Statistical Modeling I
Practical in R

Practical in R

In this practical, we will work with the Swiss dataset provided with R. We will look at different subset models and their analysis.

Looking at the Swiss data in R, we find that it contains standardized fertility measure and socio-economic indicators for each of the 47 French-speaking provinces of Switzerland in about 1888. It is composed of 47 observations on 6 variables, each of which is a percentage (i.e. in [0,100]), including:

- Fertility
- Agriculture, % of men involved in agriculture as occupation;
- Examination, % draftees receiving highest mark on army examination;
- Education, % of education beyond primary school for draftees;
- Catholic, % of catholic (as opposed to protestant);
- Infant mortality, % of live births who live less than 1 year.

We are interested in predicting the infant mortality using multiple regression models.

1. Look at the plots of these variables by using `pairs` or by using the correlation between variables.
2. For looking at the multiple regression models, we need to introduce a novel package by using the following commands:

   ```
   > install.packages("leaps")
   > library(leaps)
   ```

   By using the function `regsubsets` available in the leaps package, we can select among these predictors by using the best subset selection. This function performs the best predictor subset selection by identifying the best model that contains a given number of predictors, where the best is quantified by using the smallest residual sum of squares. In particular, one can choose the maximum number of predictor models.

   ```
   > best.subset <- regsubsets(y~., swiss, nvmax=5)
   ```

By using the $adj(R^2)$ and the Mallow's C_p metrics, say which is the best model.

3. Look at the plots of the best models and show the coefficients of the best $2-$ variable models.

4. Which model you prefer between the model with fertility and education or the one with just education?