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, considering six 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

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 in our provided variables, we plot each explanatory variable against another utilizing RStudio's `pairs` command:

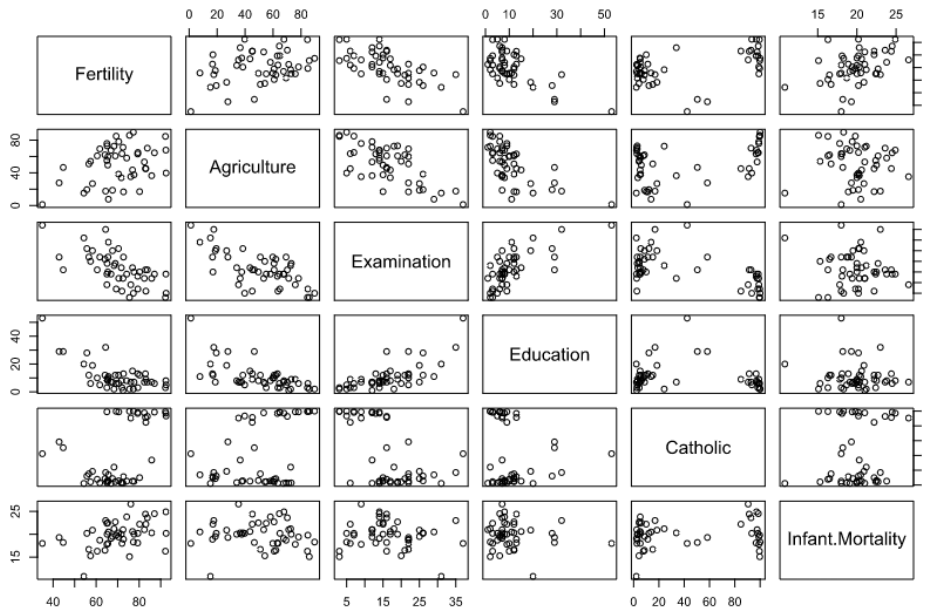

Figure 1: Pairs Plot of Explanatory Variables

Perhaps most striking is the strong bimodality of the Catholic variable across all comparisons: generally speaking, highly Catholic regions are typically less educated, less physically fit, and more agrarian. For this reason, creating a new categorical variable (catholic vs. non-catholic, say, at some arbitrary cut-off percentage) might help the fit of our model. Additionally, we note that Agriculture and Examination seem slightly collinear. Similarly, we notice that Examination and Education exhibit the same behaviour. We will keep this is mind later once we begin reducing the number of variables in our model.

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 that Examination does not immediately seem statistically significant. This makes sense, as we previously noted that Examination and Agriculture seem slightly collinear. For this reason, we remove the Examination variable from our model and see the following results:

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

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 ***
swiss$Agriculture      -0.15462    0.06819  -2.267  0.02857 *
swiss$Education        -0.98026    0.14814  -6.617 5.14e-08 ***
swiss$Catholic          0.12467    0.02889   4.315 9.50e-05 ***
swiss$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
```

Figure 3: Summary Statistics for Modified Base Model

In doing so, we ensure that all included parameters are statistically significant- that is, strengthening the global null hypothesis that all included parameters are non-zero. We do not, however, improve our adjusted R-squared value. Furthermore, removing the Agriculture variable does not marginally improve the fit of our model - our initial hypothesis that Agriculture and Education are slightly collinear is either incorrect or relatively unimportant. Even though removing Agriculture improves the significance of all remaining parameters, Agriculture is itself significant enough to warrant inclusion.

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

Residuals:
     Min       1Q   Median       3Q      Max
-14.4781  -5.4403  -0.5143   4.1568  15.1187

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            48.67707    7.91908   6.147 2.24e-07 ***
swiss$Education        -0.75925    0.11680  -6.501 6.83e-08 ***
swiss$Catholic          0.09607    0.02722   3.530  0.00101 **
swiss$Infant.Mortality  1.29615    0.38699   3.349  0.00169 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.505 on 43 degrees of freedom
Multiple R-squared:  0.6625,    Adjusted R-squared:  0.639
F-statistic: 28.14 on 3 and 43 DF,  p-value: 3.15e-10
```

Figure 4: Summary Statistics for Modified Base Model, No Agriculture

Now that we have created an otherwise satisfactory model, we consider variable transformations. By plotting residual vs. fitted values for simple linear regression models of Fertility and Education, Agriculture, Catholic, and Infant.Mortality, respectively, we see the following results:

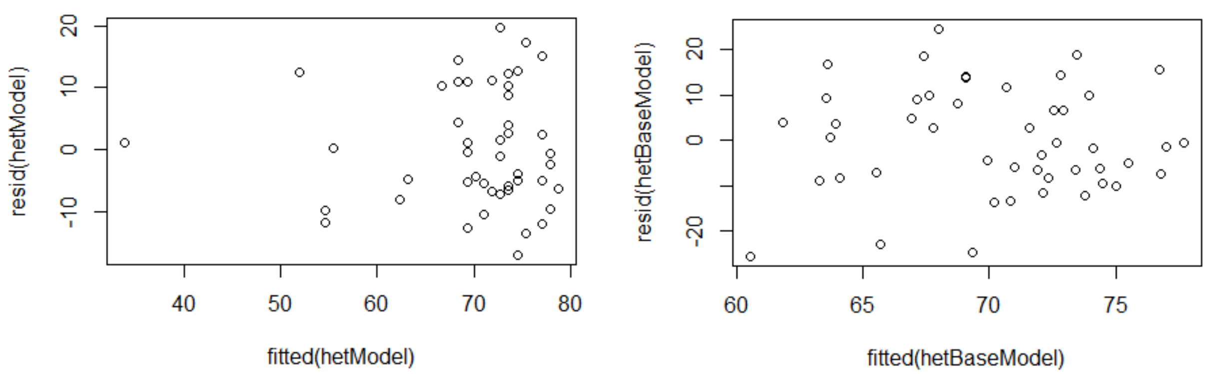

Figure 5: Checking Homoscedasticity - Education and Agriculture

Figure 6: Checking Homoscedasticity - Catholic and Infant.Mortality

Notice that all of our data is relatively homoscedastic, with the relative exception of Education and Catholic: our Education variable has one outlier, but is otherwise fine; Catholic, on the other hand, demonstrates strong bimodal behaviour towards either strongly Catholic or Protestant regions. While homoscedasticity isn't necessarily violated here, it is worthwhile to see if creation of a categorical variable (either Catholic or not) could help the situation. We arbitrarily categorize each region as Catholic if 75 percent or more of its population are Catholic. Otherwise, the region is considered not largely Catholic. This stratification yields the following results:

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

Residuals:
     Min       1Q   Median       3Q      Max
-14.7018  -5.2215   0.0323   3.0774  15.9957

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            63.20158    9.17864   6.886 2.12e-08 ***
dummy$Agriculture      -0.15542    0.06417  -2.422  0.01984 *
dummy$Education        -0.86538    0.13886  -6.232 1.84e-07 ***
dummy$CatholicCatholic 12.29424    2.48753   4.942 1.28e-05 ***
dummy$Infant.Mortality  1.00940    0.36636   2.755  0.00864 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.847 on 42 degrees of freedom
Multiple R-squared:  0.7256,    Adjusted R-squared:  0.6995
F-statistic: 27.77 on 4 and 42 DF,  p-value: 2.601e-11
```

Figure 7: Summary Statistics for Modified Catholic Variable