APPLICATIONS OF ECONOMETRICS

PROF. ANDREAS STEINHAUER

GROUP PROJECT

GROUP 144

EXAM NUMBERS:

B158012

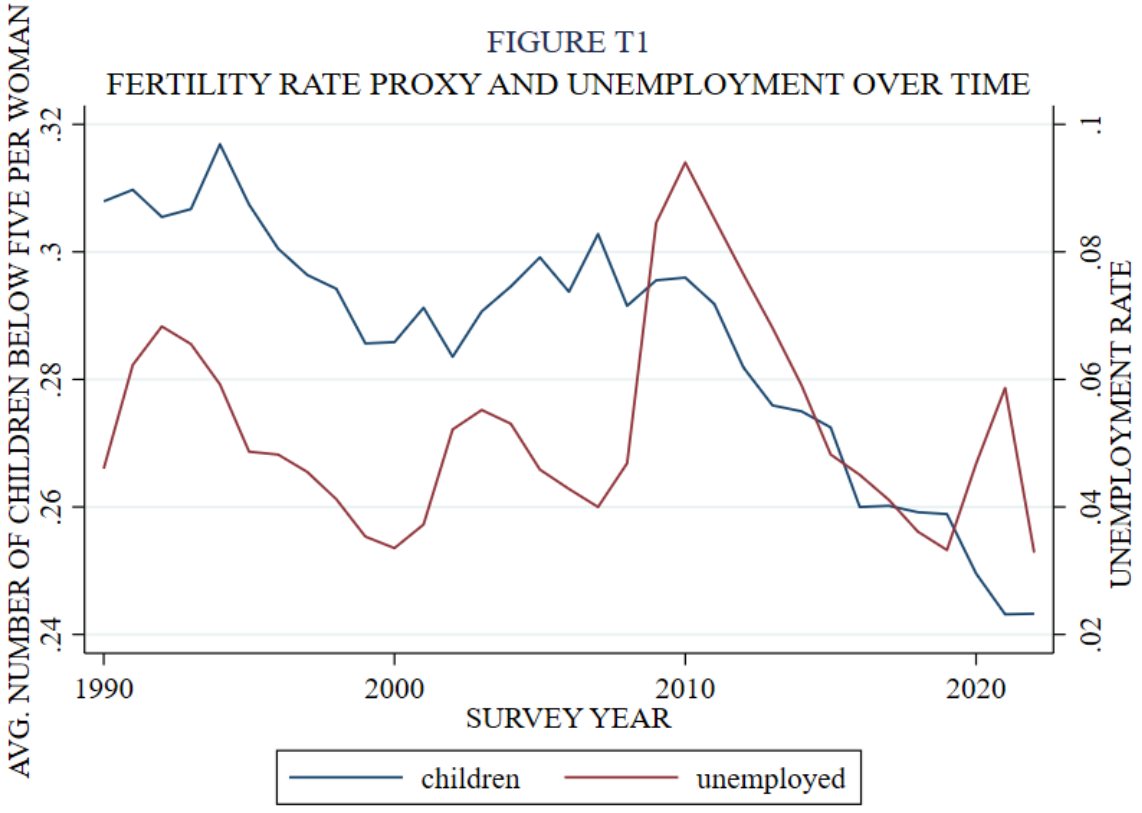
B234112

Time Series Questions

1. Introduction to the Simple Model

Plot the time series for 'children' and 'unemployed' over time. Make sure you label the axes correctly. Then run a simple regression of 'children' on 'unemployed'. This could approximately capture the relationship between fertility and the business cycle. Interpret your findings.

Using ASEC-CPS data from the United States Census Bureau, multiple key variables were extracted on an annual basis. These include "children" (The number of own children under age 5 of women age 15-45 in a household), "unemployed" (Share of unemployed individuals age 25-54 who are in the labour force), and "share_married" (Share currently married of all individuals age 18 and older), as well as "year" (The respective survey year of the aforementioned variables). By plotting "unemployed" and "children" against "year", we receive the following chart:



This could potentially present a relationship between "children" and "unemployed" – or in more general terms – the impact of the business cycle on fertility rate. To get a clearer picture, we can estimate an OLS regression of "children" on "unemployed".

Regression **T.1**: $children_t = \beta_0 + \beta_1 unemployed_t + u_t$

Where u_t is the error term with: $E[u_t] = 0$

Running the regression on STATA gives the following summary output:

VARIABLES	(1) Regression T1
unemployed	0.372*
	(0.216)
Constant	0.266***
	(0.012)
Observations	33
R-squared	0.087

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficient β_1 estimated for "unemployed" of 0.372 means that a marginal increase in the mean unemployment rate of a given year is expected to increase the mean number of young children per woman per household by 0.372. To interpret a 1 percentage point increase in unemployment – which is more in line with economic reality – "children" is to increase by the coefficient β_1 divided by 100: 0.00372.

However, this coefficient is not statistically significant, as the values given for the 95% confidence interval span across 0. More specifically, the t-statistic (1.72) is small and the p-value (0.095) > 0.05 means we fail to reject the null hypothesis ($H_0 : \beta_1 = 0$) at the 5% significance level. The constant β_0 suggests that an unemployment rate of 0 leads to a mean number of young children per woman per household of 0.266 – which as seen in Figure **T.1**, is a value at which "children" is observed despite the positive unemployment rate. All this makes inference based on this regression rather difficult.

Regardless, given the estimated coefficient of our simple regression model, unemployment appears to have a positive effect on the number of children. This runs counter to intuition and academic research which finds a negative relationship between the two variables: A worsening economic environment decreases the disposable income necessary for raising dependent children. We do not believe our counterintuitive result to arise from measurement errors as the ASEC-CPS is a reliable source. Neither is it the case that "children" and "unemployment" are poor proxies for fertility rate and the business cycle respectively. Hence, we will continue with our analysis to more accurately ascertain the effect. We estimated a plethora of similar simple regressions [Line 23 Dofile] with lagged unemployment (perhaps there is a delayed effect of the business cycle) and logged variables (to counter outliers). Logging does however not make much sense as unemployment is a figure below zero, making direct interpretation difficult. Although some regressions estimated barely negative coefficients for "unemployed" lagged one period behind "children", none of these coefficients were statistically significant.

Proceeding, we will further evaluate the recorded effect of the business cycle on the fertility rate, rather than simply improving the R^2 of the regression. To this end, we will add more control variables (Question 2.), deal with time trends (Question 3.) , and address potential nonstationarity or stationarity in the error term (Question 4.).

2. Adding Control Variables

One potentially important control variable in the relationship between fertility and the business cycle is the share married. Run a regression of 'children' on 'unemployed' including 'share_married' as a control variable. Compare your results to (1) and discuss why they might be different.

The estimated coefficients in regression (1) might suffer from an omitted variable bias. If an omitted variable X with coefficient β_2 were **i)** correlated with "employed", and said variable X is also **ii)** determinant of "children", the following bias would occur:

$$E[\beta_1] = \beta_1 + \beta_2 \frac{Cov("employed", X)}{Var("employed")}$$

In our case, "share_married" (The share of individuals currently married age 18 and above) can reasonably be assumed to meet the two requirements. Therefore, regression **T.1** is adjusted as follows:

Regression **T.2**: $children_t = \beta_0 + \beta_1 unemployed_t + \beta_2 share\_married_t + u_t$

We expect the inclusion of "share_married" to the regression **T.1** to remove negative bias from β_1, as β_2 is expected to be positive, "share_married" and "unemployed" are negatively correlated, and the variance of "unemployed" is naturally positive. Running this regression on STATA gives the following summary output:

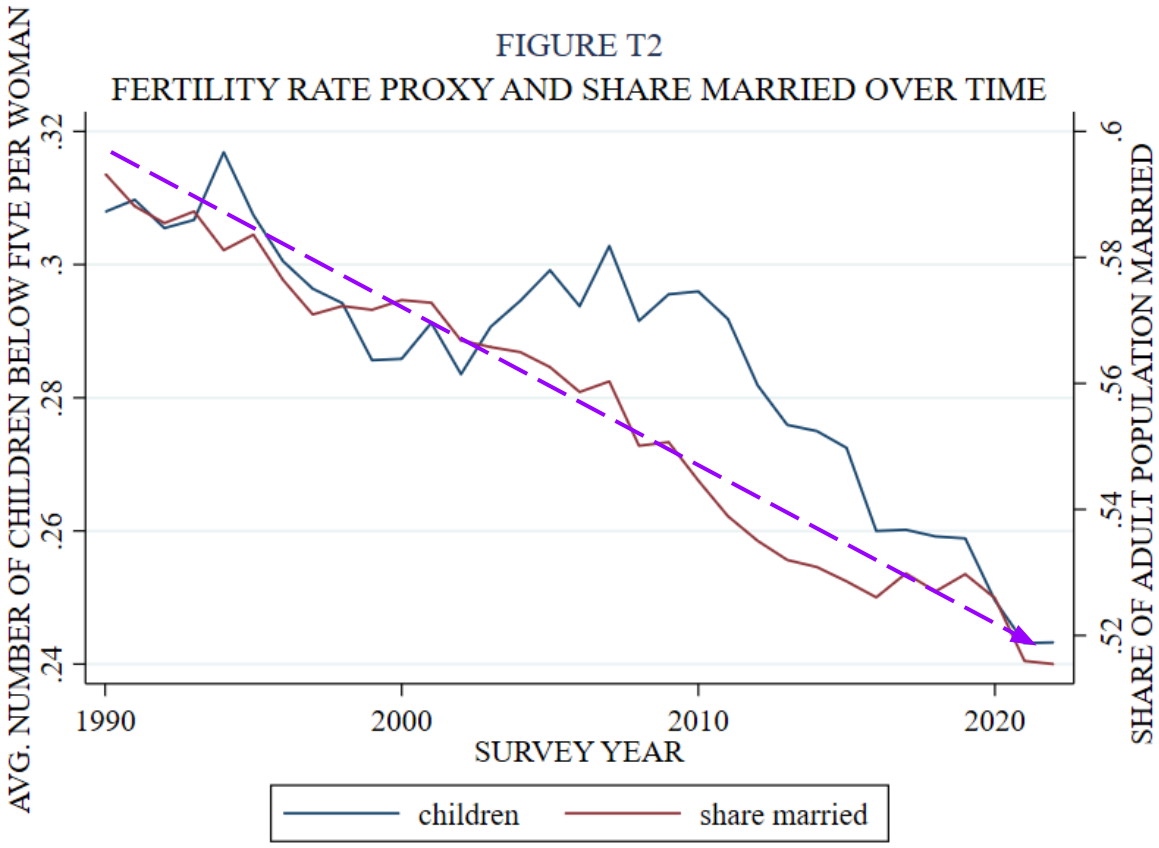
VARIABLES	(1) Regression T2
unemployed	0.389***
	(0.084)
share_married	0.746***
	(0.056)
Constant	-0.149***
	(0.032)
Observations	33
R-squared	0.867

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The effect of the coefficient β_1 has increased very slightly. The interpretation remains the same. However, the coefficient is now highly statistically significant, and we can confidently reject the null hypothesis ($H_0 : \beta_1 = 0$). The positive sign of β_1 remains however, which still runs counter to our economic intuition. This was expected however, as we presumed that β_1 suffers from negative omitted variable bias in regression **T.1**.

Regarding our control variable "share_married", its coefficient β_2 suggests that a marginal increase in the share married of the adult population leads to an increase of 0.746 mean number of young children per woman per household. β_2 is also highly statistically significant. However, there could

still be issues with endogeneity arising from an omitted time trend. Plotting "share_married" and "children" over time clearly indicates a negative time trend, indicated in purple:



We will now address this negative time trend, as we want to make sure that we do not simply have a spurious correlation at hand.

3. Issues with Trends in Variables; Controlling trend

The regression in (2) might give us biassed and inconsistent coefficients if there are trends in the variables we are using. Investigate this issue and run a regression that addresses this concern if you find that it could be important. Discuss your results.

As shown in Figure 2T, there is ample visual indication that "share_married" and "children" could be following a time trend (indicated in purple). This is naturally a violation of the assumption $Cov[u_t, x_t] = 0$, as the time trend is in the error term. There is a possibility of a spurious correlation at hand, as both variables "share_married" and "children" are both affected by a hidden time trend – giving the appearance that "share_married" is significant even though it is merely the time trend affecting "children" through "share_married". To formally test whether the time trend is significant for "children" and "share_married", we can run a simple regression on time *t*, where "ϵ_t" is the residual:

Example Regression **T3.1**: $time\ dependent\ variable_t = \gamma_0 + \gamma_1 t + \epsilon_t$

	(1)	(2)	(3)
VARIABLES	T3.1: share_married	T3.1: children	T3.1: unemployed
t	-0.002***	-0.002***	-0.000
	(0.000)	(0.000)	(0.000)
Constant	0.596***	0.317***	0.054***
	(0.001)	(0.003)	(0.006)
Observations	33	33	33
R-squared	0.972	0.779	0.003

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
t ranges from 1-33, indicating the time trend

The negative coefficients of *t* are proof of the assumption of the negative time trend in "share_married" and "children". As such, it can be said that with every subsequent period "share_married" and "children" decrease by -0.0024 and -0.0018 in absolute terms. The coefficients of *t* are also highly statistically significant, especially for "share_married". Similarly, if we assumed this estimated deterministic time trend were to hold from the observed first period value until the final (predicted) period for "share_married" and "children", the two variables would end at ca. 0.514 ("share_married") and ca. 0.248 ("children"), which almost meets the realised final values of 0.515 ("share_married") and 0.243 ("children") in 2022.

The same cannot be said for "unemployed" however. The coefficient of *t* in this regression is minute and highly statistically insignificant with a t-value of -0.28. This is in line with the economic intuition that unemployment is roughly a mean-reverting process, with the distance to the mean of natural unemployment dependent on the business cycle. Figure **T.1** nicely portrays this process.

To address the issue of this time trend in "children" and "share_married", we will simply add *t* as an additional control variable to regression **T.2**.

Regression **T.3**: $children_t = \beta_0 + \beta_1 unemployed_t + \beta_2 share\_married + \beta_3 t + \epsilon_t$

VARIABLES	Regression T3
unemployed	0.403***
	(0.092)
share_married	0.888**
	(0.369)
t	0.000
	(0.001)
Constant	-0.235
	(0.222)
Observations	33
R-squared	0.868

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Regression **T.3** could also use a quadratic time trend, but the purple line in Figure **T.2** suggests no such trend. Additionally, a quadratic trend is disregarded due to their tendency to overfit the regression. β_1, β_2 are still statistically significant on a 5% confidence level. Both coefficients have increased, especially β_2 of "share_married", jumping from 0.74 in Regression **T.2** to 0.88 in Regression **T.3**. This makes sense, as the previous omission of *t* caused a heavy negative bias.

Coefficient β_3 of *t* however does not make much intuitive sense anymore. With a minute positive coefficient of 0.003 and t-value = 0.39, *t* is not statistically significant anymore. Given the formula for the Variance of β_3, it can be inferred that the standard error of *t* is going to be high:

$$Var(\hat{\beta}_3) = \frac{\sigma^2}{(SST_3)(1-R_3^2)}$$

We can estimate the auxiliary regression:

Aux. Regression **T3.2**: $t = \beta_0 + \beta_0 share\_married_t + \beta_0 unemployed_t + u_t$

VARIABLES	Auxiliary Regression T3.2
unemployed	-39.933**
	(17.212)
share_married	-403.718***
	(11.524)
Constant	243.179***
	(6.479)
Observations	33
R-squared	0.976

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The R^2 = 0.9762 explains why the standard error is high and the resulting t-value so low of variable *t* in Regression **T.3**. With such a low t-value, we can ignore the positive sign of β_3.

The inclusion of the insignificant variable *t* in Regression **T.3** can however be easily circumvented by detrending "children" and "share_married", which allows us to omit *t* from the regression. Detrending can be achieved by storing the respective residuals of Regression **T.3.1**, and then using these detrended residuals in place of the original variables (See Regression **T.3.1**). The residuals "ϵ_t" entail the variation of the dependent variable which cannot be explained by the time trend. However, this process in the end has the same effect as controlling for t, and therefore the coefficients of interest β_1 ("unemployed") and β_2 ("share_married") and their subsequent standard errors do not change. As such the results of this alternate approach for controlling the trend is not included for brevity's sake. We will however from now on use the detrended variables "children detrended" and "share married detrended" to make comparison between regressions easier and keep the regression slim in length. These detrended variables will have a dot on them, for example: $x_t^{\bullet}$. Furthermore, Figure **T.2** exhibits no signs of seasonality, and therefore this issue will not be addressed. For the last part of the time series, lags will be introduced to improve the regression, and potential issues with nonstationarity are tackled.

4. Nonstationary and Stationary dynamics in error terms

The regressions in (2) and (3) might still give us biassed and inconsistent coefficients and/or standard errors. It could be due to the presence of nonstationary dynamics (e.g. unit roots) and stationary dynamics (e.g. autocorrelation) in the error terms. Investigate this issue and run regressions that address this concern if you find that it could be important. Discuss your results.

To further improve regression **T.3**, we will introduce lags of variables. This makes conceptual sense in our case, as contemporaneous variables might not affect our dependent variable. The decision of having a child occurs over a lengthy time, and naturally the gestation period is roughly 9 months, which is almost one period (year) earlier in our data. Therefore, it makes more sense to relate "children" to "unemployed" one period earlier. Conceptually similar, the current fertility rate might be affected by the fertility rate of the previous period. As such, regression **T.3** is updated to incorporate this:
Regression **T.4**:

$$children_t^{\bullet} = \beta_0 + \beta_1\, unemployed_{t-1} + \beta_2\, share\_married_t^{\bullet} + \beta_3 children_{t-1}^{\bullet} + u_t$$

VARIABLES	Regression T4
unemployed, lag 1	0.229***
	(0.076)
share_married•	1.010***
	(0.260)
children•, lag 1	0.692***
	(0.109)
Constant	-0.012***
	(0.004)
Observations	32
R-squared	0.769

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Detrended variables indicated with:•

This updated regression is overall confident in rejecting the null-hypotheses of the respective regressors. The coefficient β_1 has previously been overestimated, as indicated by the severe drop compared to the previous regression **T.3**. Interpreting the coefficient according to regression **T.4,** it may be expected that a marginal increase of the unemployment rate in the current period increases the mean number of young children per woman per household by 0.23. To further ensure that this is no spurious relationship, we will now test for non-stationary dynamics using the Dicky-Fueller test, and further test for stationary dynamics using the Durbin-Watson test. An Engle-Granger test for cointegration between "unemployed" and "children" has been tested but will not be presented for brevity's sake, as "unemployed" is reasonably a stationary process for the given period and as such various implementations have found no indication of cointegration.

The Dicky-Fuller test is used to test whether the regression follows a random walk, i.e whether any future variation is simply unsystematic and thus robbing our regression of any inferential or predictive power; A random walk may undergo a one-time exogenous shock and will not go back to the original relationship of dependent and independent variables, thus rendering the results of our regression useless.

Unit root in an example AR(1) process: $y_t = \rho y_{t-1} + u_t, \; s.t \; \rho = 1$

Regression **T.4.1**: AR(1) process of the fitted residuals of regression **T4**

VARIABLES	Regression T4.1: Residuals AR(1)
Residuals, 1 lag	-0.185
	(0.175)
Observations	31
R-squared	0.036

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The null-hypothesis states that the tested variable contains unit root. To this end, we store the fitted residuals of regression **T.3**. STATA offers 4 different augmented cases for the augmented Dickey-Fuller test. As we have detrended the variables, we are assuming no drift term, yet cannot infer whether there is a constant or not in the residuals. This case is the default option for augmented Dickey-Fuller test. The number of lags used for the test is determined insofar that they are significant. In our case, it is only the first lag.

		Dickey-Fuller Critical Value		
	Test Statistic	1%	5%	10%
Z(t)	-4.238	-3.716	-2.986	-2.624

MacKinnon approximate p-value for Z(t) = 0.0006

The resulting test statistic is far below any critical value, and as such, the null hypothesis of unit root can be rejected. Accordingly, the coefficients of regression **T.4** remain solid.

Having checked the non-stationary dynamics, we are now moving on to address the serial correlation of the error. Serial correlation is an issue for the strict Gauss-Markov time series assumption:

From TS5' (No serial correlation): $Cov(u_t, u_k) = 0, \forall \; t \neq s$

The Durbin-Watson test gives a test statistic between 0 - 4, where a value of 2 indicates no serial correlation in the residuals of lag 1. A statistic above 2 implies negative serial correlation and a statistic below conversely positive serial correlation. Given the value $\widehat{\rho}$ of regression **T.4.1**, we may expect a value slightly above 2.

Running the Durbin-Watson test for regression **T.4** gives the following statistic:

```
Durbin-Watson d-statistic(  4,    32) =  2.289329
```

As such, we also needn't be concerned with strong serial correlation in the residuals. The small serial correlation that there is present in the residuals of **T.4** will geometrically decay accordingly:

$$Corr(u_t, u_{t+k}) = \frac{Cov(u_t, u_{t+k})}{SD(u_t)SD(u_{t+k})} = \frac{\rho^k Var(u_t)}{SD(u_t)\,SD(u_t)} = \frac{\rho^k Var(u_t)}{Var(u_t)} = \rho^k$$

In order to immediately eliminate the serial correlation present, one could quasi-demean all variables according to the serial correlation ρ of **T.4.1**. This approach is however not presented in full as the resulting coefficients are, unsurprisingly, barely different to those of **T.4**.

Finally, we are going to assure that our inference is robust to heteroskedasticity. As such, we will run regression **T.4** with Newey-West standard errors which will be robust to heteroskedasticity and first order serial correlation (HAC). Again, we choose a lag of 1 as we are working with annual data and a rather short time series:

VARIABLES	Regression T4: HAC errors
unemployed, lag 1	0.229***
	(0.054)
share_married•	1.010***
	(0.226)
children•, lag 1	0.692***
	(0.090)
Constant	-0.012***
	(0.003)
Observations	32

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As shown by the output, the coefficients of **T.4** have not visibly changed. Notably, the robust standard errors are smaller than for the original regression, which is presumably due to the small size of the sample. Regardless, the the original interpretation of β_1 remains robust: A marginal increase in the unemployment rate is expected to increase the mean number of young children per woman per household by 0.23. Further, this implies a broader positive relationship between unemployment and the fertility rate for the given time period. These findings are at odds with the academic consensus (Kristensen and Lappegård, 2022; Aksoy, 2016). – Perhaps this idiosyncrasy arises because our data is too aggregated, the variation in unemployment is simply too slim, or our time period is too short; The series only covers two recessions entirely. Finally, we can compare the regression coefficients for all subsequent regressions on "children" for **T.1 - T.4**:

VARIABLES	(1) Regression T1	(2) Regression T2	(3) Regression T3	(4) Regression T4
unemployed	0.372*	0.389***	0.402***	
	(0.216)	(0.084)	(0.091)	
share_married		0.746***		
		(0.056)		
share_married•			0.886**	1.010***
			(0.363)	(0.260)
unemployed, lag 1				0.229***
				(0.076)
children•, lag 1				0.692***
				(0.109)
Constant	0.266***	-0.149***	-0.021***	-0.012***
	(0.012)	(0.032)	(0.005)	(0.004)
Observations	33	33	33	32
R-squared	0.087	0.867	0.403	0.769

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Detrended variables indicated with:•

The following table compare the different variations on regression **T.4**; Namely, the standard regression, the quasi-demeaned version using ρ from **T.4.1**, and finally with Newey-West standard errors:

VARIABLES	(1) Regression T4	(2) T4: Quasi-demeaned	(3) T4: HAC Errors
unemployed, lag 1	0.229***		0.229***
	(0.076)		(0.054)
share_married•	1.010***		1.010***
	(0.260)		(0.226)
children•, lag 1	0.692***		0.692***
	(0.109)		(0.090)
unemployed~, lag 1		0.225***	
		(0.068)	
share_married~•		1.066***	
		(0.235)	
Children~•, lag 1		0.759***	
		(0.095)	
Constant~		-0.036***	
		(0.011)	
Constant	-0.012***		-0.012***
	(0.004)		(0.003)
Observations	32	31	32
R-squared	0.769	0.845	N/A

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
Detrended variables indicated with: •
Demeaned variables indicated with: ~

Panel Questions

5. Overview of Fertility and Income

Plot 'children' against 'lnincome' for the year 2022 using a scatter plot where one dot is the average in a state in 2022. Try to add a regression line to highlight what the overall relationship looks like in 2022. Make sure that the axes are labelled. Briefly state what you find. Then provide some descriptive statistics for your sample by reporting the mean, minimum and maximum of key variables ('children', 'lnincome').

In this section, we will investigate whether there is a relationship between fertility rates and life-time income in the United States between 1990 and 2022.

For our panel analysis, our data set provides us with information for each of the 50 states of the United States, in addition to the District of Columbia (DC), for each year in the time period between 1990-2022. Thus, we have a strongly balanced panel data set, which includes entries for each of the same 33 years for each of the 51 state entities. This allows us to follow each entity's (state's) average household income over the time period, and control for any characteristics of the entities that do not change - or barely change - over time. In the case of states being the entity, such characteristics might be culture and religion. We include the following variables for each state and year in our analysis: 'children' - the mean number of children aged below 5 per household for women aged 15-45, 'lnincome' - the log of the median household income, 'sharewomen' - mean share of women, 'sharemarried' - mean share of people married aged over 18, 'lnpop' - the log of the state population, and 'sharehispanic' - the share of Hispanics.

We will proceed to investigate the relationship between fertility and income, using our variable 'children' as a proxy for fertility and our main explanatory variable 'lnincome' for lifetime income. First, we plot the per-household number of young children against the log of income for the year 2022 in **Figure P.1** Each dot represents the 1 of the 51 states. The fitted trend line indicates a slightly downwards-sloping trend with gradient = -0.04899, which, as seen in **Table P.I**, is not statistically significant though. We will further investigate this relationship in questions 6 and 7.

FIGURE P.1

FERTILITY RATES AND INCOME BY US-STATE IN 2022

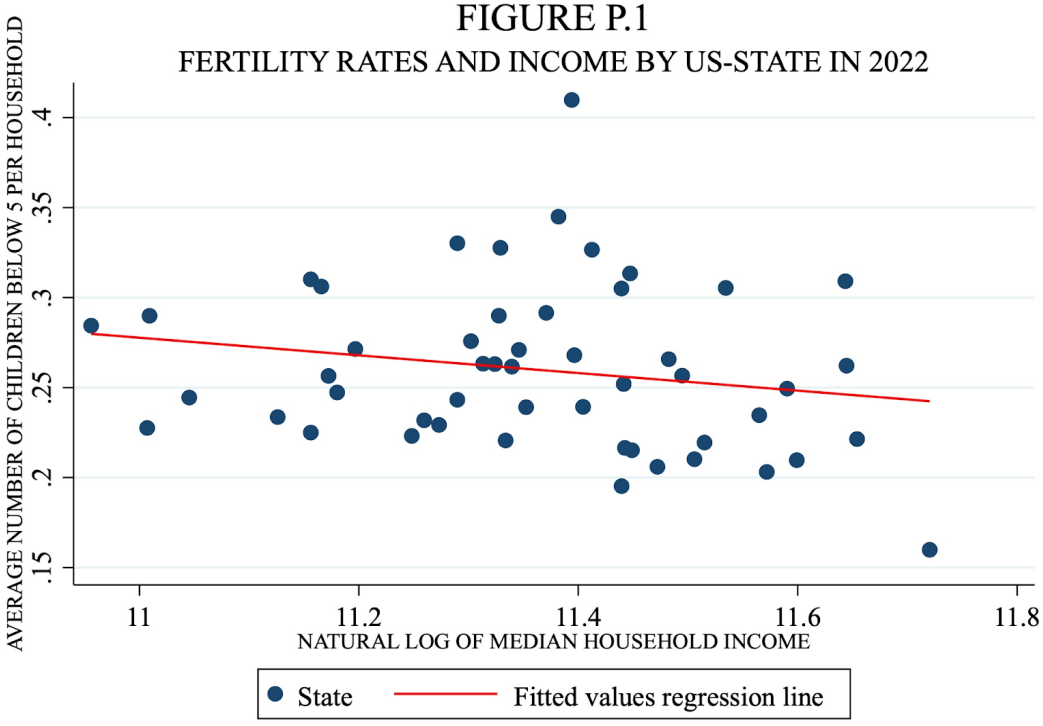


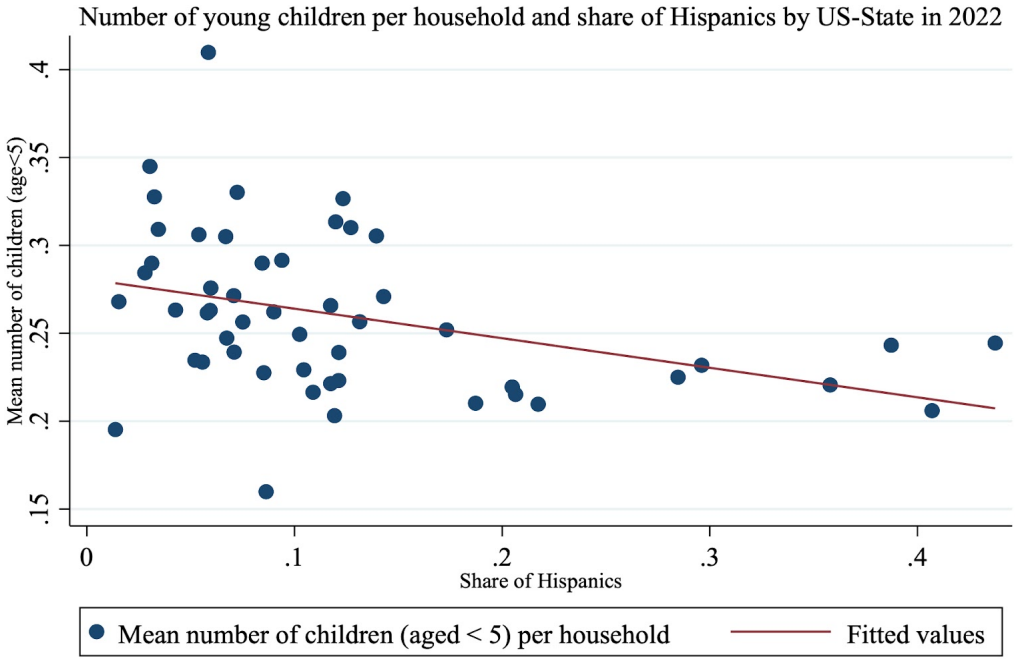
Table P.I: Linear regression of 'children' against 'lnincome' in 2022, as shown by the trendline in Fig P.1

VARIABLES	(1) children
Log of median household income	-0.0490
	(0.0360)
Constant	0.817*
	(0.410)
Observations	51
R-squared	0.036

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Further, we will add the share of Hispanics as a control variable, as there appears to be a negative relationship between the share of Hispanics and the per-household number of young children, as seen in **Figure P.2**.

Figure P.2



In **Table P.II**, we report descriptive statistics for the dependent variable and our control variables. In particular, the mean number of children under the age of 5 per household for women aged between 15-45 is 0.29. To report average income, we use the natural logarithm of the median household income in a state. It is sensible to use the median income instead of the mean to compute a representative income figure within a state because outliers like very-high-income-individuals would distort the mean. Similarly, using the natural logarithm of income de-scales the particularly high income values to ensure a small

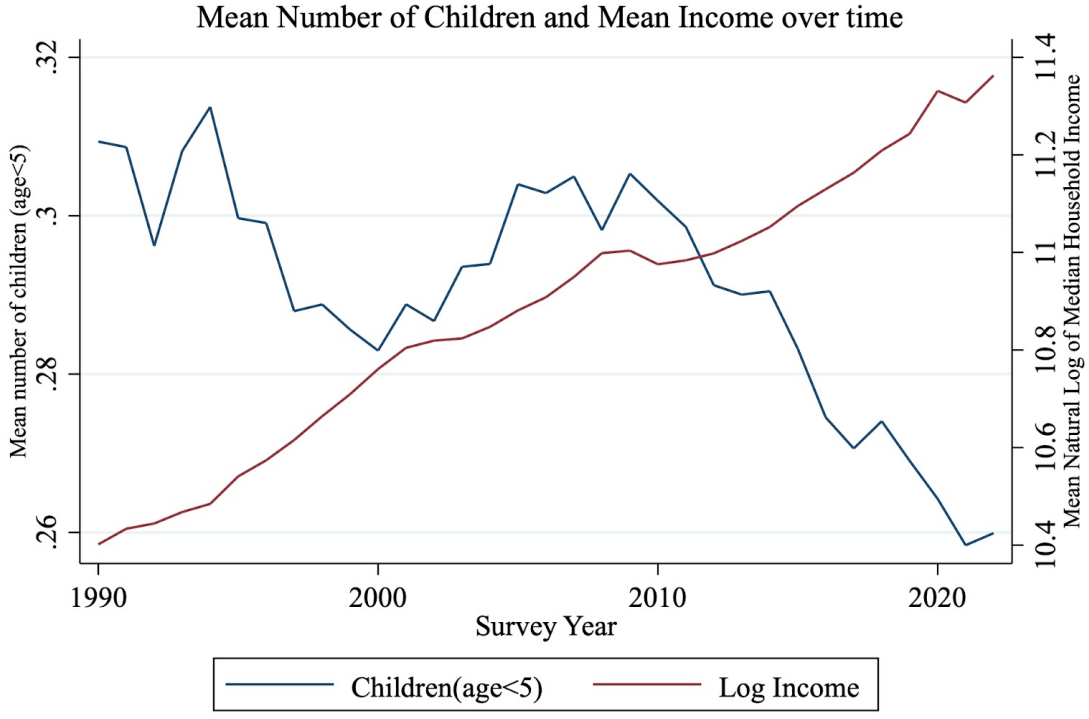
number of high-earners do not heavily affect the reported income value. For our time period, the mean log income is 10.879. The share of married people, counting anyone aged 18 and above, is on average 56.2%, the average share of women is 50.9%, and the share of Hispanics is on average 8.9%.

Table P.II: Descriptive Statistics across years 1990-2022 and the 50 US-states (+ D.C.)

Variable	**Mean**	**Min**	**Max**
Mean number of children (aged < 5) per household	0.29	0.146	0.461
Log of median household income	10.879	10.012	11.720
Share of people married (aged > 18)	0.562	0.274	0.699
Share of women	0.509	0.469	0.563
Log of State Population	7.984	6.913	9.927
Share of Hispanics	0.089	0.001	0.486

Though these descriptive statistics are helpful, it is perhaps more informative to capture the development of the number of young children and income over the time span of our sample. Figure **P.3** reports the mean value of 'children' and 'lnincome' for each year. The overall trend appears to be that the mean number of young children decreases over time, though a temporary increase can be observed between the years 2000-2008. Income consistently increased over the time of our sample.

Figure P.3



6. Pooled OLS and Added Controls

Estimate the relationship between fertility and income by pooled OLS (pooling all years) controlling only for the year of the survey. Then estimate a second pooled OLS regression adding your choice of control variables. Interpret your results and compare them to the simple regression in (5).

First, we run a pooled OLS regression to estimate the effect of income on the number of children aged below 5 per household, controlling only for the year of the survey:
$children_{it} = ß_0 + ß_1 lnincome_{it} + ß_2 year_t$, for i = 1, 2, … , 51 and t = 1, 2, … , 33 **(reg. 6.1)**

This means that we are estimating a regression model over our panel data which stacks the available data over all 51 individuals i (each state) and all 33 time periods t (between 1990-2022), creating $51 \times 33 = 1,683$ observations.

Table P.III shows that when controlling for the time variable year, the pooled OLS estimates the coefficient of the log of median household income ('lnincome') to be -0.017. This slight negative effect of income on the number of young children is similar to that estimated by the regression line added to **Figure P.1** in Question 5. The coefficient means that a 10% increase in income decreases the number of children under the age of 5 per household by approximately 0.0017. To be more precise, we can perform the following calculation:

$\Rightarrow children_{new} = ß_0 + ß_1 lnincome + ß_1 ln(1.1) + \ldots$
$\Rightarrow children_{new} - children_{old} = ß_1 ln(1.1)$

Thus, increasing income by 10% yields a change in the number of children aged below 5 per household of: $ß_1 ln(1.01) = -0.017 ln(1.1) = -0.0017$

We can see that this effect is highly statistically significant, as the reported standard error of 0.00638 is very low, and the p-value is less than 0.01. Thus, we reject the null hypothesis that income has no effect on the number of children aged below 5 per household, at a 1% significance level.

However, it is likely that the number of children a household decides to have in a given year and state also depends on other factors apart from income and year. Thus, we will now add further control variables to our regression to avoid any omitted variable bias, as seen in **(reg. 6.2).**

$children_{it} = ß_0 + ß_1 lnincome_{it} + ß_2 sharemarried_{it} + ß_3 sharewomen_{it} + ß_4 sharehispanics_{it} + ß_5 lnpop_{it} + ß_6 year_t$ for i = 1, 2, … , 51 and t = 1, 2, … , 33 **(reg. 6.2)**

We add the share of married people ('sharemarried'), the share of women ('sharewomen'), the share of hispanics ('sharehispanics'), and the state population to our original regression **(reg. 6.1)**.

The results estimated both regressions **P.1** and **P.2** using pooled OLS are also displayed in table **P.III**. For readability purposes, we have omitted the estimates for each year. Instead, the row "Time-Fixed Effects" indicates that we have controlled for the years in both regressions. Compared to regression (**P.1**), the estimated coefficient for the income variable remains negative and statistically significant. However, this coefficient $ß_1$ changes to -0.033, meaning that this regression attributes more of the change in children to income than before. This is likely because a positive coefficient is estimated for the number of shared people, thus this estimated positive effect is balanced by a larger negative effect in a different regressor (See omitted variable bias formula in TS Q2).

Table P.III Pooled OLS Regression Estimates, Children (Aged < 5) per Household

VARIABLES	(1) Pooled OLS on 'number of young children'	(2) + Control Variables
Log of median household income	-0.017***	-0.033***
	(0.006)	(0.006)
Share of people married (aged > 18)		0.297***
		(0.020)
Share of women		-1.309***
		(0.102)
Share of Hispanics		0.012
		(0.011)
Log of State Population		-0.003
		(0.002)
Constant	0.488***	1.164***
	(0.067)	(0.092)
Time-Fixed Effects	Yes	Yes
Observations	1,683	1,683
R-squared	0.115	0.379

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Increasing the share of married people by 1 is expected to increase the number of children below the age of 5 per household by 0.297. Notice that since this variable is a measure of share and the variance within the data set is 0.003, a 1-unit increase is highly unrealistic. Alternatively, this coefficient implies that a 0.1 unit increase in the share of married people would result in a 0.0297 increase in children below 5 per household. **Table P.IV** illustrates what such a change would look like.

Table P.IV Expected change in children under 5 per household

Data from 2022		Effect of a 0.1 increase in share married	
Avg. share married	Avg. children under 5 per household	Avg. share married	Avg. children under 5 per household
0.523	0.260	0.623	0.2897

Further, regression **P.2** estimates a coefficient of -1.309 for the share of women. Thus, increasing the share of women by 1 unit would decrease the number of children below 5 per household by approximately 1.309. Again, putting this into a more realistic context: a 0.1 unit increase in the share of women would decrease the number of children 0.1309. Both these coefficients are statistically significant, as the p-values are less than 0.01.

Despite observing a downwards-trend between the share of Hispanics and the number of young children in **Figure P.2**, coefficient β_4 estimated by our pooled OLS regression **P.2** is not statistically significant, as the p-value > 0.1. Thus, we fail to reject the null hypothesis that the share of Hispanics has no effect on the per-household number of children. A possible explanation for this result may be that the downwards-trend presented by **Figure P.2** did not actually capture the relationship between the variables 'sharehispanic' and 'children' but merely reflected that many southern states (which commonly have high shares of Hispanic people due to their proximity to the US-Mexican border) have lower fertility rates. Given our regression estimates for coefficient β_4 are not statistically significant, the lower fertility rates in the south might be due to other reasons than the presence of Hispanics.

Finally, we cannot assume that the state population has an effect on the number of young children per household, as the estimated coefficient = -0.003 and is not statistically significant.

7. Fixed Effects and First Differences Estimations

Estimate the relationship between fertility and income using the fixed effects estimator and the first differences estimator with the same controls you chose in (6). Interpret your results and compare them to your results in (6). Discuss whether FE or FD might be more appropriate here.

We now proceed to estimate our regression using the Fixed Effects (FE) and the First Differences (FD) estimator. These estimators are particularly useful for our case, because, unlike the pooled OLS and the Random Effects estimators, FE and FD are consistent when $E(a_i|X_i) \neq 0$. This is important to consider in our case – it seems plausible to assume that such heterogeneity denoted by a_i like geographic and cultural differences between the states affects some of our regressors such as the share of married people and the share of Hispanics.

The FE estimator captures the within-individual variation over time, creating time-demeaned variables. These are the result of the within-transformation, which first calculates the individual's mean value over all time periods. Then, the difference between the individual's time-specific values and the individual's mean value is captured as follows: $\ddot{y}_{it} = y_{it} - \overline{y_i}$

A benefit of the FE-estimator is that this within-transformation cancels out the unobserved heterogeneity effect a_i by:

$$y_{it} - \overline{y_i} = \beta_1 x_{it} + a_i + u_{it} - (\beta_i \overline{x_i} + a_i + \overline{u_i}) = \beta_1 (x_{it} + \overline{x_i}) + u_{it} - \overline{u_i}$$

However, this same property of the FE-estimator also eliminates any time-invariant variables like sex or age from the analysis. In our case, this is not a concern because all variables included in our specification change over time.

The FD-estimator also eliminates a_i as well as the time-invariant variables - though through differencing instead of within-transformation. When running the estimators on more than two time-periods, as we do here, the FE and FD estimators vary in efficiency, based on the correlation between the error terms. By assumption FE.6 presented by Wooldridge (2019), which states that the FE-estimator requires the idiosyncratic errors to be serially uncorrelated (ie. $Cov(u_{it}, u_{is}|X_i, a_i)$), FE is more efficient than the FD when we have uncorrelated errors. The FD-estimator is more efficient when the error terms u_{it} follow a random walk, i.e exhibit unit root.

To determine which estimator is more efficient, we test for serial correlation in the error term and present the results in **Table P.V**. The FE errors are positively correlated with a coefficient of 0.642 and statistically significant at the 1% significance level, meaning we observe some serial correlation and the FE estimator may not be the most efficient choice in our case. However, 0.642 does not indicate a very strong positive correlation and is not close to a unit root. Thus, we cannot assume u_{it} follows a random walk either. Further, the differenced errors present a statistically significant negative correlation with coefficient -0.348, which further indicates that we cannot confidently use the FD-estimator to be more efficient.

Table P. V: Results from testing serial correlation in the error terms.

VARIABLES	(1) Fixed Effects - Errors	(2) First Differences - Errors
Lagged Residuals,	0.642***	
	(0.019)	
Lagged Residuals,		-0.348***
		(0.023)
Constant	-0.000	0.000
	(0.001)	(0.001)
Observations	1,632	1,581
R-squared	0.413	0.123

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Therefore, we will run the FE-estimator again, with cluster-robust standard errors, to allow for correlation of errors within a cluster. This produces the same coefficient estimates but leads to larger standard errors.We will cluster the data by state, allowing for heterogeneity within a cluster, as none of the variables we are using in our regression are time-invariant within a state. The results of all three estimators are reported in **Table P.VI**.

For most of the regressors, both the FE and FD estimators produce coefficients with the same sign as the previous pooled OLS results (see **Table P.IV**), indicating the same direction of the relationship. In particular, 'children' is positively affected by 'sharemarried' ($\beta_2 > 0$) and negatively affected by 'lnincome' and 'sharewomen' ($\beta_1 > 0$ and $\beta_3 > 0$). Interestingly, the FE-estimator predicts a negative effect of the share of Hispanics ($\beta_4^{FE} = -0.148$), meaning that an increase of 1 in the share of Hispanics would decrease 'children' by approximately 0.148. This effect is statistically significant at the 1% significance level. However, when we allow for within-state correlation of the error terms, this coefficient is only significant at the 10% significance level. Similar to the results in **Table P.IV**, β_5, the coefficient for 'lnpop' is not statistically significant for any of the three estimators.

The results from FE and FD mainly differ in the scale of the effect estimated. For both income and the share of people married, the FD-estimator predicts a larger effect than the FE. Especially $\beta_2^{FD} = 0.346$, the coefficient for 'sharemarried' is nearly 3x as large as $\beta_2^{FE} = 0.148$. However, the FD estimate for the effect of the share of hispanics $\beta_4^{FD} = 0.034$ is not statistically significant and moreover indicates a positive effect, unlike $\beta_4^{FE} = -0.148$. Thus, the larger scale of the other coefficients might be because the FD estimates a smaller effect of the share of 'sharehispanic'.

Table P.VI: Fixed Effects and First Differences (+ Clustered Standard Errors) Regression Estimates, Children (Aged < 5) per Household

VARIABLES	(1) Fixed Effects	(2) First Differences	(3) FE with clustered standard errors
Log of median household income	-0.040***		-0.040
	(0.014)		(0.025)
Share of people married	0.148***		0.148**
	(0.043)		(0.071)
Share of Women	-1.086***		-1.086***
	(0.116)		(0.197)
Share of Hispanics	-0.148***		-0.148*
	(0.045)		(0.084)
Log of State Population	-0.002		-0.002
	(0.003)		(0.006)
Log of median household income = D,		-0.060***	
		(0.017)	
Share of people married = D,		0.346***	
		(0.046)	
Share of Women = D,		-0.767***	
		(0.114)	
Share of Hispanics = D,		0.034	
		(0.076)	
Log of State Population = D,		0.006	
		(0.009)	
Constant	1.215***	0.002	1.215***
	(0.157)	(0.005)	(0.265)
Time-Fixed Effect	Yes	Yes	Yes
Observations	1,683	1,632	1,683
R-squared	0.248	0.101	0.248
Number of States	51		51

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1
'Variable' = D indicates differenced variables used by the FD-estimator

Further, the statistical significance of income disappears when estimating FE with clustered standard errors, while only the share of women retains its status of high significance.

In conclusion, we cannot draw a precise conclusion on whether or not the lifetime income in a given state has an effect on fertility rates. Overall, it appears that in states and time periods where recorded median life-time income is higher, the number of children below age 5 per household is lower. This is also captured when estimating regressions using pooled OLS, Fixed-Effects, and First-Differenced estimators. However, once we allow for clustered standard errors, the statistical significance of this effect diminishes dramatically. We controlled for the share of people married, the share of women, the share of Hispanics, the population of a state and time (year).

REFERENCES

Aksoy, Cevat Giray. "The Effects of Unemployment on Fertility: Evidence from England" *The B.E. Journal of Economic Analysis & Policy*, vol. 16, no. 2, 2016, pp. 1123-1146. https://doi.org/10.1515/bejeap-2014-0127

Kristensen, A, Lappegård, T. (2022). *Unemployment and fertility: The relationship between individual and aggregated unemployment and fertility during 1994–2014 in Norway*. (4, 35) Demographic research, p. 1037-1064. DOI: 10.4054/DemRes.2022.46.35

Wooldridge, J. M. (2019) *Introductory Econometrics*. Mason, OH: Cengage.