Linear Regression Research

Yoohyun Kim

St. Paul Daechi Academy

Abstract

This paper is an in-depth examination of the fundamental concepts of linear regression and how to implement and evaluate linear regression to determine the relationship between birthrate and four factors: the ratio of female to male labor force participation rate, Gross Domestic Product, life expectancy, and Industry.

Introduction

The birth rate is one of the most fundamental and significant demographic indicators. However, its significance is not restricted to demographers alone. Birth rates influence governmental policy and financing for education and health systems, and can significantly affect the population's well-being (Zhang & Zhang, 2005). Governments, legislators, and the news media are concerned with birth rates because they are either too high or too low. Birth rates reveal a great deal about the health of a community and are a crucial metric in the field of health science as a whole. Health professionals regularly monitor the rise and fall of birth rates and measure these trends to track significant shifts and what they may indicate for the future of our civilization.

The relationship between national social and economic development, often known as "modernization," and national fertility rates has been subjected to a great deal of research in recent years. The guiding hypothesis of these studies, the Theory of the Demographic Transition, suggests that fertility rates decline in response to increasing development, whether measured by urbanization, the nonagricultural workforce, literacy, per capita income, or life expectancy (Abbasi et al., 2002). In general, the research has found substantial support for the basic thesis, although there are some disagreements over the details of the relationship such as whether cross-sectional relationships can be applied to longitudinal data and whether geographical areas or regions have unique patterns of fertility.

This paper examines in depth the fundamental concepts of linear regression and how to implement and evaluate linear regression to determine the relationship between birthrate and four factors: the ratio of female to male labor force participation rate, Gross Domestic Product, life expectancy, and Industry.

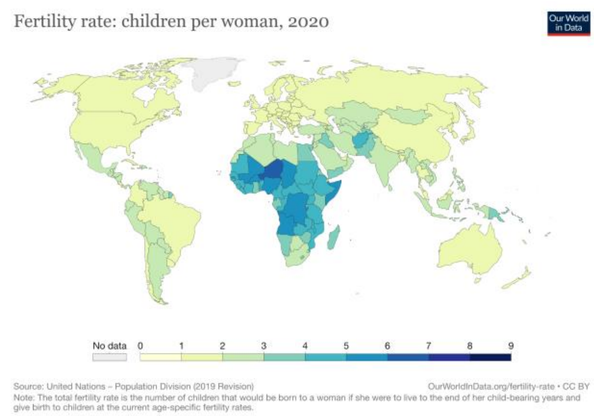
Literature Review

A. Fertility rate

This entry is concerned with the birth rate per woman in a population. Total Fertility Rate (TFR) – or simply 'fertility rate' – which estimates the average number of children per woman, is the most often employed metric. According to Roser (2017), the current average global fertility rate is just under 2.5 children per woman. In the past half-century, the worldwide fertility rate has decreased by half. And when cultures undergo modernization, the average number of children per woman reduces dramatically. Fertility rates of 4.5 to 7 offspring per woman were typical in pre-modern times. At that period, the extremely high infant mortality kept population growth low. As population health improves and mortality declines, we often observe rapid population growth. As the fertility rate drops and approaches 2 children per woman, this fast population expansion will cease to occur. The Total Fertility Rate is the metric utilized by demographers to assess offspring per parent. The TFR is defined as the average number of children a woman would have during her lifetime if she experienced the present age-specific fertility rates throughout her life. It is a metric that measures the fertility rate in a single year as opposed to throughout a woman's lifetime (Roser, 2017).

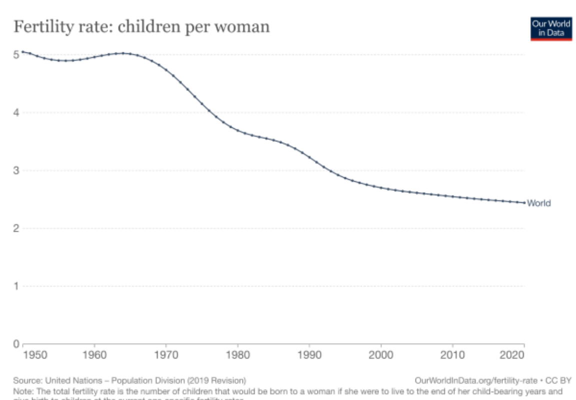
B. Decline of Fertility rate

The three primary causes are, in summary, the empowerment of women (growing access to education and labor market involvement), the drop in infant mortality, and the rising cost of raising children (to which the decline of child labor contributed). As a result of the dropping worldwide fertility rate, the global population growth rate has decreased from a high of 2.1% per year in 1968 to less than 1.1% per year presently



(Roser, 2017).

< Picture 1. 2020 Worldwide Fertility rate per woman >



< Graph 1. Fertility rate: children per woman>

An important factor for the decline in fertility is that women are delaying their first babies, thereby reducing their total childbearing years and increasing their likelihood of infertility (Jain and McDonald, 1997). While some women strive to extend their reproductive life with assisted reproductive technology, barely 2% of newborns are the result of such treatments (McDonald 2001a).

It appears that although women in their 20s tend to alter their family size preferences downward, they end up having fewer children than they desire or intend (de Vaus in this issue of Family Matters; McDonald 2001b) (Quesnel-Valee and Morgan 2002). Such patterns stand in stark contrast to those of roughly four decades ago when relatively ineffective methods of contraception frequently led to couples having more children than anticipated (Petersen 1961). The development and broad adoption of the contraceptive pill have had revolutionary implications. Several writers argue that labor market and economic changes influence fertility rates in part through their influence on life course patterns (Weston, 2022).

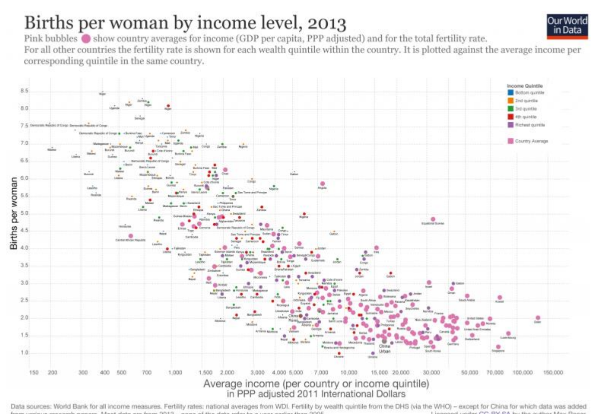
The steep decline in fertility in the 1890s is explained by the unexpected economic collapse that occurred in 1893 (Ruzicka and Caldwell, 1982). As a result of having grown up in a period of relatively continual and rapid technological, social, and cultural change, Mackay (1997) indicates that men and women born in the 1970s are more inclined to keep their options open. Ironically, a result of "drifting" is the closing of doors to having children. Teens and young adults must be well-informed on this topic. In addition, they require evidence that assures them that parenting does not preclude all other opportunities for a fulfilling life and that there will be a great deal of assistance from multiple sources (extended families, schools, neighborhoods, communities, and governments) in raising the next generation (Weston, 2022).

C. Relationship between fertility rate and economic conditions

It appears that the total fertility rate (TFR) in various OECD nations has tracked broad economic indicators such as unemployment quite closely. Due to the shift from one-earner families to dual-earner families, economic conditions now play a greater impact in determining fertility behavior than they did thirty years ago. The primary cause of this phenomenon is the decline in relative male pay, which encouraged women to participate more in the workforce. According to the New Home Economics Theory, the salary increases of women and men have distinct effects on fertility (see Becker, 1960, Mincer, 1963, Becker and Lewis, 1973, Mincer, 1962). Due to the effect of income, a rise in female earnings may boost fertility. In contrast, the substitution effect suggests that a gain in female salaries may result in a decline in fertility, as the opportunity cost of having children rises as a result of foregone earnings. In addition, if a woman is solely responsible for childrearing responsibilities, the increase in male pay has no influence on her income. Nevertheless, since women are gradually assuming a larger role as breadwinners in their households, the income effect now outweighs the substitution effect, resulting in a decline in the fertility rate (Örsal, 2010)

D. Correlation between GDP and fertility rate

The subsequent graph illustrates the close relationship between income (as defined by GDP per capita) and the total fertility rate. Not only are country averages of fertility rate and income displayed, but also the disparity within each country. Each population is divided into five quintiles, from the poorest 20% to the wealthiest 20%. When considering the fertility rate, many economists immediately refer to income as a likely driver. And indeed, across countries and throughout time, we observe a correlation between higher wages and reduced fertility. However, because richer countries are also healthier and better educated, this association between high wages and low fertility is by no means conclusive evidence that growing money is the cause of the decline in fertility. A population's higher level of education is a factor that correlates to more prosperity and fewer children. As we have seen, the second set of developments — technical change, decreased child labor, and structural change in the economy – accompany economic expansion and reduce parental demand for children. However, growing income may also have a direct effect on the reduced demand for children. Higher incomes allow for alternative, more diverse lives, which may persuade potential parents to have fewer or no children. Regarding the escalating prosperity in Europe during the previous century, historians George Alter and Gregory Clark wrote, "New products and new lifestyles in the expanding metropolitan societies spawned by the Industrial Revolution offered options." Rich families responded by consuming more of these new goods and services rather than procreating." This visualization reveals a correlation between parental income and the number of children they wish to have, which is partially driven by a direct relationship between the two variables. However, the association is also largely influenced by changes in prosperity and fertility rates (Roser, 2017).



< Graph 2: Births per woman by income level,2013>

Table 1. *Growth, fertility, school enrollment and saving in three types of countries*

Life expectancy at birth (no. of countries)	Mean of life expectancy 1960	Mean of life expectancy 1985	Fertility 1960–89	Investment/ GDP 1960–89	Secondary school enrollment 1960–89	Growth rate of per capita GDP 1960–89
≤50 (28)	43.4	55.0	6.44	14.1%	17.6%	1.4%
50 <...≤ 65 (24)	57.9	68.0	4.53	19.9%	38.8%	2.26%
>65 (24)	69.8	75.0	2.44	27.4%	71.0%	2.96%

< Table 1. Growth, fertility, school enrollment, and saving in three types of countries>

Similarly, causality may also extend from female labor force involvement to the fertility rate. In accordance with the role incompatibility theory, a rise in female labor force involvement may have a detrimental impact on reproduction. This can occur if females in the labor market delay having children due to the opportunity cost of having children, which can take numerous forms for women in full-time paid employment. First, there is a monetary cost if a woman must quit the job or convert from full-time to part-time employment to care for her kid or pay for daycare if she decides to continue working. Second, there is a less visible cost in the form of an interruption to the career path of the female, resulting in the loss of a greater possible future income stream and non-financial rewards, such as the prestige associated with a more senior position in her chosen field. A third consequence of having children is the loss of social networks in the workplace that provide an outlet outside the house if a woman leaves her job to care for her children at home(Roser 2017).

I. Linear Regression

A. Prediction and Linear Regression

General ideas of linear regression

- Regression analysis uses data to identify the correlations between variables and also it is utilized to assume predictions. We presume that the outcome we're forecasting is linearly related to the data we utilized to build it.
- The constant rate of rise of one variable in relation to another is referred to as linear dependence.

Examples

The first stage is to develop a model for determining a house's selling price based on its many qualities. The following is an example of a linear model:

Income = $\beta 0$ + $\beta 1$ (exchange rate) + $\beta 2$ (salaries) +$\beta 3$ (incentive) + $\beta 4$(taxes) + error

In this expression, $\beta 1$ represents the exchange rate for each additional dollar. B2 represents the salaries, $\beta 3$ represents the incentive that the employee earns and $\beta 4$ represents the taxes. The intercept $\beta 0$ could in theory be thought of as the constant money that the employee earns. The last term in the equation above, the "error," reflects the fact that two employees with the same position can have a different number of income sources. Even after specifying the values of a large number of variables, there is always some variability left over. An error term, which we shall regard as a random variable, captures this unpredictability. Regression gives us a method for computing estimates of the parameters $\beta 0$ and $\beta 1, \ldots, \beta 5$ from data about past incomes. Once we have these estimates, we can plug in the values of the variables for a new employee to get an estimate of one's income.

A regression model specifies a relation between a dependent variable Y and certain explanatory variables X1,...,XK. A linear model sets $Y = \beta 0 + \beta 1X1 + \cdots + \beta KXK + \varepsilon$.

Here, ε (the Greek letter epsilon) is the error term. To use such a model, we need to have data on values of Y corresponding to values of the Xi's.

Simple Linear Regression

1. A basic linear regression model has only one explanatory variable. Accordingly, a simplistic linear regression is based on the model Y = β0 + β1X + ε
2. We rarely have only one variable that can explain something, so we use multiple regression instead of simple regression. But it's easier to explain the most important ideas in a simple setting first.
3. .Let's start with a little example. A company wants to make sure that the salaries of purchasing managers in all of its divisions stay the same. As a general rule, it says that purchasing managers in different divisions who are in charge of similar budgets should get about the same pay. Figure 1 shows how much money 20 purchasing managers make and how big their budgets are.
4. There is a straight line fit to the data in Figure 1's scatter plot. The slope of this line tells you how much your salary will go up if you take on more budget responsibility.
5. Regression analysis allows us to take a more methodical approach. Regression also provides the optimal line across the data.
6. The least-squares criterion is used in regression, which we'll go through presently. Any line we come up with will have an intercept of 0 and a slope of 1. This line may pass through some of the data points, but it rarely passes through all of them. Let's name the data points based on their coordinates (X1, Y1), etc (X20, Y20). This is only a sample of the 20 pairings listed above. Our straight line indicates the wage level for the budget level Xi.

the budget level Xi, our straight line predicts the salary level : Yˆ i = β 0 + β 1 X i .

The projected value Yi will often differ from the actual value Yi unless the line passes through the point (Xi,Yi). The error, also known as the residual, is the difference between the two.

ei = Yi−Yˆi

= Yi−(β0+β1Xi)

To minimize the total squared errors, the least squares criteria picks 0 and 1.

$$\sum_{i=1}^{n} e_i^2,$$

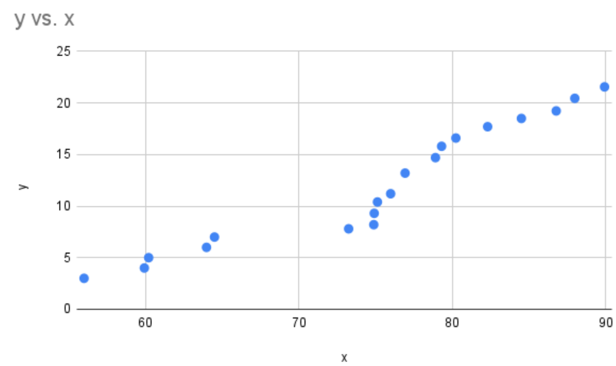
where n is the number of data points.

7. The estimated regression line always passes through the point (X,Y), and the estimated slope is given as a (non-obvious) consequence of this condition.

$$\hat{\beta}_1 = \frac{Cov[X,Y]}{StdDev[X]}.$$

Example 1. History and Scatter graph

x	y
56	3
59.93	4
60.21	5
63.98	6
64.5	7
73.243	7.8
74.88	8.2
74.91	9.3
75.12	10.4
75.98	11.2
76.921	13.2
78.9	14.7
79.3	15.8
80.23	16.6
82.3	17.7
84.5	18.5
86.77	19.23
87.98	20.45
89.92	21.56



<table 12>

Table - history

x: factor y: predict value

Y = x*a + b : linear graph

$$\sum$$

Y1 = x1*a + b, 3 = 56 * a + b

Y = ax^2+ b   y =2ax

Y = ax^2+bc^2+d

Extreme points can be found via deriviative. The value of the extreme point changes according to the coefficient at this point. If the coefficient is negative, it produces the greatest value; thus, the coefficient should be positive to bring the extreme point to the smallest value.

Minimizing the error rate is the most acceptable technique to locate a variable from an engineering viewpoint.

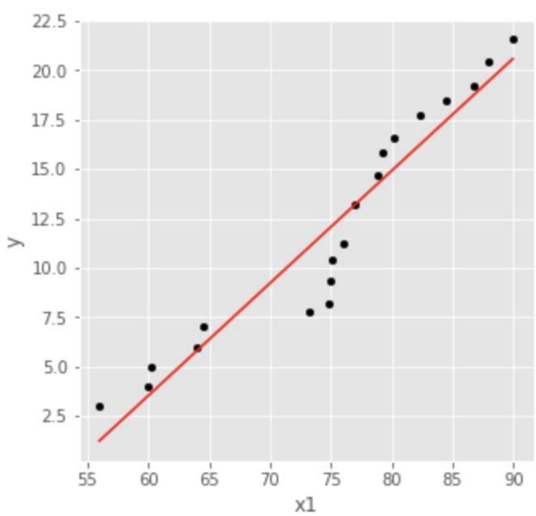
Here, we will use the sum notation since there are numerous historical events, therefore the variables are A and B, we are applying partial differentiation for A and B.

This equation stands for s, which stands for sum, while that minus y stands for error. Because there are circumstances when the value can be negative if the square is not applied, to make sure the value is positive, we apply the square.

s= sum (y hat - y) ^2

Y = xa + b yhat = x hat a + b

Example 2. Considering <table 12>, corresponding coefficient and intercept based on formula 1 and formula 2 are:



a = (1014.769354 / (108741.6266-(1425.574*1425.574)) =0.5699320879

b = ((229.64-(0.5699320879*1425.574))/19) = -30.67580876

The red line on the graph is a linear graph that has a coefficient value of 0.5699320879 and an intercept value of -30.67580876.

Y = 0.5699320879 x -30.67580876.

xample 3. Prediction

Let the factor value x =100, then the y' = 26.31740003

Example 4. Error

Let the factor value x = 56,

The history value y = 3

But the Predicted value y' = 1.2403881624

Then the error E = 1.7596118376

$\hat{yi}$ = axi + b

∑ ei^2 = ∑ (yi-axi -b) ^2

The first step: partial derivative with respect to b

(σ∑ ei^2 / σb) = 2∑ (-1) (yi - axa -b)

= -2 { ∑ yi - a∑ xi - ∑ b}

as $\sum_{i=1}^{n}$ [redacted]b = xb

= -2 { ∑ yi - a∑ xi - n b}

As we are considering (σ∑ ei^2 / σb) to be 0,

0 = { ∑ yi - a∑ xi - n b}

n b = ∑ yi - a∑ xi

∴ b = ∑ yi - a∑ xi / n ···· Formula 1

Assume that yi's average is (-y), b = (-y) - a (-x)

The second step : partial derivative with respect to a

(σ∑ ei^2 / σa) = (-1) 2∑ (y'i - axa -b) (xi)

0 = -2 { ∑ yixi - a∑ (xi)^2 -b∑ xi }

a∑ (xi)^2 = ∑ yixi- b∑ xi

a{∑ (xi^2) - (-x) ∑ xi^2} = ∑ yixi - (-y) ∑ xi

a = ∑ yixi - (-y) ∑ xi / ∑ (xi^2) - (-x) ∑ xi^2

B. Fitness test (R^2)

Goodness of fit test

Sum of Squares Total: SST

The coefficient of Determination should be between 0 to 1.

R^2 = SSR/ SST

SST = Sum of Squares Regression (SSR) + SSE

$\sum (yi - y')^2 = \sum (\hat{yi} - y')^2 + \sum (yi - \hat{yi})^2$

= 1- (SSE/SST)

= (a∑ yi+b∑ xiyi- n(y')^2)/ (∑ yi^2- n(y')^2)

= (130.879204665356 -559671.139746652444 - 2775.5015586)/(52734.5296-2775.50155860)

= -562315.762100587088 / 49959.02804

= -11.25553847

When SSE is near 0, it is ideal and it is ideal to be near 1 for R^2.

C. RMSE

Residuals and MSE

Residuals ei = yi -$\hat{y}$

Mean squared error: MSE

In order to write the equation of MSE, because Error is yi- $\hat{y}$ we should square it and sum the whole value and divide with n to solve the mean.

(∑(yi- $\hat{y}$)^2) / n

RMSE

$$\sqrt{(\sum (yi - \hat{y})^2) / n}$$

=$\sqrt{517.0762378/ 19}$

=$\sqrt{27.21453883}$

=5.216755585

We use RMSE instead of MAE because it relatively provides less error compared to MAE.

We have to reduce the influence of the error because the error value that exists outside the boundary can obscure the whole accuracy.

II. Evaluation

A. Results

X1: Ratio of female to male labor force participation rate (%) (modeled ILO estimate)

X2: GDP per capita (current US$)

X3: Life expectancy at birth, total (years)

X4: Industry (including construction), value added (% of GDP)

Y: Fertility rate, total (births per woman)

1. X1

R^2: y 0.000427

2. X2

R^2: y 0.260309

3. X3

R^2: y 0.432651

4. X4

R^2: y 0.000006

5. X1, x2

R^2: y 0.281185

6. X1, x3

R^2: y 0.432748

7. X1, x4

R^2: y 0.000459

8. X2, x3

R^2: y 0.492045

9. X2, x4

R^2: y 0.261427

10. X3, x4

R^2: y 0.435046

11. X1,x2,x3

R^2: y 0.498221

12. x1,x2,x4

R^2: y 0.281351

13. x1,x3,x4

R^2: y 0.43505

14. X2,x3,x4

R^2: y 0.495487

15. X1,x2,x3,x4

R^2: y 0.500454

The r2 value of life expectancy is 0.432651 when each component is tested with only one, indicating that life expectancy has a stronger impact on the fertility rate than other factors. And industry's contribution to the four variables is 0.000006. It is evident that it is the least significant and has no effect. When all four variables were examined, the r2 value was the greatest, which was an unexpected outcome. This figure suggests that these four factors interact with one another and influence the fertility rate.

Policy Suggestions

In developed nations with an elderly society, governments were more likely to implement measures aimed at increasing fertility.

The fertility rate has the strongest association among the four determinants with life expectancy, as shown by linear regression: R2 = y 0.432651. As a result of the advancement of contemporary medical technology, life expectancy is constantly rising. Contrary to the findings of earlier studies, the data indicated that the fertility rate increased as life expectancy increased. Consequently, it is vital to develop policies that correspond to it. Considering that the longer a person lives, the greater the expense to oneself, I deemed it prudent to provide financial assistance.

Publicly funded daycare, child or family allowances, paid or unpaid parental leave, and paid or unpaid paternity leave with job security were the four most often implemented policies. In addition, more than half of these governments offered additional incentives, such as flexible or part-time work hours for parents or child tax benefits.

Similarly, child benefits have been proven to influence fertility rates. It has been demonstrated that policies that alter the expense of having an extra kid have a causal influence on childbirth. Changes to these payments influenced the likelihood of having more than two children. Particularly, decreasing the monthly subsidy for an additional child by 2% of the average salary decreased the likelihood that a mother with two children would have a third child by approximately one percentage point every year, a pretty large reaction. The study indicated that changing the cost of having another child can be an effective method for governments to alter fertility.

A nation can provide financial assistance in the form of cash and tax credits; it can also promote employment flexibility by sponsoring parental leave and child-care programs and giving job safeguards for parents who choose part-time work. The Child Tax Credit in the American Recovery and Reinvestment Act gives the greatest Child Tax Credit ever and historic relief to the majority of working families, with the majority of families getting automatic monthly payments of $250 or $300 per child. The Child Tax Credit is available in monthly installments. For each kid aged 6 to 17, families received $250 per month, and for each child under the age of 6, families received $300 per month. Until the end of 2021, the 80% of taxpayers who get their refunds by direct deposit will continue to receive their payments on the 15th of every month. People who do not utilize direct deposit will get their payment by mail at the same time as those who do. Also, President Biden is adamant that the new Child Tax Credit should be extended for decades to come (The White House, 2022). This is what his Build Back Better Agenda promises.

Conclusion

In wealthy nations with an aging population, governments were more inclined to undertake fertility-boosting policies.

The fertility rate has the highest connection with life expectancy among the four factors, as demonstrated by linear regression: R2 = y 0.432651. It is essential to design policies that align with it. I thought it appropriate to offer financial support because the longer a person lives, the larger the expenditure. The birth rate is one of the most fundamental demographic indicators. The birth rate influences government policy and funding for education and health systems and may have considerable implications for the well-being of the population. The notion that birth rates are either too high or too low causes governments, lawmakers, and the news media to be worried about them. Birth rates disclose a great deal about a community's health and are an essential measure in the area of health research as a whole. Regularly, health experts observe the rise and fall of birth rates and quantify these patterns to detect important adjustments and to determine what they may portend for the future of our civilization.

References

1. Abbasi, M. J., Mehryar, A., Jones, G., & McDonald, P. (2002). Revolution, war and modernization: Population policy and fertility change in Iran. *Journal of population research*, *19*(1), 25-46.
2. Buckles, K., Hungerman, D., & Lugauer, S. (2021). Is fertility a leading economic indicator? *The Economic Journal*, *131*(634), 541-565.
3. Jain, S. K., & McDonald, P. F. (1997). Fertility of Australian birth cohorts: components and differentials. *Journal of the Australian Population Association*, *14*(1), 31-46.
4. McDonald, P. (2000a), 'The 'toolbox' of public policies to impact on fertility - A global view", Paper prepared for the Annual Seminar 2000 of the European Observatory on Family Matters, Low Fertility, Families and Public Policies, Sevilla (Spain), 15-16 September.
5. McDonald, P. (2001b), "Theory pertaining to low fertility", Paper presented at the International Union for the Scientific Study of Population Working Group on Low Fertility, Tokyo, 21-23 March
6. McDonald, P. (2002). Low fertility: unifying the theory and the demography.
7. Orsal, D. D. K., & Goldstein, J. R. (2010, April). The increasing importance of economic conditions on fertility. In *Annual Meeting of the Population Association of America, Dallas, Texas*.
8. Roser, M., (2017) - "Fertility Rate". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/fertility-rate' [Online Resource]
9. Ruzicka, L. T., Caldwell, J. C., & Caldwell, J. C. (1982). *The end of demographic transition in Australia* (p. 296). Canberra: Department of Demography, Australian National University.
10. The White House. (2021, June 21). *Statement by president Joe Biden on child tax credit awareness day*. https://www.whitehouse.gov/briefing-room/statements-releases/2021/06/21/statement-by-president-joe-biden-on-child-tax-credit-awareness-day/
11. Weston, R., & Parker, R. (2002). Why Is the Fertility Rate Falling?: A Discussion of the Literature. Family Matters, (63), 6–13.

https://search.informit.org/doi/10.3316/informit.751994546976074

12. Zhang, J., & Zhang, J. (2005). The effect of life expectancy on fertility, saving, schooling and economic growth: theory and evidence. *Scandinavian Journal of Economics*, *107*(1), 45-66.