Swiss Dataset exercises

Karol Topor

2023-10-30

Introduction

These exercises are based on the `swiss` dataset available in R. The `swiss` dataset contains data on fertility and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Exercises

1. Basic Statistics:

Get the structure of the swiss dataset using the str() function.

```
str(swiss)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

Calculate basic statistics for a variable (mean, median, standard deviation) using summary().

```
mean(swiss$Fertility)
```

```
## [1] 70.14255
```

```
median(swiss$Fertility)
```

```
## [1] 70.4
```

```
sd(swiss$Fertility)
```

```
## [1] 12.4917
```

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

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

2. Correlations:

Calculate the correlation between Fertility and Education.

```r
cor(swiss$Fertility, swiss$Education)
```

```
## [1] -0.6637889
```

Create a correlation matrix for all numeric variables in the dataset.

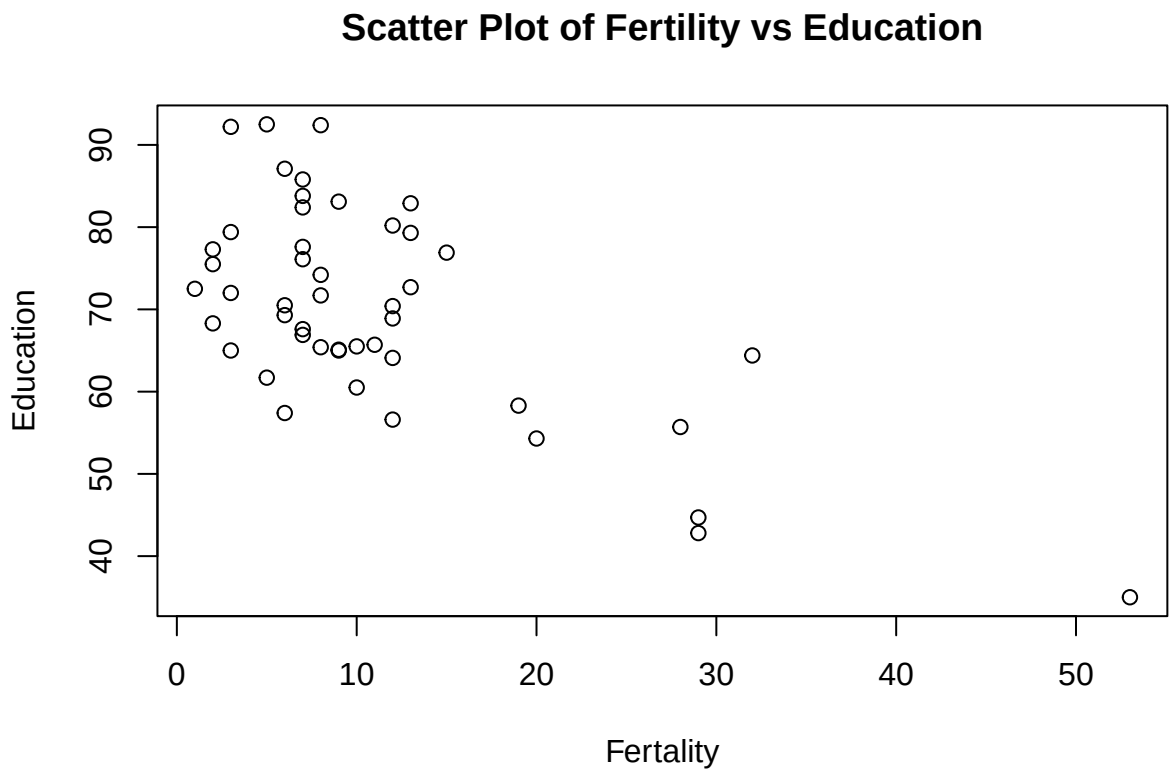
```r
cor(swiss)
```

```
##                   Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility              0.41655603
## Agriculture           -0.06085861
## Examination           -0.11402160
## Education             -0.09932185
## Catholic               0.17549591
## Infant.Mortality       1.00000000
```

3. Visualization:

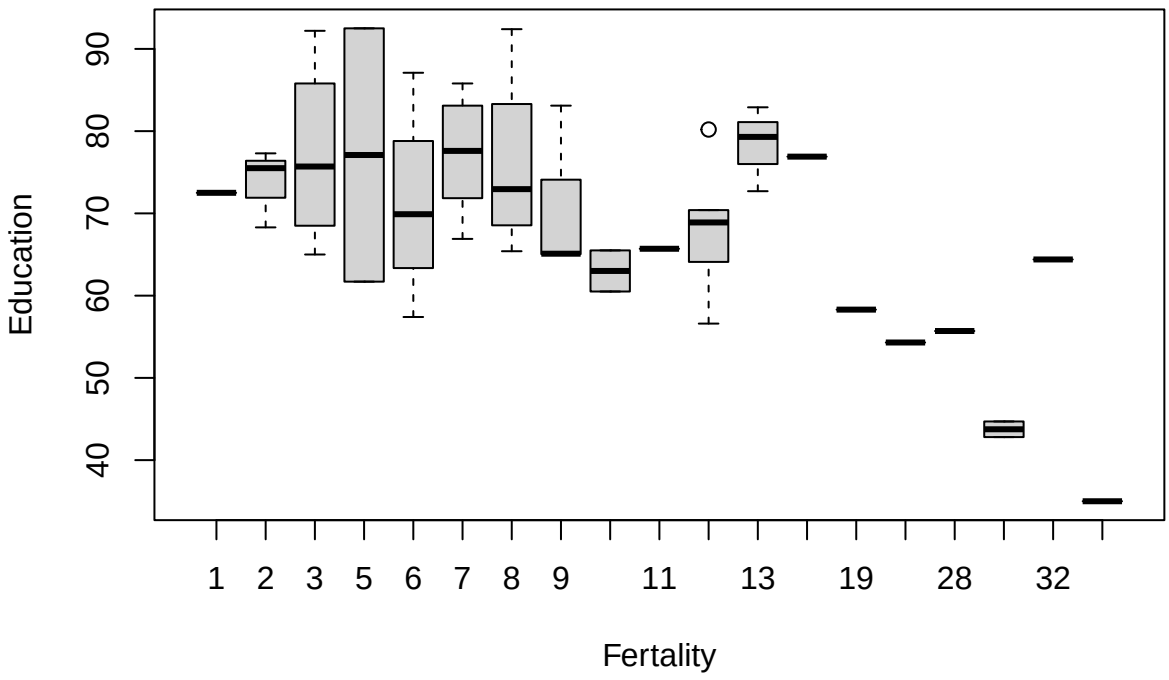
Create a scatter plot of Fertility against Education.

```r
plot(swiss$Education, swiss$Fertility,
     ylab = "Education",
     xlab = "Fertality",
     main = "Scatter Plot of Fertility vs Education"
     )
```



Create a boxplot of Fertility by Education.

```
boxplot(Fertility ~ Education, data = swiss, xlab = "Fertality", ylab = "Education")
```



4. Linear Regression:

Perform a simple linear regression to predict Fertility based on Education.

```
lm_model <- lm(Fertility ~ Education, data = swiss)
summary(lm_model)
```

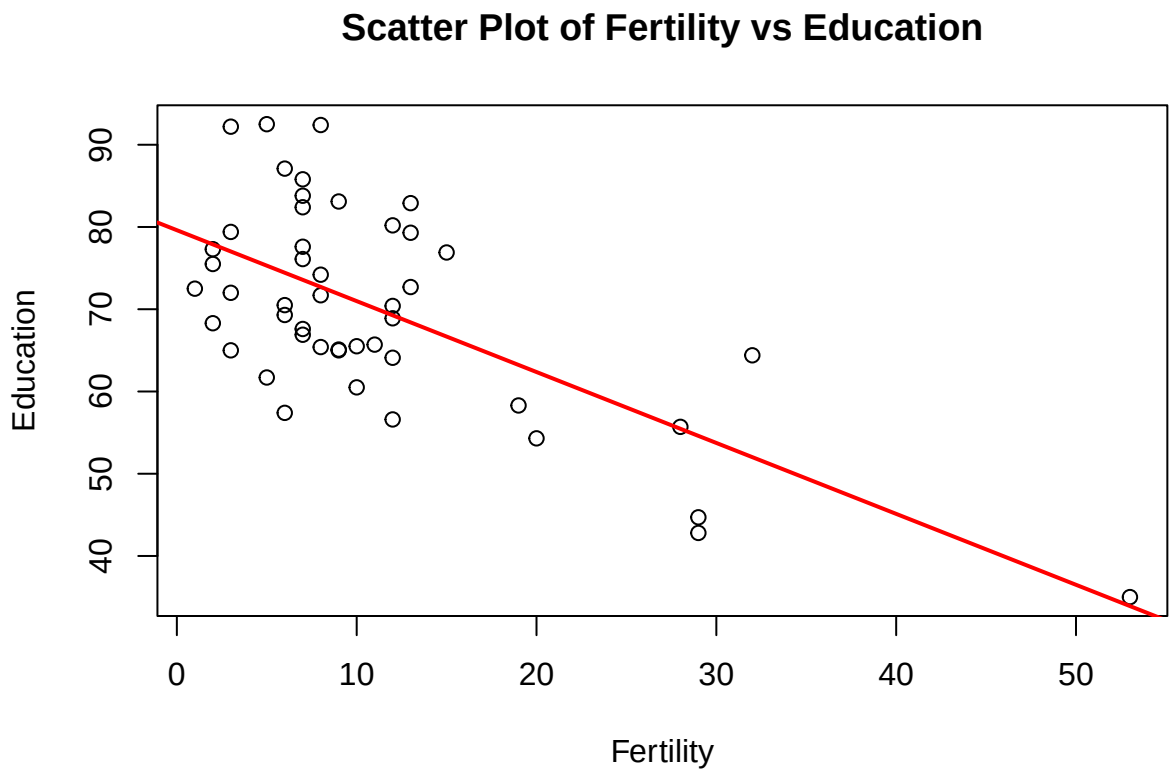
```
##
## Call:
## lm(formula = Fertility ~ Education, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.036  -6.711  -1.011   9.526  19.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## Education    -0.8624     0.1448  -5.954 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

Add the regression line to your scatter plot from the previous exercise.

```r
# Create the scatter plot with your specified labels and title
plot(swiss$Education, swiss$Fertility,
     ylab = "Education",
     xlab = "Fertility",
     main = "Scatter Plot of Fertility vs Education"
     )

# Fit the linear regression model
model <- lm(Fertility ~ Education, data = swiss)

# Add the regression line to the existing plot
abline(model, col = "red", lwd = 2)
```



Interpret the summary output of the regression model.

5. Multiple Regression:

   Perform a multiple linear regression to predict Fertility based on all other variables. Interpret the summary output of the regression model.

6. Residual Analysis:

   Create a plot of the residuals from your multiple regression model against the fitted values.

7. Subset Analysis:

Create a subset of the data including only provinces with Agriculture greater than 50. Perform a simple linear regression to predict Fertility based on Education in this subset. Compare the coefficients of this model to the coefficients of the simple linear regression model from exercise 4.