Factors affecting Swiss fertility based on socio-economic index data sets[*]

Jiayi Du

24 April 2022

Abstract

Switzerland was in a period of population transition in 1888, and its fertility began to decline. In order to study whether the decline of fertility rate is affected by socio-economic indicators, this experiment used the dataset of fertility and socio-economic indicators of 47 provinces in Switzerland named Swiss to analysis. The mathematical statistical analysis and multiple linear regression had been carried out. The results indicated that the fertility was mainly related to the proportion of men engaged in agriculture, education, religious belief and the mortality of infants.

Keywords: Swiss fertility, socio-economic, linear regression

1 Introduction

Different aspects of the Switzerland society has entered a relatively stable stage in late 19th century, such as its economy and politics. However, it does not mean that the Swiss haven't been experiencing changes that drastically affect their lifestyles. Factors including social policies and technology advancements have greatly shift the Swiss's ideology, values, and one of the most common change is their attitude towards fertility. Social development is closely related to social population structure and growth, and the development of social population will be affected by social and economic development. Fertility rate is a very important part of the population index, and it will also be affected by social and economic indicators.

The Swiss data conducted a survey on the topic of fertility in 1888. By utilizing the data collected from the survey, this paper aims to explore the different possible factors that affect Swiss fertility, such as the proportion of men engaged in agriculture, education, religious belief and the mortality of infants and so on. The survey data is thoroughly discussed in the section of 2. In the sub-section 2.1, we presented an overview of the original survey data, and explained our cleaned dataset that we'll for exploration. The methods used to collect the Swiss data as well as the strengths and weaknesses associated with these methods are outlined in the sub-sections, 2.2, and 2.3. Section 3 shows the general linear regression model which is considered in this paper. Section 4 presents a series of findings in relation to how the different indicators we picked have impacts on . In section 5, a discussion is made to elaborate on the implications of the findings we've got from the survey data. Furthermore, we talked about the possible reasons that lead to the effect of these factors on the decline of fertility.

2 Data

2.1 Dataset of interest

The survey we utilized in this paper comes from Project "16P5" in the paper by Mosteller, F. and Tukey, J. W. (1977) (Mosteller and Tukey 1977). The data collected are for 47 French-speaking provinces in Switzerland at about 1888. The dataset contains 6 variables, including Fertility, Agriculture, Examination, Education,

[*] Code and data are available at: https://github.com/dujiayi1/analysis-based-on-Swiss-fertility-and-socio-economic-index-data-sets.

Catholic and Infant.Mortality. Among them, the data of Examination and Education were the averages of 1887, 1888 and 1889. The data set is 47×6 in size, and all variables are scaled to [0, 100], where in the original, all but "Catholic" were scaled to [0, 1].

This paper focused on investigating 6 of these variables: Fertility, Agriculture, Examination, Education, Catholic and Infant.Mortality. R (R Core Team 2021), and R packages "tidyverse" (Wickham et al. 2019), "knitr" (Xie 2021), "dplyr" (Wickham et al. 2021), , "ggplot2"(Wickham 2016), "kableExtra" (Zhu 2021),"lattice"(Sarkar 2008),"PerformanceAnalytics"(Peterson and Carl 2020),"corrgram"(Wright 2021) and "glvma"(Pena and Slate 2019) are utilized to conduct linear regression on the dataset.

Table 1: Extracting the first ten rows from the cleaned Swiss dataset

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	10.0	22.2
Delemont	83.1	45.1	6	9	84.8	22.2
Franches-Mnt	92.5	39.7	5	5	93.4	20.2
Moutier	85.8	36.5	12	7	33.8	20.3
Neuveville	76.9	43.5	17	15	5.2	20.6
Porrentruy	76.1	35.3	9	7	90.6	26.6
Broye	83.8	70.2	16	7	92.8	23.6
Glane	92.4	67.8	14	8	97.2	24.9
Gruyere	82.4	53.3	12	7	97.7	21.0
Sarine	82.9	45.2	16	13	91.4	24.4

Table 1 shows the first ten rows of the Swiss dataset of interest. Variable "Fertility" indicates the fertility rate in the 47 French-speaking provinces of Switzerland. Variable "Agriculture" represents the proportion of men engaged in Agriculture. Variable "Examination" means the proportion of conscripts who got high marks in the army Examination. Variable "Education" indicates the proportion of conscripts who had Education beyond primary school. Variable "Catholic" means the proportion of Catholicism and Infant.Mortality represents the rate of infant mortality.

2.2 Strengths

The response rate of the survey is very high so that the data is relatively accurate. In addition, the value of each variable is scaled to proportion so that it's numeric and it's of benefit for us to conduct analysis.

2.3 Weaknesses

The dataset contains only 47 pieces of data, so that it may lead to some errors in the results. Besides, some questions which appeared to be quite sensitive to many respondents may result in high non-response rates. For instance, respondents were generally unwilling to provide information about their examination.

2.4 Summary Statistics

In this part, I calculated some common statistics for each variable, including minimum, maximum, first quartile, median, third quartile, mean, variance, and standard deviation. The specific results are as follows:

As can be seen from Table 2, Fertility indicates a minimum of 35, a first quartile of 64.7, a median of 70.4, a third quartile of 78.45, a maximum of 92.50, a mean of 70.40, a variance of 156.04, and a standard deviation of 12.49. The same goes for the description of other variables.

2.5 Distribution of each variable

From the Figure 1, the variables Fertility, Education, and Infant.Moritality all have outliers, which will affect the subsequent analysis to some extent, such as the accuracy of model fitting. Therefore, detection and processing of outliers are extremely important.

Then I made the histogram of each variable. From Figure 2, It's obvious that the histogram distribution of Fertility, Agriculture and Examination shows that there are more in the middle and less on both sides. The

Table 2: Common statistics of each variable

variable	min	Q1	median	Q3	max	mean	var	sd
Fertility	35.00	64.70	70.40	78.45	92.5	70.14	156.04	12.49
Agriculture	1.20	35.90	54.10	67.65	89.7	50.66	515.80	22.71
Examination	3.00	12.00	16.00	22.00	37.0	16.49	63.65	7.98
Education	1.00	6.00	8.00	12.00	53.0	10.98	92.46	9.62
Catholic	2.15	5.20	15.14	93.12	100.0	41.14	1739.29	41.70
Infant.Mortality	10.80	18.15	20.00	21.70	26.6	19.94	8.48	2.91

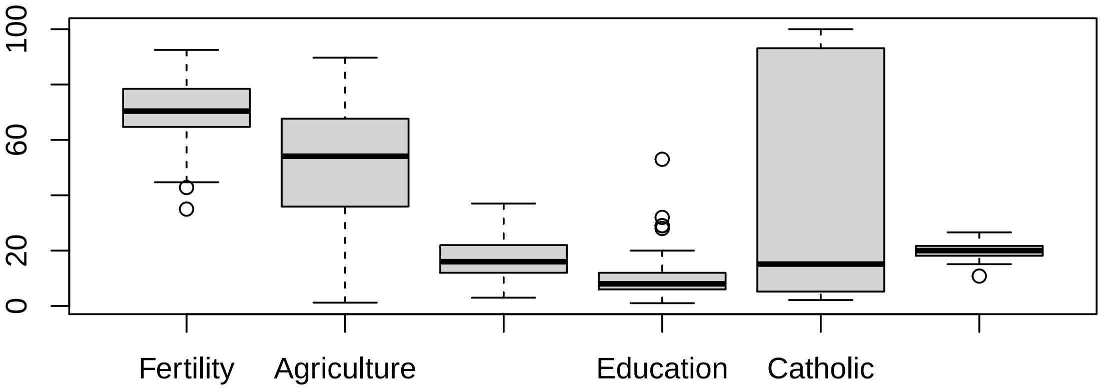

Figure 1: Boxplot of Swiss Data

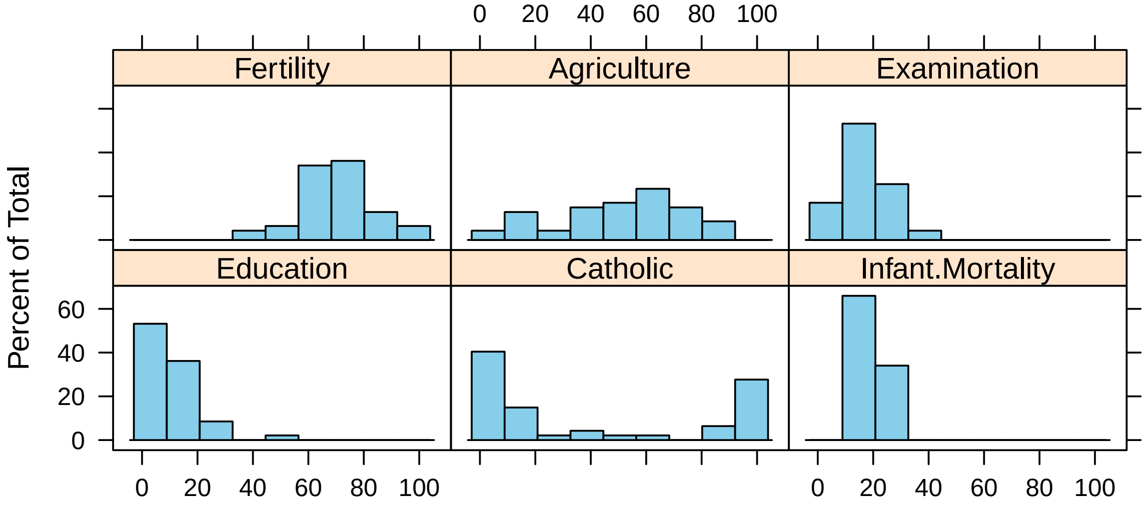

Figure 2: histogram of each variable

distribution of these three variables is close to normal distribution, but Agriculture has a high percentage in the interval [10, 20]. Therefore, Fertility and Examination are more consistent with normal distribution. Eucation and Infant.Mortality were close to skewed distribution, while Catholic was close to bimodal distribution.

In order to observe the distribution of each variable further, we draw the density function curve of each variable, and the result is shown in Figure 3.

The density function distribution curves of Fertility and Examination are most similar to the normal distribution. Secondly, Agriculture is also fairly consistent. And Infant.Mortality close to skewed distribution; Infant. Catholic has a bimodal distribution.

Figure 4 shows the empirical distribution function curve of each variable in the Swiss dataset. It can be seen from the results that the Fertility variable is closer to the normal distribution.

2.6 Relationships between the variables

In this part, I will show the relationship between each variable through different methods.

2.6.1 Pairwise scatterplot

First I performed the pairwise scatterplot of each variable.

From the results in Figure 5, Fertility is not found to have a strong linear relationship with the other five variables.

2.6.2 Pearson correlation significance test

Pearson's correlation coefficient measures the linear correlation, and the formula is as follows:

$$r = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2}}$$

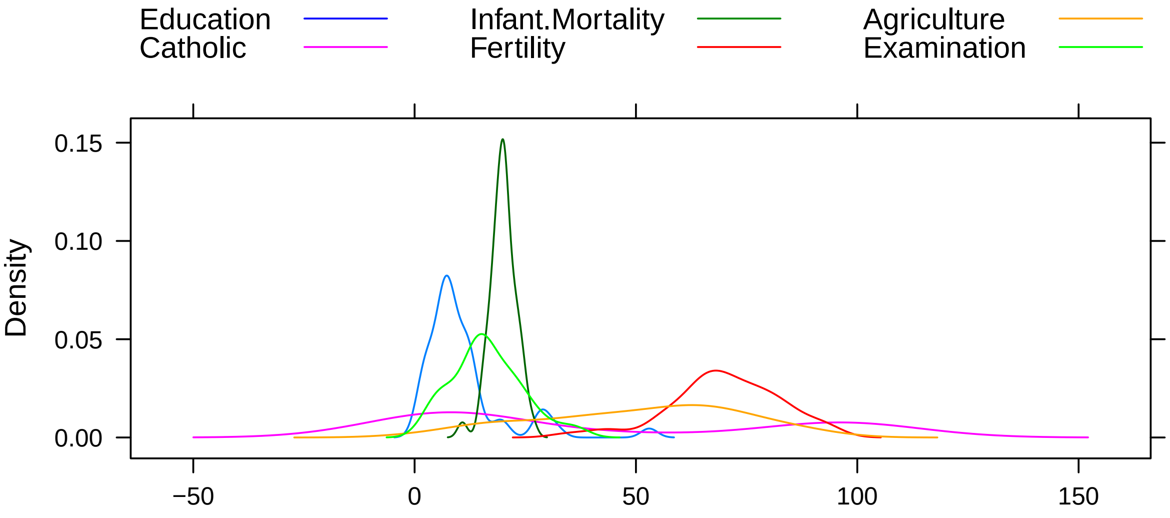

Figure 3: Density of each variable

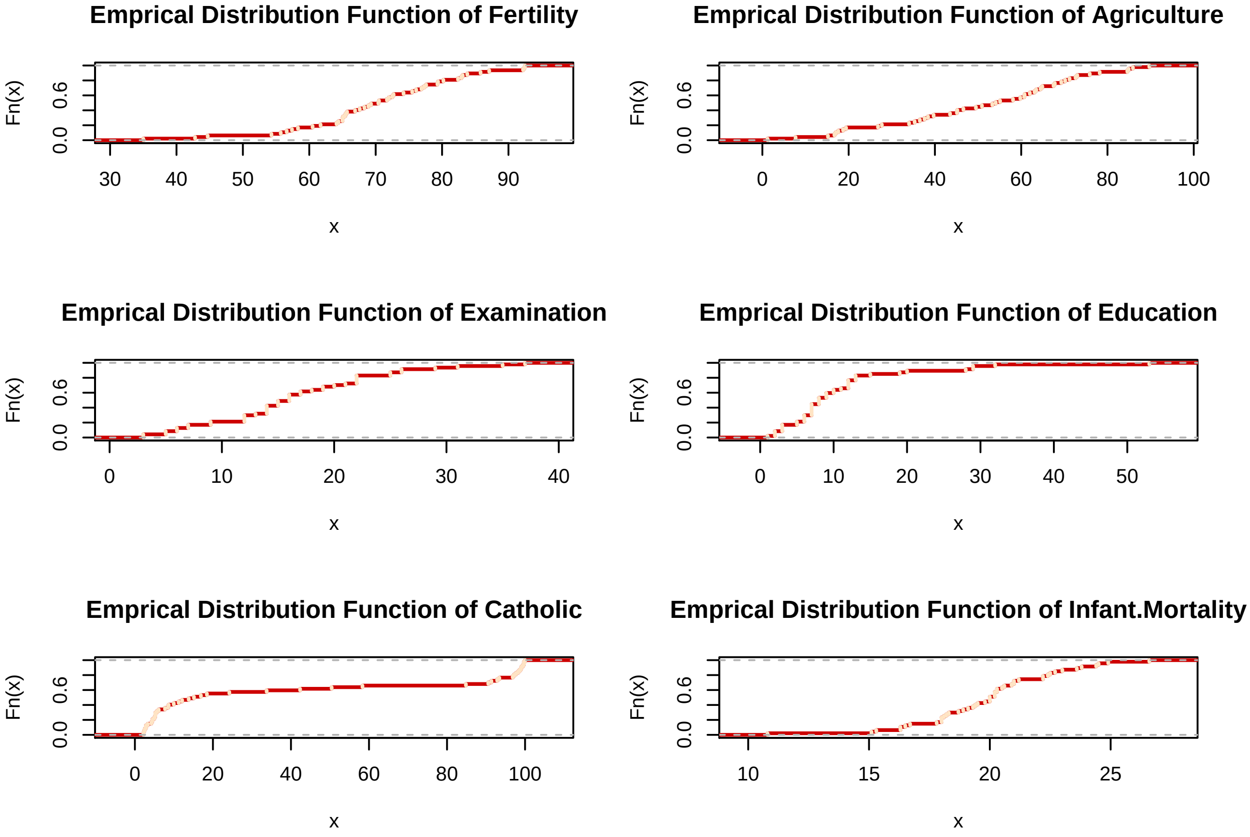

Figure 4: Empirical distribution of each variable

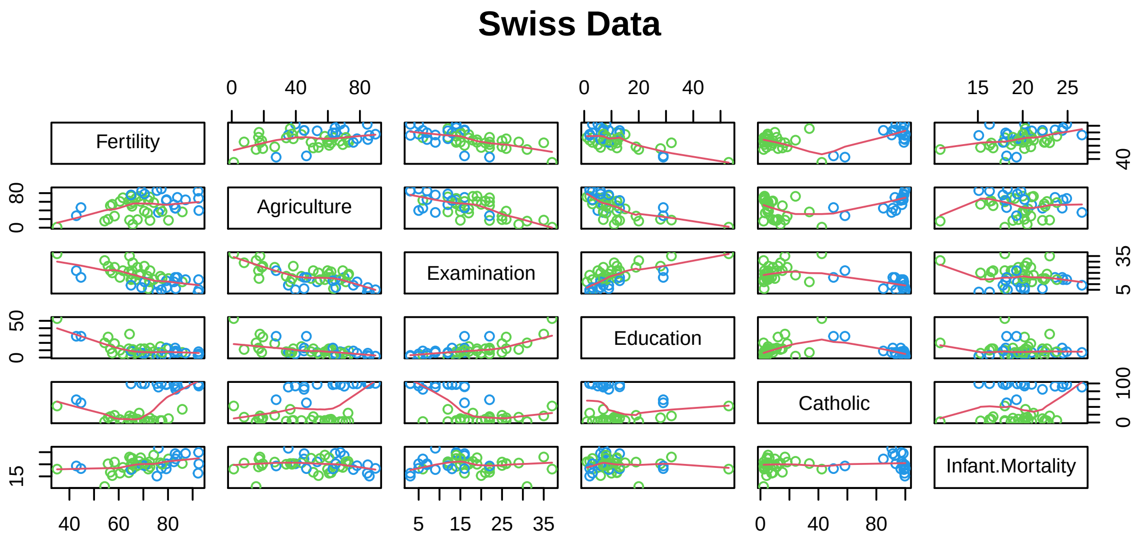

Figure 5: pairwise scatterplot of each variable

Generally, the correlation coefficient r between 0.8 and 1.0 indicates a strong correlation, 0.6 to 0.8 indicates a strong correlation, 0.4 to 0.6 indicates a moderate correlation, 0.2 to 0.4 indicates a weak correlation, and 0.0 to 0.2 indicates a very weak correlation or no correlation. When r=1, x and y are completely positively correlated; when r=-1, x and y are completely negatively correlated; when r=0, x and y are unrelated.

Pearson correlation coefficient of each variable in Swiss data set was calculated, and the results were shown in Figure 6.

2.6.3 Spearman correlation significance test

Spearman's correlation coefficient is a non-parametric index to measure the dependence of two variables. The formula is as follows:

$$\rho = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{N(\Sigma x_i - \bar{x})^2 (\Sigma y_i - \bar{y})^2}}$$

Spearman's correlation coefficient indicates the correlation direction between X (independent variable) and Y (dependent variable). When X increases, if Y tends to increase, Spearman's correlation coefficient is positive; if Y tends to decrease, Spearman's correlation coefficient is negative. When Spearman's correlation coefficient is 0, it indicates that Y has no tendency when X increases.

Spearman correlation coefficient of each variable in Swiss data set was calculated, and the results were shown in Figure 7.

2.6.4 Kendall correlation significance test

The formula of Kendall correlation coefficient is as follows:

$$\tau = \frac{(number of concordant pairs) - (number of discordant pairs)}{n(n-1)/2}$$

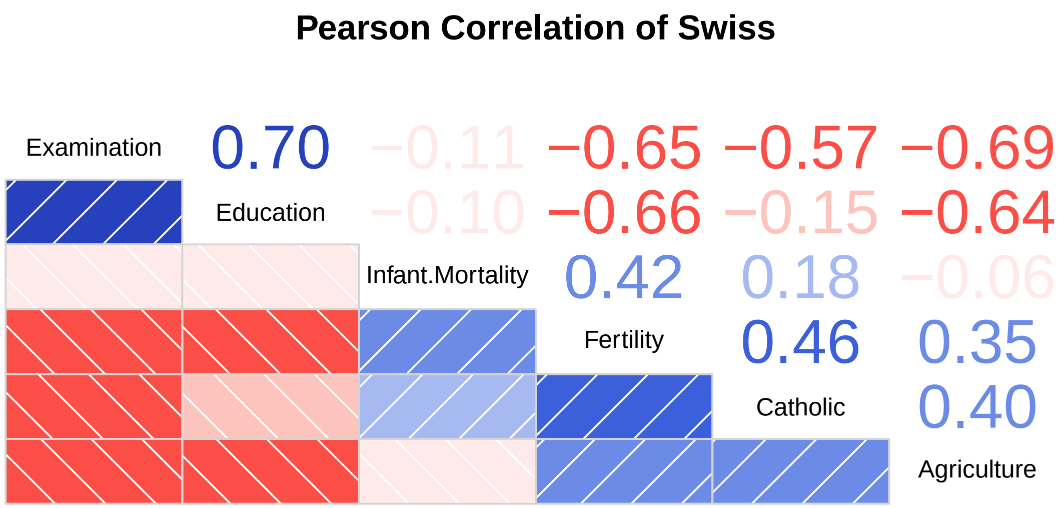

Figure 6: Pearson Correlation of each variable

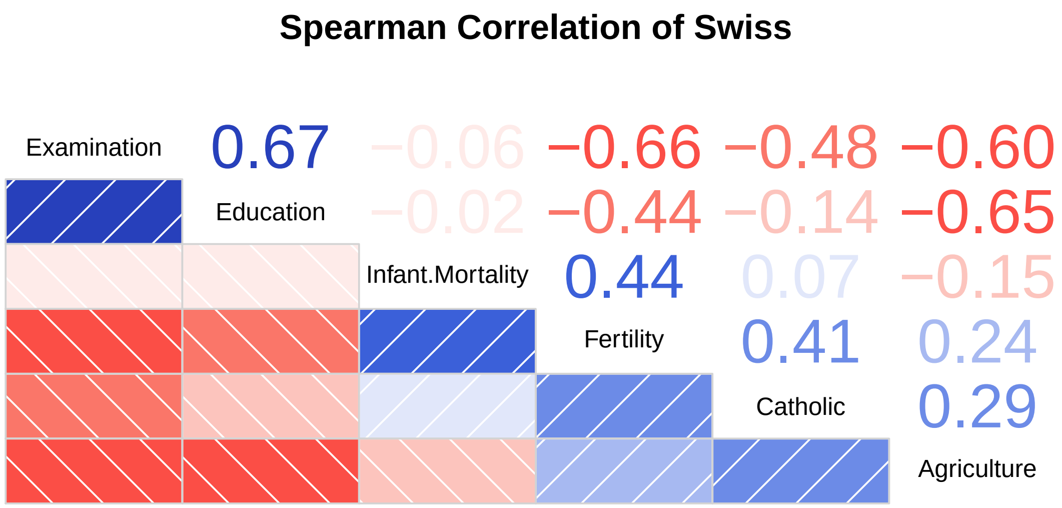

Figure 7: Spearman correlation coefficient of each variable

Kendall correlation coefficients of variables in the data set were calculated, and the results were shown in Figure 8.

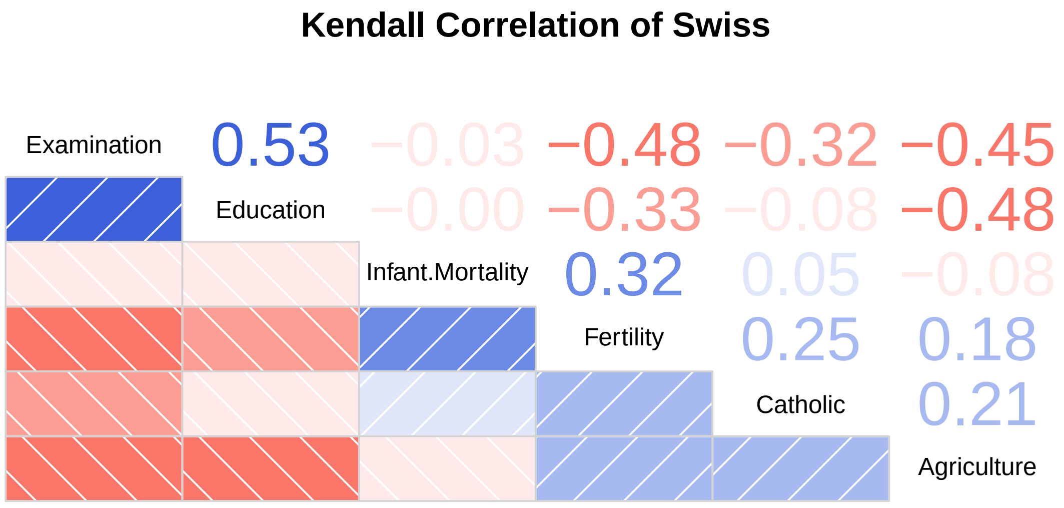

Figure 8: Kendall correlation coefficients of variables

It can be seen from the results of figures above that the positive and negative correlations and corresponding correlations among variables in the Swiss data set have little difference. Pearson correlation coefficient is suitable for calculating continuous, normally distributed and linear data. Spearman's correlation coefficient is suitable for calculating the relationship between grade data. Kendall correlation coefficient is a rank correlation coefficient, and the calculated object is the classification variable. Based on the above reasons, Pearson correlation coefficient was finally selected as the index to measure the correlation between variables in this experiment. Figure 9 shows the correlation index of variables in the Swiss data set and whether they are significant.

As shown in Figure 9, Fertility is significantly correlated with the other five variables, especially Examination and Education. Infant.Mortality showed a weak negative correlation with Agriculture, Examination and Education, but a weak positive correlation with Catholic. This correlation result indicates that all variables in the Swiss dataset are significantly correlated with Fertility, and therefore the relationship between Fertility and the other five variables can be further analyzed by regression analysis.

3 Model

In this paper, I considered the multiple linear regression model to conduct analysis. The basic formula of the model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Where y is the response variable, $x_i, i = 1, 2, \cdots, n$ is the indicator, $\beta_i, i = 1, 2, \cdots, n$ is the regression parameter, and ϵ is the residual.

Firstly, mathematical model definition is required for the selected multivariate data, followed by parameter estimation. Then significance test, residual analysis and outlier detection are carried out for the estimated parameters, and the final regression model is determined for model prediction.

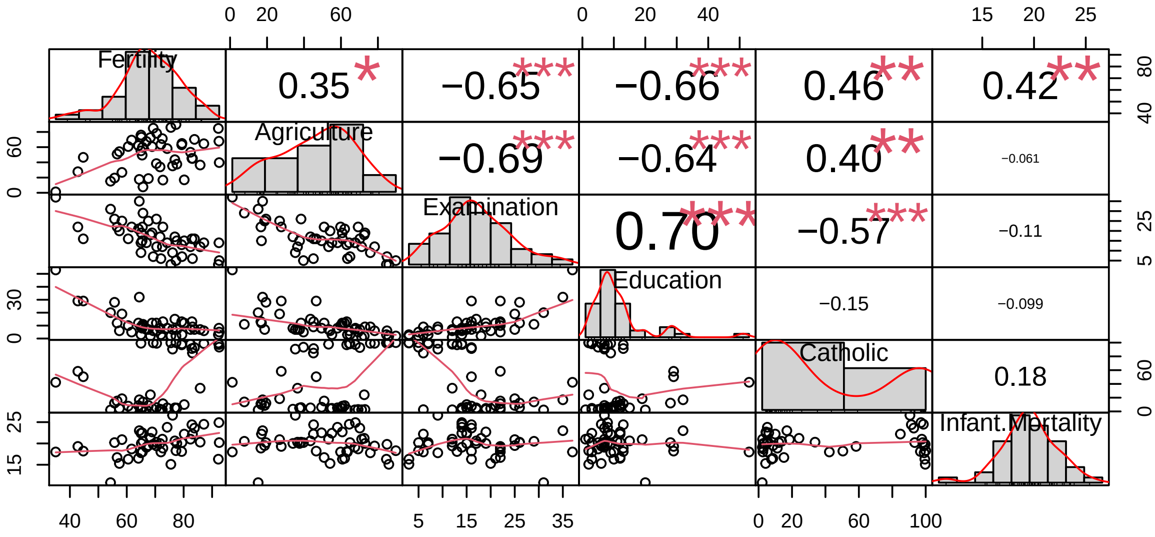

Figure 9: correlation index of variables

Table 3: Results of the full model

Variable	Estimate	Pr(>|t|)
(Intercept)	66.91518	1.91e-07
Agriculture	-0.17211	0.01873
Examination	-0.25801	0.31546
Education	-0.87094	2.43e-05
Catholic	0.10412	0.00519
Infant.Mortality	1.07705	0.00734

Combining the model with Swiss data set, the model of this experiment is obtained, and the formula is as follows:

$$Fertility = \beta_0 + \beta_1 * Agriculture + \beta_2 * Examination + \beta_3 * Education + \beta_4 * Catholic + \beta_5 * Infant.Mortality + \epsilon$$

Where β_0 represents the intercept and ϵ represents the residual, which is the synthesis of all other uncertainties.

And then I will perform the linear regression analysis.

4 Results

4.1 Full model

First I conducted linear regresson model considering all the variables, the results are as follows:

For Agriculture, the regression coefficient is -0.17, indicating that when Examination, Education, Catholic and Infant.Mortality remain unchanged, for each additional unit of Agriculture, Fertility is reduced by 0.17 units.

As can be seen from the results, at 95% confidence level, the regression coefficients of Agriculture, Education, Catholic and Infant.Mortality were significant, while the regression coefficients of Examination were not since p-value =0.32>0.05. This suggests that the linear correlation between Examination and Fertility is not significant when controlling other independent variables. The p-value under F test was 5.594e-10<0.05, which was significant.

Coefficient of Determination (R^2) is an important statistic reflecting the goodness of fit of a model. It is the ratio of regression sum of squares to total peace. R^2 is between 0 and 1, and its value reflects the relative degree of regression contribution, that is, the percentage of total variation of dependent variables that can be explained by regression relationship. R^2 is the most commonly used indicator to evaluate the pros and disadvantages of regression models. The larger R^2 is, the better the fitting degree of the model is, and the closer the regression equation is to the reality. In this experiment, adjusted R^2=0.671, indicating a good degree of model fitting.

The model needs to be tested before it can be determined. The multiple linear regression model has four assumptions: (a) there is a linear relationship between independent variable and dependent variable (numerical type). (b) Residuals are normally distributed. (c) The variance of residual is basically unchanged. (d) Residuals are independent from each other. A good multiple linear regression model should satisfy these four assumptions, so the next four assumptions will be tested.

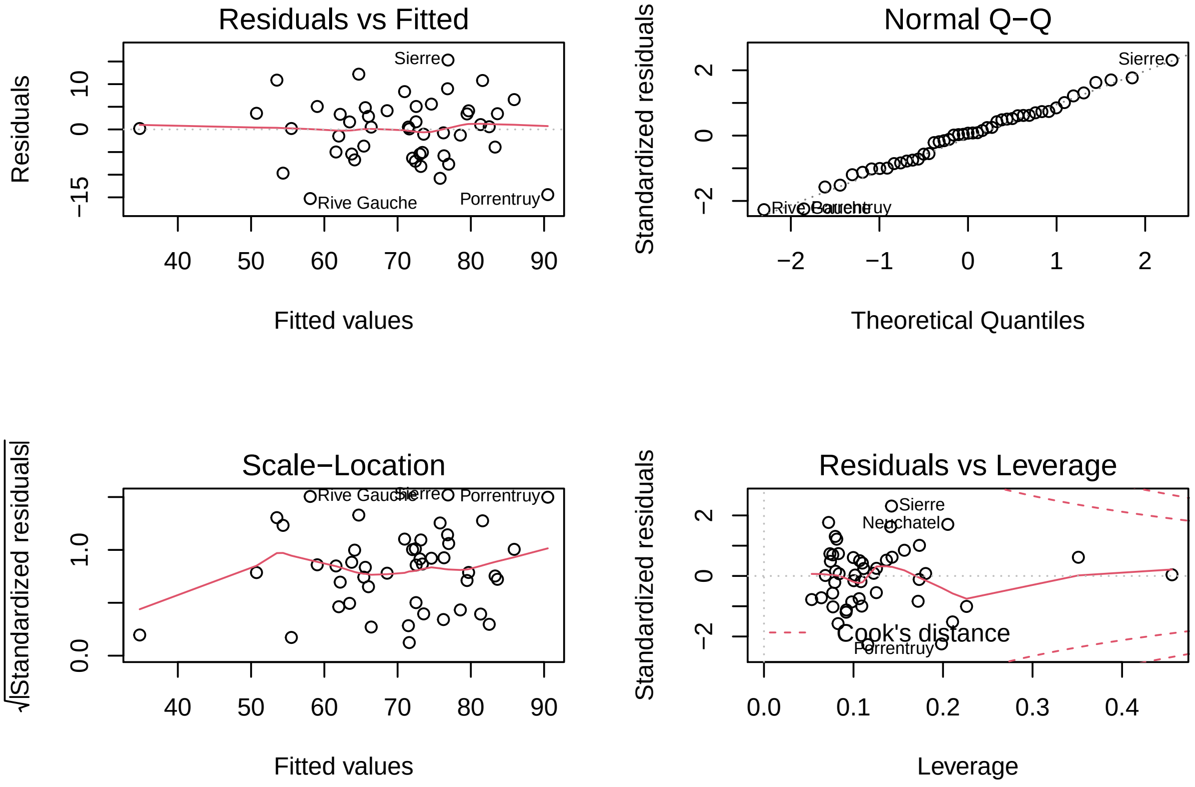

Figure 10: Residual plots of the full model

From Figure 10, the full model of the dataset conforms to the model assumptions, but there are outliers that affect the regression results.

Then I verified the kurtosis and skewness of the model, and the results were as follows:

Table 4: check assumptions of the full model

term	Value	p_value	decision
Global Stat	0.9100432	0.9231156	Assumptions acceptable.
Skewness	0.0459905	0.8301932	Assumptions acceptable.
Kurtosis	0.1091289	0.7411381	Assumptions acceptable.
Link Function	0.0867396	0.7683638	Assumptions acceptable.
Heteroscedasticity	0.6681842	0.4136854	Assumptions acceptable.

Table 5: Results of the reduced model

Variable	Estimate	Pr(>|t|)
(Intercept)	62.10131	8.49e-08
Agriculture	-0.15462	0.02857
Education	-0.98026	5.14e-08
Catholic	0.12467	9.50e-05
Infant.Mortality	1.07844	0.00722

When P >0.05, the hypothesis can pass. As can be seen from the results above, the four hypotheses of the Swiss dataset are valid.

4.2 Model optimization

From the summary of the full model, the regression coefficient of Examination was not significant for the model, and R2 of the model was 0.671, indicating that the fitting degree was not very good. Therefore, the model will be further adjusted and optimized.

First I checked the Akaike Information Criterion for each variable.

```
## Single term deletions
##
## Model:
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##                  Df Sum of Sq    RSS    AIC
## <none>                        2105.0 190.69
## Agriculture       1    307.72 2412.8 195.10
## Examination       1     53.03 2158.1 189.86
## Education         1   1162.56 3267.6 209.36
## Catholic          1    447.71 2552.8 197.75
## Infant.Mortality  1    408.75 2513.8 197.03
```

From the results above, the AIC value of Examination is the smallest. According to the AIC minimum principle, Examination is removed and then I performed the new model. The results are shown as follows.

It shows that the relationship between all variables and Fertility is all significant after removing Examination. The p-value of F test is 1.71e-10<0.05, indicating that the model is significant, and $R^2 = 0.671$ does not change. In addition, the variables can also be automatically selected according to the AIC minimum principle.

From the result of automatic selection, the model is the same as above. Although R^2 did not change after removing the variable, the regression coefficients of the respective variables became statistically significant.

In addition, I will also check the assumptions of the reduced model.

The results above show that the reduced model conforms to the model assumptions.

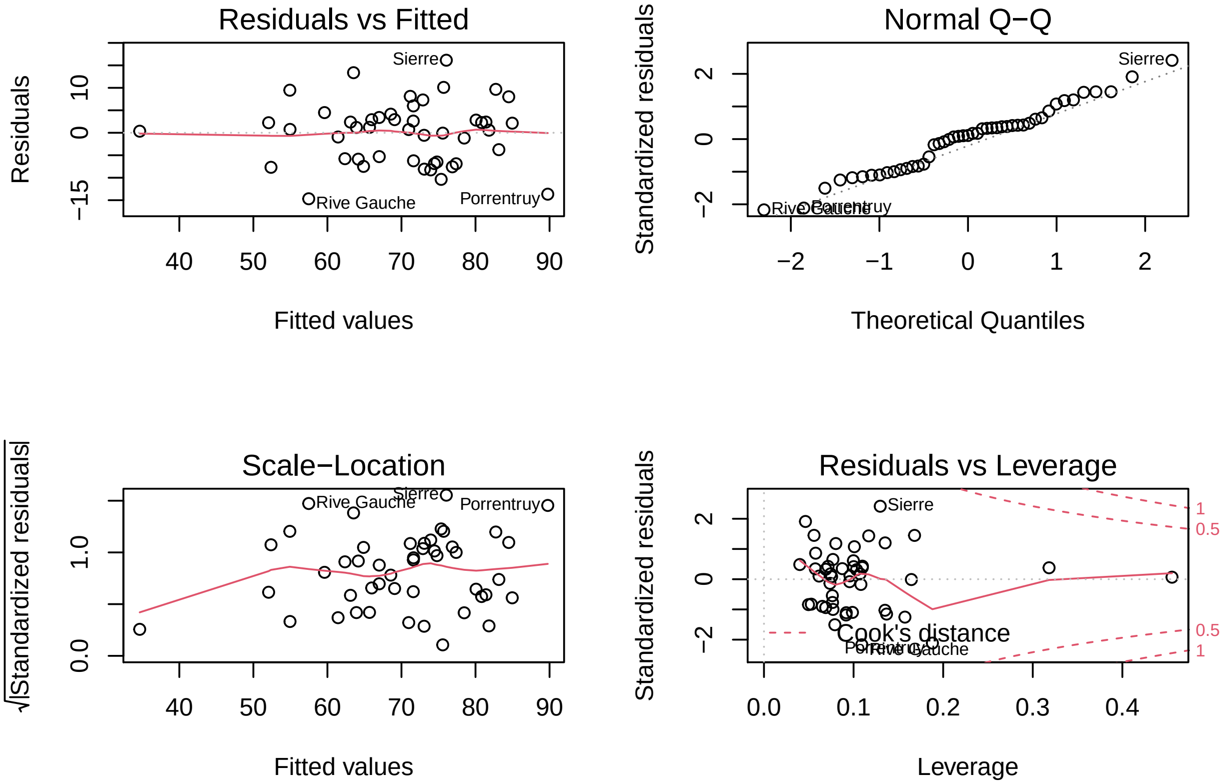

Figure 11: Residual plots of the reduced model

Table 6: check assumptions of the full model

	Value	p-value	Decision
Global Stat	0.7464017	0.9454858	Assumptions acceptable.
Skewness	0.0058959	0.9387948	Assumptions acceptable.
Kurtosis	0.2285972	0.6325659	Assumptions acceptable.
Link Function	0.2304797	0.6311683	Assumptions acceptable.
Heteroscedasticity	0.2814289	0.5957662	Assumptions acceptable.

Table 7: AIC of two models

	df	AIC
mod1	7	326.0716
mod2	6	325.2408

4.3 Model comparison

Then I used partial F test to compare the full model and the reduced model.

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
## Model 2: Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 2105.0
## 2     42 2158.1 -1   -53.027 1.0328 0.3155
```

According to the results of anova, p-value =0.3155>0.05, indicating that the test is not significant. Therefore, it can be concluded that Examination can be eliminated from the model.

Moreover, I compared the two models by the AIC minimum principle.

Table 6 indicates that the reduced model has the smaller value of AIC. So I choose the reduced model as the final model.

4.4 Final model

After the model fitting and comparison, I chose th reduced model as the final model. Its equation is :

$$Fertility = 62.1 - 0.15 * Agriculture - 0.98 * Education + 0.12 * Catholic + 1.08 * Infant.Mortality$$

5 Discussion

5.1 Conclusion

In this experiment, the Swiss dataset containing the fertility rate and socio-economic index measurement data of 47 Provinces in Switzerland was used for statistical analysis and multiple linear regression. The results show that the fertility rate is mainly related to the proportion of men engaged in Agriculture, the Education, Catholic and Infant Mortality, and the regression equation is

$$Fertility = 62.1 - 0.15 * Agriculture - 0.98 * Education + 0.12 * Catholic + 1.08 * Infant.Mortality$$

5.2 Reflection

The results are basically consistent with our common sense. It indicates that if we want to increase fertility, we have to first get people into work, then improve education system in order to raise people's knowledge level. In addition, we should improve health care so that babies can survive.

On my point of view, it's of great importance to study this topic since many countries in the world have witnessed the decline of fertility rate and even negative population growth nowadays. The analysis of this problem is of great guiding significance to today's social research. We can learn from the past experience and formulate more reasonable policies to deal with negative population growth and aging population.

5.3 Weakness and limitations

However, there are still some limitations in this experiment. For one thing, the dataset contains only 47 pieces of data, so that it may lead to some errors in the results. For another, from the result of the model, it shows that the adjusted R square isn't very close to 1 and the model assumptions are barely satisfactory. So if we have more knowledge about other models, we can take a try to see if there is a better fit.

5.4 What to proceed in the future

First of all, in response to today's declining fertility rate, we should figure out the reason behind it. Whether it is the rising cost of childbearing or people's low awareness of childbearing, we should call for corresponding policies to change this situation. Otherwise, it will only lead to the aging of the social population structure, and eventually lead to the increasing pressure of young people, which will affect the vitality of the society and the country in the future.

6 Appendix

6.1 Datasheet

6.1.1 Motivation

The dataset was created to provide data of standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The dataset is available in Project "16P5", pages 549–551 in Mosteller, F. and Tukey, J. W. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass, indicating their source as "Data used by permission of Franice van de Walle. Office of Population Research, Princeton University, 1976. Unpublished data assembled under NICHD contract number No 1-HD-O-2077."

6.1.2 Composition

The variables in the dataset are as below:

Fertility: common standardized fertility measure in 47 French-speaking provinces of Switzerland at about 1888.

Agriculture: proportion of males involved in agriculture as occupation

Examination: proportion of draftees receiving highest mark on army examination

Education: proportion of education beyond primary school for draftees.

Catholic: proportion of catholic (as opposed to 'protestant')

Infant.Mortality: live births who live less than 1 year

All variables but 'Fertility' give proportions of the population. Here, all variables are scaled to [0, 100], where in the original, all but "Catholic" were scaled to [0, 1]. In addition, variables Examination and Education are averages for 1887, 1888 and 1889.

There are 6 variables in total and 47 observations. And the dataset contains all possible instances.

In addition, the dataset doesn't contain data that, might be offensive, insulting, threatening, or might otherwise cause anxiety. The dataset also doesn't contain data that might be considered confidential. Moreover, there is no errors, sources of noise, or redundancies in the dataset.

6.2 Results of the model

In this part I will show the summary of the model, which is not shown in detail in the text.

6.2.1 Result of the full model

```
##
## Call:
## lm(formula = Fertility ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mod1)
```

```
##
##                       Value p-value                Decision
## Global Stat        0.91004  0.9231 Assumptions acceptable.
## Skewness           0.04599  0.8302 Assumptions acceptable.
## Kurtosis           0.10913  0.7411 Assumptions acceptable.
## Link Function      0.08674  0.7684 Assumptions acceptable.
## Heteroscedasticity 0.66818  0.4137 Assumptions acceptable.
```

6.2.2 Result of the reduced model

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

```
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = mod2)
##
##                        Value p-value                Decision
## Global Stat        0.746402  0.9455 Assumptions acceptable.
## Skewness           0.005896  0.9388 Assumptions acceptable.
## Kurtosis           0.228597  0.6326 Assumptions acceptable.
## Link Function      0.230480  0.6312 Assumptions acceptable.
## Heteroscedasticity 0.281429  0.5958 Assumptions acceptable.
```

6.2.3 Result of the automatic selection

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = df)
##
## Coefficients:
##      (Intercept)       Agriculture         Education          Catholic
##          62.1013           -0.1546           -0.9803            0.1247
## Infant.Mortality
##           1.0784
```

Reference

Mosteller, F., and J. W. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass.*

Pena, Edsel A., and Elizabeth H. Slate. 2019. *Gvlma: Global Validation of Linear Models Assumptions.*

Peterson, Brian G., and Peter Carl. 2020. *PerformanceAnalytics: Econometric Tools for Performance and Risk Analysis.* https://github.com/braverock/PerformanceAnalytics.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R.* New York: Springer. http://lmdvr.r-forge.r-project.org.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org,%20https://github.com/tidyverse/dplyr.

Wright, Kevin. 2021. *Corrgram: Plot a Correlogram.* https://kwstat.github.io/corrgram/.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/,%0Ahttps://github.com/haozhu233/kableExtra.