

Biostat II: Lab 8, Correlation and Linear Regression in R
Date: 5 May 2008

The data set `swiss` in R contains data on “Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.” The `swiss` dataset consists of:

A data frame with 47 observations on 6 variables,
each of which is in percent, i.e., in $[0,100]$.

```
[,1] Fertility  Ig, 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.
```

All variables but Fertility give proportions of the population.

1. Exploratory analysis and correlation

- (a) Load the `swiss` data set in R by typing:
`data(swiss)`
- (b) Type `?swiss` to read more about this dataset
- (c) Eventually, we will be using the `Fertility` variable as our outcome in a linear regression. Familiarize yourself with the distribution of this variable by creating a histogram:
`hist(swiss$Fertility)`
- (d) Create a matrix of scatter plots for all pairs of variables in this data set:
`plot(swiss[,1:6])`
- (e) The command above should have produced a grid of scatterplots, allowing us to visualize the correlation between all pairs of variables in this data set. By looking at these scatterplots, we can get a pretty good idea of the correlation coefficients corresponding to each of these pairwise relationships. Which relationships in the plot are strongest? Which are weakest? Does any scatter plot have correlation coefficient close to 1? close to zero?
- (f) To check your visual inspection of the scatter plots in part 3b), calculate the correlation coefficients for this data set:
`cor(swiss[,1:6])`

2. Now that we've looked at the relationships in the `swiss` data set using correlation, let's look these relationships more formally using a simple linear regression model with a continuous predictor

- (a) Write down the model for a simple linear regression of `Fertility` (the outcome) by `Agriculture` (the predictor). Include both the probability part of the model and the systematic part.
- (b) The `lm` command performs linear regression in R. Using the `lm` function, run your linear regression of `Fertility` on `Agriculture` :

```
model1 <- lm(Fertility ~ Agriculture, data=swiss)
summary(model1)
```

- (c) Interpret the estimates of your model coefficients (the estimated intercept and slope).
- (d) Based on your estimate of the slope, are the two variables `Fertility` and `Agriculture` positively or negatively associated? Why? Does this correspond with the correlation that you observed in Question (1)?

3. Now we will look at the relationship between `Fertility` and `Agriculture` using a binary version of `Agriculture`.

- (a) Dichotomize the `Agriculture` variable such that the new variable, called `highAgr`, takes on value '1' if `Agriculture` ≥ 50 and '0' if `Agriculture` < 50 . This variable identifies the provinces that have greater than or equal to 50% of males involved in agriculture as occupation.

```
swiss$highAgr <- ifelse(swiss$Agriculture>=50,1,0)
```

- (b) Write down the model for a simple linear regression of `Fertility` (the outcome) by `highAgr` (the predictor). Include both the probability part of the model and the systematic part.
- (c) Using the `lm` function, run your linear regression of `Fertility` on `highAgr` :

```
model2 <- lm(Fertility ~ highAgr, data=swiss)
summary(model2)
```

- (d) Interpret the estimates of your model coefficients (the estimated intercept and slope).

4. Repeat exercises 2(a-d) again but this time use `Education` as your predictor variable (x). What do you observe?