02_regression_data_modelling_workflow

May 7, 2025

1 Linear Regression - Data modelling Workflow

1. **Formulate problem:**

   What are we trying to predict? What is the purpose of the model?

   - *Prediction*: We want a model to predict a value as part of a workflow or process. Statistical validity of the model is not so important.
   - *Estimation*: We want to understand the values of the regression coefficients (β values or weights) and we should look closely at the assumptions used be the model.

   How accurately do we need the result? Consult with a statistician!

2. **Obtain data:**

   - Often the most expensive and time consuming part of a study.
   - Extract, Transform, Load (ETL):
     - Extract
       - Identify input data and place into analytics system for pre-processing (e.g. Python).
       - Visualise data. A Box and whisker plot gives a good quick overview of ranges, outliers.
     - Transform: Clean data - deal with missing values, invalid values, remove duplicates.
     - Load: Save to data store (to a database, CSV file, Pickled Pandas dataframe, HDF5 file…)

3. **Undertake the regression analysis**

   1. Load data from data storage.
   2. Data exploration and visualistation.
   3. Select features.
      - Engineer features / augment
   4. Scale, normalise, encode (if not done in ETL section).
   5. Choose validation methods (dataset split, cross validation).
   6. Train model, check assumptions and evaluate performance.
      - Are coefficients significant?
      - Are residuals: normally distributed, independant?
   7. Train alternative models and tune best resulting models.
   8. Save model and save the preprocessing methods and values.

4. **Put into production**

- Document and/or publish results.
- Use in research and business processes
- Monitor performance.

[ ]:

[ ]: