Gender Demographics and Income Inequality in the United States (1991–2017)

Ellson Lin

Student (BSc, 2023-26) Dept. of Statistical Science, UCL, UK;
ellson.lin.23@ucl.ac.uk

Abstract

This project examines the relationship between demographic structure and income inequality in the United States between 1991 and 2017. Using annual data from Our World in Data and ordinary least squares regression, income inequality is measured by the Gini coefficient and regressed on the female share of the population, controlling for education expenditure, income concentration, taxation, and labour market conditions. The results suggest that higher government expenditure on education is significantly associated with lower income inequality, while greater income concentration at the top 1% is strongly associated with higher inequality. The coefficient on female population share is negative but statistically insignificant, likely reflecting limited within-country variation and the indirect nature of the proxy. The analysis is descriptive rather than causal and is intended to demonstrate applied data handling, regression analysis, and economic interpretation.**Keywords:**
Applied Econometrics; Income Inequality; Gender Ratio; Socioeconomic Data; Regression Analysis; Education Policy; United States

1. Introduction

The relationship between gender ratio and income inequality has gained increasing interest, as the effects that gender disparities in a country's population structures can have profound economic consequences, influencing labour market dynamics, wage distributions and overall income inequality. Understanding this relationship is essential for designing policies that promote inclusive economic growth and gender equity. This project explores the effect of gender ratio on income inequality using Stata 18 and data from Our World in Data from January 01st, 1991 to December 31st, 2017.

2. Data and Variables

Empirical Strategy

	Observation	Mean	Standard Deviation	Min	Max	Type of variable
Year	27	2004	7.937254	1991	2017	
Government expenditure on education % (ES)	27	16.30276	0.7954053	15.10283	17.98815	Control variable
Gini coefficient (gini)	27	0.5536687	0.21971	0.5114148	0.5850735	Dependent variable
GDP per capita (PPP)	27	56114.51	6785.76	43742.03	66458.02	Control variable
Female_Share (FS)	27	50.27188	0.2873486	49.79057	50.8138	Independent variable
Top 1% share (PT)	27	16.95704	1.719181	13.65	19.49	Control variable
Top marginal income Tax rate (MT)	27	37.40741	2.990488	31	40	Control variable
Unemployment Rate_Female (UR_F)	27	5.802222	1.308822	4.099	8.616	Control variable
Unemployment Rate_Male (UR_M)	27	6.184333	1.812648	3.901	10.515	Control variable
Residual(R)	27	-4.85e-12	0.0018228	-0.0049804	0.0042244	Control variable

Table 1:Statistics of variables

In our regression, the dependent variable income inequality is represented using the *Gini coefficient (GC)*. This measures the uneven distribution of income among individuals or groups within a society, country, or the world. The higher the Gini coefficient from 0 to 1 suggests greater income inequality in the country. The independent variable is *Female Share (FS)* which is the proportion of women in the US.

The control variables include:

GDP per Capita (PPP) represents the standard of living in the US, which will isolate the effect of gender ratio on income inequality.

Top 1% share (PT) is the proportion of income held by the top 1% in the US, with a higher share meaning that most income is concentrated among the top 1%. Using this as a control variable will separate gender ratio from income concentration.

Top marginal income Tax rate (MT): The highest % of income tax paid that applies to the last dollar earned in a progressive tax system. This variable is included to separate the effects of the tax system from gender ratio.

Unemployment rate female (UR_F) and Unemployment rate male (UR_M): Represents the proportion of female and male who are willing and able to work and are actively looking for work, but are unable to find a job. Using this as a control variable will mitigate the influence on the different unemployment rates.

Gini coefficient	Coefficient	Standard Error	T - statistic
Female_Share(FS)	-0.0310793	0.0166772	-1.86
GDP Per Capita(PPP)	-6.67e-07	7.13e-07	-0.94
Government expenditure on education % (ES)	-0.0035421	0.0010195	-3.47
Top 1% share (PT)	0.0103767	0.0009133	11.36
Top marginal income Tax rate (MT)	0.0005078	0.0002042	2.49
Unemployment Rate_Female (UR_F)	0.0008142	0.0023022	0.35
Unemployment Rate_Male (UR_M)	-0.0002839	0.0013889	-0.20

Table 2: Results

The Multiple Regression Model

When we evaluated the data, we noticed that the distribution does not have an obvious skewness. So it was unnecessary to use a log model, as given the properties of logarithms. Given the limited sample size and the absence of strong skewness in the variables, the analysis employs a linear specification rather than a logarithmic transformation.

Therefore the regression we use is:

$$gini_i = \beta_0 + \beta_1 FS_i + \beta_2 PPP_i + \beta_3 ES_i + \beta_4 PT_i + \beta_5 MT_i + \beta_6 UR\_F_i + \beta_7 UR\_M_i + \varepsilon_i$$

The estimated regression using values from Table 2 is:

$\widehat{gini}_i$ = 2.013332 - 0.0310793FS_i - 6.67e-07PPP_i - 0.0035421ES_i + 0.0103767PT_i + 0.0005078MT_i + 0.0008142UR_F_i - *0.0002839*$UR\_M_i + \varepsilon_i$

3. Interpretation and Discussion

Cetris paribus, an additional *female share* in the population decreases the *gini coefficient* by 0.03107.This suggests that an increase in women in the population will lead to a greater female labour force participation and improved wage equality which will lead to lower income inequality, however this is statistically insignificant (p-value = 0.078) . "Furthermore, a one-unit increase in government expenditure on education decreases the Gini coefficient by 0.0035421. This effect is statistically significant (p =0.003), indicating strong evidence that increased investment in education is associated with reduced income inequality."This could mean that greater investment in education could mean greater social mobility, which could reduce income inequality. While the *GDP per capita* coefficient of -6.67e-07 suggests a negligible effect on income inequality, it is also statistically insignificant (p=0.361), indicating that per capita income alone does not have a meaningful impact on inequality in this model. When the *top 1% share* increases, income inequality rises by 0.0103767. This suggests that the concentration is a strong predictor of income inequality requiring policies which target wealth accumulation to reduce income inequality. However, higher *top marginal income tax rate* has a positive coefficient, which suggests that the wealthy will attempt to find tax loopholes to reduce the amount of tax to pay. Finally, *unemployment rate male* and *unemployment rate female* are both statistically insignificant, meaning both figures have no clear impact on inequality.

Data Visualization

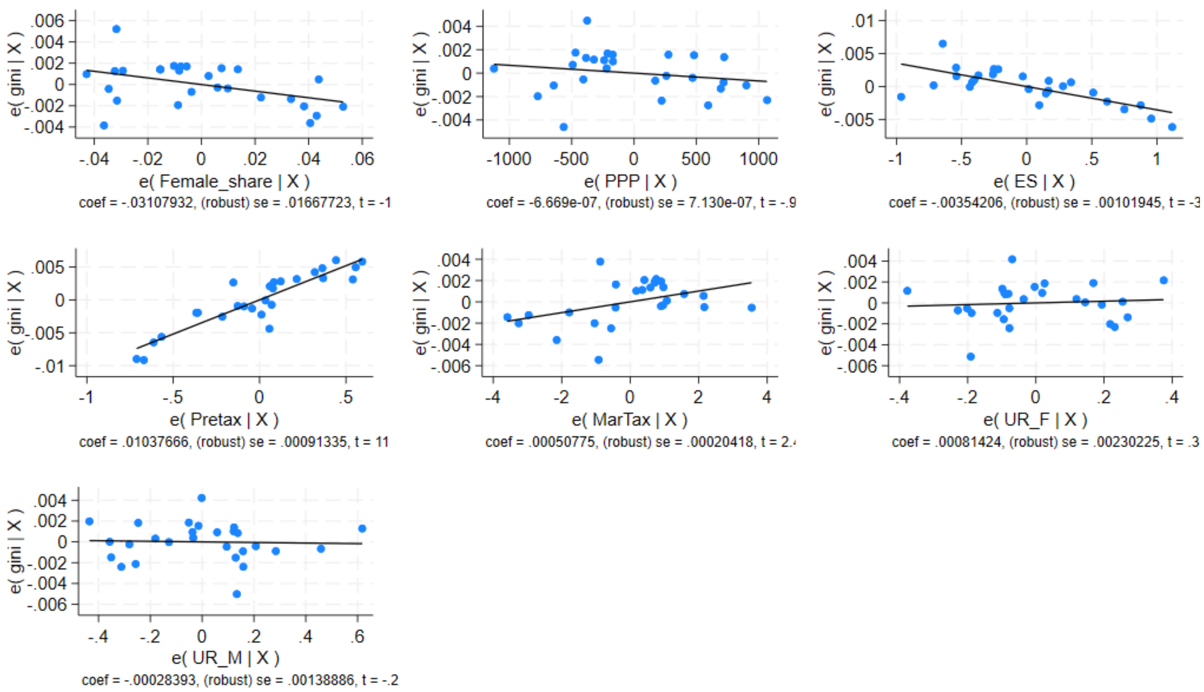
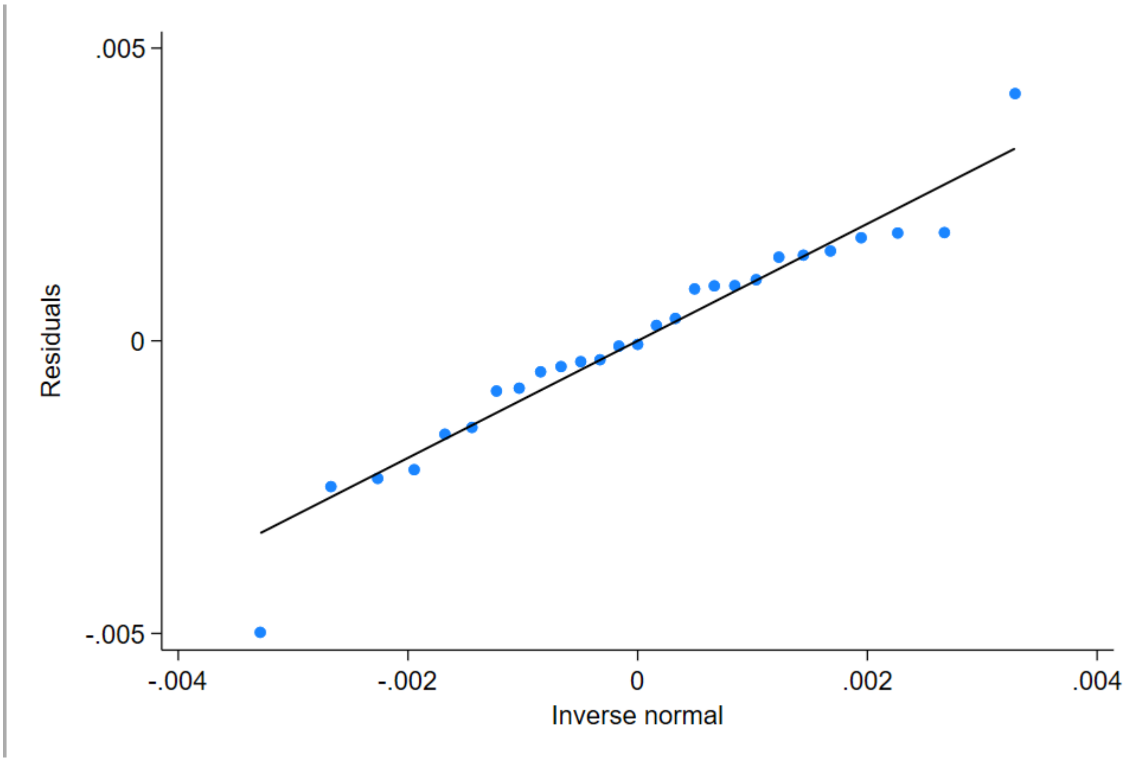
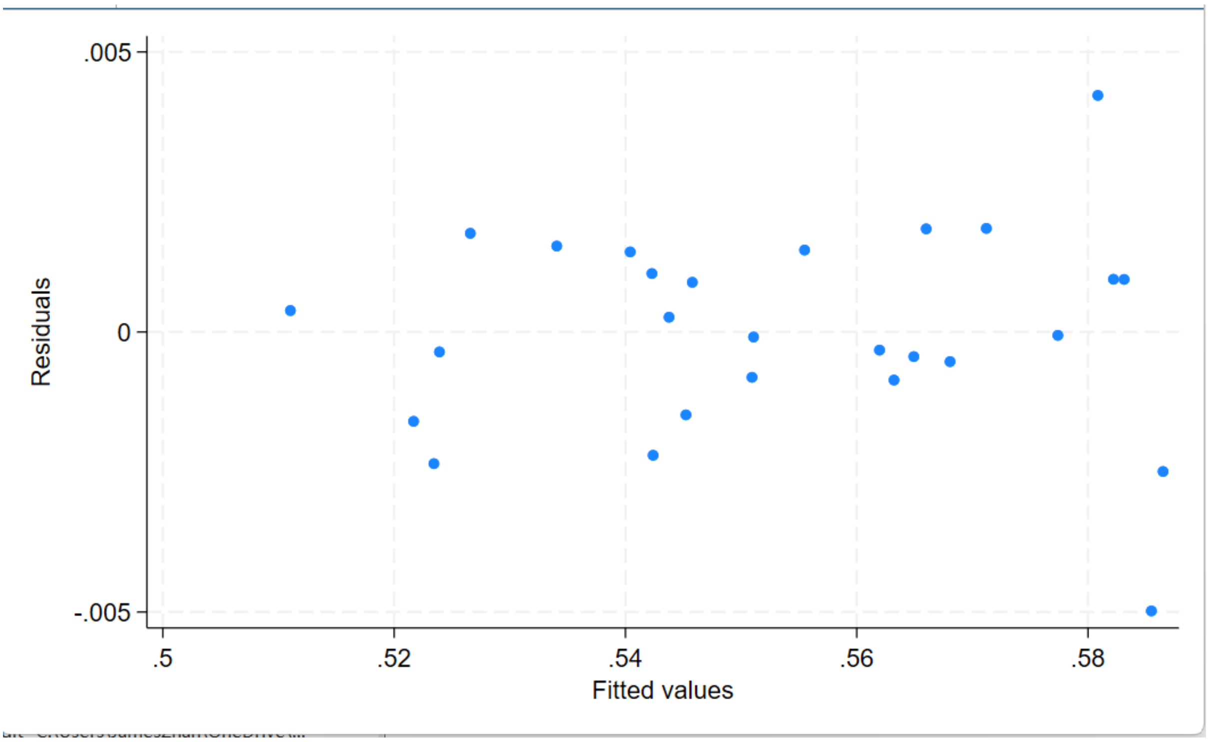


Figure 1: Relationships between gini coefficient and independent variables

The R-Squared value we obtained is 99.31%, indicating 99.31% of the variance in gini coefficient is explained by the independent variable, though this high R^2 should be interpreted cautiously given the small sample size and potential overfitting. However there is limitation on the model, as the gini coefficient could possibly be impacted by other factors outside of the model, such as technological advancement, as if the country is highly developed, low-skill jobs could be replaced by AI and automation, which would cause more inequality. Another limitation is that we have only got 27 observations that fit our model, this is indicating less significance on our data as low observation could lead to inaccurate data.



As shown by the QQ plot above, it fits almost perfectly, indicating the data is normally distributed and no significant skewness arriving. However, the low observation is still challenging the accuracy of the data.



As shown by the plot above,The residual plot below shows that the assumption of linearity is mostly satisfied as the standardised residuals are randomly dispersed around 0. However, the value of residuals somehow expands as fitted values get bigger, this could lead to an issue of homoscedasticity.

4. Conclusion

This study examines the relationship between *gender ratio* and *income inequality* from 1991 to 2017 using multiple regression analysis. Our findings indicate that a greater female share in the population decreases the *gini coefficient* by 0.03107, suggesting that it could lead to greater female participation in the labour force and improved wage equality. However this effect is marginally significant ($p = 0.078$) suggesting that other factors influence this relationship. Other factors to be considered include *government expenditure on schooling* which has a statistically significant negative effect on inequality, which suggests greater investment on education will increase social mobility while greater concentration of 1% share in income is a strong indicator of inequality.

Given the R^2 value is 99.31% this suggests that 99.31% of the income inequality is predicted by the independent variable, suggesting that our data fits the model significantly well.

Appendix

Variable name	Storage type	Display format	Value label	Variable label
Country	str13	%13s		Country
Code	str3	%9s		Code
Year	int	%10.0g		Year
ES	double	%10.0g		ES
gini	double	%10.0g		gini
PPP	double	%10.0g		PPP
E	byte	%10.0g		
Female_share	double	%10.0g		Female_share
Pretax	double	%10.0g		Pretax
MarTax	byte	%10.0g		MarTax
F	byte	%10.0g		
UR_F	double	%10.0g		UR_F
UR_M	double	%10.0g		UR_M

```
Time variable: Year, 1972 to 2020
        Delta: 1 unit

. *select the year
. keep if Year >= 1991 & Year <= 2017
(22 observations deleted)

. *mutiple regression model
. regress gini Female_share PPP ES Pretax MarTax UR_F UR_M, robust

Linear regression                               Number of obs     =         27
                                                F(7, 19)          =    1663.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.9931
                                                Root MSE          =     .00213
```

gini	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
Female_share	-.0310793	.0166772	-1.86	0.078	-.0659852	.0038265
PPP	-6.67e-07	7.13e-07	-0.94	0.361	-2.16e-06	8.25e-07
ES	-.0035421	.0010195	-3.47	0.003	-.0056758	-.0014083
Pretax	.0103767	.0009133	11.36	0.000	.008465	.0122883
MarTax	.0005078	.0002042	2.49	0.022	.0000804	.0009351
UR_F	.0008142	.0023022	0.35	0.727	-.0040044	.0056329
UR_M	-.0002839	.0013889	-0.20	0.840	-.0031909	.002623
_cons	2.013332	.8913219	2.26	0.036	.1477737	3.87889

Variable name	Storage type	Display format	Value label	Variable label
Country	str13	%13s		Country
Code	str3	%9s		Code
Year	int	%10.0g		Year
ES	double	%10.0g		ES
gini	double	%10.0g		gini
PPP	double	%10.0g		PPP
E	byte	%10.0g		
Female_share	double	%10.0g		Female_share
Pretax	double	%10.0g		Pretax
MarTax	byte	%10.0g		MarTax
F	byte	%10.0g		
UR_F	double	%10.0g		UR_F
UR_M	double	%10.0g		UR_M

Reference

Our World in Data (2021). GDP per capita. [online] Our World in Data. Available at: https://ourworldindata.org/grapher/gdp-per-capita-worldbank.

Ortiz-Ospina, E., Tzvetkova, S. and Roser, M. (2018). Women's Employment. [online] Our World in Data. Available at: https://ourworldindata.org/female-labor-supply.

Ortiz-Ospina, E. and Roser, M. (2023). Government Spending. [online] Our World in Data. Available at: https://ourworldindata.org/government-spending.

Ortiz-Ospina, E. and Roser, M. (2013). Taxation. [online] Our World in Data. Available at: https://ourworldindata.org/taxation.