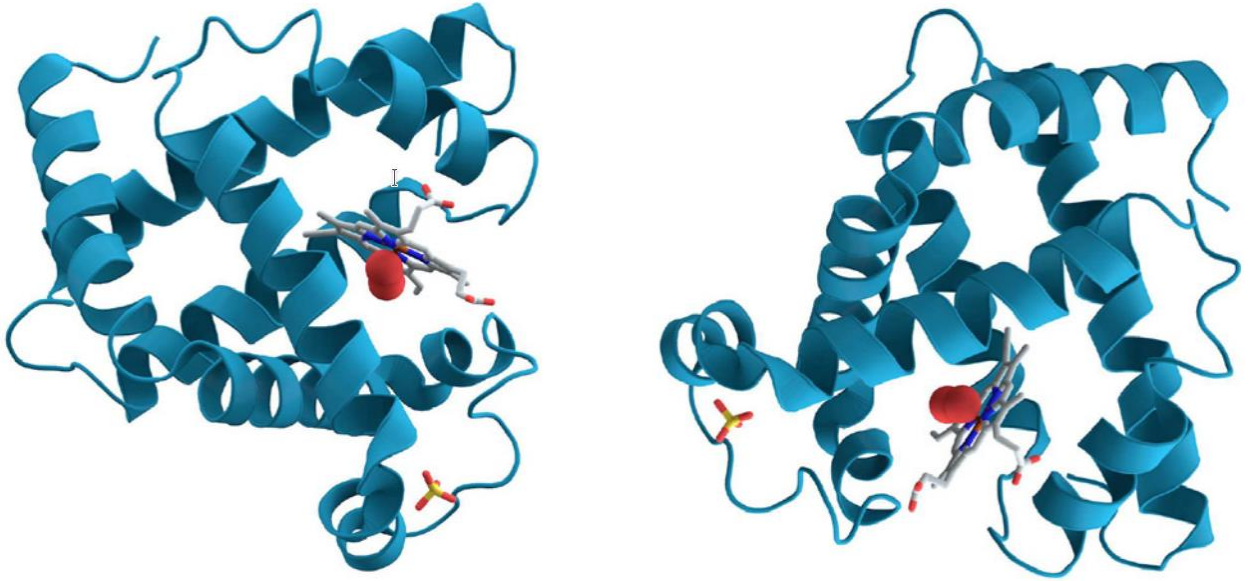
Protein Tertiary Structure Prediction

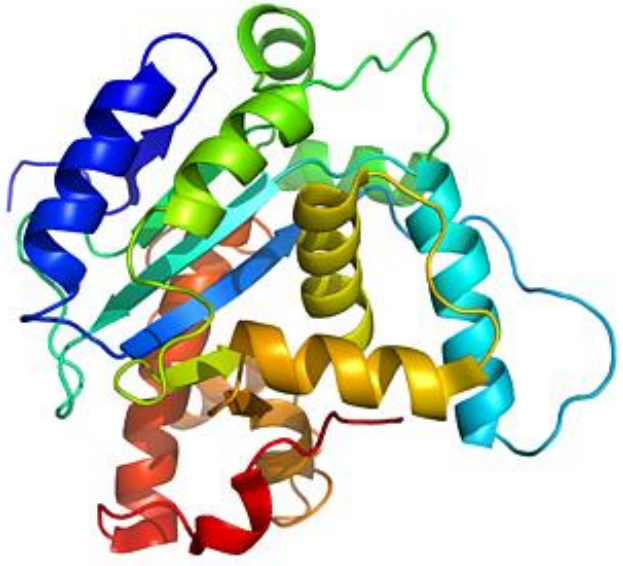
(Factors that Influence Protein Folding)



Ravinarayan Raghupathi, MSc PhD

Background

- Proteins are naturally occurring complex biological molecules that perform a variety of essential functions in living organisms.
- They are polymers of amino acids joined together by chemical bonds.
- The sequence of the amino acid residues determine each protein's unique structure and function.
- The tertiary structure of a protein is its three-dimensional (3D) folded shape.



- The prediction of a protein's tertiary structure is crucial in understanding protein-drug, protein-protein interactions, etc.
- It is one of the most challenging exercises in bioinformatics, because of the number of variables involved.
- Machine-learning and other soft-computing techniques have become popular tools in protein structure prediction.

Overview

Scientific problem

Will regression analysis be able to determine which physicochemical factors are important in predicting the tertiary structure of a protein?

Strategy

- Use multilinear regression analysis of a freely available dataset that includes 9 physicochemical properties (the independent variables) to assess their effect on predicting the dependent variable RMSD (Root Mean Square Deviation) of residue size.
- All the variables are numerical and continuous and there are no categorical variables present in the dataset.
- The data was originally obtained from the Critical Assessment of protein Structure Prediction (CASP) 5 to 9 experiments.

Source of the data

(http://archive.ics.uci.edu/dataset/265/physicochemical+properties+of+protein+tertiary+structure)

Methodology

Exploratory Data Analysis

⬇

Model Iteration 1 (test assumptions of regression)

⬇

Check for multicollinearity and transform independent variables

⬇

Model Iteration 2 (test assumptions of regression)

⬇

Any further transformation of data and new Model iteration(s), and test assumptions of regression

⬇

Final model validation and conclusions

Model 1 (Baseline)

The baseline model was created using all the available independent variables after removing outliers.

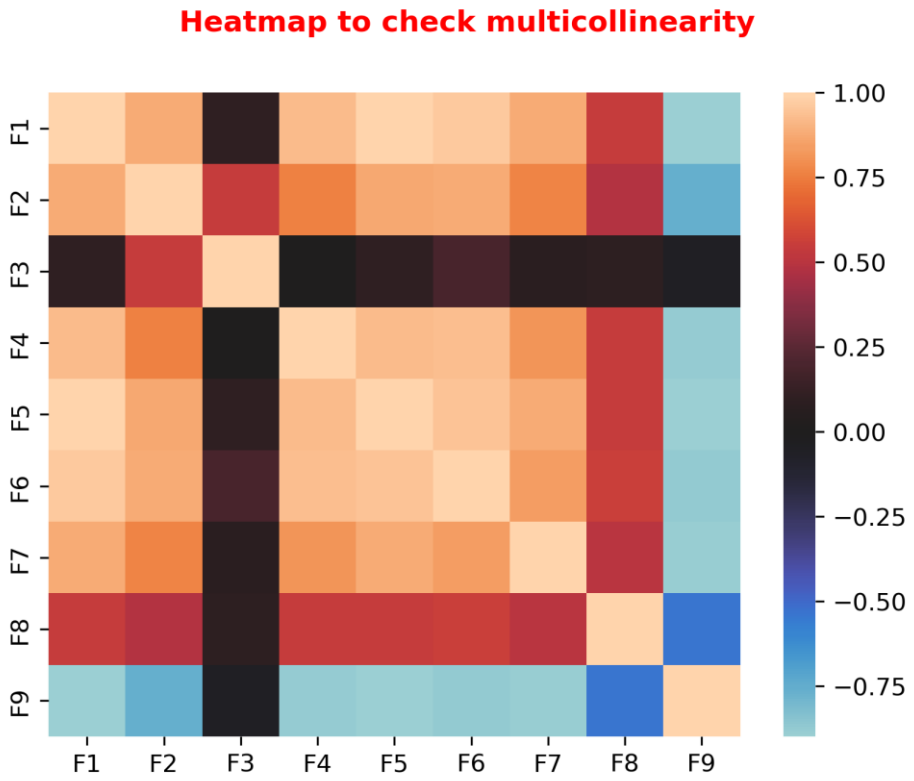
Observations

- The independent variables were more normally distributed after the removal of outliers.
- The adjusted R-squared value was very low (0.29).
- The assumptions for linearity, homoskedasticity and normality appeared to be satisfied in this model.
- The values for kurtosis and skewness were within acceptable limits but the high Condition number suggested multicollinearity.

Model 2 (Iteration 2)

This model was created after I checked for multicollinearity and transformed variables.

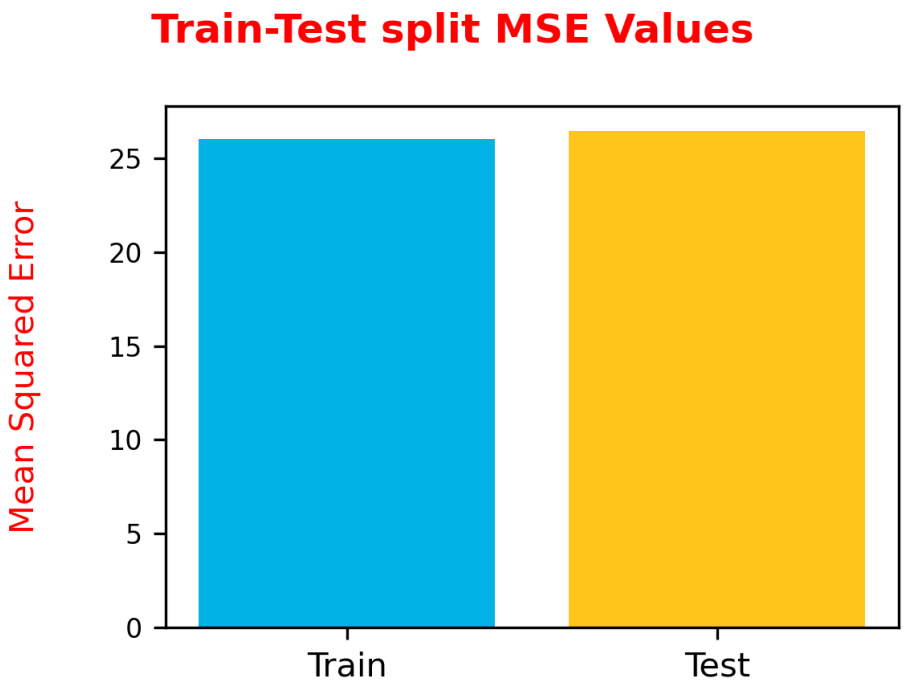
Observations

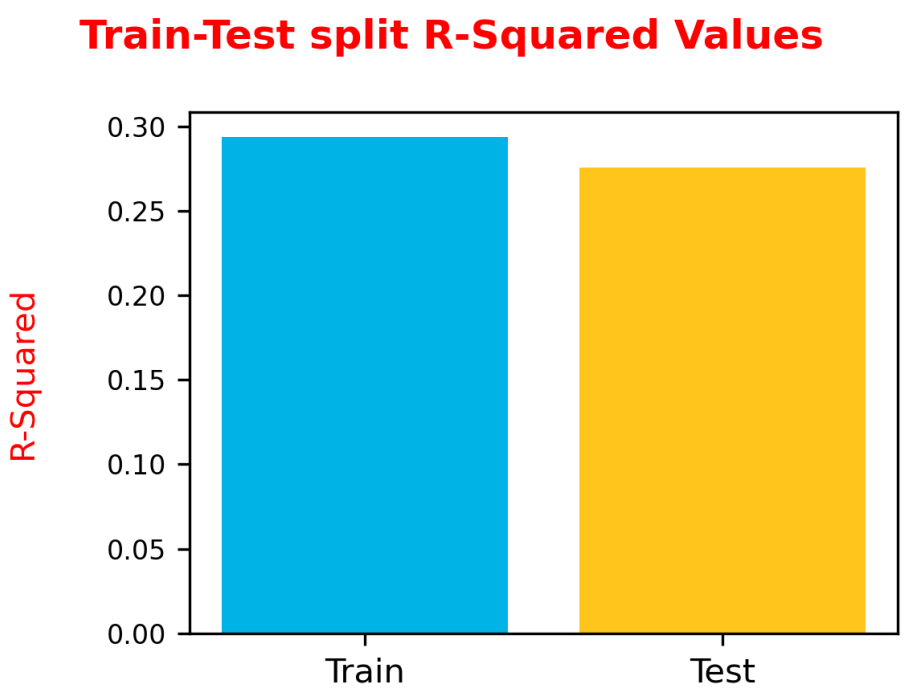


- There was no major difference between Models 1 and 2, in fact the statistics seemed to have worsened.
- The normality plot had also worsened, indicating that Model 1 is better for further analysis.
- No further tweaking of the variables will improve the regression modelling.

Model validation

Model 1 was evaluated using Train-Test splits and the prediction accuracy was computed.



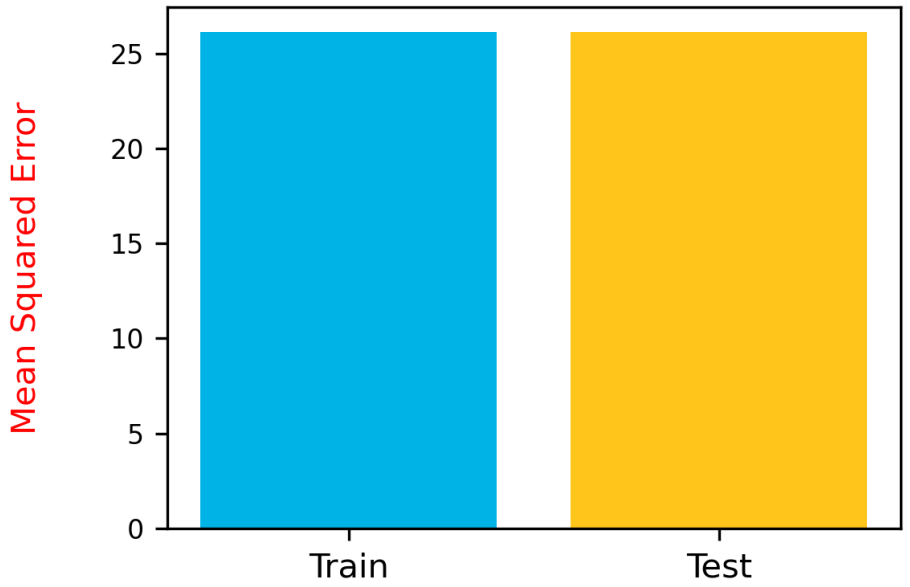


Computed accuracy of the model: 0.27546723825388264

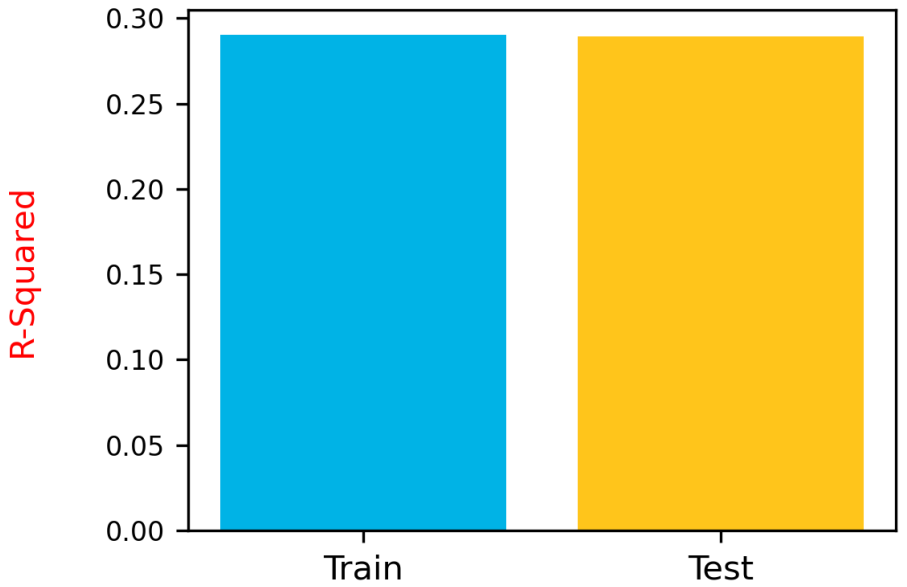
Model validation

Model 1 was also evaluated using Cross validation and the Random Forest Regressor.

Average Cross Validation MSE Scores (10 splits)



Average Cross Validation R-Squared Values (10 splits)



Random Forest Regressor Result: Mean Squared Error is 12.012106346138834

Conclusions

1. The multilinear regression model appears to be fitted well but the R-Squared value is too low to be considered statistically useful and the MSE values are still very high.

2. The model will predict correctly around 27% of the time, which is very low for the prediction accuracy that was the goal of this regression analysis.

3. The MSE by reduced by around 50% by using the Random Forest Regressor. This might be the best approach to modelling data of the kind studied here.

Actionable insight

Multilinear regression analysis of physicochemical measurements might not be the best approach to predicting the tertiary structure of proteins, given that many of these measurements are likely to be highly correlated.

Acknowledgments

Academy Xi

Hardik Idnani

and of course…..

All my wonderful classmates!