Model Evaluation

The task of this exercise is to evaluate predictive performance of several linear models and choose the best one for predicting medical costs billed by the health insurance. The prediction influences monthly payments of the clients. Therefore, a high predictive accuracy is a priority. The file `insurance-all.Rdata` contains information on the covariates and dependent variable which are described in detail in Table 1. The simulated data set was created using demographic statistics from the U.S. Census Bureau for the covariates, and thus approximately reflect real-world conditions.[1]

Table 1: Description of the Variables

Variable	**Description**
age	age of primary beneficiary (numeric)
sex	insurance contractor gender (factor: “female”, “male”)
bmi	body mass index, ideally 18.5 to 24.9 (numeric)
children	number of children covered by health insurance / number of dependents (numeric)
smoker	smoking (factor: “yes”, “no”)
region	the beneficiary’s residential area in the US (factor: “northeast”, “southeast”, “southwest”, “northwest”)
charges	individual medical costs billed by health insurance

Group Home Assignment (max. 4 points)

The mandatory group home assignment has to be submitted before 12:00 o’clock prior PC-session 2. It is obligatory to solve the assignment in R Markdown with `echo = TRUE` option for every code chunk and generate a PDF file with the solution. Generating an HTML file from the R Markdown and converting it into the PDF is also possible. Make sure that the final PDF file has no readability issues. The PDF with the answers to the six questions below as well as the file with the R Markdown code has to be submitted via Canvas.

Download the data set `insurance-all.Rdata` from Canvas. Load the data into R. Install and load the package `ggplot`.

1. How many observations are in the whole dataset? How many covariates were collected? (0.5 points)

2. What is the highest number of children who are covered by one health insurance? (0.5 points)

3. Look at the percentage shares of smokers and non-smokers within each region, e.g. in one region there could be 80% smokers and 20% non-smokers. Which region has the lowest share of smokers? What is the share of smokers there? (0.5 points)

[1] Lantz, B. (2013) Machine Learning with R. Packt Publishing.

4. Create a scatter plot with `charges` on the y-axis and `age` on the x-axis. Distinguish in the plot by color which data points belong to smokers and non-smokers (coded in covariate `smoker`). Describe the patterns in the data. (0.5 points)

5. Write a function that has a data argument and three string arguments `x.variable, y.variable, color.variable` which are used to generate a scatter plot of two variables against each other and use the third to color the points. Use this function to plot `x.variable="bmi", y.variable="charges", color="sex"` and interpret the results. (1 point)
   Hint: the aesthetic mapping `aes()` in `ggplot` does not interpret string inputs. Search for an alternative.

6. Write a function that creates a boxplot of the `bmi` variable split by another variable that is passed as the argument. Do you see any difference in the `bmi` by region? (1 point)

We will solve the following exercises during the PC session. They are not part of the group home assignment.

For the prediction task you consider the following three sets of predictors for a linear model:

Model 1: `smoker`, `age` and `age*smoker`,
Model 2: all covariates and all two-way interaction terms with the variable `smoker`,
Model 3: all covariates with all two-, three- and four-way interactions with all the variables.

Exercise 1: Prepare Data

1. Install and load the package `caret` needed for Exercise 4.

2. Plot a histogram for the variable `charges`. Set the bandwidth of the histogram bins to 1000.

3. Generate new variable `logcharges` = log(`charges`). Plot the histogram of the variable `logcharges`. Set the bandwidth of the histogram bins to 0.1.

4. Split the data set into a training set and test set. Use 80% of the sample for the training set and 20% for the test set.

Exercise 2: In-Sample MSE

1. Take `logcharges` as the dependent variable and estimate the three models by OLS using the training sample. Compute the in-sample MSEs. Which model has the smallest in-sample MSE and why? Plot the in-sample MSEs in a barplot.

2. Check the in-sample fit of the three models by:

- creating a “fit plot”, i.e. plot `logcharges` on the y-axis and fitted OLS values on the x-axis.
- creating a “residual scatter plot”, i.e. plot residuals on the y-axis and fitted OLS values on the x-axis.

Exercise 3: Training and Test Data Sets

1. Name advantages and disadvantages of a Training-Test split for the out-of-sample model evaluation.

2. Estimate the three models on the training set and calculate the out-of-sample MSE on the test set. Which model has the smallest out-of-sample MSE? Plot the out-of-sample MSEs in a barplot. Name a couple of alternatives to MSE that can be used for model evaluation.

Exercise 4: Cross-Validation

1. Name advantages and disadvantages of Cross-Validation for the out-of-sample model evaluation.

2. Estimate the out-of-sample MSE by K-fold Cross-Validation for $K = \{5, 10, N\}$ for all 3 models. Does the choice of K matter here? Which model has the smallest CV error? Plot all the out-of-sample MSEs in a barplot.

Useful links:

- Description of the `caret` package: `http://topepo.github.io/caret/index.html`.