Analysis of Swiss Fertility Concerning Socio-economic Factors

Jacob Ratzlaff, Joseph Hunt

Colorado School of Mines

Abstract

In this research paper, we fit a linear model for Fertility rates among populations of 40 regions within Switzerland in the year 1888, considering five possible explanatory variables. Provided data is visualized and discussed, leading to how our transformations of given data are considered and why categorical variables are introduced.A finalized model is presented in which all included variables are statistically significant. Influentiality and leverage of certain data are considered. Assumptions regarding our model are checked and validated. Lastly, the impact each remaining explanatory variable has on fertility rates is analyzed.

Overview of Provided Data

To begin, we first describe our provided dataset. Our given data describes fertility rates among Swiss families as a response to six numeric variables. Below is provided a table describing the nature of these six variables in three columns: variable name, type (Either "N" for numeric, or "C" for categorical), and a brief description.

Variables	**Var**	**Description**
Fertility	N	Common standardized fertility measure
Agriculture	N	% of males involved in agriculture as occupation
Examination	N	% of draftees receiving highest mark on army examination
Education	N	% with education beyond primary school for draftees
Catholic	N	% Catholic (as opposed to Protestant)
Infant Mortality	N	% of live births who lived less than one year

Table 1: Provided Variables

Initial predictions of the variables effects included increases in fertility from an increase in any of the Agriculture, Examination, Catholic or Infant Mortaliity variables. We expected a decrease in fertility associated with an increase in education. As the proportion of individuals involved in agriculture increased we expected families to want to grow to increase the number of hands available to assist with farming tasks. The examination variable corresponds to the health of fathers in the community, we expected that the birth of healthy children would lead to an increased fertility rate in a community. We expect an increase in fertility as Catholic percentage increases, this assumption comes from annecdotal evidence. As infant mortality increases in a community we expected an increase in fertility as families will have another child. Finally we expected an increase in education to correspond to a decrease in the fertility of a community. This expectation follows similar reasoning to the expected increase in

fertility with agriculture, a more educated family won't need a large number of young helpers with their work.

Intuitively, one could argue that Education, Examination, and Infant Mortality are most likely to impact Fertility the most: greater education is often associated with greater earnings potential, leading to improved medical care; healthier men are more likely to bond with women and form families, including fathering healthy children; and lastly, a low infant mortality rate is proportional with greater fertility. We can also intuitively hypothesize that infant mortality may be confounded by education and examination - poorly educated people may not be able to afford effective medical care, for example - and agriculture may be confounded by education. It will be worthwhile to study the colinearity of such variables later.

Additionally, consideration for the Catholicism of a region garners special attention: to what degree does an exact percentage impact fertility rates? Are Catholicism and educational achievement colinear?

Pair-Plot of Provided Variables

To investigate colinearity and confounding, we plot each explanatory variable against another utilizing RStudio's `pairs` command:

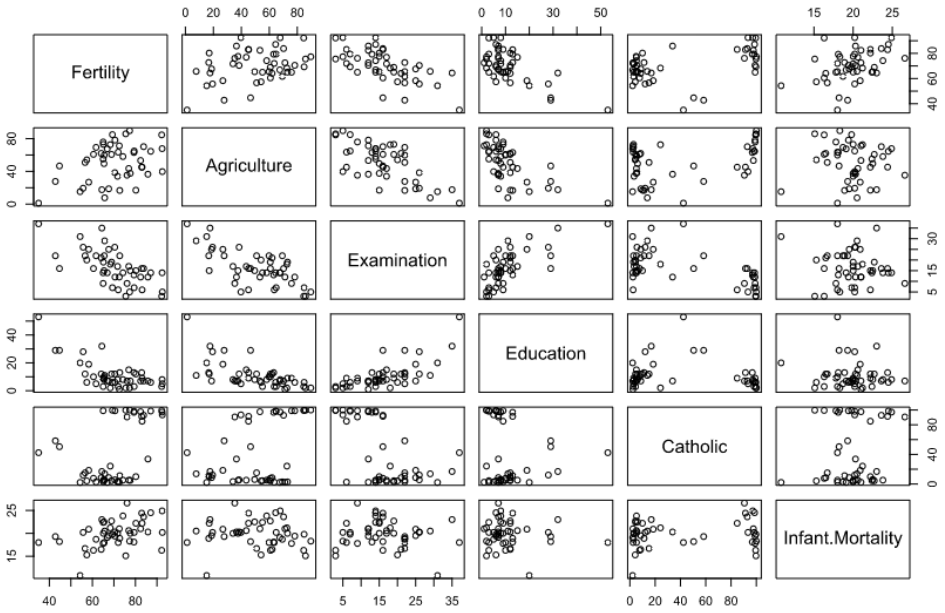

Figure 1: Pairs Plot of Explanatory Variables

The Catholic variable is highly bimodal, with only a few points lying between the extremes of a majority Catholic or majority Protestant. Highly Catholic regions are typically less educated, score lower on the physical examinations, and are more agrarian. Given these extremes we decided to translate the Catholic variable into a categorical variable.

From the pairs plot we can also see that the examination parameter is not linear with fertility, this is a transformation we will consider if the variable has low significance.

Fitting A Linear Model

In order to determine possible data transformations, interactions, and variable selection, an initial linear fit of all explanatory variables is necessary. A multiple linear regression in RStudio yields the following output:

```
Call:
lm(formula = swiss$Fertility ~ swiss$Agriculture + swiss$Examination +
    swiss$Education + swiss$Catholic + swiss$Infant.Mortality)

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 ***
swiss$Agriculture   -0.17211    0.07030  -2.448  0.01873 *
swiss$Examination   -0.25801    0.25388  -1.016  0.31546
swiss$Education     -0.87094    0.18303  -4.758 2.43e-05 ***
swiss$Catholic       0.10412    0.03526   2.953  0.00519 **
swiss$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
```

Figure 2: Summary Statistics for Base Model

We can see from our resulting fit that Examination does not immediately seem statistically significant - that is, we cannot confidently state that it's parameter value in a linear model is not zero. Given this low significance we decided to consider a transformation to the Examination variable. The plot suggests an inverse relationship between examination and fertility, as such we fit another linear model with `1/Examination`.

Figure 3: Summary Statistics for Transformed Model