ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Bernward Asprion
29.11.2024

MCR DAS WS2024 Homework1 – v02

V02: 29.11.2024: Task description has been adjusted. In 2e, lm02 has been replaced by lm04.

General conditions:

Deadline	04.12. 08:50
Cooperation	The tasks can be worked on together (up to 3 students). This applies above all to the code, or when discussing the procedure together. Text passages must either be written in your own words (adding an adjective or changing the sentence structure does not fall under "own words") or, if external content is used, a reference must be provided. In the case of obvious copies of text passages between students, no points are awarded for the corresponding subtasks. External text content can be added to support your own explanations. However, this is generally not necessary - only the texts you have produced yourself will be assessed.
Submission	Please submit only one Jupyter notebook and no other files.
Programming language	Python
Grading	This submission accounts for 25% of DAS's final score.

Guidelines for the Jupyter Notebook:

- File name **HW1-{UID}.ipynb**, e.g. **HW1-ds123456.ipynb**
- Do not include your own name in the notebook (anonymized evaluation)
- Place all answers under the appropriate headings,
  e.g.: # 2.task, ## 2a, ## 2b
- Use of external content:
  For both code snippets and text: include references, e.g.:
  # https://stackoverflow.com/questions/7695982/what-does-pythons-dir-function-stand-for
  # ChatGPT
- Graphics: Insert (meaningful) headings and label axes (correctly), insert legend if necessary.
- Please answer the questions in a maximum of 5 sentences each.

Please notice:
There are many more points for single answers than for single code snippets!

The aim is to estimate doctor's fees.
The data is contained in the file doctors_fees.csv and contains the following information:

Column	Description
age	Age of the person
sex	Gender (female, male)
bmi	Body mass index
children	Number of children
smoker	Smoking
region	Residential area in US
charges	Billed doctor 's fee in USD

Task 1: Data Analysis

a) Import the data from the "doctors_fee.csv" file into a dataframe.
b) How many rows and how many columns does the data frame have?
c) Delete the column id from the dataframe.
d) Change the column names to lower case.
e) Change the column names "sex" to "gender".
f) In which columns are there how many missing values?
g) If there are missing values, delete the corresponding rows from the dataframe.
h) Replace „female" with 0, and „Male" with 1.
i) Have you found incorrect values (which ones)?
j) Estimate the distribution of age with a visualization. Comment on this distribution.
k) What is the distribution of gender? What is the distribution on smoker? What is the distribution on region? Comment these distributions.
l) Estimate the distribution of "bmi" with a visualization. Comment on this distribution.
m) What is the distribution of the charges?
n) Create a correlation matrix (with the numeric variables including gender). Comment all correlation values. E.g.: on average, do women or mean have higher charges? Which correlations look strange?
o) Create a scatterplot with „charges" vs age. Comment this plot.
p) Create a „side-by-side-boxplot" of your choice. Comment this plot.
q) Create a new column „bmi_age" – as the multiplication of bmi and age.
r) Delete faulty data records - see task (i).

Task 2: Linear Regression

a) Split the dataframe from Task 1 into a training part and a test part in the ratio 75% to 25%.
b) Create and fit the following linear regression models:
   lm01: Target = Charges, features = bmi
   lm02: Target = Charges, features = bmi, sex
   lm03: Target = Charges, features = bmi, smoker
   lm04: Target = Charges, features = age, sex, bmi, children, smoker, region, bmi_age
c) For lm02:
   What are values for R2? Comment these values.
   Which features are significant?
   Write the regression equation as a formula (rounding is allowed)
d) For lm04:
   What are values for R2? Comment these values.
   Which features are significant?
   Write the regression equation as a formula (rounding is allowed).
   How big is the expected increase in charges if a person has one more child?
e) Read the second note in the summary table of lm04.
   What does that mean and what can we do?
f) Create the fitted vs target plot for lm02. Comment this plot.
g) Create the fitted vs target plot for lm04. Comment this plot.
h) lm04: comment the values of R2 with the training data and the test data, and their relationship.
i) Lm04: plot the residual plot. Comment the plot.
j) Predict the charges of one person with the following code:

```
data = {'age': [10], 'sex':[0], 'bmi': [32], 'children': [0], 'smoker': ['no'],
'region': ['southeast'], 'bmi_age': [320]}
df1 = pd.DataFrame(data)
lm04.predict(df1)
```