The impact of human development on fertility rates

An empirical analysis of fertility rates amoung the regions of Europe

Marcell Granát

April 12, 2021

Contents

Abstract

If one attends to the extremely large literature of demographic trends in the developed world, then the uncertainty about the effect of economic and human development factors on the fertility rate cannot be covered for a long time. Several empirical studies argue for the existence of the J-shaped effect of the development, but many papers come up with statements to the opposite. The goal of this paper is to contribute to the literature with an advanced panel econometric model based on regional observations. Beyond the human development factors (living standard, education and health) I extend my analysis by using youth unemployment and family benefit indicators as dependent variables. Important to note that statistics about unemployment are available only for a critically short period in the case of many regions. To manage this highly unbalanced nature of the dataset – while not rejecting the possibility to control for youth unemployment – I estimate the model with two different modeling frames: one without youth unemployment and another one with it. As a result, the paper confirms the empirical evidence that increasing human development in developed countries has a positive effect on total fertility rates, and income is the most important component. This finding is robust to the mentioned two frameworks. In contrast, the research come up only with week evidence for the significant effect of expenditure on family on total fertility rates on the long run.

Keywords— fertility rates, human development

List of Tables

List of Figures

Introduction

Total fertility rates have decreased significantly over the past few decades, and in most of the developed countries, they are far below the minimum level, which would ensure the reproducibility of the population (2.1 children/woman). The motivation why this demographic pattern frequently becomes the focus of scientific and political discussions is that it has a radical impact on the economy of the future.

The one which must be mentioned is that low fertility is one of the key determinants that lead a country to a high old-age dependency ratio. In extreme cases, today's poor childbearing tendency may put huge economic weight on the future working generation. An interesting fact is that this risk is typically faced by developed economies. Moreover, historical data show a clear correlation between the increasing trend of human development and falling childbearing willingness. The first research question I aim to answer in this study is the following: Does human development truly contribute to the fall in fertility rates, and if it does, are the three components (healthy life; access to knowledge; decent standard of living) included with the same weight?

The related literature contains many uncertainties about this question. Not even the existence of the relation, but the correct direction of the effect is also unclear. An important milestone is a piece of empirical evidence that the association between development and fertility might turn from negative to positive above a given threshold of development [Myrskylä et al., 2009].

Detecting the trend of falling fertility rates is not new. It is well documented that this issue was a key question in politics in many countries after the second world war, and so is nowadays. Governments around the world have already implemented several policies to increase fertility: Abortion ban, campaigns and various financial incentives. But the effectiveness (and their social benefit in many other important aspects) of these are often questioned. Since only the last one is frequently applied (and can be observed accurately) in the developed countries, my second research question is whether spending on family support increases the childbearing willingness significantly.

The contribution of the current paper to the literature is that it applies longitudinal econometrics on NUTS-2 level regional annual observations. Eurostat is the main source of the data. I use fixed effect panel regression to estimate the effect of the mentioned predictors on fertility. Important to note that statistics about unemployment are available only for a critically short period in the case of many regions. To manage this highly unbalanced nature of the dataset – while not rejecting the possibility to control for youth unemployment – I estimate the model with two different modeling frames: one without youth unemployment and another one with it. Figure 1 visualize the main steps of the current paper.

As a result, the paper confirms the *empirical evidence that increasing human development in developed countries has a positive effect on total fertility rates, and income is the most important component.* This finding is robust to the mentioned two frameworks. In contrast, the research come up *only with week evidence for the significant effect of expenditure on family on total fertility rates on the long run.*

To ensure fully reproducibility I enclose the used codes in the Appendix, but these are also available on the following GitHub repository: https://github.com/MarcellGranat/fertilityEU.

Theoretical consideration

To give a detailed answer to our research question I introduce the discussed indicators and the possible causal mechanisms among them.

Political debates and unprofessional public discourses usually focus on the total number of livebirth per year. If the target is to describe the effect of demographic changes on the economy, this perspective is usually right. However, the situation is different when the focus is on the causal mechanisms that affect the level of childbearing willingness. The reason is that the total number of births relies explicitly on the demographic trends in the past. This leads to the problem that the number of births decreases if the number of women at childbearing age reduces. Measuring the efficiency of family support expenditures and demographic policies requires eliminating these effects. The total fertility rate (TFR) is a suitable indicator to manage this.

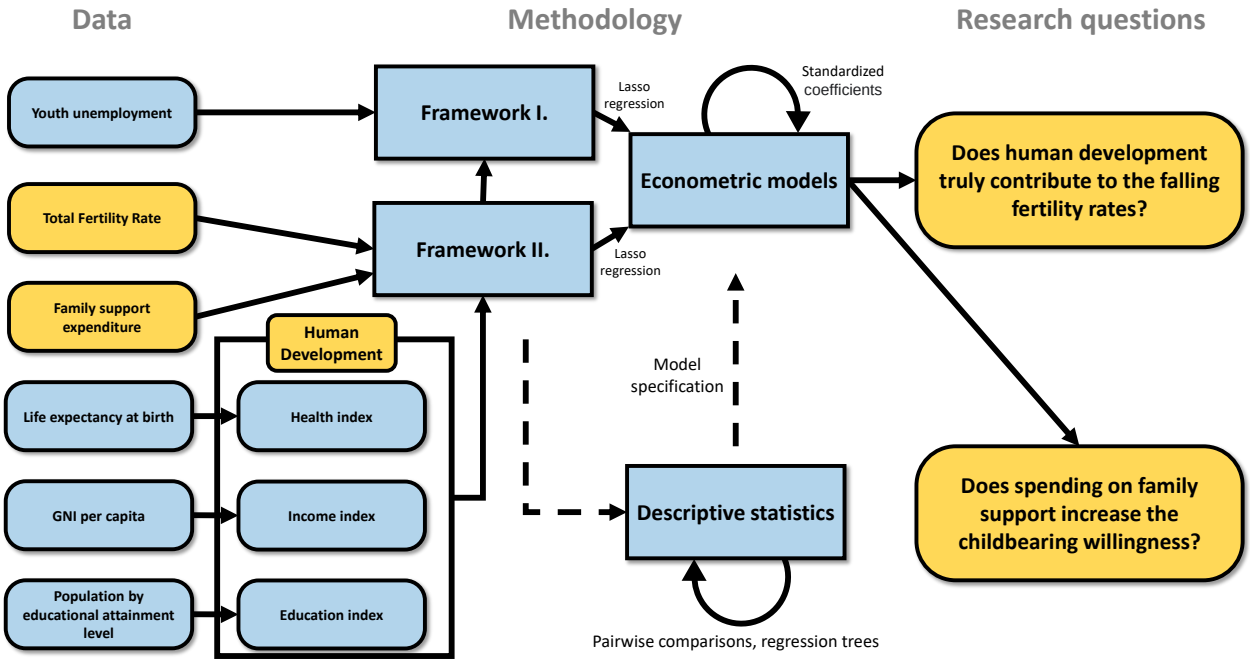


Figure 1: Main steps of the study.

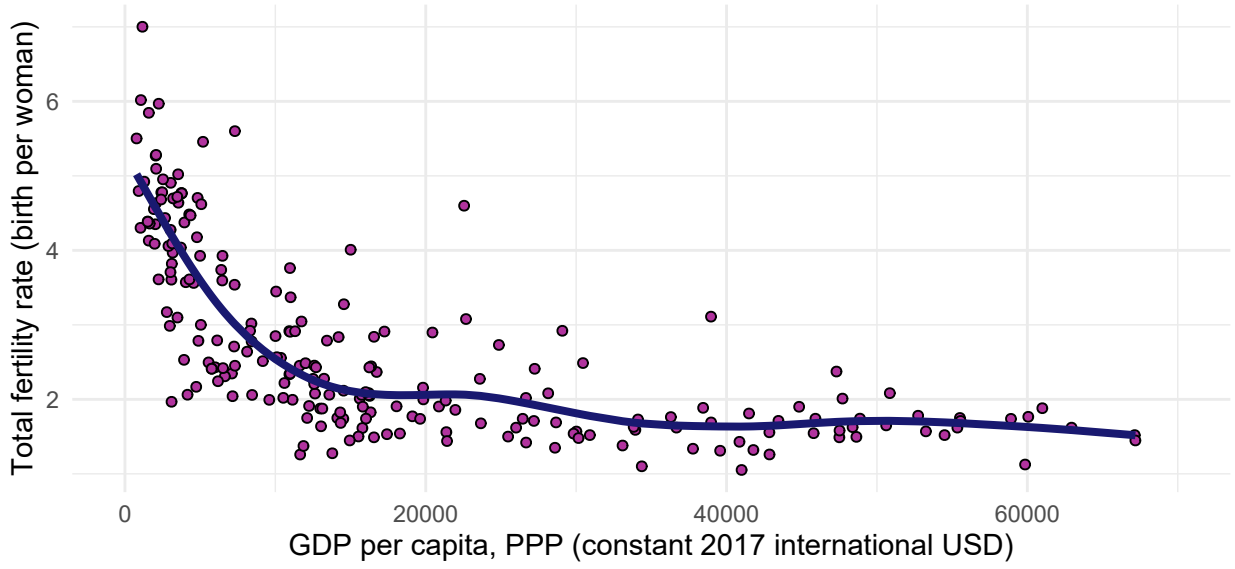
Total fertility rate (TFR) is "the average number of children born per woman over a lifetime given current age-specific fertility rates and assuming no female mortality during reproductive years. TFRs are computed as the sum of age-specific fertility rates defined over five-year intervals" [oec, 2018]. This calculation method ensures that TFR is insensitive to the existing demographic characteristics (robust to the change in the number of childbearing-aged women).

Total fertility rates have dropped significantly in the last decades. More precisely, in the OECD countries, the average has decreased from 2.8 children to 1.7 in the last 50 years [oec, 2019]. Another well-known fact is that fertility is lower in the more developed countries. This correlation is visualized in figure 2.

Figure 1 shows a clear pattern between income and fertility, but national incomes correlate with hundreds of indicators which can be the true reason for failing fertility. The presented relationship was an unquestionable rule of demography a few decades ago, but the empirical evidence of the "J-shaped" fertility trend changed this [eco, 2009]. Myrskyl et al. (2009) reported in their study that the relationship between development and fertility is reversing. The authors explain this with the innovation in family behavior and government policies that improve the compatibility between economic success and family life. This argument confirms the relevance of investigating the combined effect of human development and family support expenditures on fertility.

It is important to note that the mentioned study refers to the Human Development Index, which is a composite index of three key dimensions of human development. "Human development encompasses more than just economic development. The concomitant construction of the HDI offered a simple, yet multidimensional approach to comparatively evaluate the human development of various countries." [Sagar and Najam, 1998]

To describe the motivation behind the decomposition of human development I refer to the findings of the closely related literature. A sizeable number of studies focus on the effect of education, health, income and fertility. The range of methodologies is wide: several studies use cross-sectional or longitudinal national-level indicators, while others refer to data from questionnaires (Generations- and Gender Surveys is a frequently used source). This paper belongs to the first group of studies but extends the research with regional level observations.



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 2: Negative correlation between gross domestic product and fertility rates based on nation level observations (2017)

Harttgen and Vollmer (2014) write that pairwise investigation among the components of human development confirms that development leads to higher fertility in countries where the HDI is higher than 0.86. The authors used country-level data and they highlighted that proving the robustness of their findings with a different sample is needed. The current study performs this.

As mentioned previously, Myrskl et al. (2009) argue that innovations related to feasible work-life balance may be the reason for the increasing fertility in highly developed countries. Taking that into account, I extend the set of explanatory variables with employment statistics. The effect of unemployment on fertility is already investigated and proved using the observations of OECD nations [Adser, 2004].

The perception of rapidly declining fertility in Europe during deteriorated labor market conditions provides further evidence for the relevance of unemployment in the model framework [Matysiak et al., 2020]. Since the former findings state that the working and financial condition affects young people more sensitively [Frejka et al., 2016], I focus on the effect of youth unemployment rates on fertility.

Data

Eurostat is an abundant and reliable source of NUTS-2 level statistics, so it is a reasonable choice to use[1]. However, human development is a hardly available indicator. "The Human Development Index (HDI) is a summary measure of achievements in three key dimensions of human development: a long and healthy life, access to knowledge and a decent standard of living." [UNITED NATIONS DEVELOPMENT PROGRAMME, 2020] The disadvantage of the HDI and its components defined by the United Nations Development Programme is that they are reported officially only on the national level. To ensure comparability with studies that used national-level HDI, I aim to obtain or calculate equivalent indicators.

[1]Technical note: To ensure easy reproducibility data was download with the dedicated Eurostat R package [Lahti et al., 2017].

Human Development

A possible source for this target is the database provided by the Global Data Lab. [Global Data Lab - Innovative Instruments for Turning Data into Knowledge] This website and the data are created by the Institute of Management Research at Radboud University. Their reported values for the national level are almost fully equivalent to the ones published by UNDP. Table 1 figures the similarity between the two datasets.

The reason why this dataset does not fully meet the requirements to use it in this research is that the territorial units do not completely fit with the ones reported by Eurostat, so merging the two data sources is only possible with a high rate of mismatching observations. Although I do not use these data for model building, the matching observations are useful as benchmark points for calculating the index based on the available data from the Eurostat database. This usability is confirmed by its similarity to the officially reported national-level dataset by UNDP.

Table 1: Indicators of similiarity 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%

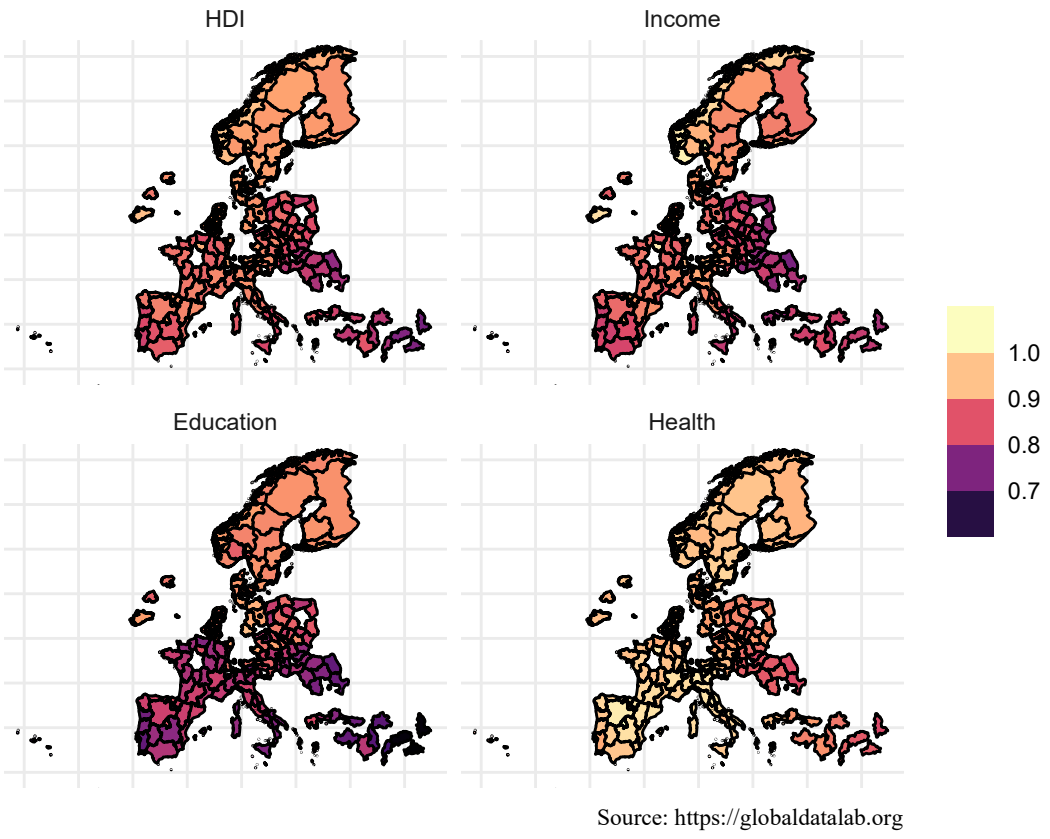


Figure 3: HDI and its components based on the dataset from Global Data Lab (2017)

A decent standard of living

To determine the standard of living dimension of the development, the gross national income (GNI) per capita would be required in PPP terms (constant 2017 PPP$) according to the technical note of UNDP [UNITED NATIONS DEVELOPMENT PROGRAMME, 2020]. To transform it into an index, the income is put into the following equation:

$$\text{Income index } = \frac{\ln(\text{GNI}) - \ln(100)}{\ln(75,000) - \ln(100)} \tag{1}$$

This normalization process keeps the value between 0 (at 100 dollars annual income) and 1 (at 75,000 dollars annual income). Eurostat reports GNI in PPP terms (constant 2020 PPP) in euros for regional observations. To eliminate this difference, I divide the given GNI value by the corresponding annualized value of the EUR/USD exchange rate (similarly downloaded from the Eurostat database). Table 2 figures the similarity to the values reported by Global Data Lab for the matching observations (where the territorial unit corresponds with the one used by Eurostat). In this aspect, I would like to highlight the very high Spearman R^2 between them. It shows that it hardly ever happens that a region has a different rank in the two datasets.

Table 2: Indicators of similiarity 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

The health dimension of HDI is determined based on the life expectancy at birth. The logic of the variable transformation is the same as previously written at the income index:

$$\text{Health index } = \frac{\text{life expectancy at birth } - 20}{85 - 20} \tag{2}$$

The natural minimum of life expectancy is at 20 years, while the maximum is estimated at 85 years, and with the given formula, the range of health index is also set between 0 and 1. Since life expectancy is reported for NUTS 2 levels by Eurostat, this transformation can be performed without any additional calculation.

Knowledge

Determining the education index is more complex compared to the previous ones. The required expected years of schooling are not available at a regional level. However, Eurostat reports several pieces of information on the topic of education. In this study, I substitute the original formula with an index based on the available variables. For this purpose, I use data about the population by educational attainment level and NUTS 2 regions. The available data are also grouped by age. Considering that the discussed decline in fertility rates is mainly related to women under the age of 30 [oec, 2019], statistics about younger generations seem more relevant. To manage this, I only use the age class 20-24 and 25-34.

The educational dimension is based on the International Standard Classification of Education (ISCED 2011). This means the following categories: Less than primary, primary and lower secondary education belongs to

levels 0-2, upper secondary and post-secondary non-tertiary education is labeled with levels 3 and 4 and tertiary education with levels 5-8.

These two dimensions lead to eight possible variables, and the target is to define one index from their values. To manage this dimensionality reduction, I use principal component analysis (PCA). PCA is a useful tool to identify independent sources of variance, and in many cases, economic interpretation can be found in the loadings [Maddala and Lahiri, 1992]. For this step, I impute missing values based on auxiliary regressions using mice R package, and I normalize the variables. Figure 4 shows the result.

The next step is to find the relevant principle component or components. In addition to the interpretability, I also focus on how well this principle describes the education index reported by Global Data Lab. Table 3 contains the similarity indicators.

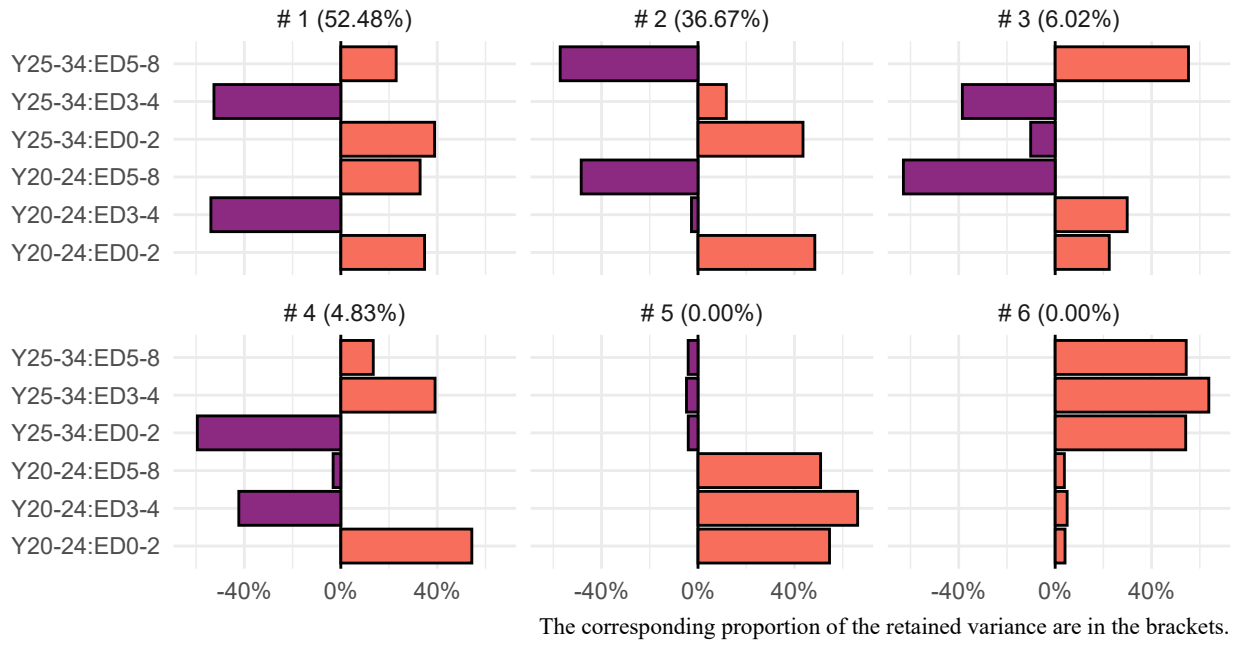


Figure 4: PCAs and the explained variance

The next step is to find the relevant principle component or components. In addition to the interpretability, I also focus on how well this principle describes the education index reported by Global Data Lab. Table 3 contains the similarity indicators.

Table 3: Indicators of similiarity 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
R^2	5.73%	31.18%	17.31%	12.33%	0.00%	0.05%
Spearman R^2	2.72%	26.35%	16.43%	11.23%	6.59%	0.86%

Since the R^2 and Spearman R^2 are high between the second principal component and the education index from Global Data Lab, and the component has good interpretability, I use the second principal component in the following. Based on the loading, its value increases if the proportion of lower educated people increases and decrease if the share of highly educated people increases. This tells the opposite of what the education index means. For this purpose, I determine the education index as the normalized value of the given PCA

score multiplied by minus one.

$$\text{Education index} = \frac{-\text{PCA score} + |\min(-\text{PCA score})|}{\max(-\text{PCA score})} \tag{3}$$

Total fertility rates, family benefit and unemployment statistics

Family benefit data can be found in the Eurostat database. It is reported on the national level, and in this paper, I choose its unit as a percentage of gross national income. The national-level family support corresponds to all regions of a country. The highly unbalanced distribution of family support within countries can lead to incorrect results. However, better statistics are not available at present. In this paper, I systematically refer to this indicator (expenditure on family/children benefits as a percentage of gross national income) as a family benefit or family support.

Youth unemployment rate and fertility statistics are reported for NUTS-2 levels by Eurostat. To perform this empirical analysis, the transformation of these variables is not required.

Explore the data

This section contains an exploratory analysis of the calculated indices and used variables to describe the easily identifiable patterns before the model building. For this purpose, I report distributions, pairwise comparisons and regression trees about the variables.

Pairwise comparison

Figure 5 shows the relationship between human development indices and total fertility rates. All the pairwise linear correlation coefficients are positive and significant, so it is reasonable to assume that spurious correlations exist in this framework. (An index can explain a significant proportion of the variance of TFR, but the indices also explain each other. As a consequence, the correct mechanism has to be identified with a multivariate model).

Similarly, correlations among youth unemployment statistics and fertility are shown in figure 6. Unemployment rates show a negative correlation with TFR, but the pattern here is noisy as well. In contrast, the unemployment statistics seem to move strongly together, foreshadowing the existence of high multicollinearity in our regression.

Unlike the previous relations, the share of family support in gross national income shows a clear pattern with the fertility rate. This is visualized in figure 7. The value of the linear correlation coefficient between the two indicators is 0.51, and one percentage point increase in the support goes with a 0.138 children/woman increase in the TFR. T-statistic is 16.0 in this bivariate regression, which indicates high statistical significance. Since values of family benefit have a wide range in the sample (from 0 to 4.5), the mentioned slope of the regression line shows significance from the economic aspect as well.

Regression tree

A regression tree is a useful statistical tool, where the process behind the regression is the sorting of the observations into as homogenous groups as possible concerning the response variable (total fertility rates in our case) [James et al., 2013]. This model does not have any longitudinal characteristics, and is by far not sufficient to answer the research question[2], but it can visualize and describe much descriptive statistical information with simple interpretations. To find the optimal cut-points, I use the CART algorithm.

The first step is to determine the set of predictors. At this point, it is important to note again that unemployment rates are available only for a critical short period in most cases. That being the case, I decide

[2]Hence I do not focus on tuning the hyperparameter of the models. I set the complexity parameter to return a well-visualisable number of nodes.

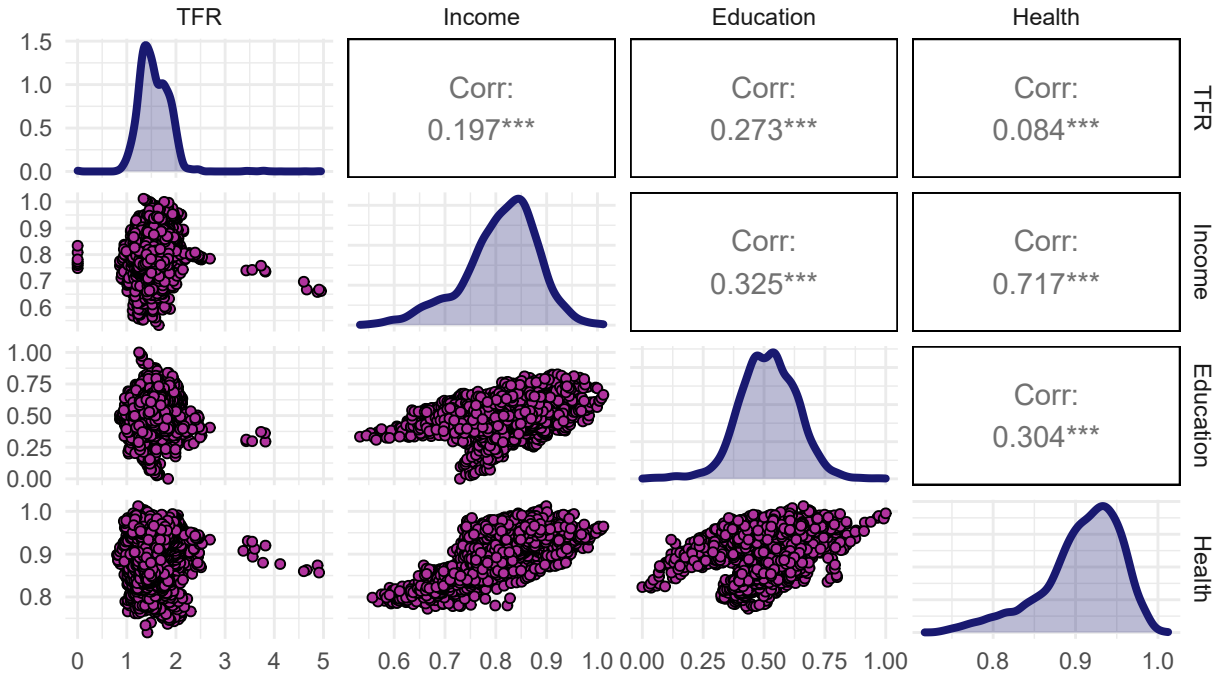


Figure 5: Pairwise correlation among TFR and calculated human development indices

to apply two different frameworks: one without unemployment, and one including it. The model contains each of the observations regardless of the time dimension. The result of the first frame is shown in figure 8.

Including every datapoint leads to 3773 observations (see the top of the tree). Based on the regression tree, the first logical statement made to generate homogenous groups is that if the share of family benefit is below 1.85. This sorts the observations into two subgroups. The TFR in the group where the statement holds (the family benefit is higher) is higher by 0.2 children/woman on average. The second statement is whether the education index is higher than 0.51 (this generates again two similar-sized subgroups), and fertility is higher in the higher educated subgroup. The third pivotal question is also about family benefit. In the subgroup where the family support as a percentage of GDP is higher than 3.4 percentage, the average TFR is 1.8, and 1.5 in the subgroup where this statement does not hold. At this node, the model separates a group of homogeneous datapoints: where the health index is above 0.9, the TFR is extremely high. These few observations come from France.

These lead to the following interpretation: (1) Even if TFR is lower in low educated regions, there are observations with high family benefit rate and high fertility among these regions. (2) Childbearing willingness tends to increase related to a higher share of family support as a percentage of GDP, but under the mentioned circumstances, one can observe extremely high TFR data with modest family benefits. Integrating youth unemployment statics into the modeling framework decreases the number of complete observations significantly. The extended regression tree contains only 1039 data points. On the other hand, variable importance indicators assign youth unemployment as an important explanation to the variance of fertility among the regions. 49 percentage of prediction error reduction is related to the family benefit, but 16 percentage to the unemployment among young people aged 15 to 19.

The effect of youth unemployment is interesting in this regression tree. TFR is higher in the category where the unemployment rate among young people is lower at node 2, but at node 3 the conclusion is the opposite. The cutting-points has two main differences: observations, where the family benefit is below 2.1, belongs to node 2, and that cutting-point is based on the unemployment rate among young people aged 15-19, while node 3 belongs to the data points with family support above 2.1, and the model finds youth unemployment

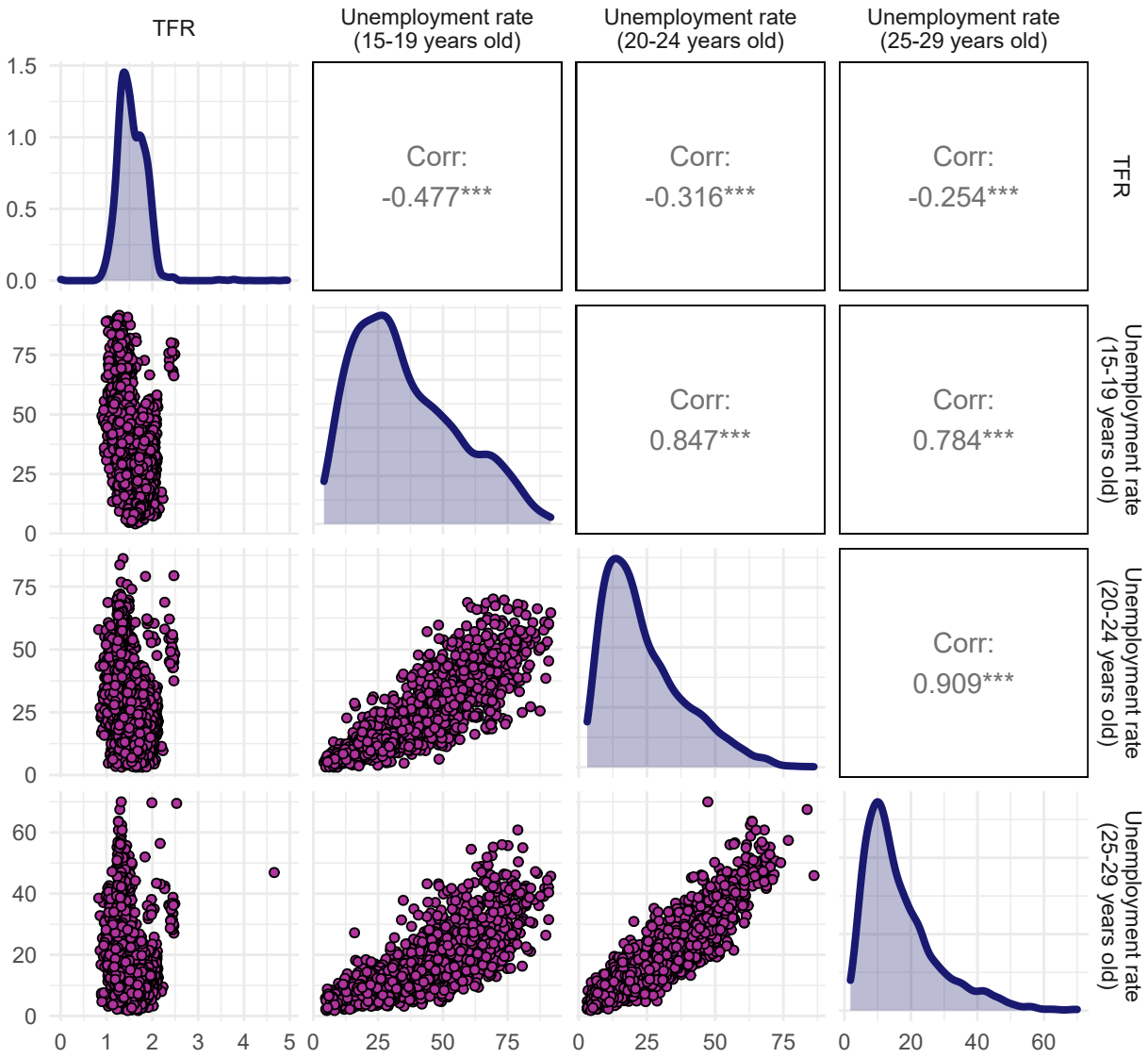


Figure 6: Pairwise correlation among TFR and youth unemployment rates

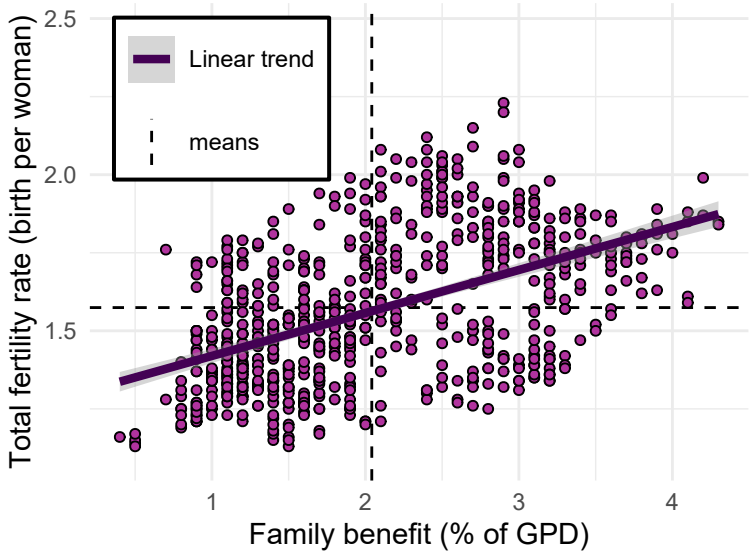


Figure 7: Correlation between TFR and family benefits

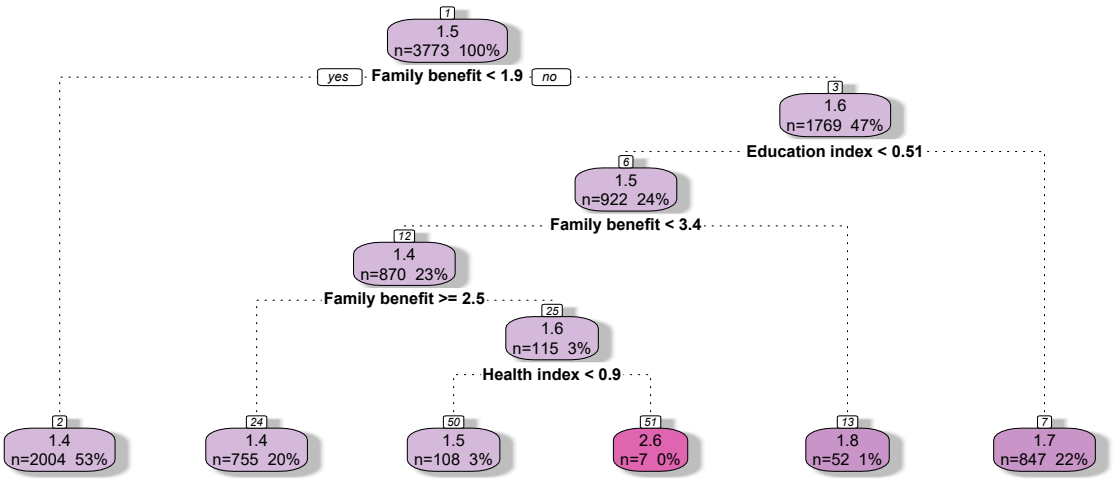


Figure 8: Regression tree explaining the TFR excluding youth unemployment rates (cp = 0.02)

rate among 20-24 years olds more important. Therefore, the difference in the direction of the effect can be explained by the amount of family support or the fact, that the effect-mechanism of unemployment differs in the two age categories.

Based on the literature and theoretical considerations, the first alternative is more plausible. In countries where the family support is higher, young people are probably less sensitive to their working situation, when they decide about childbearing. A high value of family support-to-GDP ratio can eliminate the aforementioned mechanism, that some young parents postpone their childbearing because they cannot afford it. Based on these findings, I extend the design-matrix with the interactions of youth unemployment rates and family benefits.

Model building

Following the findings in the previous chapters, I highlight the main ideas for the model building: (1) Youth unemployment is a significant factor in explaining the variance of fertility, but including it in the model reduces the number of complete observations, so I report one model with the youth unemployment included and one without that. (2) Extending the design matrix with the interactions of youth unemployment rates and fertility is justified. (This extension only concerns the framework which contains unemployment).

Additionally, concerning the nature of fertility (duration of pregnancy) and childbearing decisions (parents may interpolate their expected socio-economic situation from their past) extending the model with the lagged value of the predictors is also reasonable. This raises the issue that the model contains too many predictors and variable selection becomes difficult. I manage variable selection based on lasso selection.

The first step to estimate panel regression models is to identify the appropriate model type. Choosing among the pooled, within and random-effects model requires performing the Chow test and Hausman test. The null hypothesis in the former one is that whether a significant difference among the individual intercepts exists, and the test is performed based on an F-test. If the given H_0 cannot be rejected at any standard significance level, estimating a pooled model is suggested. In other cases, the outcome depends on the result of the Hausman test. The Hausman test (or Durbin-Wu-Hausman test) is more complex from mathematical aspects, but the interpretation is simple: if the given H_0 cannot be rejected, the within (fixed-effects) model is not as efficient as the random-effects, and estimating random-effects model is suggested. If the null hypothesis is rejected at any standard significance level, the fixed-effects model will be suitable [Wooldridge, 2016].

The optimal model may differ if the set of predictors change, but as a starting point, I estimate the one-one simplest model for the two mentioned model frameworks (including unemployment statistics or not). These models do not contain interaction but lagged effect and quadratic terms are included. The performed tests show that the fixed-effects model is optimal (see the results in Table 5), and the estimated coefficients from the fixed-effects models are presented in figure 10.

Table 4: Models

Indicator	Model I.	Model II.
Pooltest	0.00%	0.00%
Phtest	0.00%	0.00%
Adjusted R^2	20.19%	29.24%
Observations	699	3257

In the following, I perform lasso variable selection on these model frameworks to find the most relevant effects.

Framework I: with unemployment

As repeatedly mentioned in the previous section, including unemployment statistics in the regression analysis drastically reduces the number of complete observations. Moreover, there is a good reason to believe that

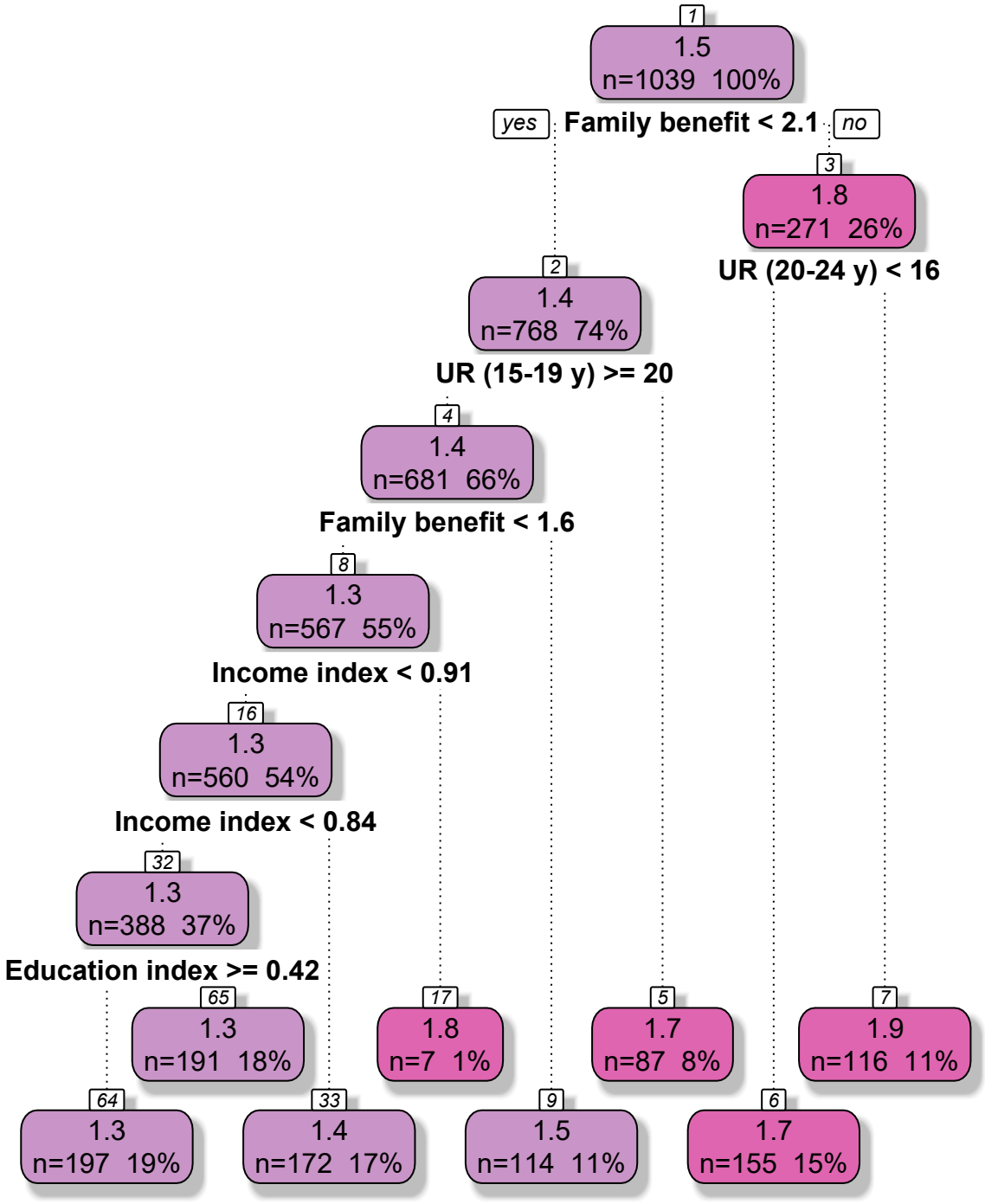


Figure 9: Regression tree explaining the TFR using all the mentioned explanatory variables (cp = 0.01)

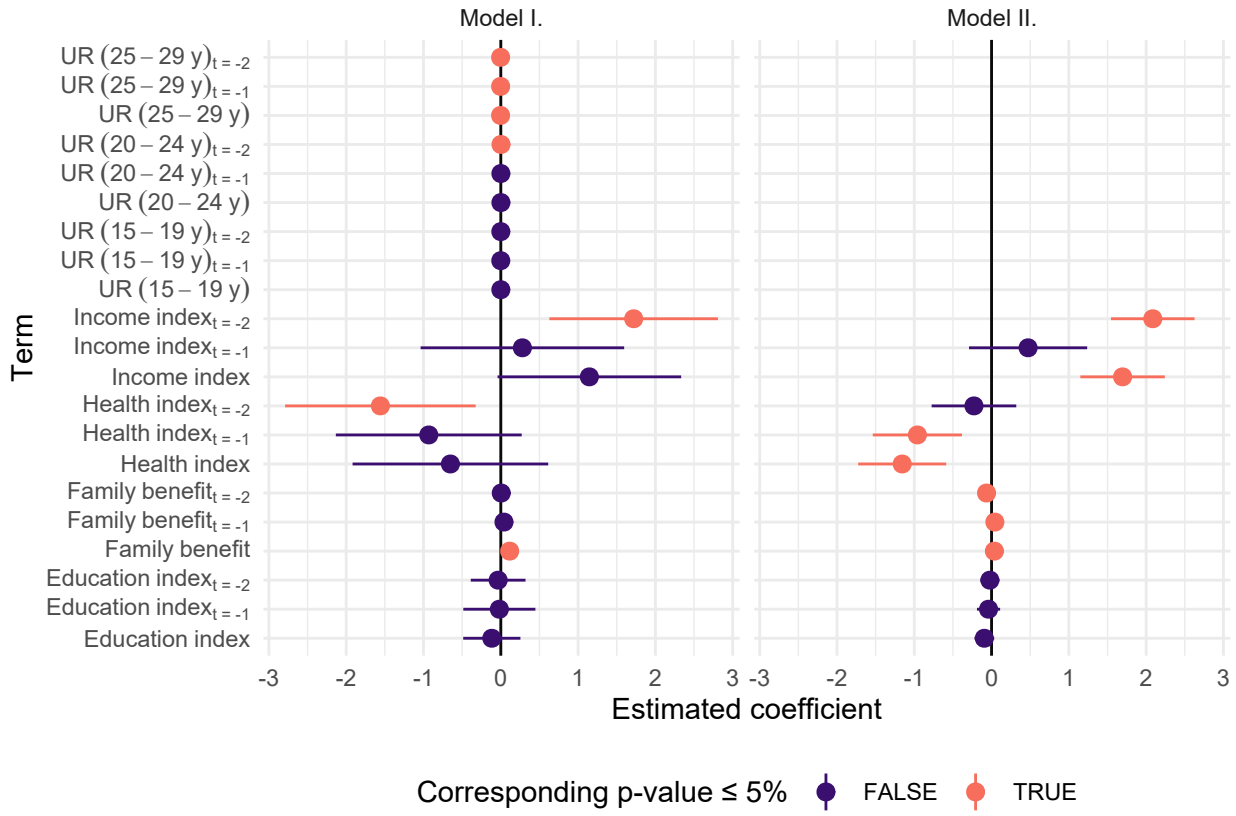


Figure 10: Panel models on the total fertility rates

this selection method is not random, and omitting the incomplete data points may cause bias in the results. Figure 10 shows the number of complete observations by territorial units related to the two frameworks.

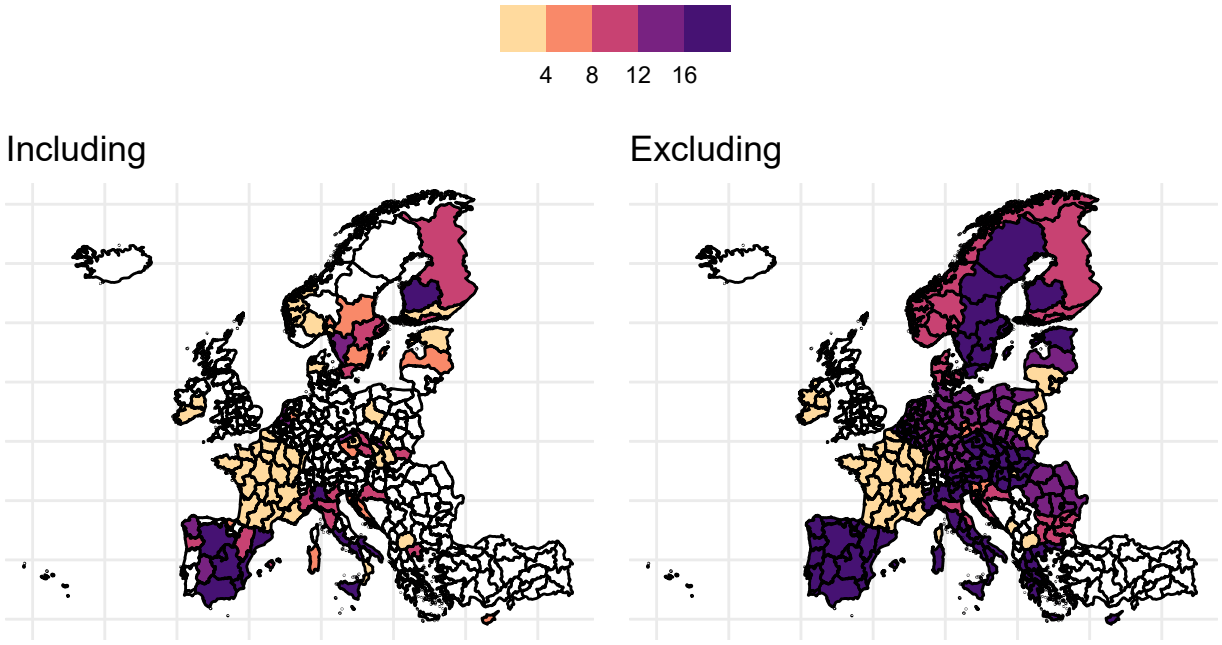


Figure 11: Number of complete observations by countries when the model includes or excludes unemployment statistics

Figure 10 reveals that there are unbalances among the regions concerning their representation in the dataset. A prominent amount of incomplete and thus unusable observations concerns the Central European region and the UK. In this paper, I do not impute these missing values, and therefore, the limited generalizability of the results is taken into account.

To measure the bias, I calculate the mean of the given variables in the total sample (including incomplete observations), and the sample that is used in this framework (excluding incomplete observations). The result is reported in table 5. The table describes that the health index and income index are higher, while the family benefit is lower in the used sample compared to the dataset containing the incomplete observations as well.

Table 5: Comparison of average values of the variables for incomplete and complete observations (Framework I)

Variable	Mean in total sample	Mean in used sample	Number of observations in the total sample
TFR	1.5575	1.4801	7249
Education index	0.5152	0.5088	5373
Health index	0.9089	0.9375	6983
Income index	0.8120	0.8298	4291
Family benefit	2.0759	1.6302	6155
UR (15-19 y)	36.6883	44.4313	1713
UR (20-24 y)	24.3994	25.5202	2935
UR (25-29 y)	17.0107	17.0332	2664

Methodology of model estimating

As described above, an optimal statistical tool for a high-dimensional dataset is the lasso regression. From mathematical aspect lasso regression means to add a $\lambda \sum_{j=1}^{p} |\beta_j|$ term to the target function of regression [James et al., 2013]. Intuitively, with this additional term, the value of the target function is lower (which has to be minimized), if more parameters are equal to zero, but the prediction error does not decrease significantly.

Lasso regression has a hyperparameter λ. Setting its value to zero leads to the unmodified OLS model. In contrast, a lasso regression with $\lambda = 1$ would lead to an empty model. Finding the optimal λ hyperparameter requires estimating the model with different parameters. In each case, the model contains a different number of variables. To determine the optimal value of λ, leave-one-out cross-validation is performed with 10 folds, then the model having the lowest mean squared error on the validation set is chosen. This process is visualized in figure 11.

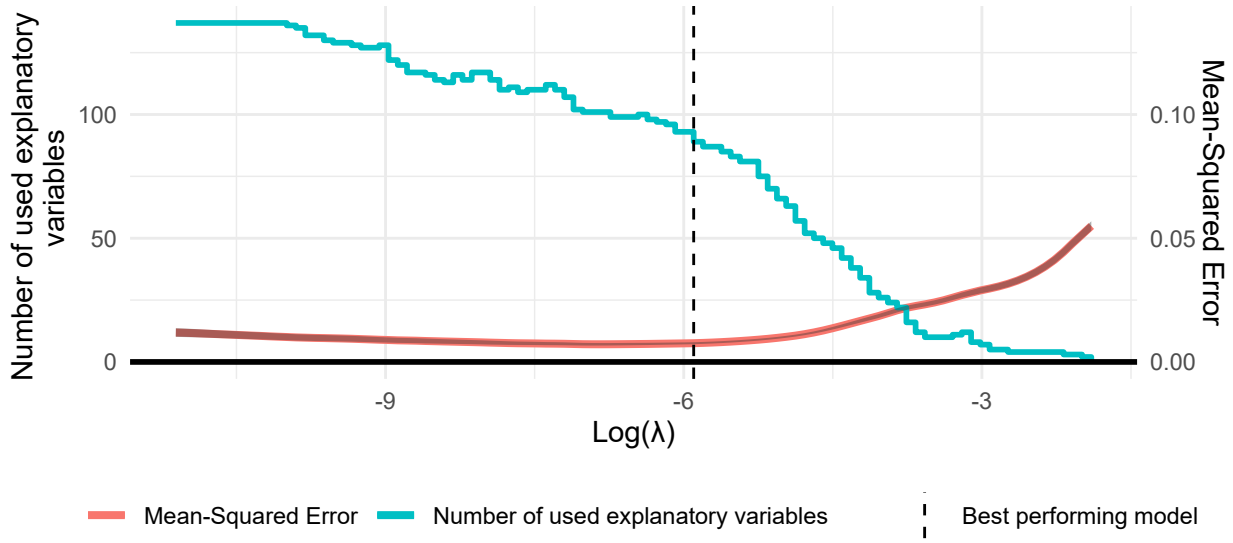


Figure 12: Performance of lasso regression models with different parameters

It is important to note that the above-described algorithm eliminated the insignificant individual intercept terms from the model. To adjust this property, I reestimate the within model including all the individual intercept terms and predictor variables from the best-performing lasso regression model.

In addition, the difference in the measurement of the variables causes complexity in interpretation. Interpreting the direct effect of a predictor is simple, but the coefficients are not sufficient to describe the importance of the variables, because they are measured on different scales (one percentage point change in the family benefit-to-GDP ratio would be extreme, but not as outstanding as one percentage point change in the youth unemployment rate). To manage this, I reestimate the model with standardized variables. The benefit of this model is that the explained variance of the regression model can be decomposed with it:

$$\mathrm{R}^2 = \sum_{j=1}^{p} r_j \times \beta_{standardized,j} \tag{4}$$

Based on equation 4, the coefficient multiplied by the linear correlation coefficient (r) can be interpreted as the contribution to the explained variance. The result of this and the above-mentioned computations are reported in figure 12.

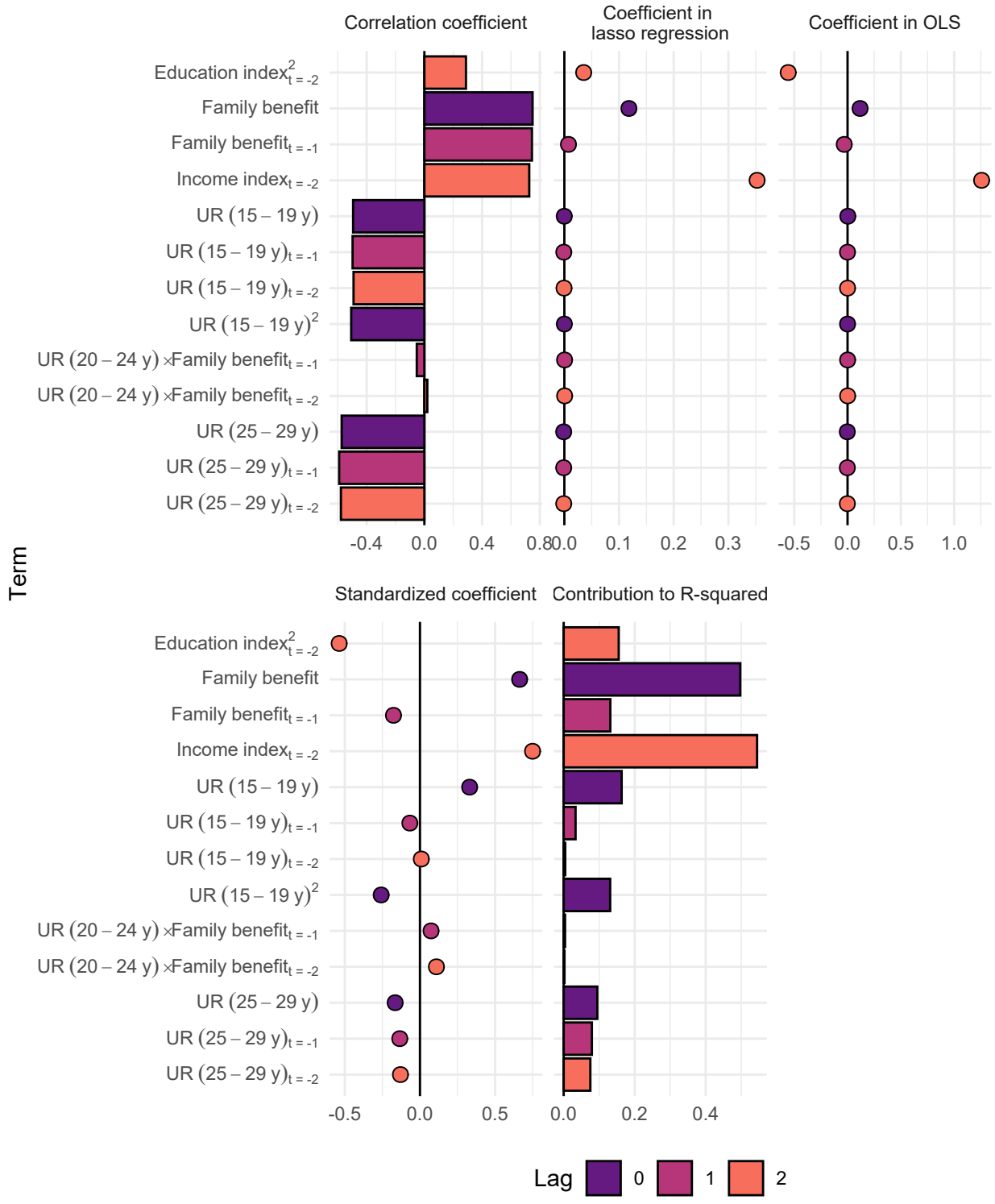


Figure 13: Estimated coefficient of the fixed panel model controlling for youth unemployment indicators

Interpreting the results

The high contribution of income index and family benefit to the explained variance revealed by figure 12. Both of them seem to have a positive effect on fertility. Increasing income per capita leads to higher total fertility rate based on the model parameters (positive coefficient correspond to each lagged variable). In contrast, coefficients related to the different lagged values of family support indicate a complex mechanism: family support has a high instantaneous effect on fertility, but the lagged negative effect implies that this birth-surplus disappears in the following years.

Based on the results the answer to our research question is that in the developed world (1) *income is far the most important component of human development influencing fertility (highest contribution to* R^2) and (2) *family support also has a significant instantaneous effect on childbearing willingness, but it seems weaker on the long run.*

Youth unemployment rates among 25-29 year-olds have a negative effect on fertility, but its total effect is lagged. Among 15-19 year-olds this effect is different. In their case, the increase in unemployment causes an instantaneous increase in fertility. But in the case of a permanent increase in unemployment, this increase disappears (ceteris paribus). Extending the model with the interactions of youth unemployment and family benefit was truly beneficial comparing the standardized effect of the unemployment rates and the interactions.

Education index is also detected as a significant explanation of the variance of fertility, but its interpretation is more complex. The lagged quadratic terms are represented in the model with negative coefficients. This reflects that the higher education index leads to lower fertility, and this effect is stronger in the case of those regions, where the education index is higher. This confirms the former findings in the literature that highly educated women tend to have less children [Martin, 1995], but recent research found empirical evidence, that "higher educated German women, who already decided to have a child despite their high opportunity costs are more family oriented" [König, 2011].

Framework II: without unemployment

I continue my study reporting the results from the model excludes unemployment statistics. This framework omits significantly fewer data points, so the probability of contra selection-caused bias is reduced. Table 6 describes the average difference comparing the used sample and the values from the incomplete observations.

The methodology of the model estimation is equivalent to the one described in the first model framework. The results are presented in figure 15. The H_0 of the Chow test ($p = 0.00$) and the Hausman test ($p = 0.00$) are rejected, so fixed effect model is adequate. The R^2 of this model is 14.85%[3].

Table 6: Comparison of average values of the variables for incomplete and complete observations (Framework II)

Variable	Mean in total sample	Mean in used sample	Number of observations in total sample
TFR	1.5575	1.4935	7249
Education index	0.5152	0.4971	5373
Health index	0.9089	0.9202	6983
Income index	0.8120	0.8228	4291
Family benefit	2.0759	1.9699	6155

Interpreting the results

Figure 13 shows that the outstandingly high contribution to the R^2 of the income index did not change, so the answer for my first research question is robust to the framework: *income index explains significantly more*

[3]The heterogenity of the contries containing complete observation is higher in this setup, that is comparing the R^2 of the two frameworks is not suggested.

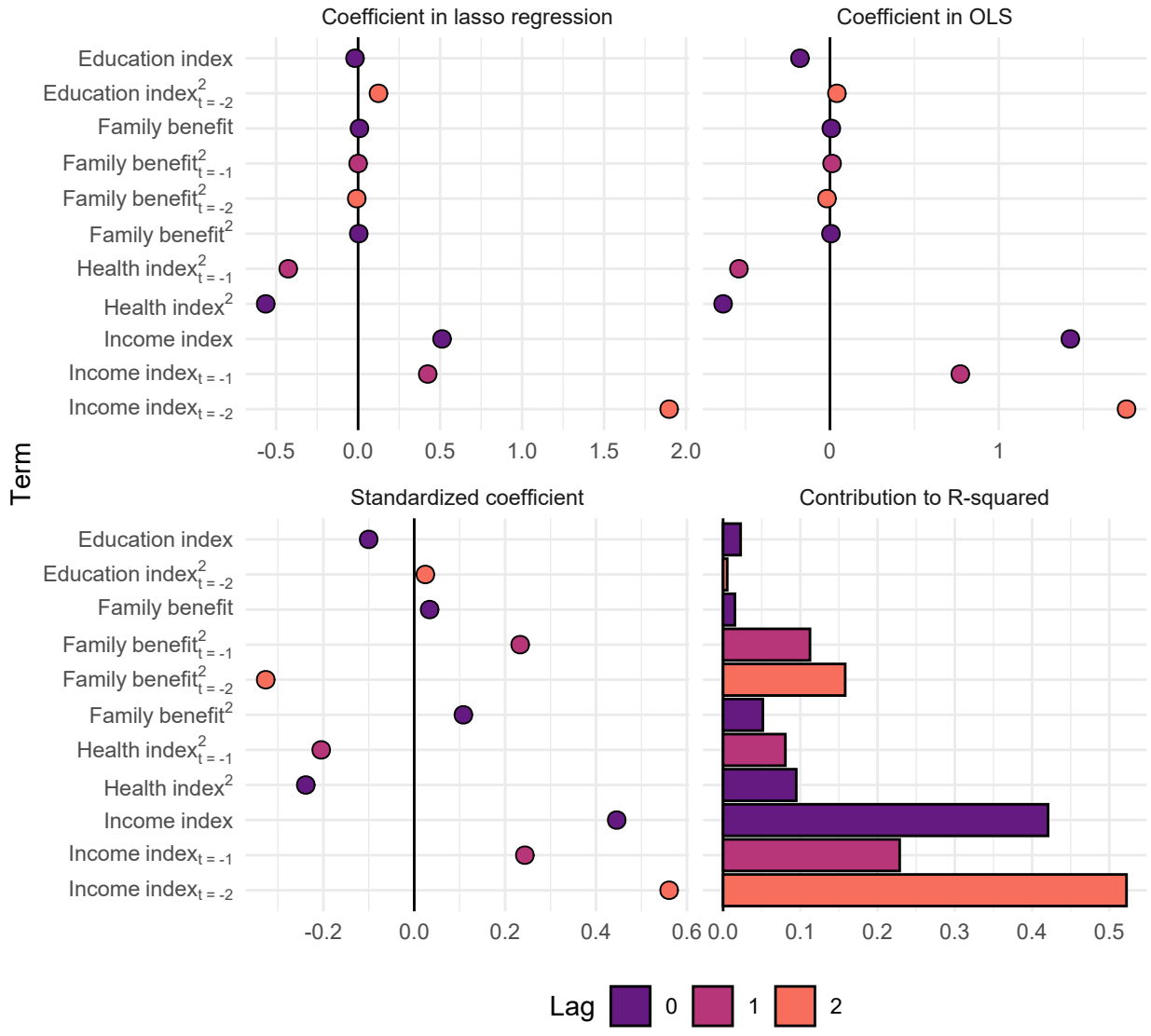


Figure 14: Estimated coefficient of the fixed panel model omitting youth unemployment indicators

of the variance of the fertility than the other components of human development, and it has a positive effect on childbearing willingness.

The estimated effect of family benefit differs in this framework compared to the one including unemployment statistics (and excluding observations where youth unemployment is not available). The new observations come from Central-European regions and the estimated structure of the effect of family benefit became significantly different. The reversal effect in this framework is close to the instantaneous effect of the family spending. This leads to the interpretation that family support has only an instantaneous impact on fertility, but on the long run it can not significantly influence the fertility.

The main difference between the results reported by the two frameworks is that the health index is signed as a significant variable in the second one. Since all of its transformed terms has a negative coefficient, a higher health index leads to lower fertility. The frequently mentioned reason for this is the changing lifestyle and women in the EU are having their first child later. One possible explanation why health index was not relevant in the previous framework is that the share of observations from Central-European countries are much higher in this model (due to the lack of unemployment statistics from the early 2000s). Extension of average childbearing age led to failing fertility rates in this area [Berde and Németh, 2014]. But many articles suggest that an adjusted TFR should be considered [Bongaarts and Feeney, 1998], because the drastically low fertility during the time of this mechanism. However, these indices are currently not available for regional datasets, but this could be a possible further research direction.

Conclusion

This paper focused on the effect of two key factors on fertility rates in Europe: human development and expenditures on family support. To quantify the relationships, econometric modeling on regional dataset were performed. This requires to generate or estimate the human development indices for regional level. From statistical aspect the key issue is the trade of between limited complete observations or omitting possibly important explanatory variables. This is due to the lack of data about youth unemployment. To manage this a I set up two model frameworks: one including unemployment statistics and one excluding it. Based on theoretical considerations and the previously performed analysis the design-matrix was extended with lagged variables (duration of pregnancy), quadratic terms (Theory of "J-shaped" effect) and interactions between youth unemployment and family benefit (based on the previously performed regression tree). This lead to a high-diemansonal dataset, so lasso-based feature selection was the bases of the reported longitudinal econometric models.

As a result, the paper confirms the *empirical evidence that increasing human development in developed countries has a positive effect on total fertility rates, and income is the most important component.* This finding is robust to the framework. In contrast, the research come up *only with week evidence for the significant effect of expenditure on family on total fertility rates.*

References

[eco, 2009] (2009). The best of all possible worlds? page 1.

[oec, 2018] (2018). *SF2.1 Fertility rates*. OECD.

[oec, 2019] (2019). *Society at a Glance 2019*. OECD.

[Adser, 2004] Adser, A. (2004). Changing fertility rates in developed countries. the impact of labor market institutions. *Journal of Population Economics*, 17(1):17–43.

[Adsera, 2004] Adsera, A. (2004). Changing fertility rates in developed countries. the impact of labor market institutions. *Journal of population economics*, 17(1):17–43.

[Berde and Németh, 2014] Berde, É. and Németh, P. (2014). Az alacsony magyarországi termékenység új megközelítésben. *Statisztikai Szemle*, 92(3):253–274.

[Bongaarts and Feeney, 1998] Bongaarts, J. and Feeney, G. (1998). On the quantum and tempo of fertility. *Population and development review*, pages 271–291.

[Frejka et al., 2016] Frejka, T., Gietel-Basten, S., Abolina, L., Abuladze, L., Aksyonova, S., Akrap, A., Antipova, E., Bobic, M., Čipin, I., Fakeyeva, L., et al. (2016). Fertility and family policies in central and eastern europe after 1990. *Comparative Population Studies*.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

[König, 2011] König, S. (2011). Higher order births in germany and hungary: Comparing fertility intentions in a national context. *Sozialforschung MZES Mannheim*, pages 1–18.

[Lahti et al., 2017] Lahti, L., Huovari, J., Kainu, M., and Biecek, P. (2017). eurostat r package. Version 3.6.84.

[Maddala and Lahiri, 1992] Maddala, G. S. and Lahiri, K. (1992). *Introduction to econometrics*, volume 2. Macmillan New York.

[Martin, 1995] Martin, T. C. (1995). Women's education and fertility: results from 26 demographic and health surveys. *Studies in family planning*, pages 187–202.

[Matysiak et al., 2020] Matysiak, A., Sobotka, T., and Vignoli, D. (2020). The great recession and fertility in europe: A sub-national analysis. *European Journal of Population*, 37(1):29–64.

[Myrskylä et al., 2009] Myrskylä, M., Kohler, H.-P., and Billari, F. C. (2009). Advances in development reverse fertility declines. *Nature*, 460(7256):741–743.

[Sagar and Najam, 1998] Sagar, A. D. and Najam, A. (1998). The human development index: a critical review. *Ecological economics*, 25(3):249–264.

[UNITED NATIONS DEVELOPMENT PROGRAMME, 2020] UNITED NATIONS DEVELOPMENT PROGRAMME (2020). Technical notes: Calculating the human development indices.

[Wooldridge, 2016] Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

Appendix: R codes

```r
# Set up ------------------------------------------------------------------------------

## Packages ==========================================================================

library(tidyverse)
library(patchwork)
library(knitr)
library(broom)
library(eurostat)

## Gg theme ==========================================================================

update_geom_defaults("point", list(fill = "#B1339E",
                                   shape = 21,
                                   color = "black",
                                   size = 1.4))
update_geom_defaults("line",
                     list(color = "midnightblue", size = 1.4))

update_geom_defaults("smooth", list(color = "red4", size = 1.4))

update_geom_defaults("density",
                     list(color = "midnightblue", fill =  "midnightblue", alpha = .3,
                          size = 1.4))

extrafont::loadfonts(device="win")

theme_set(theme_minimal() + theme(
  legend.direction = "vertical",
  # text = element_text(family = "Impact"),
  plot.caption = element_text(family = "serif"),
  legend.key=element_blank()
))

# https://data.worldbank.org/indicator/SP.DYN.TFRT.IN

WB_fertility <- read_csv("WB_fertility.csv", skip = 4)

# https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.KD

WB_GDP <- read_csv("WB_GDP.csv", skip = 4)

merge(WB_fertility %>%
        select('Country Name', '2017') %>%
        rename(tfr = '2017'),
      WB_GDP %>%
        select('Country Name', '2017') %>%
        rename(GDP = '2017')) %>%
  ggplot(aes(GDP, tfr)) + geom_point() +
  ggformula::geom_spline() +
  scale_x_continuous(limits = c(0, 7e+4)) +
  labs(y = "Total fertility rate (birth per woman)",
```

```
       x = "GDP per capita, PPP (constant 2017 international USD)",
       caption = "Own editing based on the Figure 5-2. from Kreiszné Hudák (2019).
       The trend is drawn via splines.
       Source of the data: World Bank."
  )

# Global Data Lab ============================================================================

### Nation-level data ##########################################################################

GDL_nat <- read_csv("GDL-Sub-national-HDI-data.csv") %>%
  filter(Level == "National") %>%
  select(Country, 6:34) %>% pivot_longer(-1, names_to = "time", values_to = "values") %>%
  mutate(time = as.numeric(time)) %>%
  na.omit()

### Data from UNDP #############################################################################

# source: http://hdr.undp.org/

HDI_UNDP <- read_csv("Human Development Index (HDI).csv",
                     skip = 5) %>%
  select(!starts_with("X"), - 'HDI Rank') %>%
  mutate_at(-1,
            function(x) {as.numeric(ifelse(x == "..", NA, x))}) %>%
  pivot_longer(-1, names_to = "time", values_to = "values") %>%
  mutate(time = as.numeric(time)) %>%
  na.omit()

### Comparison ---> Table 1 ####################################################################

merge(GDL_nat %>% rename(GDL = values),
      HDI_UNDP %>% rename(UNDP = values)) %>%
  {
    c(
      scales::percent(cor(x = .$GDL, y = .$UNDP)^2, accuracy = .01),
      scales::percent(cor(x = .$GDL, y = .$UNDP, method = "spearman")^2, accuracy = .01),
      as.character(format(mean(abs(.$GDL - .$UNDP)), digits = 1)),
      scales::percent(mean(abs(.$GDL - .$UNDP)/.$UNDP), accuracy = .01)
    )
  } %>%
  {tibble(
    Indicator = c("$R^2$", "Spearman $R^2$",
                  "Mean absolute deviation", "Mean absolute percentage deviation"),
    Value = .
  )} %>%
  kable(
    caption = "Indicators of similiarity between the Human Development Indices
    provided by UNDP and GDL"
  )

### Sub-national data ##########################################################################

```

```r
# source of csv files: https://globaldatalab.org/

GDL_import <- function(x) {
  get_eurostat_geospatial(nuts_level = 2) %>%
    data.frame() %>%
    tibble %>%
    mutate(
      ISO_Code = countrycode::countrycode(CNTR_CODE, origin = "iso2c", "iso3c"),
      ISO_Code = ifelse(CNTR_CODE == "UK", "GBR", ISO_Code),
      ISO_Code = ifelse(CNTR_CODE == "EL", "GRC", ISO_Code),
    ) %>%
    select(ISO_Code, NUTS_NAME, geo) %>%
    merge(read_csv(x), by = "ISO_Code") %>%
    mutate(
      z = stringdist::stringsim(NUTS_NAME, Region)
    ) %>%
    arrange(desc(z)) %>%
    filter(!duplicated(Region)) %>%
    filter(!duplicated(NUTS_NAME)) %>%
    filter((z > .5 | Country %in% c("Greece", "Turkey", 'Romania',
                                    'Malta', 'Italy')) & NUTS_NAME != "Dresden") %>%
    select(geo, '1990':'2018') %>%
    pivot_longer(-1, names_to = "time", values_to = "values") %>%
    mutate(time = as.numeric(time))
}

GDL_subnat <- GDL_import("GDL-Sub-national-HDI-data.csv") %>% rename(HDI = values) %>%
  merge(
    GDL_import("GDL-Educational-index--data.csv") %>% rename(education = values)
  ) %>%
  merge(
    GDL_import("GDL-Health-index-data.csv") %>% rename(health = values)
  ) %>%
  merge(
    GDL_import("GDL-Income-index-data.csv") %>% rename(income = values)
  )

plot_NUTS2 <- function(df, viridis_c = T, ..., all.x = F) {
  p <- df %>%
    {merge(eurostat::get_eurostat_geospatial(nuts_level = 2), ., all.x = all.x)} %>%
    ggplot(aes(fill = values)) +
    geom_sf(color = "black") +
    theme(
      axis.text = element_blank()
    ) +
    xlim(c(-30, 44)) +
    ylim(c(35, 70)) +
    labs(fill = NULL)

  if (viridis_c) {
    p <- p + scale_fill_viridis_c(option = "magma", ...,
                                  guide = guide_colorbar(frame.colour = "black",
                                                         ticks.colour = "black"),
```

```r
                                    na.value = "white")
  }
  p
}

GDL_subnat %>%
  filter(time == 2017) %>%
  select(-time) %>%
  pivot_longer(-1, names_to = "var", values_to = "values") %>%
  filter(!is.na(var)) %>%
  mutate(
    var = str_to_title(var),
    var = str_replace(var, "Hdi", "HDI"),
    var = factor(var,
                 levels = c("HDI", "Income", "Education", "Health"),
                 ordered = T)
  ) %>%
  plot_NUTS2() + facet_wrap(~ var, ncol = 2) +
  labs(caption = "Source: https://globaldatalab.org") +
  scale_fill_viridis_c(guide = guide_colorsteps(), option = 'magma')

GDP_index <- get_eurostat("nama_10r_2gdp", time_format = "num") %>%
  filter(unit == "PPS_EU27_2020_HAB") %>%  # Purchasing power standard (PPS) per inhabitant
  select(-unit) %>%
  rename(GDP = values) %>% merge(
    get_eurostat("ert_bil_eur_a", time_format = "num") %>% # EUR/USD annual avg exc r
      filter(currency == "USD" & statinfo == "AVG") %>%
      select(time, e = values)
  ) %>%
  {
    GDP <- .$GDP/.$e # mutate to USD
    mutate(.,
           GDPindex = (log(GDP) - log(100))/
             (log(75000) - log(100))
    )
  }

merge(GDL_subnat, GDP_index) %>%
  {
    c(
      scales::percent(cor(x = .$income, y = .$GDPindex)^2, accuracy = .01),
      scales::percent(cor(x = .$income, y = .$GDPindex, method = "spearman")^2,
                      accuracy = .01),
      as.character(format(mean(abs(.$income - .$GDPindex)), digits = 1, nsmall = 4)),
      scales::percent(mean(abs(.$income - .$GDPindex)/.$GDPindex), accuracy = .01)
    )
  } %>%
  {tibble(
    Indicator = c("$R^2$", "Spearman $R^2$",
                  "Mean absolute deviation", "Mean absolute percentage deviation"),
    Value = .
  )} %>%
  kable(
```

```
    caption = "Indicators of similiarity between the income component of the
    Human Development Indices provided by GDL and the estimation based on regional GDP"
  )

health_index <- get_eurostat("demo_r_mlifexp", time_format = "num") %>%
  filter(age == "Y_LT1" & sex == "T") %>%
  select(geo, time, le = values) %>%
  mutate(
    health_index = (le - 20) / (85 - 20)
  )

edu_wide <- get_eurostat("edat_lfse_04", time_format = "num") %>%
  filter(sex == "T" & !str_detect(isced11, "GEN") &
           !str_detect(isced11, "VOC") & isced11 != "ED3-8" & age != 'Y25-64'&
           age != 'Y30-34' & !str_detect(geo, "TR")
  ) %>%
  mutate(
    var = str_c(age, ":", isced11)
  ) %>%
  select(geo, time, var, values) %>%
  pivot_wider(names_from = var, values_from = values) %>%
  {
    x <- .
    names(x) <- letters[1:length(x)]
    x <- cbind(
      .[, 1:2],
      mice::complete(mice::mice(select(x, -a,-b), printFlag = F))
    )
    names(x) <- names(.)
    x
  }

edu_comps <- edu_wide%>%
  select(-time, -geo) %>%
  na.omit() %>%
  {princomp(scale(.))}

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

edu_comps %>%
  .$loadings %>%
  unclass() %>%
  data.frame() %>%
  rownames_to_column() %>%
  pivot_longer(-1) %>%
  mutate(
    name = as.numeric(str_remove(name, 'Comp.')),
  ) %>%
  arrange(name) %>%
```

```r
  mutate(
    name = edu_comps %>%
      summary() %>%
      {.$sdev^2/sum(.$sdev^2)} %>%
      scales::percent(accuracy = .01) %>%
      {str_c("# ", 1:length(.), " (", ., ")")} %>%
      .[name],
    rowname = str_replace_all(rowname, '_', '-')
  ) %>%
  ggplot +
  aes(rowname, value, fill = value < 0) +
  geom_hline(yintercept = 0) +
  geom_col(color = 'black') +
  coord_flip() +
  scale_fill_viridis_d(guide = F, option = "magma", begin = .4,
                       end= .7, direction = -1) +
  scale_y_continuous(labels = scales::percent) +
  facet_wrap(~name, ncol = 3) +
  labs(x = NULL, y = NULL, caption =
         "The corresponding proportion of the retained variance are in the brackets.")

edu_comps %>% .$scores %>%
  cbind(edu_wide) %>% merge(GDL_subnat) %>%
  select(starts_with('Comp'), education) %>%
  {
    x <- .$education
    apply(select(., -education), 2, function(y) {
      c(
        scales::percent(cor(x = x, y = y)^2, accuracy = .01),
        scales::percent(cor(x = x, y = y, method = "spearman")^2, accuracy = .01)
      )
    })
  } %>%
  data.frame() %>%
  mutate(Indicator = c("$R^2$", "Spearman $R^2$")) %>%
  rename_all(function(x) str_replace(x, "p.", "p ")) %>%
  select(Indicator, everything()) %>%
  kable(
    caption = "Indicators of similiarity between the knowledge component of Human
    Development Indices provided by UNDP and the calculated principal components using
    educational attainment level"
  )

edu_index <- edu_comps %>%
  .$scores %>%
  data.frame() %>%
  select(2) %>%
  cbind(edu_wide) %>%
  select(geo, time, edu_index = Comp.2) %>%
  mutate(
    edu_index = -edu_index,
    edu_index = edu_index + abs(min(edu_index)),
    edu_index = edu_index/max(edu_index)
```

```r
  )

# Variables without transformation ====================================================

f_data <- get_eurostat("demo_r_find2", time_format = "num") %>% # TFR
  select(geo, time, var = indic_de, values) %>%
  filter(!str_detect(geo, "TR"))

FAM_df <- get_eurostat("spr_exp_sum", time_format = "num") %>% # Family benefit
  filter(spdeps == "FAM" & unit == "PC_GDP") %>%
  rename(FAM = values, country = geo) %>%
  select(-(spdeps:unit))

yth_empl_byage <- get_eurostat("yth_empl_110", time_format = "num") %>%
  # youth unemployment
  filter(unit == "PC" & sex == "F") %>%
  filter(age %in% c("Y15-19", "Y20-24", "Y25-29")) %>%
  select(-unit, -sex) %>%
  pivot_wider(names_from = age, values_from = values) %>%
  rename(
    "uY15"  = "Y15-19",
    "uY20"  = "Y20-24",
    "uY25"  = "Y25-29"
  )

# Merging the data.frames ============================================================

dat <- f_data %>%
  pivot_wider(names_from = var, values_from = values) %>%
  merge(edu_index, all = T) %>%
  merge(health_index, all = T) %>%
  merge(GDP_index, all = T) %>%
  filter(!str_detect(geo, "TR"))%>%
  mutate(country = str_sub(geo, end = 2)) %>%
  merge(FAM_df, all.x = T, all.y = F) %>%
  merge(yth_empl_byage, all.x = T, all.y = F) %>%
  filter(!str_detect(geo, "TR"))

f.clean_names <- function(v, Tosparse = F) {
  v <- str_replace_all(v, "GDPindex", "Income index") %>%
    str_replace_all("health_index", "Health index") %>%
    str_replace_all("edu_index", "Education index") %>%
    str_replace_all("TOTFERRT", "TFR") %>%
    str_replace_all("GDPindex", "Income index") %>%
    str_replace_all("FAM", "Family benefit") %>%
    str_replace_all("uY15", "UR (15-19 y)") %>%
    str_replace_all("uY20", "UR (20-24 y)") %>%
    str_replace_all("uY25", "UR (25-29 y)") %>%
    str_replace_all("Fu15", "UR (15-19 y)*Family benefit") %>%
    str_replace_all("Fu20", "UR (20-24 y)*Family benefit") %>%
    str_replace_all("Fu25", "UR (25-29 y)*Family benefit")

  if(Tosparse) v <- str_replace_all(v, " ", "~") %>%
```

```r
      str_replace_all('\\*', '%\\*%')
  v
}


# Explore the data ------------------------------------------------------------

# Pairwise correlations ========================================================

dat %>%
  filter(str_length(geo) == 4) %>%
  select(TOTFERRT, GDPindex, edu_index, health_index) %>%
  {set_names(., f.clean_names(names(.)))} %>%
  rename_all(.funs = function(x) str_remove_all(x, ' index')) %>%
  GGally::ggpairs()

dat %>%
  filter(str_length(geo) == 4) %>%
  select(TOTFERRT, uY15, uY20, uY25) %>%
  set_names("TFR", "Unemployment rate\n(15-19 years old)",
            "Unemployment rate\n(20-24 years old)",
            "Unemployment rate\n(25-29 years old)") %>%
  GGally::ggpairs()

dat %>%
  filter(str_length(geo) == 2) %>%
{ggplot(., aes(FAM, TOTFERRT)) +
  geom_vline(aes(xintercept = mean(.$FAM, na.rm = T), linetype = "means")) +
  geom_hline(yintercept = mean(.$TOTFERRT, na.rm = T), linetype = 2) +
  geom_point() +
  geom_smooth(method = "lm", aes(color = "Linear trend"), size = 1.5) +
  scale_color_viridis_d() +
  scale_linetype_manual(values = c("means" = 2)) +
  labs(y = "Total fertility rate (birth per woman)",
       x = "Family benefit (% of GPD)",
       linetype = NULL, color = NULL
  ) +
    theme(
      legend.position = c(.2, .8),
      legend.box.background = element_rect(
        colour = "black",
        size = .7,
        fill = "white"
      ),
      legend.spacing.y = unit(0.05, 'cm')
    )
}

dat %>%
  filter(str_length(geo) == 2) %>%
  select(FAM, TOTFERRT) %>%
  na.omit() %>%
  {print(cor.test(.$FAM, .$TOTFERRT)); broom::tidy(lm(TOTFERRT ~ FAM, data = .))}
```

```r

# Regression trees ============================================================================

m_part <- dat %>%
  filter(str_length(geo) == 4) %>%
  select(TOTFERRT, GDPindex, edu_index, health_index, FAM) %>%
  {set_names(., f.clean_names(names(.)))} %>%
  na.omit() %>%
  rpart::rpart(formula = TFR ~ ., cp = .02)

rattle::fancyRpartPlot(m_part, palettes = 'PuRd', sub = NULL)

summary(m_part)


m_part2 <- dat %>%
  filter(str_length(geo) == 4) %>%
  select(TOTFERRT, GDPindex, edu_index, health_index, uY15, uY20, uY25, FAM) %>%
  {set_names(., f.clean_names(names(.)))} %>%
  na.omit() %>%
  rpart::rpart(formula = TFR ~ ., cp = .01)

rattle::fancyRpartPlot(m_part2, palettes = 'PuRd', sub = NULL)

summary(m_part2)

# Model building ------------------------------------------------------------------------------

# Transform the data for panel modeling ======================================================

dat_plm <- dat %>%
  select(
    geo, time, TOTFERRT, edu_index, health_index, GDPindex, FAM, uY15, uY20, uY25
  ) %>%
  mutate( # TODO new element
    Fu15 = FAM*uY15,
    Fu20 = FAM*uY20,
    Fu25 = FAM*uY25
  ) %>% # TODO end of new
  filter(str_length(geo) == 4 & !is.na(TOTFERRT))

dat_plm <- dat_plm %>%
  select(-TOTFERRT) %>%
  mutate(time = time + 1) %>%
  {
    set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
                        paste0(names(.), '_l')))
  } %>%
  merge(dat_plm, all.x = F, all.y = T)

dat_plm <- dat_plm %>%
  select(!ends_with("_l")) %>%
  select(-TOTFERRT) %>%
  mutate(time = time + 2) %>%
```

```r
  {
    set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
                        paste0(names(.), '_l_l')))
  } %>%
  merge(dat_plm, all.x = F, all.y = T)

dat_plm <- dat_plm %>%
  select(-TOTFERRT) %>%
  mutate_at(-(1:2), function(x) x^2) %>%
  {
    set_names(., ifelse(names(.) == 'geo' | names(.) == 'time', names(.),
                        paste0(names(.), '_2')))
  } %>%
  merge(dat_plm, all.x = F, all.y = T) %>%
  select(-c("Fu15_l_l_2", "Fu20_l_l_2", "Fu25_l_l_2", "Fu15_l_2", "Fu20_l_2",
            "Fu25_l_2", "Fu15_2", "Fu20_2", "Fu25_2"))

## Initial models ==============================================================

library(plm)

m_panels <- c(
  'TOTFERRT ~ edu_index_l_l + health_index_l_l + GDPindex_l_l + FAM_l_l +
  uY15_l_l + uY20_l_l + uY25_l_l + edu_index_l + health_index_l + GDPindex_l + FAM_l +
  uY15_l + uY20_l + uY25_l + edu_index + health_index + GDPindex + FAM + uY15 + uY20 +
  uY25',
  'TOTFERRT ~ edu_index_l_l + health_index_l_l + GDPindex_l_l + FAM_l_l + edu_index_l +
  health_index_l + GDPindex_l + FAM_l + edu_index + health_index + GDPindex + FAM'
) %>%
  lapply(function(formula) {
    pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
    within <- plm(eval(formula), data = dat_plm, model = "within")
    random <- plm(eval(formula), data = dat_plm, model = "random")

    list(
      tests = c(
        pooltest(pooling, within)$p.value,
        phtest(within, random)$p.value,
        plm::r.squared(within, dfcor = T)),
      model = within,
      OLS = formula %>%
  str_replace_all('\\~', ',') %>%
  str_replace_all('\\+', ',') %>%
  str_split(',') %>%
  .[[1]] %>%
  trimws() %>%
  {c(., 'geo')} %>%
  {select(dat_plm, .)} %>%
  na.omit() %>%
  group_by(geo) %>%
  summarise_all(.funs = function(x) mean(x)) %>%
  merge(
    plm::fixef(within) %>%
```

```
      {tibble(geo = names(.), a = .)}
  ) %>%
  mutate(
    TOTFERRT = TOTFERRT - a
  ) %>%
  lm(formula = formula)
    )
  })

### Plot coefficients ##############################################################################

m_panels %>%
  lapply(function(output) {
    output$model %>% broom::tidy(conf.int = T) %>%
      rownames_to_column()
  }) %>%
  reduce(rbind) %>%
  mutate(
    rowname = paste0("Model ", as.roman(cumsum(rowname == 1)), "."),
    term = f.clean_names(term, Tosparse = T),
    term = gsub("_2", "^2", term),
    term = gsub("_l_1", '["t = -2"]', term),
    term = gsub("_l", '["t = -1"]', term),
  ) %>%
  ggplot() +
  aes(estimate, term, color = p.value <= .05) +
  geom_vline(xintercept = 0, color = "gray4") +
  geom_point() +
  geom_pointrange(aes(xmin = conf.low, xmax = conf.high)) +
  facet_wrap(~rowname, nrow = 1) +
  labs(x = "Estimated coefficient", y = "Term",
       color = "Corresponding p-value \u2264 5%") +
  scale_color_viridis_d(option = "magma", begin = .2, end = .7) +
  scale_y_discrete(labels=scales::parse_format()) +
  theme(
    legend.position = "bottom",
    legend.direction = "horizontal"
  )

### Model descriptions #############################################################################

m_panels %>%
  lapply(function(output) {
    c(output$tests, nrow(augment(output$model)))
  }) %>%
  reduce(rbind) %>%
  t() %>%
  data.frame() %>%
  mutate_all(function(x) c(scales::percent(x[1:3], accuracy = .01),
                           as.character(x[4]))) %>%
  {set_names(., paste0("Model ", as.roman(1:ncol(.)), "."))} %>%
  mutate(
    Indicator = c("Pooltest", "Phtest", "Adjusted $R^2$", "Observations")
```

```
  ) %>%
  select(Indicator, everything()) %>%
  knitr::kable(caption = "Models", align = c("l", "c", "c", "c"))

theme_update(legend.direction = 'horizontal', legend.position = 'bottom')

ggpubr::ggarrange(
  dat_plm %>%
    na.omit() %>%
    group_by(geo) %>%
    summarise(
      values = n()
    ) %>%
    plot_NUTS2(all.x = T) +
    scale_fill_viridis_b(option = "magma", direction = -1, begin = .2,
                         na.value = "white") +
    ggtitle('Including'),
  dat_plm %>%
    select(!starts_with('u') & !starts_with('Fu')) %>%
    na.omit() %>%
    group_by(geo) %>%
    summarise(
      values = n()
    ) %>%
    plot_NUTS2(all.x = T) +
    scale_fill_viridis_b(option = "magma", direction = -1, begin = .2,
                         na.value = "white") +
    ggtitle('Excluding'),
  common.legend = T
)
# Framework I. ================================================================

## Bias of framework ##########################################################

names(dat_plm) %>%
  { ifelse(
    . %in% c("geo", "time") | str_detect(., "_1") | str_detect(., "_2") |
      str_detect(., 'Fu')
    , NA, .
  )} %>%
  na.omit() %>%
  lapply(function(variable){
    x <- pull(dat_plm, variable) %>% na.omit()
    y <- pull(na.omit(dat_plm), variable)
    tibble(
      Variable = f.clean_names(variable),
      'Mean in total sample' = mean(x),
      'Mean in used sample' = mean(y),
      'Number of observations in the total sample' = length(x)
    )
  }
  ) %>% reduce(rbind) %>%
  knitr::kable(caption =
```

```
                  'Comparison of average values of the variables for incomplete
                and complete observations (Framework I)', digits = 4,
                align = c('l', 'c', 'c', 'c', 'c'))

### Run lasso regression ##########################################################

library(glmnet)

y <- na.omit(dat_plm)$TOTFERRT
X <- model.matrix(TOTFERRT ~ ., data = select(na.omit(dat_plm), -time))
LASSO <- cv.glmnet(X, y)

tidy(LASSO) %>%
  ggplot() +
  aes(log(lambda), ymin = conf.low*1000, ymax = conf.high*1000) +
  geom_line(aes(log(lambda), estimate*1000, color = "Mean-Squared Error")) +
  geom_step(aes(y = nzero, color = "Number of used explanatory variables"),
            size = 1) +
  geom_ribbon(alpha = .4) +
  geom_hline(yintercept = 0, size = 1) +
  geom_vline(aes(xintercept = log(LASSO$lambda.1se), linetype = "Best performing model"),
             color = 'black') +
  scale_y_continuous(
    name = "Number of used explanatory \n variables",
    sec.axis = sec_axis( trans=~./1000, name = "Mean-Squared Error")
  ) +
  labs(y = "Mean-Squared Error", x = "Log(\u03BB)", color = NULL, linetype = NULL) +
  scale_linetype_manual(values = c(2)) +
  theme(legend.position = "bottom", legend.direction = "horizontal")

lasso_coefs <- capture.output(
  coef(LASSO, LASSO$lambda.1se)
) %>%
  .[-(1:2)] %>%
  {tibble(x = .)} %>%
  mutate(term = gsub(" .*", "", x), coef = gsub(".* ", "", x)) %>%
  select(-x) %>%
  filter(!str_detect(term, "geo") & coef != "" & term != "(Intercept)")

m_panels2 <- paste("TOTFERRT ~", paste(lasso_coefs$term, collapse = " + ")) %>%
  lapply(function(formula) {
    pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
    within <- plm(eval(formula), data = dat_plm, model = "within")
    random <- plm(eval(formula), data = dat_plm, model = "random")
    list(
      tests = c(
        pooltest(pooling, within)$p.value,
        phtest(within, random)$p.value,
        plm::r.squared(within, dfcor = T)),
      model = within,
      OLS = formula %>%
  str_replace_all('\\~', ',') %>%
  str_replace_all('\\+', ',') %>%
```

```
  str_split(',') %>%
  .[[1]] %>%
  trimws() %>%
  {c(., 'geo')} %>%
  {select(dat_plm, .)} %>%
  na.omit() %>%
  group_by(geo) %>%
  summarise_all(.funs = function(x) mean(x)) %>%
  merge(
    plm::fixef(within) %>%
      {tibble(geo = names(.), a = .)}
  ) %>%
  mutate(
    TOTFERRT = TOTFERRT - a
  ) %>%
  lm(formula = formula)
    )
  }
  )

m_panels2 %>%
  lapply(function(output) {
    output$tests
  })

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

standard_beta <- augment(m_panels2[[1]]$OLS) %>%
  select(TOTFERRT:.fitted) %>%
  select(-.fitted) %>%
  cor() %>%
  data.frame() %>%
  select(1) %>%
  rownames_to_column() %>%
  rename(term = rowname) %>%
  merge(standard_beta) %>%
  mutate(explain = abs(TOTFERRT*beta)) %>%
  select(term, cor = TOTFERRT, standard_beta = beta, explain)

lasso_coefs %>%
  rename(lasso = coef) %>%
  merge(tidy(m_panels2[[1]]$OLS, conf.int = T)) %>%
  merge(standard_beta) %>%
  mutate_at(-1, function(x) as.numeric(x)) %>%
  pivot_longer(c(lasso:estimate, cor:explain)) %>%
  mutate(
    conf.low = ifelse(name != "estimate", NA, conf.low),
    conf.high = ifelse(name != "estimate", NA, conf.high),
    lag = ifelse(str_detect(term, "_l"),
                 ifelse(str_detect(term, "_l_l"), 2, 1), 0),
```

```r
    bar = ifelse(name %in% c("cor", "explain"), value, NA),
    value = ifelse(!(name %in% c("cor", "explain")), value, NA),
    term = f.clean_names(term, Tosparse = T),
    term = gsub("_2", "^2", term),
    term = gsub("_l_l", '["t = -2"]', term),
    term = gsub("_l", '["t = -1"]', term),
    name = case_when(
      name == "cor" ~ 'Correlation coefficient',
      name == "estimate" ~ 'Coefficient in OLS',
      name == "lasso" ~ 'Coefficient in \nlasso regression',
      name == "standard_beta" ~ 'Standardized coefficient',
      name == "explain" ~ 'Contribution to R-squared'
    ),
    name = factor(
      name, levels = c(
        'Correlation coefficient',
        'Coefficient in \nlasso regression',
        'Coefficient in OLS',
        'Standardized coefficient',
        'Contribution to R-squared'
      )
    ),
  ) %>%
  {
    ggplot(.) +
      geom_vline(xintercept = 0) +
      geom_point(aes(value, term, fill = factor(lag)), size = 3) +
      geom_col(aes(bar, term, fill = factor(lag)), color = "black") +
      scale_fill_viridis_d(option = "magma", begin = .3, end = .7) +
      scale_y_discrete(labels=scales::parse_format(), limits = rev) +
      facet_wrap(~ name, scales = "free_x") +
      labs(x = NULL, y = "Term", fill = "Lag") +
      theme(legend.position = "bottom", legend.direction = "horizontal")
  }

dat_plm <- dat_plm %>%
  select(!starts_with("uY") &! starts_with("Fu"))

names(dat_plm) %>%
  {ifelse(
    . %in% c("geo", "time") | str_detect(., "_l") | str_detect(., "_2"), NA, .
  )} %>%
  na.omit() %>%
  lapply(function(variable){
    x <- pull(dat_plm, variable) %>% na.omit()
    y <- pull(na.omit(dat_plm), variable)
    tibble(
      Variable = f.clean_names(variable),
      'Mean in total sample' = mean(x),
      'Mean in used sample' = mean(y),
      'Number of observations in total sample' = length(x)
    )
  }
```

```
  ) %>% reduce(rbind) %>%
  knitr::kable(caption = 'Comparison of average values of the variables for incomplete and complete obs
               align = c('l', 'c', 'c', 'c'))

LASSO <- cv.glmnet(model.matrix(TOTFERRT ~ ., data = select(na.omit(dat_plm), -time)),
                   na.omit(dat_plm)$TOTFERRT)

lasso_coefs <- capture.output(
  coef(LASSO, LASSO$lambda.1se)
) %>%
  .[-(1:2)] %>%
  {tibble(x = .)} %>%
  mutate(term = gsub(" .*", "", x), coef = gsub(".* ", "", x)) %>%
  select(-x) %>%
  filter(!str_detect(term, "geo") & coef != "" & term != "(Intercept)")

m_panels2 <- paste("TOTFERRT ~", paste(lasso_coefs$term, collapse = " + ")) %>%
  lapply(function(formula) {
    pooling <- plm(eval(formula), data = dat_plm, model = "pooling")
    within <- plm(eval(formula), data = dat_plm, model = "within")
    random <- plm(eval(formula), data = dat_plm, model = "random")
    list(
      tests = c(
        pooltest(pooling, within)$p.value,
        phtest(within, random)$p.value,
        plm::r.squared(within, dfcor = T)),
      model = within,
      OLS = formula %>%
  str_replace_all('\\~', ',') %>%
  str_replace_all('\\+', ',') %>%
  str_split(',') %>%
  .[[1]] %>%
  trimws() %>%
  {c(., 'geo')} %>%
  {select(dat_plm, .)} %>%
  na.omit() %>%
  group_by(geo) %>%
  summarise_all(.funs = function(x) mean(x)) %>%
  merge(
    plm::fixef(within) %>%
      {tibble(geo = names(.), a = .)}
  ) %>%
  mutate(
    TOTFERRT = TOTFERRT - a
  ) %>%
  lm(formula = formula)
    )
  }
  )

m_panels2 %>%
  lapply(function(output) {
    output$tests
```

```r
  })

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

standard_beta <- augment(m_panels2[[1]]$OLS) %>%
  select(TOTFERRT:.fitted) %>%
  select(-.fitted) %>%
  cor() %>%
  data.frame() %>%
  select(1) %>%
  rownames_to_column() %>%
  rename(term = rowname) %>%
  merge(standard_beta) %>%
  mutate(explain = abs(TOTFERRT*beta)) %>%
  select(term, cor = TOTFERRT, standard_beta = beta, explain)

lasso_coefs %>%
  rename(lasso = coef) %>%
  merge(tidy(m_panels2[[1]]$OLS, conf.int = T)) %>%
  merge(standard_beta) %>%
  mutate_at(-1, function(x) as.numeric(x)) %>%
  pivot_longer(c(lasso:estimate, cor:explain)) %>%
  mutate(
    conf.low = ifelse(name != "estimate", NA, conf.low),
    conf.high = ifelse(name != "estimate", NA, conf.high),
    lag = ifelse(str_detect(term, "_l"),
                 ifelse(str_detect(term, "_l_l"), 2, 1), 0),
    bar = ifelse(name %in% c("cor", "explain"), value, NA),
    value = ifelse(!(name %in% c("cor", "explain")), value, NA),
    term = f.clean_names(term, Tosparse = T),
    term = gsub("_2", "^2", term),
    term = gsub("_l_l", '["t = -2"]', term),
    term = gsub("_l", '["t = -1"]', term),
    name = case_when(
      name == "cor" ~ 'Correlation coefficient',
      name == "estimate" ~ 'Coefficient in OLS',
      name == "lasso" ~ 'Coefficient in lasso regression',
      name == "standard_beta" ~ 'Standardized coefficient',
      name == "explain" ~ 'Contribution to R-squared'
    ),
    name = factor(
      name, levels = c(
        'Correlation coefficient',
        'Coefficient in lasso regression',
        'Coefficient in OLS',
        'Standardized coefficient',
        'Contribution to R-squared'
      )
    )
  ) %>%
```

```r
  filter(name != 'Correlation coefficient') %>%
  {
    ggplot(.) +
      geom_vline(xintercept = 0) +
      geom_point(aes(value, term, fill = factor(lag)), size = 3) +
      geom_col(aes(bar, term, fill = factor(lag)), color = "black") +
      scale_fill_viridis_d(option = "magma", begin = .3, end = .7) +
      scale_y_discrete(labels=scales::parse_format(), limits = rev) +
      facet_wrap(~ name, scales = "free_x") +
      labs(x = NULL, y = "Term", fill = "Lag") +
      theme(legend.position = "bottom", legend.direction = "horizontal")
  }
```