Fragile Fertility: Examining the effect of Socioeconomic factors on Swiss Fertility

CSC 433 Project 8

Sarah Cummings

Introduction:

Given our planet's limited space and resources, our rising population is something of which to be concerned. It's no secret that throughout history, underdeveloped countries have had the strongest tendency towards overpopulation, but why? Do these countries also have higher fertility rates? This is a very interesting question which lead me to pick the Swiss dataset in R's built in dataset library.

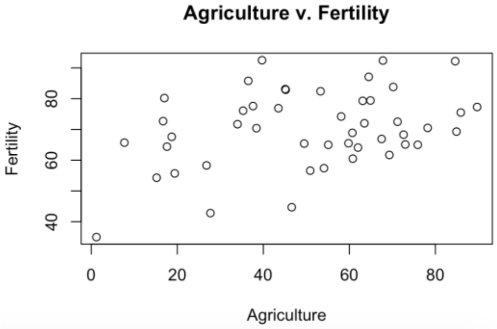
Switzerland had a shift in fertility rates in the late 1800's, just as it was transitioning to a more-developed country. In researching this shift in 1888, data was collected across provinces in Switzerland. The R library datafile "Swiss" contains a subset of the original data, with 47 Swiss provinces represented. Our dataset allows us to examine the relationship between fertility rates and five different socioeconomic factors (as percentages [1-100]) described as follows:

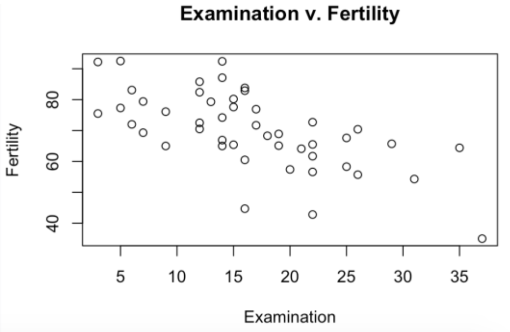
1. Fertility: common standardized fertility measure
2. Agriculture: % of males involved in agriculture as occupation
3. Examination: % draftees receiving highest mark on army examination
4. Education: % education beyond primary school for draftees.
5. Catholic: % 'catholic' (as opposed to 'protestant').
6. Infant Mortality: live births who live less than 1 year.

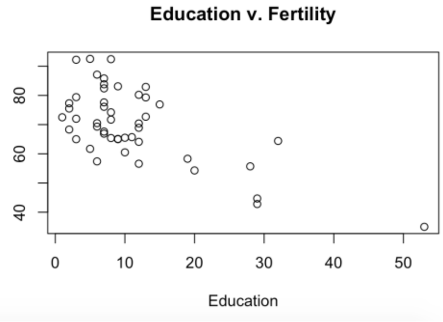
Since we are expected less developed areas to have higher fertility, I have come up with the following hypotheses: Fertility and Agriculture will have a positive relationship; Fertility and Examination, Education, and Catholic will have a negative relationship; and Fertility and Infant Mortality will have a negative relationship. I have formed these hypothesis because higher socio-economic areas tend to have higher intelligence (thus higher education and examination scores). There also tends to be more religion in higher class society (thus a higher scrore for the catholic variable). I predicted agriculture and fertility, and infant mortality and fertility, to have a negative relationship because lower socioeconomic areas often have more farm based jobs and a lower income (thus higher scores for the agriculture and infant mortality variables).

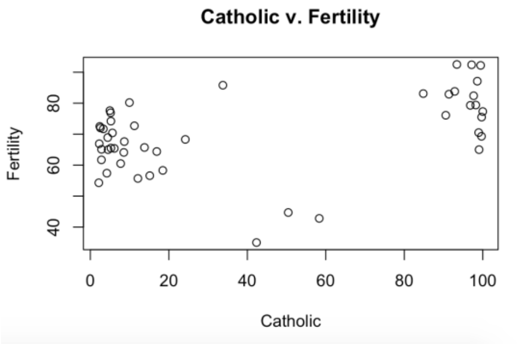
Exploration:

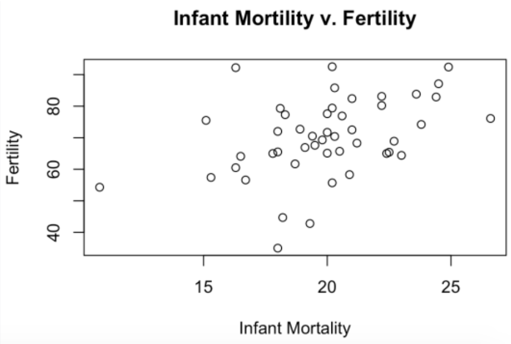
I will start by making a quick scatterplot of fertility versus each socioeconomic variable:



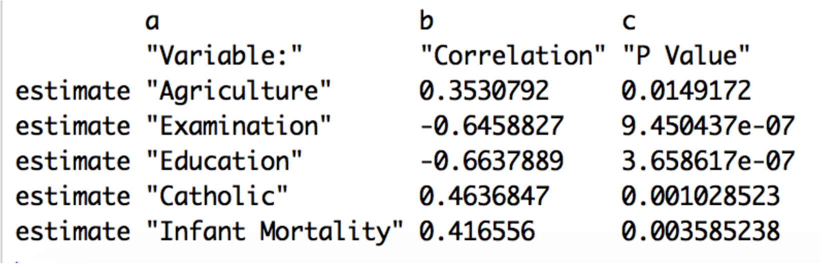








Based on the plots above, I decided to create a table of correlations of these variables with fertility:

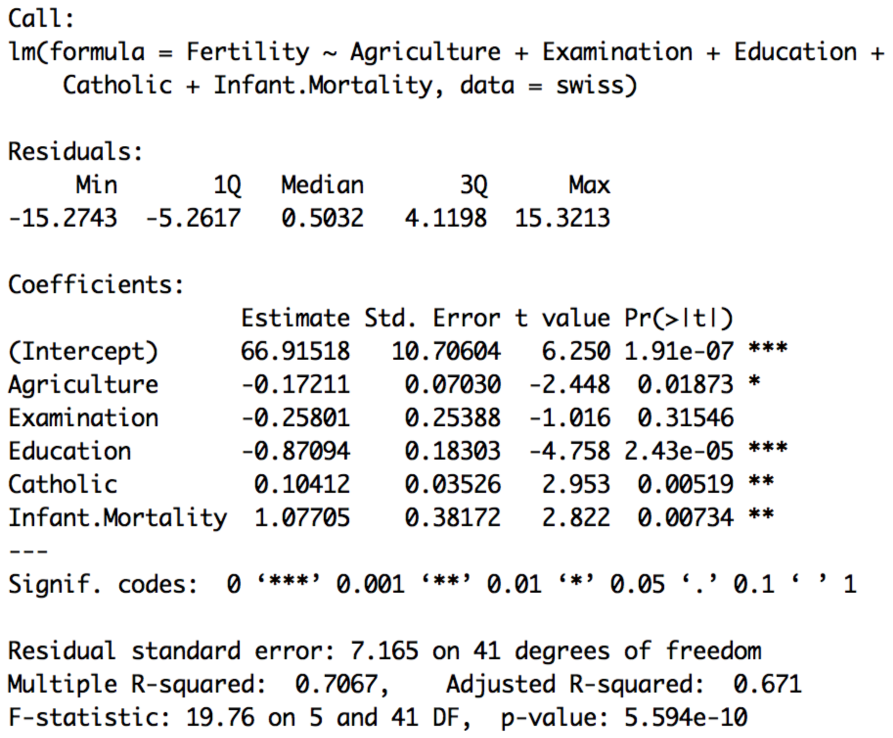


```
         a                   b            c
         "Variable:"         "Correlation" "P Value"
estimate "Agriculture"       0.3530792    0.0149172
estimate "Examination"       -0.6458827   9.450437e-07
estimate "Education"         -0.6637889   3.658617e-07
estimate "Catholic"          0.4636847    0.001028523
estimate "Infant Mortality"  0.416556     0.003585238
```

As you can see, some of our correlation hypothesis were confirmed. Fertility and Education, and Fertility and Examination, both have a statistically significant negative correlation. Our other three variables, Agriculture, Catholic and Infant mortality, all have a statistically significant positive correlation with Fertility.

Modeling:

Given the significant correlations we have found, I decided it would be interesting to run a multiple linear regression on this dataset, with hopes of modeling fertility with agriculture, education, examination, catholic, and infant mortality as predictors.



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

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
```

Our model provides the following regression:
FERTILITY= 66.91518 - 0.17211(AGRICULTURE) -0.25801(EXAMINATION) - 0.87094(EDUCATION) + 0.10412(CATHOLIC) +1.07705(INFANT MORTALITY)

Note that though the Examination variable is not statistically significant on its own, the overall model is proven significant and useful with F statistic= 19.76 having p=5.594x10^-10. While we cannot generalize our model to further data for prediction at this point, we now have a significant understanding of the relationship between Fertility and the provided predictor variables for Swiss provinces in

Conclusion:

In conclusion, we found that the Fertility of Swiss provinces in 1888 have a statistically significant correlation with Agriculture, Education, Catholic, and Infant.Morality. These variables, in combination with the examination variable, also can be used as predictors to form a statistically significant multiple linear regression for Fertility. Finally, provinces with higher Education and Examination scores have lower fertility. Provinces with higher agriculture jobs and higher higher infant mortality rates have higher fertility. Overall, there is a tendency of more socioeconomically developed areas to have lower fertility.

Code:

```
#view data
print(swiss)
#make scatterplots
plot(x=swiss$Agriculture,y=swiss$Fertility,xlab="Agriculture",ylab = "Fertility",main = "Agriculture v. Fertility”)
plot(x=swiss$Examination,y=swiss$Fertility,xlab="Examination",ylab = "Fertility",main = "Examination v. Fertility")
plot(x=swiss$Education,y=swiss$Fertility,xlab="Education",ylab = "Fertility",main = "Education v. Fertility”)
plot(x=swiss$Catholic,y=swiss$Fertility,xlab="Catholic",ylab = "Fertility",main = "Catholic v. Fertility”)
plot(x=swiss$Infant.Mortality,y=swiss$Fertility,xlab="Infant Mortality",ylab = "Fertility",main = "Infant Mortility v. Fertility”)
#run correlation tests with each variable and store in a new object
```

```
c1<-cor.test(swiss$Fertility,swiss$Agriculture)
c2<-cor.test(swiss$Fertility,swiss$Examination)
c3<-cor.test(swiss$Fertility,swiss$Education)
c4<-cor.test(swiss$Fertility,swiss$Catholic)
c5<-cor.test(swiss$Fertility,swiss$Infant.Mortality)
#create vectors so we can show these correlations more easily
a<-c(“Variables”,”Agriculture","Examination","Education","Catholic","Infant Mortality”)
b<-c(“Correlation”,c1[4],c2[4],c3[4],c4[4],c5[4])
c<-c(“PVal”,c1[3],c2[3],c3[3],c4[3],c5[3])
cbind(a,b,c)
#create a predictive model for the data
model=lm(Fertility~Agriculture+Examination+Education+Catholic+Infant.Mortality, data=swiss)
summary(model)
```