A regression model to shorten dissolution test

Samd Guizani

Abstract

- Among many tests, pharmaceutical oral drug products must be tested for their capability to release their active ingredient (API).

- To do so, the product is placed in an agitated vessel, containing a specific aqueous medium, and the dissolved API amount is recorded over the testing time. A product would be declared "***good***" if the released API amount reaches pre-defined levels at given time points. **The dissolution test is often time-consuming**!!! (API release is followed over several hours)

- The studied question is to evaluate if the early measurements of API release can be used to reliably predict the future dissolution profile. **If so, the dissolution test can be shortened**!!!

- The dataset used in the study is an extract of dissolution test results performed on a specific pharmaceutical product between Jan-2019 and Apr-2020.

- The method used relies on **regressions** to develop a model allowing to predict the future dissolution results, based on the onset of API release during the test.

- It was found that **dissolution results recorded within the first 15 min can be used to forecast the dissolution results at 90 min**.

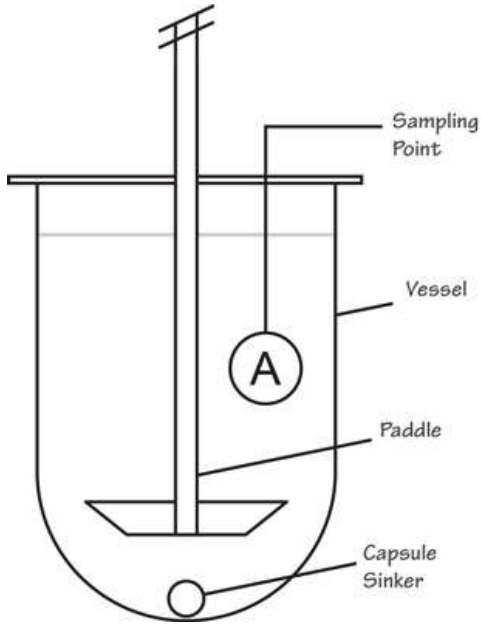
Motivation

- Oral drug product active ingredient (API) release must be tested before any batch is released on the market.
  The manufacturer is responsible for meeting pre-defined approved acceptance criteria.

- API release is evaluated using a Dissolution Test (refer to US or European Pharmacopeia). The testing equipment is presented on the picture on the right. For some products the test can be very long because the product is designed to slowly release the API: the dissolved API amount has to be recorded during several hours.

- A predictive model, using the onset of the API dissolution to reliably predict the future dissolution behavior can help shorten the dissolution test.

- Such model could be useful to Quality Control lab to optimize the use of their resources (equipment, employees, material…) by stopping tests when the early dissolution test results predict with confidence that the rest of the dissolution profile will match the target.



https://www.labhut.com/education-centre/about-dissolution-testing/apparatus-2-paddle-test.html

Dataset(s)

- The dataset is an extract of dissolution test results on a specific product, recorded between Jan-2019 and Apr-2020. It’s a private company dataset.
- It contains 26818 rows and 14 columns. A large amount of data is provided in the dataset (e.g.: dissolution results, technician name, batch numbers, equipment used…)
- Not all the data are relevant for the study. We’re interested in:

1. Dissolution results at 15 min (to be used as predictor)
2. Dissolution results at 90 min (to be used as predicted response)

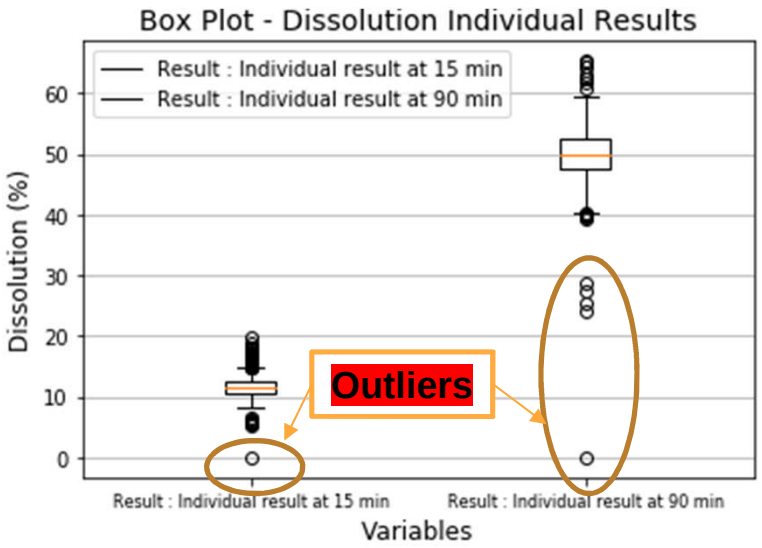
Data Preparation and Cleaning

- Data preparation included:
  - Filtering of the variables of interest (dissolution results 15 and 90 min)
  - Removal of missing data rows and outliers (erroneous entries, non plausible results…)
- Some challenges were encountered:
  - Data format conversion (e.g. dealing with dates)
  - Keeping track of the sample ID's and test replicates → a solution was found using *pandas.pivot_table* method

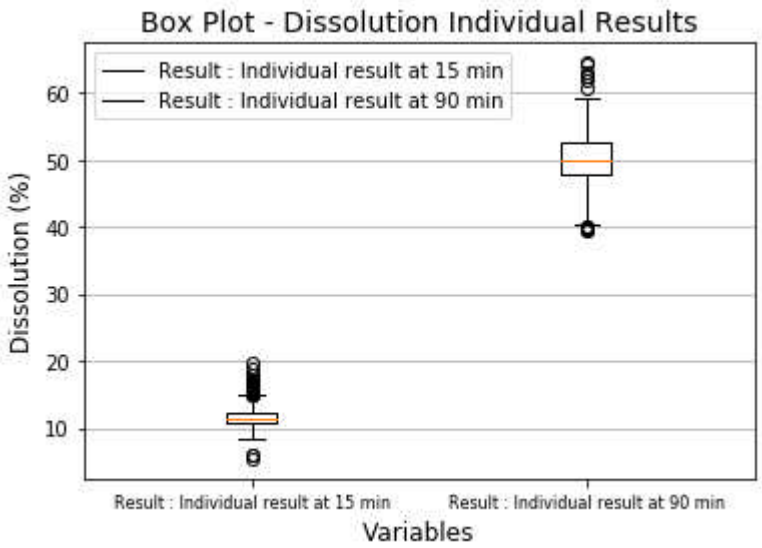
Data Preparation and Cleaning

- Example of outlier removal

Before cleaning: 3555 observations



After cleaning: 3548 observations



Research Question(s)

- 2 questions have been studied:

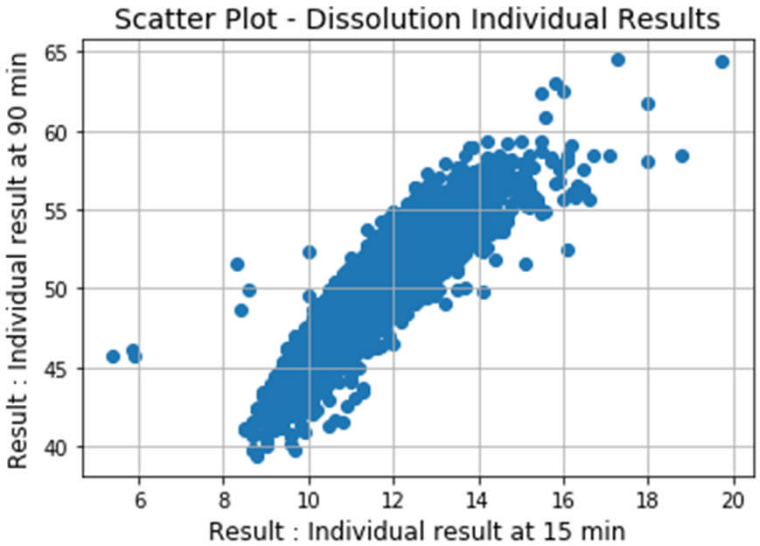
1. Can a predictive model be developed to forecast the dissolution result at 90 min based on the recorded dissolution result at 15 min?

2. How does the predicted uncertainty compares to the measurement method precision?

Methods

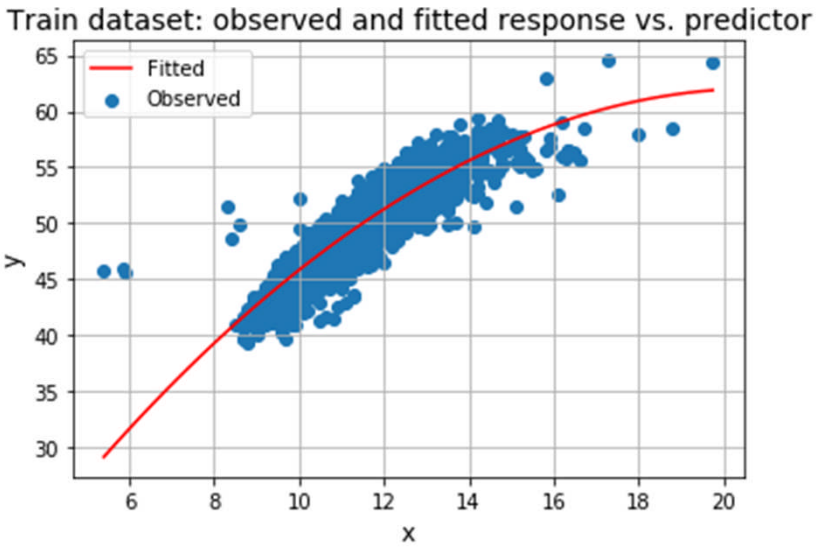
- Explore visually the relationship between dissolution results at 15 and 90 min
- Split the dataset in training and testing data set: data from 2019 are use to train the model, data from 2020 are used to assess the model performance
- Apply regression methods on the training dataset with 1 predictor ***x*** (dissolution 15 min) and 1 response ***y*** (dissolution 90 min)
- ***Statsmodels*** Python library was used to fit the regression models and evaluate their statistical properties

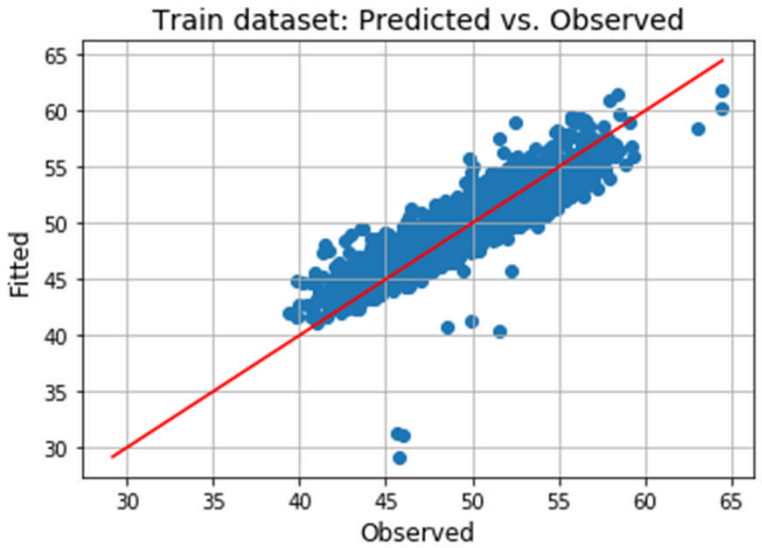
Findings (1): Visual exploration



- Visual exploration (scatter plot) of relationship between dissolution 15 min and 90 min
- A **strong correlation** can be seen between the 2 variables
- A "**curvature**" is observed → quadratic model maybe of interest to evaluate
- Some outliers may still exist in the dataset (they are kept in the next analysis)

Findings (2): Regression model on train dataset

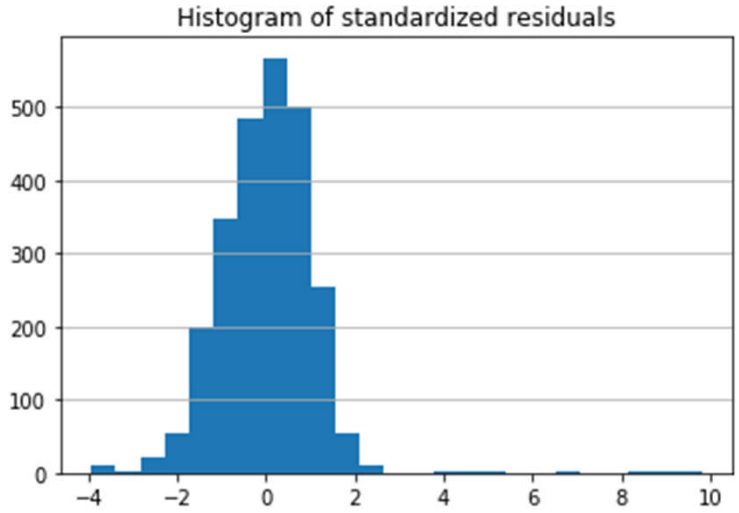




- Several models (linear and quadratic regression) have been tested
- Best performance is found with Ordinary Least Square model:

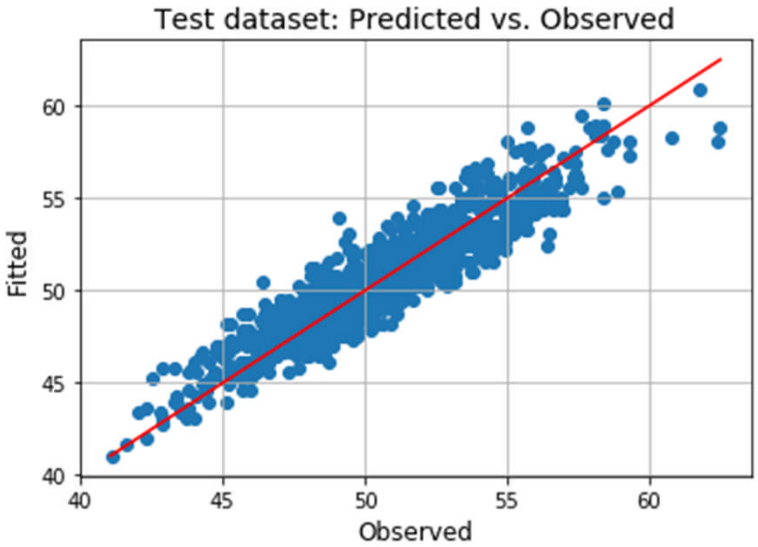
$$y = ax^2 + bx + intercept$$

Findings (3): Model error on train dataset



- Response residual errors (*observed – fitted*) are close to normally distributed and centered on 0. This is a desirable property of the model!!
- As expected, outliers are not predicted well (large error)

Findings (4): Model performance




```
Model Root Mean Squared Error =  1.6900083129222072
Model Root Mean Squared Error of Prediction =  1.2805745936994357
```

- On test dataset, predicted and observed results are in very good agreement
- Model error on train and test dataset are close to each other
- Prediction error is in same magnitude as measurement method precision

Limitations

- This modelling approach may be specific to the studied product and generalization to other product can possibly be unsuccessful

- Using only the predicted value to estimate the product dissolution at later time point may be insufficient

- A further development could be to use the model to estimate a *prediction interval* instead. The limits of this interval would be compared to the product specifications to have more confidence in the product dissolution profile

Conclusions

- A **quadratic regression model** can be established to predict dissolution at 90 min based on the recorded results at 15 min
- Such model could be used to shorten the test duration when the early results allow to project future results to be within the acceptance criteria of the product. However, such change in the laboratory practices needs to be officially approved.
- The **prediction error was found to be similar to the measurement error**. Hence, the model is expected to marginally contribute to the results uncertainty

Acknowledgements

I’d like to thank the Quality Control lab personnel for providing the dataset and helping to assess the model performance compared to dissolution measurement method precision.

Note to reader: the dataset and the Jupyter Notebook printouts cannot be shared for confidentiality reasons.

References

- European Pharmacopeia Dissolution test
  https://pdf4pro.com/view/2-9-3-dissolution-test-for-solid-dosage-2bd45.html
- Python library Pandas 1.0.5 online documentation
  https://pandas.pydata.org/
- Python library Matplotlib 3.2.2 online documentation
  https://matplotlib.org/index.html
- Python library Statsmodels 0.11.1 online documentation
  https://www.statsmodels.org/stable/index.html