Pair Assignment 1

Tori Dykes and Mansi Wadhwa

Tuesday, October 04, 2016

The two datasets are: Swiss Fertility and Socio-economic indicators (1888) Data and the New York Air Quality Measurements from among the built-in R datasets.

Swiss Fertility and Socio-economic indicators (1888) Data

```
data(swiss)
names(swiss)
```

```
## [1] "Fertility"        "Agriculture"      "Examination"
## [4] "Education"        "Catholic"         "Infant.Mortality"
```

The first dataset 'swiss' is the Swiss Fertility and Socio-economic indicators (1888) Data from R itself that provides standardized fertility measure and socioeconomic indicators for each of the 47 French-speaking provinces of Switzerland. It contains a total of 47 observations and 6 variables (all in percent).

Exploring the relationship between the variables Fertility and Catholic:

The variable Fertility gives the standardized fertility measure and the variable Catholic gives us the percentage of Catholics in the population of each province. Quite some research has been focussed on exploring the impact of socio-economic, cultural and religious conditions on fertility levels. The case of Switzerland is interesting due to its cultural diversity and the fact that its population declined greatly in 1885.

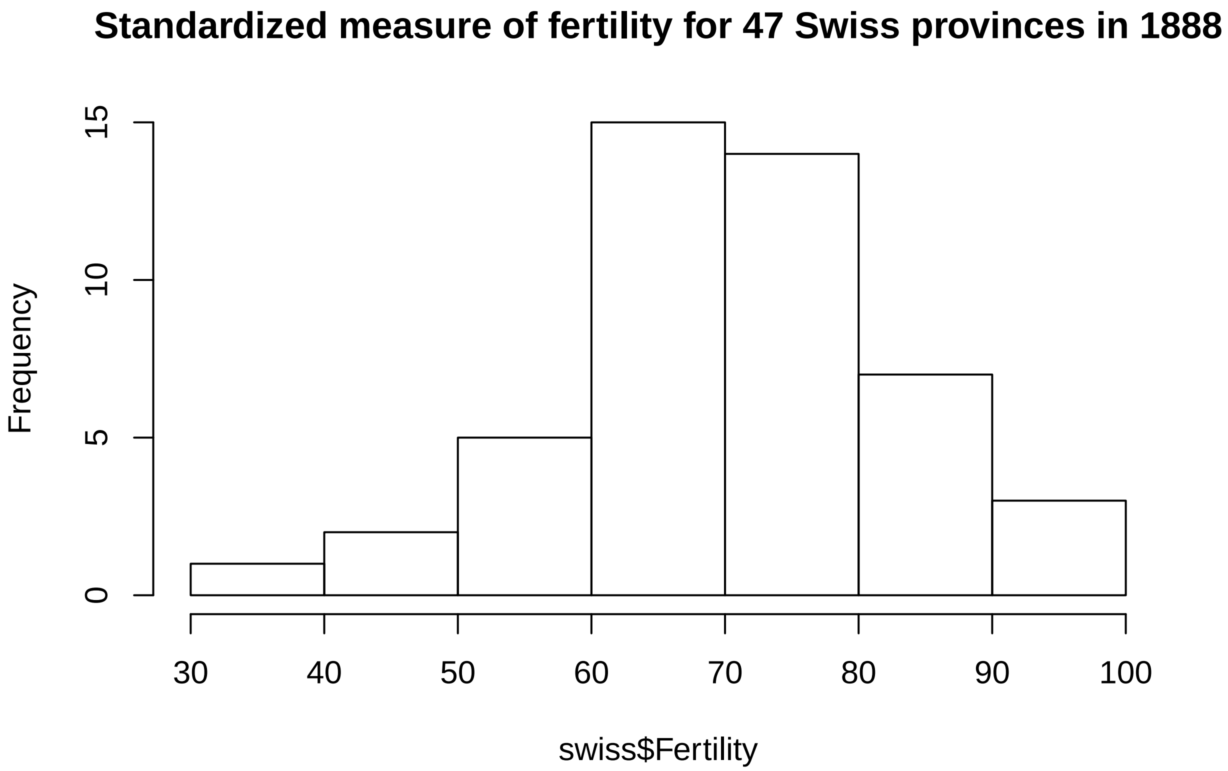
```
summary(swiss$Fertility)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   64.70   70.40   70.14   78.45   92.50
```

```
summary(swiss$Catholic)
```

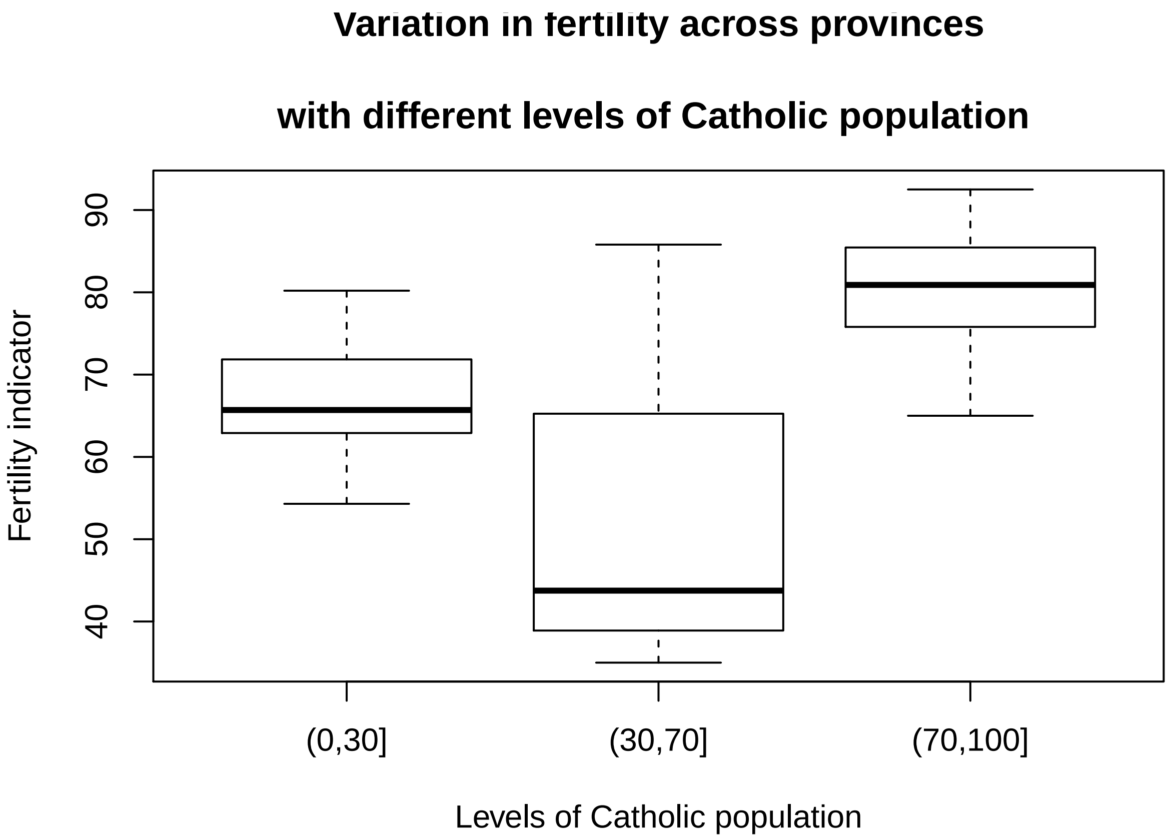
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.150   5.195  15.140  41.140  93.120 100.000
```

```
hist(swiss$Fertility, main = 'Standardized measure of fertility for 47 Swiss provinces in 1888')
```



To understand the relationship between fertility and Catholic better, we create a factor variable ‘factor’ that creates 3 categories for the variable Catholic.Variable factor takes 3 levels 0 to 30, 30 to 70 to 100. Level 0 to 30 represents those provinces where the Catholic population is less than or equal to 30% and so on.

```
swiss$factor <- cut(swiss$Catholic, c(0, 30, 70, 100))
boxplot(swiss$Fertility ~ swiss$factor, main = 'Variation in fertility across provinces
\nwith different levels of Catholic population \n', xlab = 'Levels of Catholic population',
ylab = 'Fertility indicator')
```



The boxplot shows that provinces with a Catholic population between 30 and 70 percent have a substantially lower ferility level than the other two categories.

```
cor.test(swiss$Fertility, swiss$Catholic)
```

```
##
##  Pearson's product-moment correlation
##
## data:  swiss$Fertility and swiss$Catholic
## t = 3.5107, df = 45, p-value = 0.001029
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2036326 0.6626204
## sample estimates:
##       cor
## 0.4636847
```

The Pearson's correlation for this pair of variables is 0.46 and hence, shows a positive but moderately strong relationship.

We can also further explore the data by way of a simple regression exploring the relationship between fertility and the percentage of Catholics living in a province.

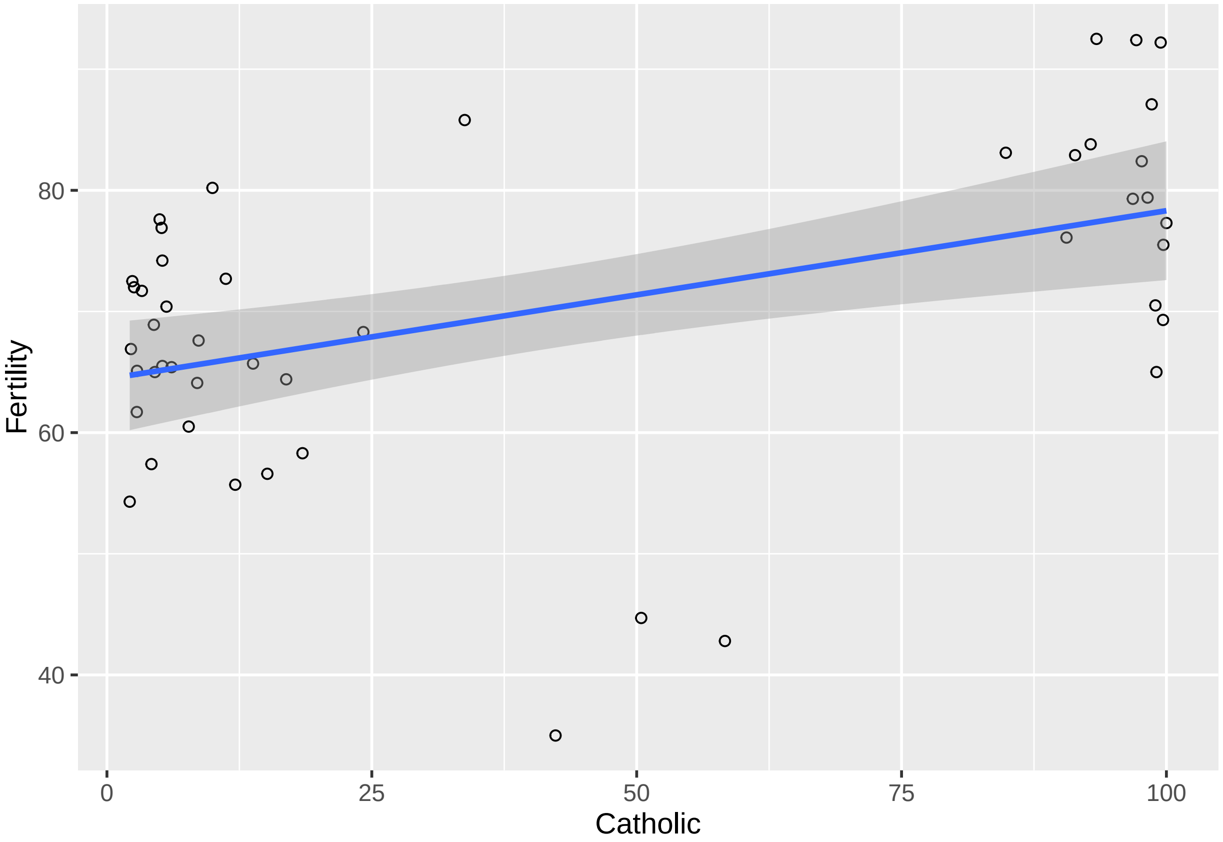
```
fertilreg <- lm(swiss$Fertility ~ swiss$Catholic)

stargazer(list(fertilreg), header = F, float = F, single.row = T)
```

	Dependent variable:
	Fertility
Catholic	0.139*** (0.040)
Constant	64.428*** (2.305)
Observations	47
R^2	0.215
Adjusted R^2	0.198
Residual Std. Error	11.190 (df = 45)
F Statistic	12.325*** (df = 1; 45)
Note:	*p<0.1; **p<0.05; ***p<0.01

Finally, we can plot the values and apply a regression line to get a further visual perspective:

```
ggplot(swiss, aes(x=Catholic, y=Fertility)) +
    geom_point(shape=1) +     # Use hollow circles
    geom_smooth(method=lm)
```



II) New York Air Quality Measurements

```
data(airquality)
names(airquality)
```

```
## [1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

The dataset contains 154 observations and 6 variables. These are daily readings of the given air quality measures from May 1, 1973 to September 30, 1973.The mean temperature over this period was 77.88 F while the mean wind speed was 9.958 mph. The months are taken as numerical variales and the Day is also numeric (1 to 30/31 for each month).

```
summary(airquality)
```

```
##      Ozone           Solar.R           Wind             Temp      
##  Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00  
##  1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00  
##  Median : 31.50   Median :205.0   Median : 9.700   Median :79.00  
##  Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88  
##  3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00  
##  Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00  
##  NA's   :37       NA's   :7                                       
##      Month            Day      
##  Min.   :5.000   Min.   : 1.0  
##  1st Qu.:6.000   1st Qu.: 8.0  
##  Median :7.000   Median :16.0  
##  Mean   :6.993   Mean   :15.8  
##  3rd Qu.:8.000   3rd Qu.:23.0  
##  Max.   :9.000   Max.   :31.0  
## 
```