Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Applied Statistics
June 18th, 2021

Problem n.3

The data contained in the file `students.txt` were collected from 148 students enrolled in an introductory statistics course at a large university. The data consist of: gender, age [years], height [inches], distance from home town of student to the university [miles], number of siblings, number of hours spent on computer per week, number of hours spent exercising per week, number of music CDs owned, number of hours spent playing games per week, number of hours spent watching TV per week.

a) Formulate a linear regression model for the number of hours spent watching TV per week, as a function of all the other variables. Include in the model a possible dependence of the number of hours spent watching TV per week on the categorical variable `gender`, but only in the intercept. Report the estimates of the 10 parameters of the model (the coefficients $\beta_0, \dots, \beta_9$ and the errors' standard deviation σ). Analyze the model residuals and verify the assumptions of the model.

b) Perform a variable selection through a Lasso method, by setting the parameter controlling the penalization to $\lambda = 0.3$. Report the significant coefficients.

c) Optimize the parameter λ within the range [0.01; 1] via cross-validation. Report the optimal λ and the corresponding estimated coefficients.

d) Using the linear model found in the previous point, predict, with a point-wise estimator, the number of hours spent watching TV per week by a new student whose characteristics are: gender=male, age=21, height=73, distance=100, siblings=1, computertime=10, exercisehours=2, musiccds=35, playgames=4.

Upload your results here:
`https://forms.office.com/Pages/ResponsePage.aspx?id=K3EXCvNtXUKAjjCd8ope612LHtvIHvFEsEi2L6mhPglUOUJNRlJWNVNIMU5PRkxJUlpFSTMzQk5LVS4u`