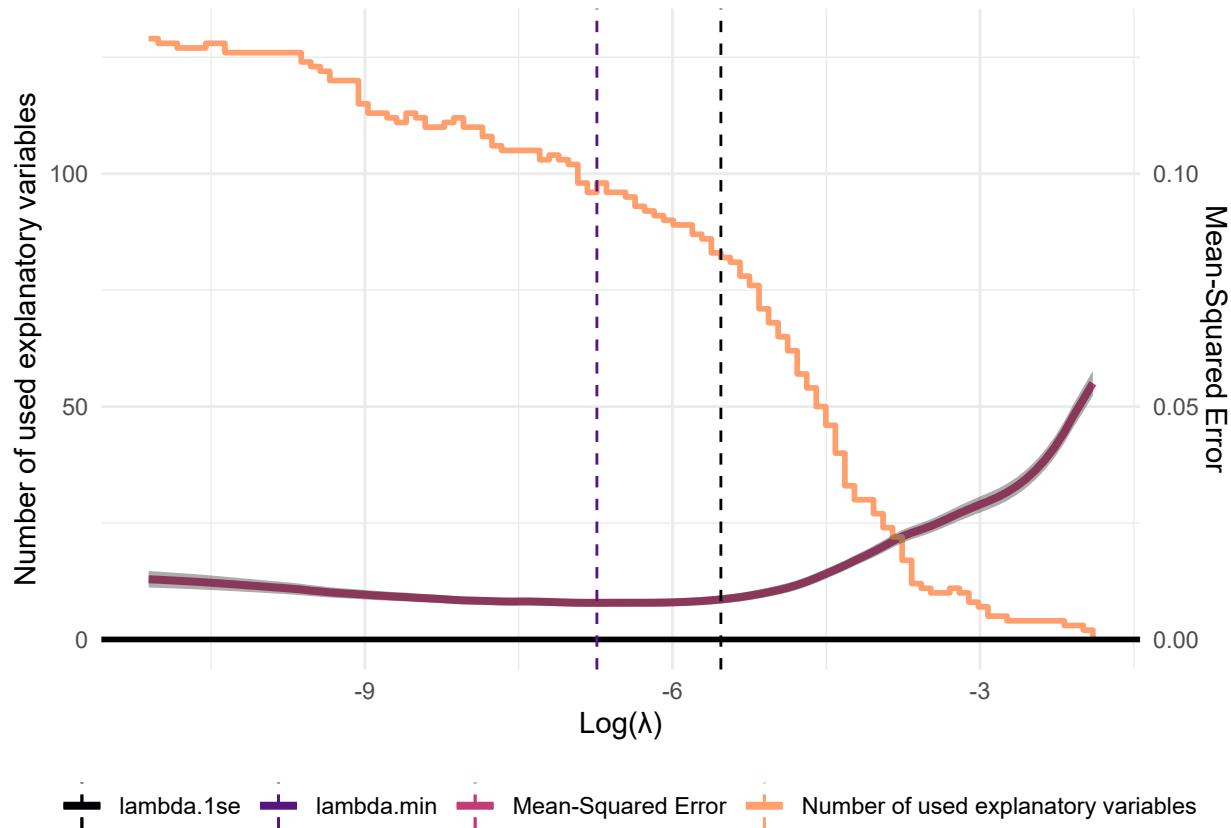


Figure 11: Number of used observations by countries when the model contains youth unemployment



Framework II: without unemployment

Table 7: T-tests

Variable	Mean in total sample	Mean in used sample	Number of observations in total sample	T-statistic	P-value
TFR	1.5575	1.4935	7249	11.7945	0.00%
Education index	0.4246	0.4082	5373	5.8006	0.00%
Health index	0.9089	0.9202	6983	-12.2876	0.00%
Income index	0.8120	0.8228	4291	-7.0235	0.00%
Family benefit	2.0759	1.9699	6155	5.6539	0.00%

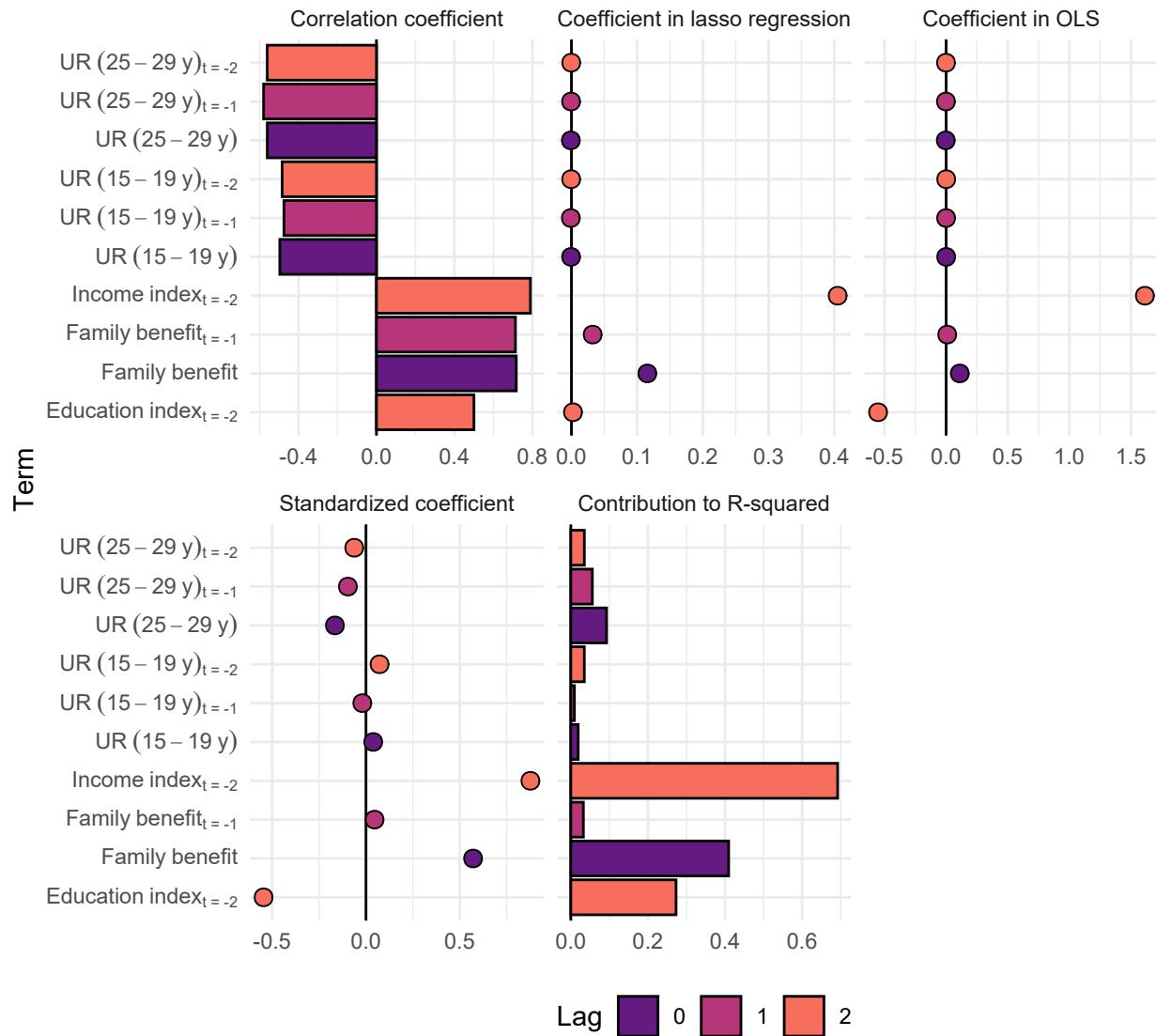


Figure 12: Estimated coefficient of the fixed panel model controlling for youth unemployment indicators

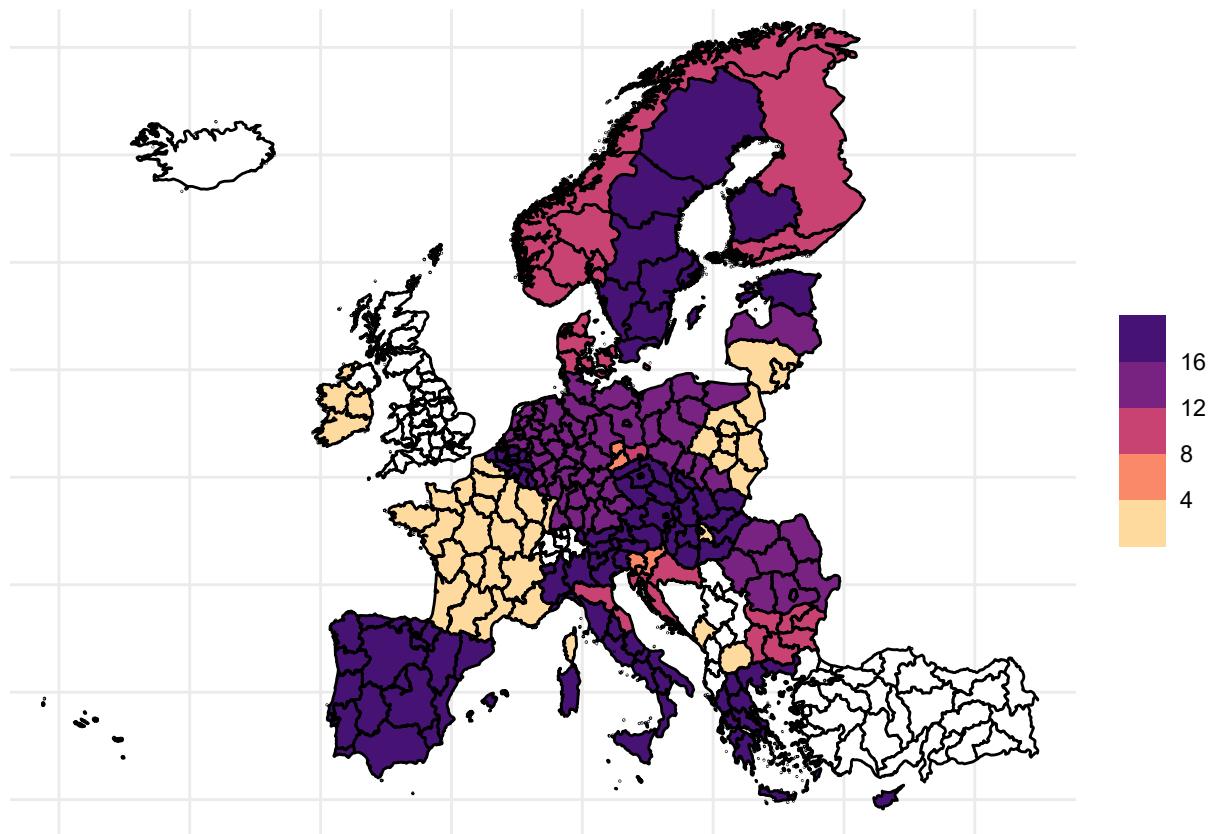


Figure 13: Number of used observations by countries when the model does not contain youth unemployment

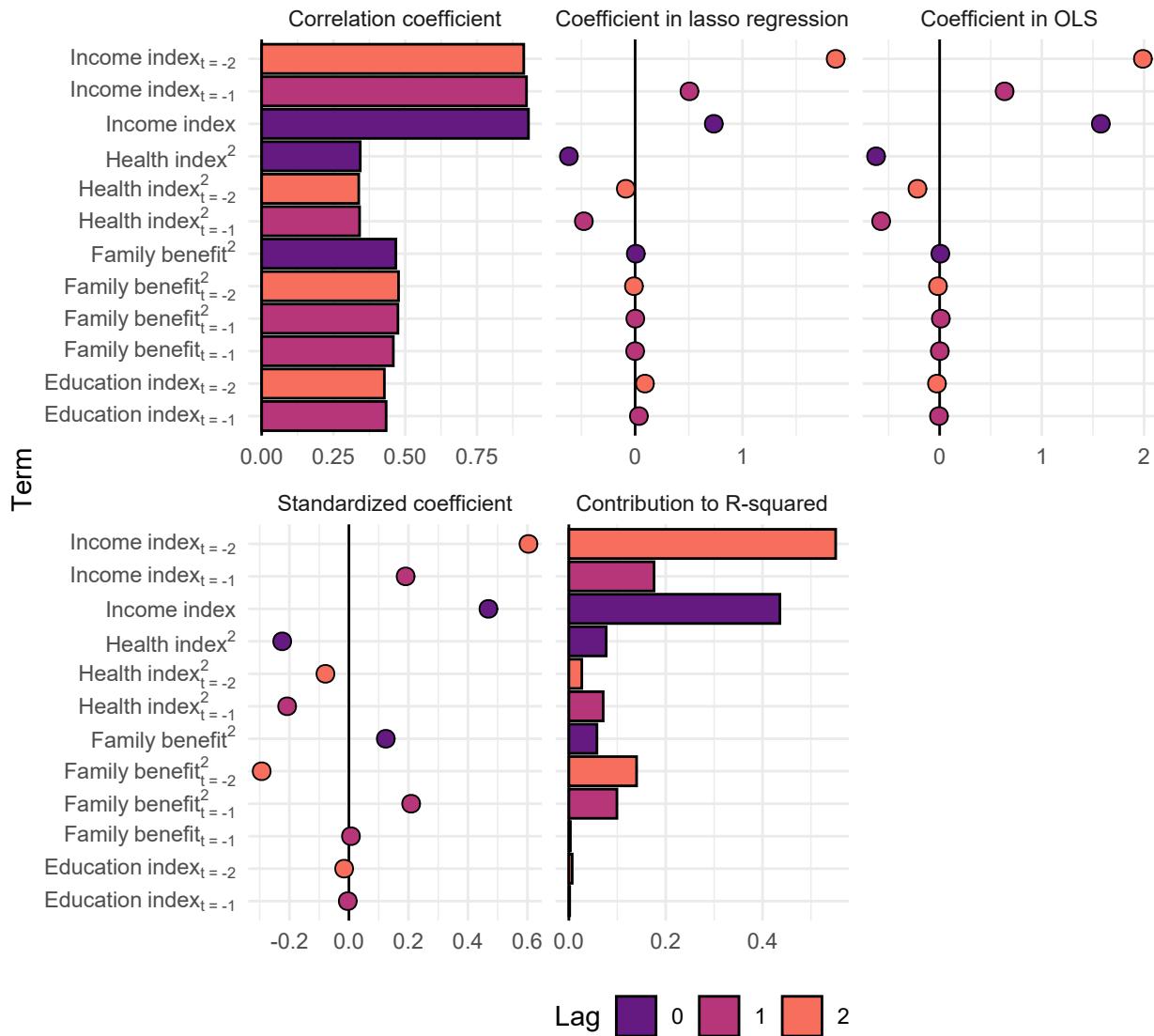


Figure 14: Estimated coefficient of the fixed panel model omitting youth unemployment indicators

Appendix: R codes

```

1 # Set up -----
2
3 ## Packages =====
4
5 library(tidyverse)
6 library(patchwork)
7 library(knitr)
8 library(broom)
9 library(eurostat)
10
11 ## Gg theme =====
12
13 update_geom_defaults("point", list(fill = "#B1339E",
14                               shape = 21,
15                               color = "black",
16                               size = 1.4))
17 update_geom_defaults("line",
18                      list(color = "midnightblue", size = 1.4))
19
20 update_geom_defaults("smooth", list(color = "red4", size = 1.4))
21
22 update_geom_defaults("density",
23                      list(color = "midnightblue", fill = "midnightblue", alpha = .3,
24                           size = 1.4))
25
26 extrafont::loadfonts(device="win")
27
28 theme_set(theme_minimal() + theme(
29   legend.direction = "vertical",
30   # text = element_text(family = "Impact"),
31   plot.caption = element_text(family = "serif"),
32   legend.key=element_blank()
33 ))
34
35 # https://data.worldbank.org/indicator/SP.DYN.TFRT.IN
36
37 WB_fertility <- read_csv("WB_fertility.csv", skip = 4)
38
39 # https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD
40
41 WB_GDP <- read_csv("WB_GDP.csv", skip = 4)
42
43 merge(WB_fertility %>%
44       select('Country Name', '2017') %>%
45       rename(tfr = '2017'),
46       WB_GDP %>%
47       select('Country Name', '2017') %>%
48       rename(GDP = '2017')) %>%
49 ggplot(aes(GDP, tfr)) + geom_point() +
50   ggfformula::geom_spline() +
51   scale_x_continuous(limits = c(0, 7e+4)) +
52   labs(y = "Total fertility rate (birth per woman)",
```

```

53     x = "GDP per capita, PPP (constant 2017 international USD)",
54     caption = "Own editing based on the Figure 5-2. from Kreiszné Hudák (2019).
55     The trend is drawn via splines.
56     Source of the data: World Bank."
57   )
58
59 plot_NUTS2 <- function(df, viridis_c = T, ..., all.x = F) {
60   p <- df %>%
61     {merge(eurostat::get_eurostat_geospatial(nuts_level = 2), ., all.x = all.x)} %>%
62     ggplot(aes(fill = values)) +
63     geom_sf(color = "black") +
64     theme(
65       axis.text = element_blank()
66     ) +
67     xlim(c(-30, 44)) +
68     ylim(c(35, 70)) +
69     labs(fill = NULL)
70
71   if (viridis_c) {
72     p <- p + scale_fill_viridis_c(option = "magma", ...,
73                                   guide = guide_colorbar(frame.colour = "black",
74                                              ticks.colour = "black"),
75                                   na.value = "white")
76   }
77   p
78 }
79
80 # Data import -----
81
82 # Fertility: source: Eurostat database =====
83
84 f_data <- get_eurostat("demo_r_find2", time_format = "num") %>%
85   select(geo, time, var = indic_de, values) %>%
86   filter(!str_detect(geo, "TR"))
87
88 #### Map of TFR #####
89
90 f_data %>%
91   filter(var == "TOTFERRT" & str_length(geo) == 4 & time == 2017) %>%
92   plot_NUTS2(limits = c(1, 3))
93 # Data from Global Data Labor =====
94
95 #### Sub-national data #####
96
97 # source of csv files: https://globaldatalab.org/
98
99 GDL_import <- function(x) {
100   get_eurostat_geospatial(nuts_level = 2) %>%
101     data.frame() %>%
102     tibble %>%
103     mutate(
104       ISO_Code = countrycode::countrycode(CNTR_CODE, origin = "iso2c", "iso3c"),
105       ISO_Code = ifelse(CNTR_CODE == "UK", "GBR", ISO_Code),

```

```

106     ISO_Code = ifelse(CNTR_CODE == "EL", "GRC", ISO_Code),
107 ) %>%
108 select(ISO_Code, NUTS_NAME, geo) %>%
109 merge(read_csv(x), by = "ISO_Code") %>%
110 mutate(
111   z = stringdist::stringsim(NUTS_NAME, Region)
112 ) %>%
113 arrange(desc(z)) %>%
114 filter(!duplicated(Region)) %>%
115 filter(!duplicated(NUTS_NAME)) %>%
116 filter((z > .5 | Country %in% c("Greece", "Turkey", 'Romania',
117                           'Malta', 'Italy')) & NUTS_NAME != "Dresden") %>%
118 select(geo, '1990':'2018') %>%
119 pivot_longer(-1, names_to = "time", values_to = "values") %>%
120 mutate(time = as.numeric(time))
121 }
122
123 GDL_subnat <- GDL_import("GDL-Sub-national-HDI-data.csv") %>% rename(HDI = values) %>%
124 merge(
125   GDL_import("GDL-Educational-index--data.csv") %>% rename(education = values)
126 ) %>%
127 merge(
128   GDL_import("GDL-Health-index-data.csv") %>% rename(health = values)
129 ) %>%
130 merge(
131   GDL_import("GDL-Income-index-data.csv") %>% rename(income = values)
132 )
133
134 #### National data #####
135
136 GDL_nat <- read_csv("GDL-Sub-national-HDI-data.csv") %>%
137   filter(Level == "National") %>%
138   select(Country, 6:34) %>% pivot_longer(-1, names_to = "time", values_to = "values") %>%
139   mutate(time = as.numeric(time)) %>%
140   na.omit()
141
142 # Data from UNDP =====
143
144 # source: http://hdr.undp.org/
145
146 HDI_UNDP <- read_csv("Human Development Index (HDI).csv",
147   skip = 5) %>%
148   select(!starts_with("X"), - 'HDI Rank') %>%
149   mutate_at(-1,
150     function(x) {as.numeric(ifelse(x == "...", NA, x))}) %>%
151   pivot_longer(-1, names_to = "time", values_to = "values") %>%
152   mutate(time = as.numeric(time)) %>%
153   na.omit()
154
155 merge(GDL_nat %>% rename(GDL = values),
156       HDI_UNDP %>% rename(UNDP = values)) %>%
157 {
158   c(

```

```

159 scales::percent(cor(x = .\$GDL, y = .\$UNDP)^2, accuracy = .01),
160 scales::percent(cor(x = .\$GDL, y = .\$UNDP, method = "spearman")^2, accuracy = .01),
161 as.character(format(mean(abs(.\$GDL - .\$UNDP)), digits = 1)),
162 scales::percent(mean(abs(.\$GDL - .\$UNDP)/.\$UNDP), accuracy = .01)
163 )
164 } %>%
165 {tibble(
166   Indicator = c("$R^2$", "Spearman $R^2$",
167                 "Mean absolute deviation", "Mean absolute percentage deviation"),
168   Value = .
169 )} %>%
170 kable(
171   caption = "Indicators of similiarity between the Human Development Indices
172 provided by UNDP and GDL"
173 )
174
175 GDL_subnat %>%
176   filter(time == 2017) %>%
177   select(-time) %>%
178   pivot_longer(-1, names_to = "var", values_to = "values") %>%
179   filter(!is.na(var)) %>%
180   mutate(
181     var = str_to_title(var),
182     var = str_replace(var, "Hdi", "HDI"),
183     var = factor(var,
184                   levels = c("HDI", "Income", "Education", "Health"),
185                   ordered = T)
186   ) %>%
187   plot_NUTS2() + facet_wrap(~ var, ncol = 2) +
188   labs(caption = "Source: https://globaldatalab.org")
189
190 GDP_index <- get_eurostat("nama_10r_2gdp", time_format = "num") %>%
191   filter(unit == "PPS_EU27_2020_HAB") %>% # Purchasing power standard (PPS) per inhabitant
192   select(-unit) %>%
193   rename(GDP = values) %>% merge(
194     get_eurostat("ert_bil_eur_a", time_format = "num") %>% # EUR/USD annual avg exc r
195       filter(currency == "USD" & statinfo == "AVG") %>%
196       select(time, e = values)
197   ) %>%
198   {
199     GDP <- .\$GDP/.\$e # mutate to USD
200     mutate(.,
201       GDPindex = (log(GDP) - log(100))/(
202         log(75000) - log(100))
203     )
204   }
205
206 merge(GDL_subnat, GDP_index) %>%
207   {
208     c(
209       scales::percent(cor(x = .\$income, y = .\$GDPindex)^2, accuracy = .01),
210       scales::percent(cor(x = .\$income, y = .\$GDPindex, method = "spearman")^2,
211                     accuracy = .01),

```

```

212     as.character(format(mean(abs(.income - .GDPindex)), digits = 1, nsmall = 4)),
213     scales::percent(mean(abs(.income - .GDPindex)/.GDPindex), accuracy = .01)
214   )
215 } %>%
216 {tibble(
217   Indicator = c("$R^2$", "Spearman $R^2$",
218                 "Mean absolute deviation", "Mean absolute percentage deviation"),
219   Value = .
220 )} %>%
221 kable(
222   caption = "Indicators of similiarity between the income component of the
223   Human Development Indices provided by GDL and the estimation based on regional GDP"
224 )
225
226 health_index <- get_eurostat("demo_r_mlifexp", time_format = "num") %>%
227   filter(age == "Y_LT1" & sex == "T") %>%
228   select(geo, time, le = values) %>%
229   mutate(
230     health_index = (le - 20) / (85 - 20)
231   )
232
233 merge(GDL_subnat, health_index) %>%
234 {
235   c(
236     scales::percent(cor(x = .$health, y = .$health_index)^2, accuracy = .01),
237     scales::percent(cor(x = .$health, y = .$health_index, method = "spearman")^2,
238                     accuracy = .01),
239     as.character(format(mean(abs(.health - .$health_index)), digits = 1, nsmall = 4)),
240     scales::percent(mean(abs(.health - .$health_index)/.$health_index), accuracy = .01)
241   )
242 } %>%
243 {tibble(
244   Indicator = c("$R^2$", "Spearman $R^2$",
245                 "Mean absolute deviation", "Mean absolute percentage deviation"),
246   Value = .
247 )} %>%
248 kable(
249   caption = "Indicators of similiarity between the health component of the
250   Human Development Indices provided by GDL and the estimation based on regional life
251   expectancy"
252 )
253
254 edu_wide <- get_eurostat("edat_lfse_04", time_format = "num") %>%
255   filter(sex == "T" & !str_detect(isced11, "GEN") &
256         !str_detect(isced11, "VOC") & isced11 != "ED3-8" & !str_detect(geo, "TR"))
257 ) %>%
258   mutate(
259     var = str_c(age, ":", isced11)
260   ) %>%
261   select(geo, time, var, values) %>%
262   pivot_wider(names_from = var, values_from = values) %>%
263   {
264     x <- .

```

```

265     names(x) <- letters[1:length(x)]
266     x <- cbind(
267       .[, 1:2],
268       mice::complete(mice::mice(select(x, -a,-b), printFlag = F))
269     )
270     names(x) <- names(.)
271     x
272   }
273
274 edu_comps <- edu_wide%>%
275   select(-time, -geo) %>%
276   na.omit() %>%
277   {princomp(scale(.))}

278
279 edu_comp_vars <- edu_comps %>%
280   summary() %>%
281   {$.sdev^2/sum($.sdev^2)} %>%
282   scales::percent(accuracy = .01) %>%
283   {str_c("# ", 1:length(.), " (", ., ")")}

284
285 edu_comps %>%
286   .$loadings %>%
287   unclass() %>%
288   data.frame() %>%
289   rownames_to_column() %>%
290   pivot_longer(-1) %>%
291   mutate(
292     name = as.numeric(str_remove(name, 'Comp.')),
293   ) %>%
294   arrange(name) %>%
295   mutate(
296     name = edu_comps %>%
297       summary() %>%
298       {$.sdev^2/sum($.sdev^2)} %>%
299       scales::percent(accuracy = .01) %>%
300       {str_c("# ", 1:length(.), " (", ., ")")}) %>%
301     .[name]
302   ) %>%
303   ggplot +
304   aes(rowname, value, fill = value < 0) +
305   geom_hline(yintercept = 0) +
306   geom_col(color = 'black') +
307   coord_flip() +
308   scale_fill_viridis_d(guide = F, option = "magma", begin = .4,
309                         end= .7, direction = -1) +
310   scale_y_continuous(labels = scales::percent) +
311   facet_wrap(~name, ncol = 3) +
312   labs(x = NULL, y = NULL, caption =
313         "The corresponding proportion of the explained variance are in the brackets.")

314
315 edu_comps %>% .$scores %>%
316   cbind(edu_wide) %>% merge(GDL_subnat) %>%
317   select(3:11, education) %>%

```

```

318 {
319   x <- .$education
320   apply(select(., -education), 2, function(y) {
321     c(
322       scales::percent(cor(x = x, y = y)^2, accuracy = .01),
323       scales::percent(cor(x = x, y = y, method = "spearman")^2, accuracy = .01)
324     )
325   })
326 } %>%
327 data.frame() %>%
328 mutate(Indicator = c("$R^2$", "Spearman $R^2$")) %>%
329 rename_all(function(x) str_replace(x, "p.", "p ")) %>%
330 select(Indicator, 1:9) %>%
331 kable(
332   caption = "Indicators of similarity between the knowledge component of Human
333   Development Indices provided by UNDP and the calculated principal components using
334   educational attainment level"
335 )
336
337 edu_index <- edu_comps %>%
338   .$scores %>%
339   data.frame() %>%
340   select(2) %>%
341   cbind(edu_wide) %>%
342   select(geo, time, edu_index = Comp.2) %>%
343   mutate(
344     edu_index = -edu_index,
345     edu_index = edu_index + abs(min(edu_index)),
346     edu_index = edu_index/max(edu_index)
347   )
348
349 dat <- f_data %>%
350   pivot_wider(names_from = var, values_from = values) %>%
351   merge(edu_index, all = T) %>%
352   merge(health_index, all = T) %>%
353   merge(GDP_index, all = T) %>%
354   filter(!str_detect(geo, "TR"))
355
356 FAM_df <- get_eurostat("spr_exp_sum", time_format = "num") %>%
357   filter(spdeps == "FAM" & unit == "PC_GDP") %>%
358   rename(FAM = values, country = geo) %>%
359   select(-(spdeps:unit))
360
361 yth_empl_byage <- get_eurostat("yth_empl_110", time_format = "num") %>%
362   filter(unit == "PC" & sex == "F") %>%
363   filter(age %in% c("Y15-19", "Y20-24", "Y25-29")) %>%
364   select(-unit, -sex) %>%
365   pivot_wider(names_from = age, values_from = values) %>%
366   rename(
367     "uY15" = "Y15-19",
368     "uY20" = "Y20-24",
369     "uY25" = "Y25-29"
370   )

```

```

371
372 dat <- dat %>%
373   mutate(country = str_sub(geo, end = 2)) %>%
374   merge(FAM_df, all.x = T, all.y = F) %>%
375   merge(yth_empl_byage, all.x = T, all.y = F) %>%
376   filter(!str_detect(geo, "TR"))
377
378 f.clean_names <- function(v, Tosparse = F) {
379   v <- str_replace_all(v, "GDPindex", "Income index") %>%
380     str_replace_all("health_index", "Health index") %>%
381     str_replace_all("edu_index", "Education index") %>%
382     str_replace_all("TOTFERRT", "TFR") %>%
383     str_replace_all("GDPindex", "Income index") %>%
384     str_replace_all("FAM", "Family benefit") %>%
385     str_replace_all("uY15", "UR (15-19 y)") %>%
386     str_replace_all("uY20", "UR (20-24 y)") %>%
387     str_replace_all("uY25", "UR (25-29 y)")
388   if(Tosparse) v <- str_replace_all(v, " ", "~")
389   v
390 }
391
392 # Explore the data -----
393
394 # Pairwise correlations =====
395
396 dat %>%
397   filter(str_length(geo) == 4) %>%
398   select(TOTFERRT, GDPindex, edu_index, health_index) %>%
399   {set_names(., f.clean_names(names(.)))} %>%
400   GGally::ggpairs()
401
402 dat %>%
403   filter(str_length(geo) == 4) %>%
404   select(TOTFERRT, uY15, uY20, uY25) %>%
405   set_names("TFR", "Unemployment rate\n(15-19 years old)",
406             "Unemployment rate\n(20-24 years old)",
407             "Unemployment rate\n(25-29 years old)") %>%
408   GGally::ggpairs()
409
410 dat %>%
411   filter(str_length(geo) == 2) %>%
412 {ggplot(., aes(FAM, TOTFERRT)) +
413   geom_vline(aes(xintercept = mean(..$FAM, na.rm = T), linetype = "means")) +
414   geom_hline(yintercept = mean(..$TOTFERRT, na.rm = T), linetype = 2) +
415   geom_point() +
416   geom_smooth(method = "lm", aes(color = "Linear trend"), size = 1.5) +
417   scale_color_viridis_d() +
418   scale_linetype_manual(values = c("means" = 2)) +
419   labs(y = "Total fertility rate (birth per woman)",
420        x = "Family benefit (% of GPD)",
421        linetype = NULL, color = NULL
422      ) +
423   theme(
424     legend.position = "bottom"

```

```

425   )
426 }
427
428 # Regression trees =====
429
430 m_part <- dat %>%
431   filter(str_length(geo) == 4) %>%
432   select(TOTFERRT, GDPindex, edu_index, health_index, FAM) %>%
433   {set_names(., f.clean_names(names(.)))} %>%
434   na.omit() %>%
435   rpart::rpart(formula = TFR ~ ., cp = .01)
436
437 m_part %>% rattle::fancyRpartPlot(palettes = 'PuRd', sub = NULL)
438
439 m_part %>% summary()
440
441 m_part2 <- dat %>%
442   filter(str_length(geo) == 4) %>%
443   select(TOTFERRT, GDPindex, edu_index, health_index, uY15, uY20, uY25, FAM) %>%
444   {set_names(., f.clean_names(names(.)))} %>%
445   na.omit() %>%
446   rpart::rpart(formula = TFR ~ ., cp = .01)
447
448 m_part2 %>% rattle::fancyRpartPlot(palettes = 'PuRd', sub = NULL)
449
450 # Model building -----
451
452 # Transform the data for panel modeling =====
453
454 dat_plm <- dat %>%
455   select(
456     geo, time, TOTFERRT, edu_index, health_index, GDPindex, FAM, uY15, uY20, uY25
457   ) %>%
458   filter(str_length(geo) == 4 & !is.na(TOTFERRT))
459
460 dat_plm <- dat_plm %>%
461   select(-TOTFERRT) %>%
462   mutate(time = time + 1) %>%
463   {
464     set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
465                           paste0(names(.), '_1')))
466   } %>%
467   merge(dat_plm, all.x = F, all.y = T)
468
469 dat_plm <- dat_plm %>%
470   select(!ends_with("_1")) %>%
471   select(-TOTFERRT) %>%
472   mutate(time = time + 2) %>%
473   {
474     set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
475                           paste0(names(.), '_1_1')))
476   } %>%
477   merge(dat_plm, all.x = F, all.y = T)

```

```

478
479 dat_plm <- dat_plm %>%
480   select(-TOTFERRRT) %>%
481   mutate_at(-(1:2), function(x) x^2) %>%
482   {
483     set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
484                   paste0(names(.), '_2')))
485   } %>%
486   merge(dat_plm, all.x = F, all.y = T)
487
488 ## Initial models =====
489
490 library(plm)
491
492 m_panels <- c(
493   'TOTFERRRT ~ edu_index_l_1 + health_index_l_1 + GDPindex_l_1 + FAM_l_1 +
494   uY15_l_1 + uY20_l_1 + uY25_l_1 + edu_index_l + health_index_l + GDPindex_l + FAM_l +
495   uY15_l + uY20_l + uY25_l + edu_index + health_index + GDPindex + FAM + uY15 + uY20 +
496   uY25',
497   'TOTFERRT ~ edu_index_l_1 + health_index_l_1 + GDPindex_l_1 + FAM_l_1 + edu_index_l +
498   health_index_l + GDPindex_l + FAM_l + edu_index + health_index + GDPindex + FAM'
499 ) %>%
500   lapply(function(formula) {
501     pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
502     within <- plm(eval(formula), data = dat_plm, model = "within")
503     random <- plm(eval(formula), data = dat_plm, model = "random")
504
505     list(
506       tests = c(
507         pooltest(pooling, within)$p.value,
508         phtest(within, random)$p.value,
509         plm::r.squared(within, dfcor = T)),
510       model = within,
511       OLS = formula %>%
512         str_replace_all('\\~', ' ', ' ') %>%
513         str_replace_all('\\+', ' ', ' ') %>%
514         str_split(' ', ' ') %>%
515         .[[1]] %>%
516         trimws() %>%
517         {c(., 'geo')} %>%
518         {select(dat_plm, .)} %>%
519         na.omit() %>%
520         group_by(geo) %>%
521         summarise_all(.funs = function(x) mean(x)) %>%
522         merge(
523           plm::fixef(within) %>%
524             {tibble(geo = names(.), a = .)}
525         ) %>%
526         mutate(
527           TOTFERRT = TOTFERRT - a
528         ) %>%
529         lm(formula = formula)
530       )
531     })

```

```

532
533 ### Plot coefficients #####
534
535 m_panels %>%
536   lapply(function(output) {
537     output$model %>% broom::tidy(conf.int = T) %>%
538       rownames_to_column()
539   }) %>%
540   reduce(rbind) %>%
541   mutate(
542     rowname = paste0("Model ", as.roman(cumsum(rowname == 1)), "."),
543     term = f.clean_names(term, Tosparse = T),
544     term = gsub("_2", "^2", term),
545     term = gsub("_1_1", '[t = -2]', term),
546     term = gsub("_1", '[t = -1]', term),
547   ) %>%
548   ggplot() +
549   aes(estimate, term, color = p.value <= .05) +
550   geom_vline(xintercept = 0, color = "gray4") +
551   geom_point() +
552   geom_pointrange(aes(xmin = conf.low, xmax = conf.high)) +
553   facet_wrap(~rowname, nrow = 1) +
554   labs(x = "Estimated coefficient", y = "Term", color = "Corresponding p-value \u2264 5%") +
555   scale_color_viridis_d(option = "magma", begin = .2, end = .7) +
556   scale_y_discrete(labels=scales::parse_format()) +
557   theme(
558     legend.position = "bottom",
559     legend.direction = "horizontal"
560   )
561
562 ### Model descriptions #####
563
564 m_panels %>%
565   lapply(function(output) {
566     c(output$tests, nrow(augment(output$OLS)))
567   }) %>%
568   reduce(rbind) %>%
569   t() %>%
570   data.frame() %>%
571   mutate_all(function(x) c(scales::percent(x[1:3], accuracy = .01),
572                           as.character(x[4]))) %>%
573   {set_names(., paste0("Model ", as.roman(1:ncol(.)), "."))} %>%
574   mutate(
575     Indicator = c("Pooltest", "Phtest", "Adjusted $R^2$", "Observations")
576   ) %>%
577   select(Indicator, everything()) %>%
578   knitr::kable(caption = "Models", align = c("l", "c", "c", "c"))
579
580 # Framework I. =====
581
582 ## Bias of framework #####
583
584 names(dat_plm) %>%
585   { ifelse(

```

```

586     . %in% c("geo", "time") | str_detect(., "_1") | str_detect(., "_2"), NA, .
587   )} %>%
588   na.omit() %>%
589   lapply(function(variable){
590     x <- pull(dat_plm, variable) %>% na.omit()
591     y <- pull(na.omit(dat_plm), variable)
592     t <- t.test(x, y)
593     tibble(
594       Variable = f.clean_names(variable),
595       'Mean in total sample' = mean(x),
596       'Mean in used sample' = mean(y),
597       'Number of observations in the total sample' = length(x),
598       'T-statistic' = t$statistic,
599       'P-value' = scales::percent(t$p.value, accuracy = .01)
600     )
601   }
602   ) %>% reduce(rbind) %>%
603   knitr::kable(caption = 'T-tests', digits = 4,
604                 align = c('l', 'c', 'c', 'c', 'c', 'c', 'c'))
605
606 dat_plm %>%
607   na.omit() %>%
608   group_by(geo) %>%
609   summarise(
610     values = n()
611   ) %>%
612   plot_NUTS2(all.x = T) +
613   scale_fill_viridis_b(option = "magma", direction = -1, begin = .2,
614                         na.value = "white")
615
616 #### Run lasso regression #####
617
618 library(glmnet)
619
620 y <- na.omit(dat_plm)$TOTFERRT
621 X <- model.matrix(TOTFERRT ~ ., data = select(na.omit(dat_plm), -time))
622 LASSO <- cv.glmnet(X, y)
623
624 tidy(LASSO) %>%
625   ggplot() +
626   aes(log(lambda), ymin = conf.low*1000, ymax = conf.high*1000) +
627   geom_line(aes(log(lambda), estimate*1000, color = "Mean-Squared Error")) +
628   geom_step(aes(y = nzero, color = "Number of used explanatory variables"),
629             size = 1) +
630   geom_ribbon(alpha = .4) +
631   geom_hline(yintercept = 0, size = 1) +
632   geom_vline(aes(xintercept = log(LASSO$lambda.1se), color = "lambda.1se"),
633              # TODO name the line
634              linetype = 2) +
635   geom_vline(aes(xintercept = log(LASSO$lambda.min), color = "lambda.min"),
636              linetype = 2) +
637   scale_y_continuous(
638     name = "Number of used explanatory variables",

```

```

639     sec.axis = sec_axis( trans=~./1000, name = "Mean-Squared Error")
640   ) +
641   labs(y = "Mean-Squared Error", x = "Log(\u03bb)", color = NULL) +
642   scale_color_viridis_d(option = "magma", end = .8) +
643   theme(legend.position = "bottom", legend.direction = "horizontal")
644
645 lasso_coefs <- capture.output(
646   coef(LASSO, LASSO$lambda.1se)
647 ) %>%
648   .[-(1:2)] %>%
649   {tibble(x = .)} %>%
650   mutate(term = gsub(" .*", "", x), coef = gsub(".* ", "", x)) %>%
651   select(-x) %>%
652   filter(!str_detect(term, "geo") & coef != "" & term != "(Intercept)")
653
654 m_panels2 <- paste("TOTFERRT ~", paste(lasso_coefs$term, collapse = " + ")) %>%
655   lapply(function(formula) {
656     pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
657     within <- plm(eval(formula), data = dat_plm, model = "within")
658     random <- plm(eval(formula), data = dat_plm, model = "random")
659     list(
660       tests = c(
661         pooltest(pooling, within)$p.value,
662         phtest(within, random)$p.value,
663         plm::r.squared(within, dfcor = T)),
664       model = within,
665       OLS = formula %>%
666         str_replace_all('\\~', ' ', ' ') %>%
667         str_replace_all('\\+', ' ', ' ') %>%
668         str_split(',') %>%
669         .[[1]] %>%
670         trimws() %>%
671         {c(., 'geo')} %>%
672         {select(dat_plm, .)} %>%
673         na.omit() %>%
674         group_by(geo) %>%
675         summarise_all(.funs = function(x) mean(x)) %>%
676         merge(
677           plm::fixef(within) %>%
678             {tibble(geo = names(.), a = .)}
679         ) %>%
680         mutate(
681           TOTFERRT = TOTFERRT - a
682         ) %>%
683         lm(formula = formula)
684       )
685     }
686   )
687
688 m_panels2 %>%
689   lapply(function(output) {
690     output$tests
691   })

```

```

692 standard_beta <- m_panels2[[1]]$OLS %>%
693   QuantPsyc::lm.beta() %>%
694   {tibble(term = names(.), beta = .)} %>%
695   filter(!str_detect(term, "geo"))

696 standard_beta <- augment(m_panels2[[1]]$OLS) %>%
697   select(TOTFERRT:.fitted) %>%
698   select(-.fitted) %>%
699   cor() %>%
700   data.frame() %>%
701   select(1) %>%
702   rownames_to_column() %>%
703   rename(term = rowname) %>%
704   merge(standard_beta) %>%
705   mutate(explain = abs(TOTFERRT*beta)) %>%
706   select(term, cor = TOTFERRT, standard_beta = beta, explain)

707 lasso_coefs %>%
708   rename(lasso = coef) %>%
709   merge(tidy(m_panels2[[1]]$OLS, conf.int = T)) %>%
710   merge(standard_beta) %>%
711   mutate_at(-1, function(x) as.numeric(x)) %>%
712   pivot_longer(c(lasso:estimate, cor:explain)) %>%
713   mutate(
714     conf.low = ifelse(name != "estimate", NA, conf.low),
715     conf.high = ifelse(name != "estimate", NA, conf.high),
716     lag = ifelse(str_detect(term, "_1"),
717                  ifelse(str_detect(term, "_1_1"), 2, 1), 0),
718     bar = ifelse(name %in% c("cor", "explain"), value, NA),
719     value = ifelse(!(name %in% c("cor", "explain")), value, NA),
720     term = f.clean_names(term, Tosparse = T),
721     term = gsub("_2", "^2", term),
722     term = gsub("_1_1", '[t = -2]', term),
723     term = gsub("_1", '[t = -1]', term),
724     name = case_when(
725       name == "cor" ~ 'Correlation coefficient',
726       name == "estimate" ~ 'Coefficient in OLS',
727       name == "lasso" ~ 'Coefficient in lasso regression',
728       name == "standard_beta" ~ 'Standardized coefficient',
729       name == "explain" ~ 'Contribution to R-squared'
730     ),
731     name = factor(
732       name, levels = c(
733         'Correlation coefficient',
734         'Coefficient in lasso regression',
735         'Coefficient in OLS',
736         'Standardized coefficient',
737         'Contribution to R-squared'
738       )
739     ),
740   ) %>%
741   {
742     ggplot(.) +

```

```

746     geom_vline(xintercept = 0) +
747     geom_point(aes(value, term, fill = factor(lag)), size = 3) +
748     geom_col(aes(bar, term, fill = factor(lag)), color = "black") +
749     scale_fill_viridis_d(option = "magma", begin = .3, end = .7) +
750     scale_y_discrete(labels = scales::parse_format()) +
751     facet_wrap(~ name, scales = "free_x") +
752     labs(x = NULL, y = "Term", fill = "Lag") +
753     theme(legend.position = "bottom", legend.direction = "horizontal")
754   }
755
756 dat_plm <- dat_plm %>%
757   select(!starts_with("uY"))
758
759 LASSO <- cv.glmnet(model.matrix(TOTFERRT ~ ., data = select(na.omit(dat_plm), -time)),
760                     na.omit(dat_plm)$TOTFERRT)
761
762 names(dat_plm) %>%
763   {ifelse(
764     . %in% c("geo", "time") | str_detect(., "_1") | str_detect(., "_2"), NA, .
765   )} %>%
766   na.omit() %>%
767   lapply(function(variable){
768     x <- pull(dat_plm, variable) %>% na.omit()
769     y <- pull(na.omit(dat_plm), variable)
770     t <- t.test(x, y)
771     tibble(
772       Variable = f.clean_names(variable),
773       'Mean in total sample' = mean(x),
774       'Mean in used sample' = mean(y),
775       'Number of observations in total sample' = length(x),
776       'T-statistic' = t$statistic,
777       'P-value' = scales::percent(t$p.value, accuracy = .01)
778     )
779   })
780 } %>% reduce(rbind) %>%
781 knitr::kable(caption = 'T-tests', digits = 4,
782               align = c('l', 'c', 'c', 'c', 'c', 'c', 'c'))
783
784 dat_plm %>%
785   na.omit() %>%
786   group_by(geo) %>%
787   summarise(
788     values = n()
789   ) %>%
790   plot_NUTS2(all.x = T) +
791   scale_fill_viridis_b(option = "magma", direction = -1, begin = .2,
792                         na.value = "white")
793
794 lasso_coefs <- capture.output(
795   coef(LASSO, LASSO$lambda.1se)
796 ) %>%
797   .[-(1:2)] %>%
798   {tibble(x = .)} %>%

```

```

799   mutate(term = gsub(" .*", "", x), coef = gsub(".* ", "", x)) %>%
800   select(-x) %>%
801   filter(!str_detect(term, "geo") & coef != "" & term != "(Intercept)")
802
803 m_panels2 <- paste("TOTFERRT ~", paste(lasso_coefs$term, collapse = " + ")) %>%
804   lapply(function(formula) {
805     pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
806     within <- plm(eval(formula), data = dat_plm, model = "within")
807     random <- plm(eval(formula), data = dat_plm, model = "random")
808     list(
809       tests = c(
810         pooltest(pooling, within)$p.value,
811         phtest(within, random)$p.value,
812         plm::r.squared(within, dfcor = T)),
813       model = within,
814       OLS = formula %>%
815       str_replace_all('\\~', ' ', ' ') %>%
816       str_replace_all('\\+ ', ' ', ' ') %>%
817       str_split(' ', ' ') %>%
818       .[[1]] %>%
819       trimws() %>%
820       {c(., 'geo')} %>%
821       {select(dat_plm, .)} %>%
822       na.omit() %>%
823       group_by(geo) %>%
824       summarise_all(.funs = function(x) mean(x)) %>%
825       merge(
826         plm:::fixef(within) %>%
827         {tibble(geo = names(.), a = .)}
828       ) %>%
829       mutate(
830         TOTFERRT = TOTFERRT - a
831       ) %>%
832       lm(formula = formula)
833     )
834   }
835 )
836
837 m_panels2 %>%
838   lapply(function(output) {
839     output$tests
840   })
841
842 standard_beta <- m_panels2[[1]]$OLS %>%
843   QuantPsyc::lm.beta() %>%
844   {tibble(term = names(.), beta = .)} %>%
845   filter(!str_detect(term, "geo"))
846
847 standard_beta <- augment(m_panels2[[1]]$OLS) %>%
848   select(TOTFERRT:.fitted) %>%
849   select(-.fitted) %>%
850   cor() %>%
851   data.frame() %>%

```

```

852   select(1) %>%
853   rownames_to_column() %>%
854   rename(term = rowname) %>%
855   merge(standard_beta) %>%
856   mutate(explain = abs(TOTFERRT*beta)) %>%
857   select(term, cor = TOTFERRT, standard_beta = beta, explain)
858
859 lasso_coefs %>%
860   rename(lasso = coef) %>%
861   merge(tidy(m_panels2[[1]]$OLS, conf.int = T)) %>%
862   merge(standard_beta) %>%
863   mutate_at(-1, function(x) as.numeric(x)) %>%
864   pivot_longer(c(lasso:estimate, cor:explain)) %>%
865   mutate(
866     conf.low = ifelse(name != "estimate", NA, conf.low),
867     conf.high = ifelse(name != "estimate", NA, conf.high),
868     lag = ifelse(str_detect(term, "_l"),
869                  ifelse(str_detect(term, "_l_l"), 2, 1), 0),
870     bar = ifelse(name %in% c("cor", "explain"), value, NA),
871     value = ifelse(!(name %in% c("cor", "explain")), value, NA),
872     term = f.clean_names(term, Tosparse = T),
873     term = gsub("_2", "^2", term),
874     term = gsub("_l_l", '[t = -2]', term),
875     term = gsub("_l", '[t = -1]', term),
876     name = case_when(
877       name == "cor" ~ 'Correlation coefficient',
878       name == "estimate" ~ 'Coefficient in OLS',
879       name == "lasso" ~ 'Coefficient in lasso regression',
880       name == "standard_beta" ~ 'Standardized coefficient',
881       name == "explain" ~ 'Contribution to R-squared'
882     ),
883     name = factor(
884       name, levels = c(
885         'Correlation coefficient',
886         'Coefficient in lasso regression',
887         'Coefficient in OLS',
888         'Standardized coefficient',
889         'Contribution to R-squared'
890       )
891     )
892   ) %>%
893 {
894   ggplot(.) +
895   geom_vline(xintercept = 0) +
896   geom_point(aes(value, term, fill = factor(lag)), size = 3) +
897   geom_col(aes(bar, term, fill = factor(lag)), color = "black") +
898   scale_fill_viridis_d(option = "magma", begin = .3, end = .7) +
899   scale_y_discrete(labels=scales::parse_format()) +
900   facet_wrap(~ name, scales = "free_x") +
901   labs(x = NULL, y = "Term", fill = "Lag") +
902   theme(legend.position = "bottom", legend.direction = "horizontal")
903 }

```