Data Analysis Project

Yi Chen

Teachers College, Columbia University

Linear Model & Experimental Design

Data Analysis Project

# Introduction

The purpose of this project is to apply the linear regression and experimental design methodologies to explore the meaningful research questions with a real data set. In this proposal paper, I will describe the data set and primary research questions for this project. Meanwhile, I will briefly describe the research hypotheses, visualization tools, and statistical methods I will use.

# Data Set

The data set that I will use for this project is based on the study of Mauldin and Berelson (1978). In their study, they did a macro-analysis of the correlates of fertility decline in developing countries during the period of 1965 to 1975. Geographically, this dataset covered 94 developing countries in Asia, Americas, and North Africa, where there was fertility decline to different extent. The original dataset is one of the most popular and best available data sources for the research of worldwide fertility rate.

To simplify the analysis and focus on my research question, I did not contain the whole data source in this project, which seems to be too complex and even misleading. Instead, I pick 20 countries in Latin America as representatives. I have to recognized that small sample size could impact the robustness and reliable of the results from statistic models. But this relatively small data set is already good enough for illustrating the statistic models, and providing insightful findings. The reproduced dataset is shown in Table 1. The variable in this table including:

* Index of social setting. The index of social setting combines seven social indicators, namely literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15–64 in the non-agricultural labor force, gross national product per capita and percent of population living in urban areas. Higher scores represent higher socio-economic levels.
* The percent decline in the crude birth rate (CBR). It is based on the number of births per thousand population between 1965 and 1975. This data mainly come from the United Nations sources and from vital statistics for individual countries. In this project, we take it as continuous variable. Mauldin and Berelson (1978) motioned in their paper that we can transfer it into a binary variable based on whether the value is bigger than 5. Because the crude birth rate is flawed by the normal technical difficulties of data collection, less than 5 percent should be taken to indicate essentially no change.
* The index of family planning effort. It combines 15 different program indicators, including the existence of an official family planning policy, the availability of contraceptive methods, and the structure of the family planning program. An index of 0 denotes the absence of a program, 1–9 indicates weak programs, 10–19 represents moderate efforts and 20 or more denotes fairly strong programs.

In this project, I will treat the percent decline in the CBR as a continuous response and the indices of social setting and family planning effort as predictors.

Table 1: The Fertility Decline Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Countries | Settinga | Effortb | CBR Declinec | Location |
| Bolivia | 46 | 0 | 1 | South America |
| Brazil | 74 | 0 | 10 | South America |
| Chile | 89 | 16 | 29 | South America |
| Colombia | 77 | 16 | 25 | South America |
| Costa Rica | 84 | 21 | 29 | Central America |
| Cuba | 89 | 15 | 40 | Caribbean |
| Dominican Rep | 68 | 14 | 21 | Caribbean |
| Ecuador | 70 | 6 | 0 | South America |
| El Salvador | 60 | 13 | 13 | Central America |
| Guatemala | 55 | 9 | 4 | Central America |
| Haiti | 35 | 3 | 0 | Caribbean |
| Honduras | 51 | 7 | 7 | Central America |
| Jamaica | 87 | 23 | 21 | Caribbean |
| Mexico | 83 | 4 | 9 | North America |
| Nicaragua | 68 | 0 | 7 | Central America |
| Panama | 84 | 19 | 22 | Central America |
| Paraguay | 74 | 3 | 6 | South America |
| Peru | 73 | 0 | 2 | South America |
| Trinidad-Tobago | 84 | 15 | 29 | South America |
| Venezuela | 91 | 7 | 11 | South America |

Data source: Mauldin and Berelson (1978).

a  Index of social setting; b  The index of family planning effort; c  The percent decline in the crude birth rate.

# Research Questions & Hypothesis

Research question in this project is: how much of the fertility decline is associated with the social-economics context (index of social setting) and family background (index of family planning effort). My hypothesis is both index of social setting and index of family planning effort has a significant main effect on fertility decline. While index of social setting is more significant. There is also an interaction effect between index of social setting and index of family planning effort.

# Research Result

In order to explore different statistical methods, in a first approach to the data, I will treat the predictors as continuous covariates with linear effects (for linear regression). Later I will group them into categories and treat them as discrete factors (for ANOVA and ACOVA)

This data analysis will cover the components including:

* Explorational data analysis and descriptive data summary. For example, correlation analysis and boxplot for each variable.
* OLS regression. This section will also include OLS estimation, model diagnosis, and model comparations.
* ANOVA analysis. This section will also include dummy coding, estimate interaction, simple, and main effect.
* ACOVA analysis.
* For the constrains: I will compare the CBR Decline for the countries in the south America and the countries not in south America (north America, Caribbean, and central America)

# Conclusion

References

Mauldin, W. P., Berelson, B., & Sykes, Z. (1978). Conditions of Fertility Decline in Developing Countries, 1965-75. *Studies in Family Planning*. https://doi.org/10.2307/1965523

# Appendix