Advanced methods of statistical learning

Multiple regression. Multiple testing

1. **Influence of correlation**.

   a) Generate the matrix $X_{100\times 2}$ such that its rows are iid random vectors from the multivariate normal distribution $N(0, \Sigma/100)$, where

   $$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix} .$$

   Then generate the vector of response variable as $Y = \beta_1 X_1 + \epsilon$, where $\beta_1 = 3$, X_1 is the first column of X and $\epsilon \sim N(0, I)$.

   b) Construct the 95% confidence interval for β_1 and perform the 0.05 significance level t-test for the hypothesis $\beta_1 = 0$ using the model of the simple linear regression $Y = \beta_0 + \beta_1 X_1 + \epsilon$ and using the model with both explanatory variables $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Compare and explain the results.

   c) Calculate the standard deviation of the estimate of β_1 and the power of identification of X_1 in both models.

   d) Generate 1000 independent copies of the vector of errors ϵ and 1000 related copies of the vector of response variable. For each of such data sets estimate β_1 and perform the test for the significance of β_1 in both models (with one and two explanatory variables). Estimate the standard deviation of β_1 and the power of the test and compare these values to the theoretical results obtained in point c).

2. **Influence of dimension.** Generate the design matrix $X_{1000\times 950}$ 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 $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$.

   a) Estimate values of regression coefficients and use t-tests at the significance level 0.05 to identify true regressors when the model is build using first

      i) 1
      ii) 2
      iii) 5
      iv) 10
      v) 50
      vi) 100
      vii) 500
      viii) 950

      columns of the design matrix. For each of these models report residual sum of squares $||Y - \hat{Y}||^2$, square error of the estimation of the vector of expected values of Y : $||X\beta - \hat{Y}||^2$, estimate of the prediction error PE, p-values corresponding to the first two explanatory variables and the number of false discoveries. Which model would be selected based on the estimate of PE (C_p criterion) ?

   b) Repeat point a) when the models are build using variables with the largest estimated regression coefficients. Compare values of calculated statistics with those obtained in point a). Which model would be selected based on the estimate of PE ?

c) For each of dimensions specified in point a) calculate the expected value of the elements on the diagonal of $(X_s^T X_s)^{-1}$, where X_s contains first s columns of X, the power of identification of X_1 and the expected number of false discoveries produced by 0.05 significance t-tests.

d) Repeat the generation of X and Y and the steps a) and b) 1000 times. For each of the sub-problems estimate the power of identification of X_1 and the expected number of the false discoveries. Compare these results to the theoretical values calculated in point c). Additionally, compare the average model size selected by the C_p criterion for points a) and b).

Malgorzata Bogdan