

Statistical Modeling I

Practical in R

Practical in R

In this practical, we will work with the dataset on presidential elections in US in year 2000 (on the <https://electionlab.mit.edu/data> is possible to found other data). We will look at how to select the best model by using the AIC and other measures.

In the file USElection.csv, we have different variables of interest, such as the fraction of the state's total counted vote for George W. Bush, which is the response variable. In the file, we find the following eleven columns for each of the US states:

- $Y = \%Bush$ which is the percentage of votes for G.W. Bush;
- $X_1 = UnEmpR$ which is the unemployment rate;
- $X_2 = Pop$ is the total population of the state;
- $X_3 = \%Male$ is the percentage of male;
- $X_4 = \%Pop > 65$ is the percentage of population older than 65;
- $X_5 = \%NonMetr$ is the percentage of rural (nonmetro) population;
- $X_6 = \%PopPov$ is the percentage of population below the poverty level;
- $X_7 = NuHouse$ is the total number of households;
- $X_8 = \%Inc > 50$ is the percentage of house income bigger than \$50000;
- $X_9 = \%Inc > 75$ is the percentage of house income bigger than \$75000;
- $X_{10} = \%Inc > 100$ is the percentage of house income bigger than \$100000.

Note to find the VIF values, we should use the command `vif(model)` and we should install and load the correct library:

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

1. Fit the full model for the response variable by including all the explanatory variables and then comment on the values of VIF.
2. Use R to find the best reduced model using the AIC procedure by using

- the backwards elimination procedure if we have defined the full model with all the variables

```
> reduced.model <- step(mody, direction="backward")
```

- the forward fit model if we start by the null model

```
aic.forward.model <- step(modyn, scope=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, direction="forward")
```

3. Fit the best reduced model (add the regressors in the same order as found in the previous point) and comment on
 - (a) the model assumptions;
 - (b) the significance of the overall regression and of the individual explanatory variables in presence of other variables;
 - (c) the values of VIF;
 - (d) the values of adjusted R^2 ;
 - (e) the leverage values and the Cook's distance.