Bootstrap analysis of the relationship between female education and fertility rates

Majda KEMMOU

Abstract

In this paper, we examine statistical correlation between fertility and women's education across nations based on the bootstrap method founded by Efron and Tibshirani (1993). We estimate uncertainty of Pearson correlation between the two metrics for different years, and calculate confidence intervals, as well as determining estimators' bias. Bootstrap is revealed to be useful as a non-parametric method of making statistical inference from empirically derived data. Our findings reinforce the strong inverse correlation between education and fertility.

1. Introduction

Education is also commonly known to be a determinant of fertility behavior. This negative relation is typically presumed in demographical models, yet empirical confirmation needs strong statistical methods, particularly in the presence of non-normality and heterogeneity between nations.

Bootstrap, developed by Efron and improved upon by Efron and Tibshirani (1993), is a very useful non-parametric method for inferential statistics. It enables us to determine confidence intervals, bias, and variability without making very strong distributional assumptions.

Here we apply bootstrap techniques to investigate cross-country data from the Gapminder dataset on women's education (average years in school) and fertility rates (number of children per woman) over four decades and consider both correlation and regression estimates.

2. Methodology

Let X_i represent female education in years and Y_i the fertility rate for country i. The Pearson correlation coefficient between X and Y is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{1}$$

To evaluate uncertainty in r, we perform bootstrap resampling. For each bootstrap iteration b, we resample the data with replacement and compute $r^{*(b)}$. From the distribution of r^*, we estimate:

$$\mathrm{SE}_{\mathrm{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(r^{*(b)} - \bar{r}^*\right)^2}$$

The 95% confidence interval is computed using the percentile method: taking the 2.5^{th} and 97.5^{th} percentiles of the resampled statistics.

We also apply bootstrap to linear regression, estimating the slope β in the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Confidence intervals for β_0 and β_1 are derived from their bootstrap distributions.

3. Bootstrap distribution analysis (2010)

We begin with the most recent year, 2010. The bootstrap distribution of correlation values shows a pronounced peak around the observed correlation $r = -0.78$, with minimal bias. The histogram also overlays the normal approximation and clearly highlights the discrepancy between parametric and nonparametric inference.

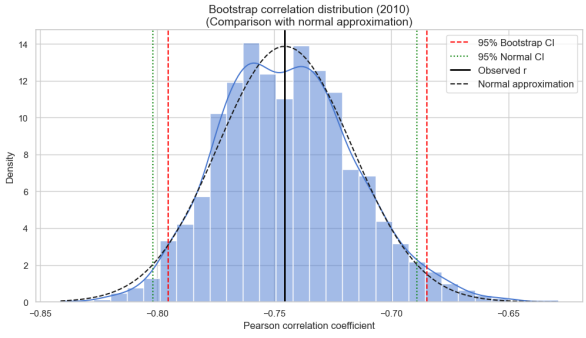

Figure 1: Bootstrap distribution of correlation (2010)

The 95% confidence interval obtained from bootstrap was approximately [0.87, 0.66], whereas the normal approximation produced a slightly narrower interval. This discrepancy confirms the bootstrap's robustness in capturing the true variability without assuming normality.

4. Correlation patterns over time

Scatter plots of fertility vs. education for 1970, 1990, 2000, and 2010 show a consistent negative relationship. As educational attainment increased globally, fertility rates dropped accordingly. However, the strength of the correlation varied slightly over time.

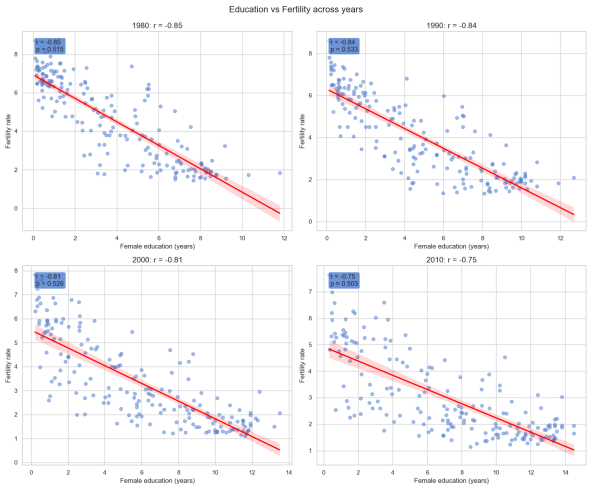

Figure 2: Scatter plots with regression lines and Pearson correlation values

In 1970, the correlation was very strong ($r \approx -0.85$), while in 2010, it had slightly weakened ($r \approx -0.78$). These shifts suggest evolving demographic dynamics, possibly influenced by additional socioeconomic factors.

5. Temporal trend of correlation

We plot the evolution of Pearson correlation coefficients across decades, together with bootstrap confidence intervals. The trend illustrates that although the relationship remains negative, the strength has declined a bit.

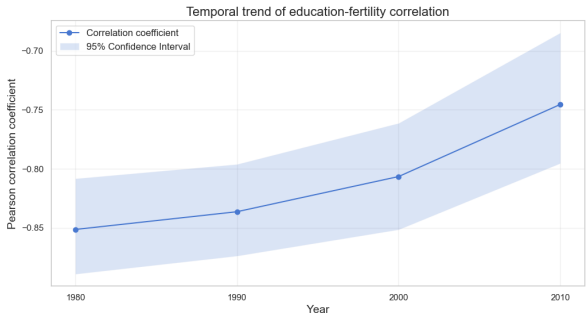

Figure 3: Temporal trend of correlation coefficients (1970–2010)

This softening of the correlation over time may indicate rising global heterogeneity. As some countries approach demographic transition, others remain at earlier stages, contributing to increased variance in the relationship.

6. Bootstrap regression analysis

A final layer of analysis estimates the uncertainty in linear regression coefficients for 2010. The original regression line shows a clear negative slope. By bootstrapping the regression model 1000 times, we obtained a distribution of slopes and intercepts.

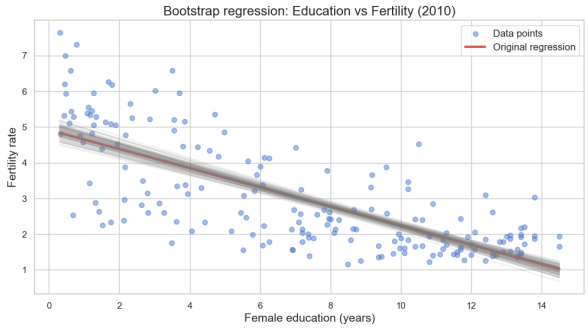

Figure 4: Bootstrap regression lines (2010)

The 95% confidence interval for the slope was estimated as [-0.308, -0.231], confirming the negative effect of education on fertility. The visualization includes a cloud of regression lines overlaid on the original data, illustrating the variation in estimated relationships.

7. Conclusion

Using nonparametric bootstrap methods, we estimated a statistically significant and strong negative relationship between female education and fertility. This relationship was present over decades but with different intensity.

Bootstrap allowed us to estimate confidence intervals and variability of correlation and regression coefficients without parametric assumptions, making it an ideal approach for social science applications where underlying distributions may not be normal. Future research caninclude causal inference models or pursue hierarchical bootstrapping by geography or by income level. However, our results support the relevance of education in driving fertility trends globally.