Project 2: Heating in Sønderborg II

Formalities, structure and expectations – second mandatory project

In this project, we'll formulate and select a suitable multiple linear regression model which describes the heat consumption during winter for four houses in Sønderborg. The assignment must be solved using the software program Python. Some code suggestions are provided but, in addition, it's a good idea to take a look at the Python code from project 1, as well as chapter 5 and 6 of the book.

The results of your analysis must be documented in a report with tables, figures, appropriate mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in an appendix. Present the results of your analysis as you would when explaining them to one of your peers. Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. Code should not be included in the report itself but must be handed in as an appendix (a .py or .ipynb file). The report and appendix must be handed in under Assignments on Learn at: `Projekt 1: Varmeforbrug i Sønderborg II`

The report should not exceed 6 pages (excluding figures, tables, and the appendix). A page contains 2400 characters.

It's important that you describe and explain the Python output in words – figures and tables cannot stand alone.

When you're asked to state a formula, insert numbers, and then perform certain computations, it's important to show that you've done this by including your intermediate

results. (In these cases, it's not enough to report results obtained directly from Python). Furthermore, remember that when performing a hypothesis test, you must go through the following steps: State the hypothesis and significance level (α), compute the test statistic and state it's distribution, compute the p-value, and summarize your findings.

Figures and tables are not included in the assessment of the length of the report. However, it's not in itself an advantage to include many figures, if they aren't relevant!

You may work in groups, but the report must be written individually. Questions may be addressed to the teaching assistants, see the guidelines on the *Projects* page of the course website.

Data

Read the dataset `soenderborg2_data.csv` into Python. As with the first project, a Jupyter notebook is given to you as assitance with the analysis. See the section "Read Data" for code on this first step.
In this project, we'll analyse data from the Sønderborg database, which consists of measurements of the heat consumption of buildings together with climatic measurements like outdoor temperature, solar irradiance, and wind speed. The buildings are typical single-family houses built with bricks during the 1950s and 60s. The houses are located in the area around Borgmester Andersens vej in Sønderborg and are of varying sizes. The climatic measurements were recorded by a weather station located at the local district heating plant.

The dataset for the project consists of observations of five variables:

- `t`: Date of observation
- `houseId`: House id number
- `Q`: Heat consumption (kW)
- `Ta`: Outdoor temperature (°C)
- `G`: Solar irradiance (W/m^2)

In this project, we'll only analyse the data from the winter of 2009/2010, defined as the period from 15 October 2009 to 15 April 2010. Furthermore, we'll only consider the houses with id numbers 3, 5, 10, and 17, respectively. We need to extract the relevant subset of data for the analysis. This may, for example, be done using the code in the section "Filter data" in the notebook.

The dataset has many observations which have one or more missing values (NAs). These observations must be removed from the dataset prior to the statistical analysis. This is also handled in the "Filter data" section in the notebook.

Statistical analysis

In the following, use the dataset which contains observations from houses 3, 5, 10, and 17 during the time period 15 October 2009 to 15 April 2010, which don't have missing values for any of the variables.

a) Present a short descriptive analysis and summary of the data for the variables `Q`, `Ta`, and `G`. Include scatter plots of the heat consumption against the two other variables, as well as histograms and box plots of all three variables. Present a table containing summary statistics, which includes the number of observations, and the sample mean, standard deviation, median, and 0.25 and 0.75 quantiles for each variable.

b) Formulate a multiple linear regression model with the heat consumption as the dependent/outcome variable (Y_i), and outdoor temperature and solar irradiance as the independent/explanatory variables ($x_{1,i}$ and $x_{2,i}$, respectively). Remember to state the model assumptions. (See Equation (6-1) and Example 6.1).

c) Estimate the parameters of the model. These consist of the regression coefficients, which we denote by β_0, β_1, β_2, and the variance of the residuals, σ^2. You may use the Python code from section "Fitting the model - Estimating Parameters" in the notebook.
Give an interpretation of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, explaining what they tell us about the relation between the heat consumption and the model's explanatory variables. (See Remark 6.14). Furthermore, present the estimated standard deviations of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, the degrees of freedom used for the estimated residual variance $\hat{\sigma}^2$, and the explained variation, R^2.

d) Perform model validation with the purpose of assessing whether the model assumptions hold. Use the plots, which can be made using the Python code from the section "Model validation" in the notebook, as a starting point for your assessment. (See section 6.4 on residual analysis).

e) State the formula for a 95% confidence interval for the outdoor temperature coefficient, here denoted by β_1. (See Method 6.5). Insert numbers into the formula, and compute the confidence interval. Use the Python code in section "Confidence interval" to check your result, and to determine confidence intervals for the two other regression coefficients.

f) It is of interest whether β_1 might be -0.25. Formulate the corresponding hypothesis. Use the significance level $\alpha = 0.05$. State the formula for the relevant test statistic (see Method 6.4), insert numbers, and compute the test statistic. State the distribution of the test statistic (including the degrees of freedom), compute the p-value, and write a conclusion.

g) Use backward selection to investigate whether the model can be reduced. (See Example 6.13). Remember to estimate the model again, if it can be reduced. State the final model, including estimates of its parameters.

Now, use the code in section "Test-subset" in the notebook to make a subset of the original data, which contains one observation for each of the four houses, from the winter of 2008/2009. Note that this subset does not contain any of the observations that were used in the statistical analysis above. We'll use this subset to evaluate the prediction capabilities of the final model.

h) Use your final model from the previous question as a starting point. Determine predictions and 95% prediction intervals for the heat consumption, for each of the four observations in the validation set (`D_test`). See Example 6.8, Method 6.9 and the Python code in section "Predictions" in the notebook. Compare the predictions to the observed heat consumptions in the validation set and make an assessment of the prediction capabilities of the final model.
Hence, don't write the formulas in the report, but instead refer to that the Python function `get_predictions` was used for the calculations. The formulas requires a matrix formulation, which are out of the curriculum (to derive the formulas use Equations (6-48) and (6-49) together with the derivations leading to Equations (5-57) and (5-58)).