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. Influential points and the 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 five 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 Mortality 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 anecdotal 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. 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 collinearity 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:

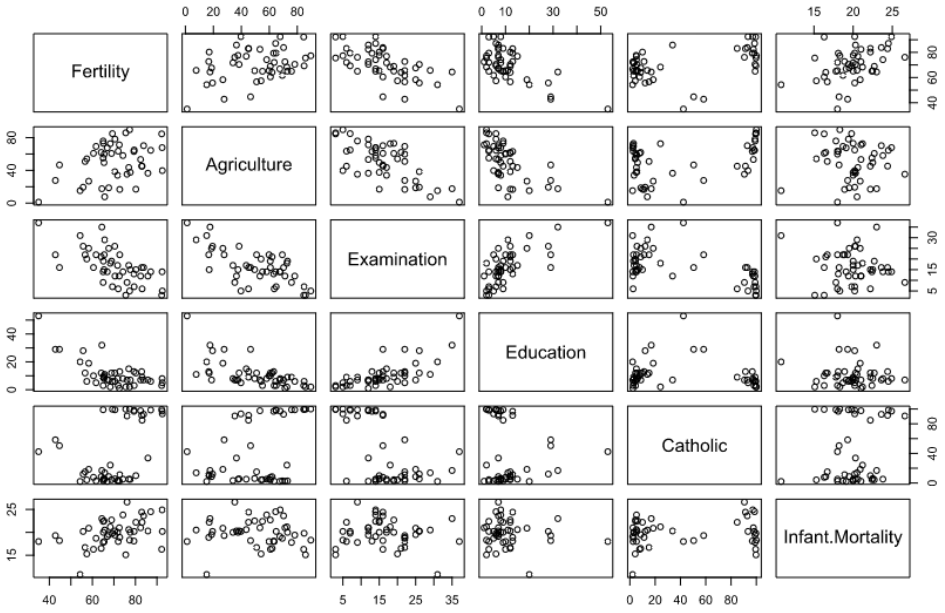

Figure 1: Pairs Plot of Explanatory Variables

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.

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 Examination 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:

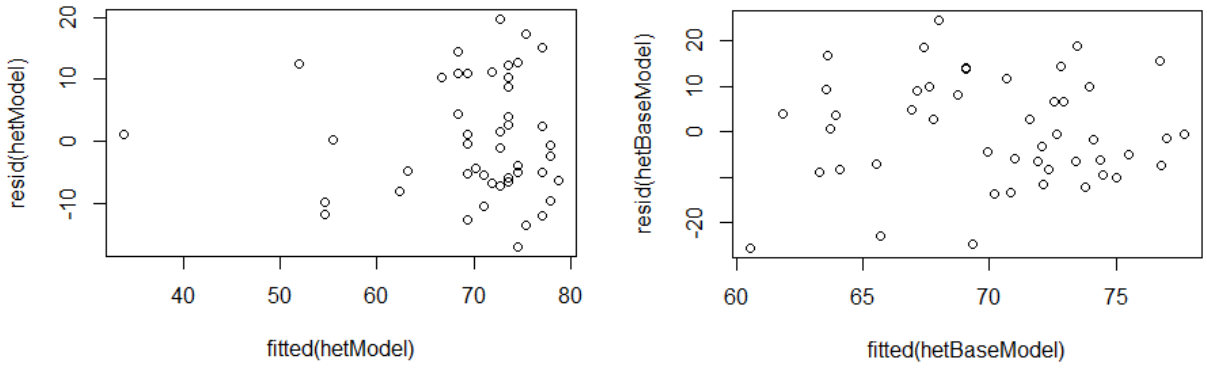

Figure 5: Checking Homoscedasticity - Education and Agriculture

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 categorize each region as Catholic if 75 percent or more of its population are Catholic. Otherwise, the region is considered not largely Catholic. The point 75% was chosen to caputre all of the extremely Catholic in one group and the remaining groups into another. 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

Leverage and Influence

To determine if our model includes any high leverage points, we utilize RStudio's `plot()` function to create a Residuals vs. Leverage plot, shown below. Since all points are within $\frac{1}{2}$ of Cook's distance, no points are influential in our model. Note however that two points have relatively high leverage compared to the rest of our data. Since both of these points agree on the fit of the model we can assume that they follow the same distribution of the rest of the data, so we can leave them in the model.

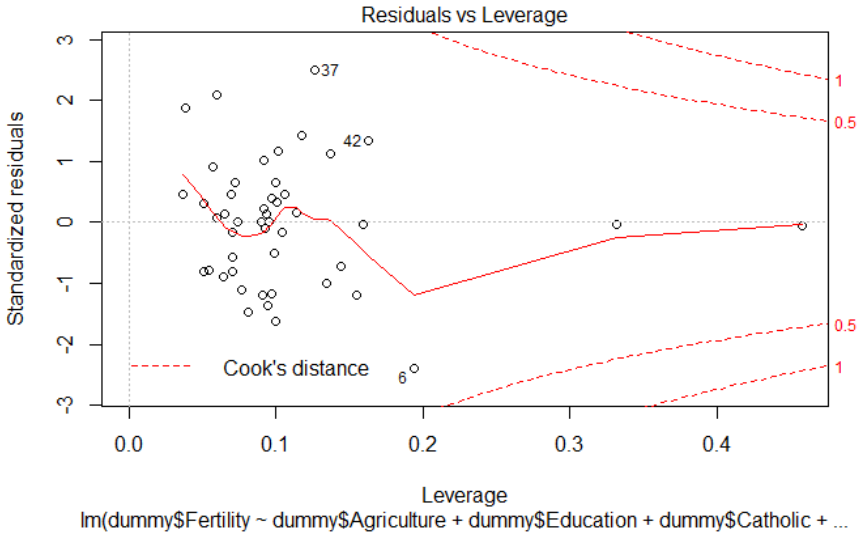

Figure 8: Influence of data points

Summary

From our final model we can see that increase in agriculture corresponds to a small decrease in fertility. Education leads to a large decrease in fertility. In communities that we considered majority Catholic we saw a huge increase in fertility. Finally we see a increase in fertility with a near one-to-one scaling with infant mortality - clearly parents have a target number that they will reach, regardless of the failure rate.

Outside of the agriculture variable we see expected results. The agriculture variable was slightly collinear with examination. We believe that this discrepancy with our expectations comes from eligible males being drafted into the army given their high examination scores, leaving them unable to produce progeny during their prime years.

Appendix of Code

[1] Code for Base Model Creation

```
#generate linear model using all explanatory variables
baseModel = lm(swiss$Fertility ~ swiss$Agriculture + swiss$Examination
    + swiss$Education + swiss$Catholic + swiss$Infant.Mortality)
summary(baseModel)
#create residual plots for Swiss Fertility and all variables in base model
hetBaseModel=lm(swiss$Fertility ~ swiss$Catholic)
```

[2] Code for Reduced Model Creation

```
#generate linear model with reduced parameters
removedBaseModel = lm(swiss$Fertility ~ swiss$Education + swiss$Catholic
    + swiss$Infant.Mortality)
summary(removedBaseModel)
```

[3] Creation of Categorical Catholic Variable

```
#creates categorical catholic var
dummy = swiss
x = c()
for(i in 1:47) {
  if (dummy$Catholic[i] >= 75) {
    x[i] = "Catholic"
  }
  else {
    x[i] = "Average"
  }
}
dummy$Catholic = x
```

[4] Generation of Final Model and Leverage Plot

```
#Code for End Model Generation
EndModel = lm(dummy$Fertility ~ dummy$Agriculture + dummy$Education
    + dummy$Catholic + dummy$Infant.Mortality )
summary(EndModel)
#Decided this was the best music
hetEndModel=lm(dummy$Fertility ~ dummy$Catholic)
plot(fitted(hetEndModel),resid(hetEndModel))
plot(EndModel)
```