Homework 2

Using classical methods for feature selection for linear models. Analyzing fit for possible improvements by feature selection

The data for homework is in the file `hw2_studybookno.csv` (where `studybookno` is your study book number), which is attached to the assignment in Moodle. The aim is to fine a good linear model for predicting the variable `y`. For the current assignment use the data as it is (no elimination of possible outliers, no transformations except adding dummies for nominal variables).

I remind you that you are expected to solve the homework problems by yourself, sharing solutions (and copying parts of other student's solutions) is not allowed.

Please submit your .R or .Rmd file and and html/pdf file which shows you code together with output and comments

Problem 1

Transform your data so that each level of a nominal variables (except one) is replaced by a separate dummy variable (`step_dummy()` with default options). In further model fitting use this transformed data set.

Problem 2

Use full transformed data set to find best models with two-way stepwise search with the command `stats::step()`, starting from the simplest model (which predicts a constant value), by using BIC for model selection. Also find the best model by backward search, starting with a model with all variables and using BIC for model selection. Are the obtained models the same? If not, then which one should be preferred for future predictions if we can assume that the model used for likelihood computations is adequate for the data? Show a table with information about the coefficients of the best model (estimated values,standard errors,p_values)

Problem 3

Use exhaustive search with `leaps::regsubsets()` to find the best model according to BIC which uses up to 10 variables. Show a table with information about the coefficients of the best model (estimated values,standard errors,p_values). Is it the same as the best model from the previous step? If not, then which of the models considered so far is the best according to BIC?

Problem 4

Analyze the fit of the best linear model found in previous problems for possible bias in the predictions with respect to numerical variables in the model (plots about residuals with respect to each of the continuous variable in the model, together with a curve showing approximate average value of residuals for each value of the variable under consideration, added with `geom_smooth()` command). Is there any evidence that some nonlinear terms with respect to the variables should be added to the model (or to the data before model fitting)?