# Duże bazy danych

## Estymacja błędu predykcji i kryteria informacyjne

Generate the design matrix $X_{1000\times950}$ such that its elements are iid random variables from $N(0, \sigma = 0.1)$. Then generate the vector of the response variable according to the model

$$Y = X\beta + \epsilon$$

, where $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$ and $\epsilon \sim N(0, I)$.

Perform the following analyses using the model with

i) 5 first variables

ii) 10 first variables

iii) 20 first variables

iv) 100 first variables

v) 500 first variables

vi) all 950 variables.

- For each of the considered models

  a) Estimate $\beta$ with the Least Squares method and calculate residual sum of squares and the true expected value of the prediction error

  $$PE = E||X(\beta - \hat{\beta}) + \epsilon^\star||^2 \ ,$$

  where $\epsilon^\star \sim N(0, I)$ is a new noise vector, independent on the training sample.

  b) Use the residual sum of squares to estimate $PE$ assuming that $\sigma$ is known and replacing $\sigma$ with its regular unbiased estimator.

  c) Estimate $PE$ using leave-one-out crossvalidation (do not perform analysis 1000 times but apply the formula for leave-one-out cross-validation error provided in class).

- Select the optimal model using two versions of AIC: for known and unknown $\sigma$.

- Repeat the above calculations 100 times and

  – for each of the considered models compare the boxplots of $\hat{PE} - PE$ for three estimates of $PE$, mentioned above.

  – Provide histograms of the number of false negatives and false positives produced by both versions of AIC (with known and unknown $\sigma$).

2. Use BIC, AIC, RIC, mBIC i mBIC2 (you can use *bigstep* library in R) to identify important covariates when the search is performed over the data base date consisting of

   i) 20 first variables

   ii) 100 first variables

   iii) 500 first variables

   iv) all 950 variables.

a) Report the number of false and true discoveries and the square error of the estimation of the vector of expected values of $Y : ||X\beta - \hat{Y}||^2$.

b) Repeat point a) 100 times and report the estimated power, FDR and mean squared error of the estimation of expected values of Y for all criteria considered above. Critically summarize the results.

3. Compare RIC, mBIC and mBIC2 using example vi) of Problem 1 when the vector of true regression coefficients contains 50 nonzero entries, i.e. $\beta_i = 3$ for $i = 1, \ldots, 50$ and $\beta_i = 0$ for $i = 51, \ldots, 950$.

4. Generate the vector of the response variable according to the model

$$Y = X\beta + \epsilon \ ,$$

where $\beta_1 = \ldots = \beta_{30} = 10$, $\beta_{31} = \ldots = \beta_{950} = 0$ and $\epsilon_1, \ldots, \epsilon_n$ are iid from a

a) shifted exponential distribution with $\lambda = 1$

b) Cauchy distribution.

i) Use mBIC, mBIC2, rBIC and rBIC2 to identify important covariates. Report the number of true and false discoveries.

ii) Use variables selected by rBIC2 and estimate the corresponding regression coefficients using least squares as well as the robust regression based on the Huber and Bi-square objective functions. Report the mean square error of estimation of regression coefficients.

iii) Repeat this experiment 100 times and report estimated FDR and Power. For this part you do not need to estimate regression coefficients (i.e. you do not need to perform step ii).

Malgorzata Bogdan