Protein Folding Problem

Naymul Hossain

December 11, 2020

Summary:
The objective of my analysis is to create a model that determines the accuracy of a computer-generated protein structure, compared with a known benchmark structure. By getting rid of the initial multicollinearity by using VIF(Variance Inflation Factor), I developed 3 models with AIC(Akaike information criterion) (Forward, Backward and Forward-Backward) and 3 more models with BIC(Bayesian information criterion) (Forward, Backward and Forward-Backward) as the criterion. Though the dataset was not initially splitted to training and validation for model selection, I implemented $k = N$ cross-validation(C.V.) to try compensate for that. Following that, I compared the 6 models and chose the one with the the model 4 (BIC Forward after $k = N$ C.V.), since all the other RMSPE(Root Mean Square Prediction Error) were similar, and this one had the least difference in error compared to the validation set, I chose this with a RMSPE value of 0.5448438. Following that, I came up with the Box-Cox Transformation stabilize the variance. My model 4 contains 89 predictors, and also follows the MLR model assumptions.

Exploratory Analysis of Dataset:
Initially I started off playing with the data by gathering some variables, and also angles, since it seemed unique compared to others. I found out that there is a trend for angles when the scatterplot is observed. The plot below shows some randomly picked variables and angles to understand the dataset.
What I found out was there were many variables which has no relation with accuracy(such as the 2^{nd} to 4^{th} scatterplots shown below. In particular, I found out that the variables *scArgN_bbC_ medshort* (column 478) and *scArgN_bbO_short* (column 483), formed exact multicollinearity. Therefore, even before I started getting rid of multicollinearity in the next section using VIF, I got rid of them from the dataset. If this was not done, I would not have been able to find the VIF, since X^TX would *not* have existed

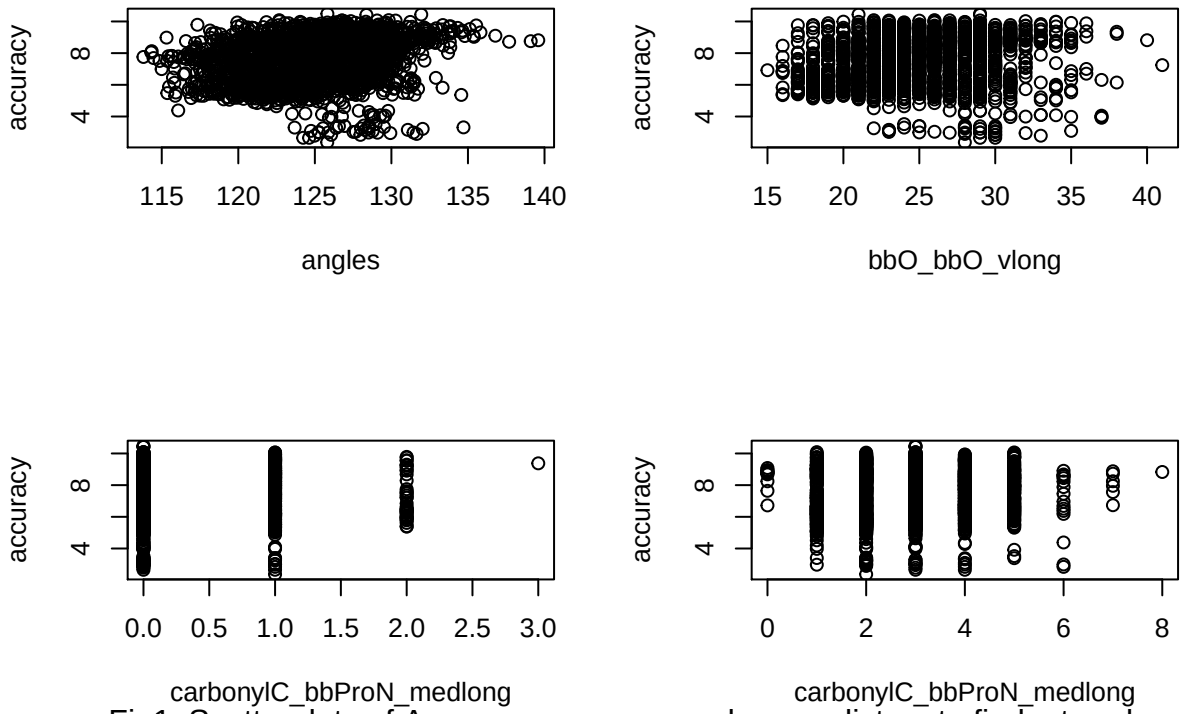


Fig1: Scatterplots of Accuracy vs some random predictors to find a trend

Methods:

- *Getting rid of Multicollinearity*:

Firstly, getting rid of multicollinearity was important, so as to make sure we get rid of as many predictors, which does the same contribution in building our model. Therefore, any predictor with a value ≥ 10 was removed. So, now we have 569 variables to work with. Reducing the number of variables also contributed in lower computation time, and also helps in interpretability of our analysis.

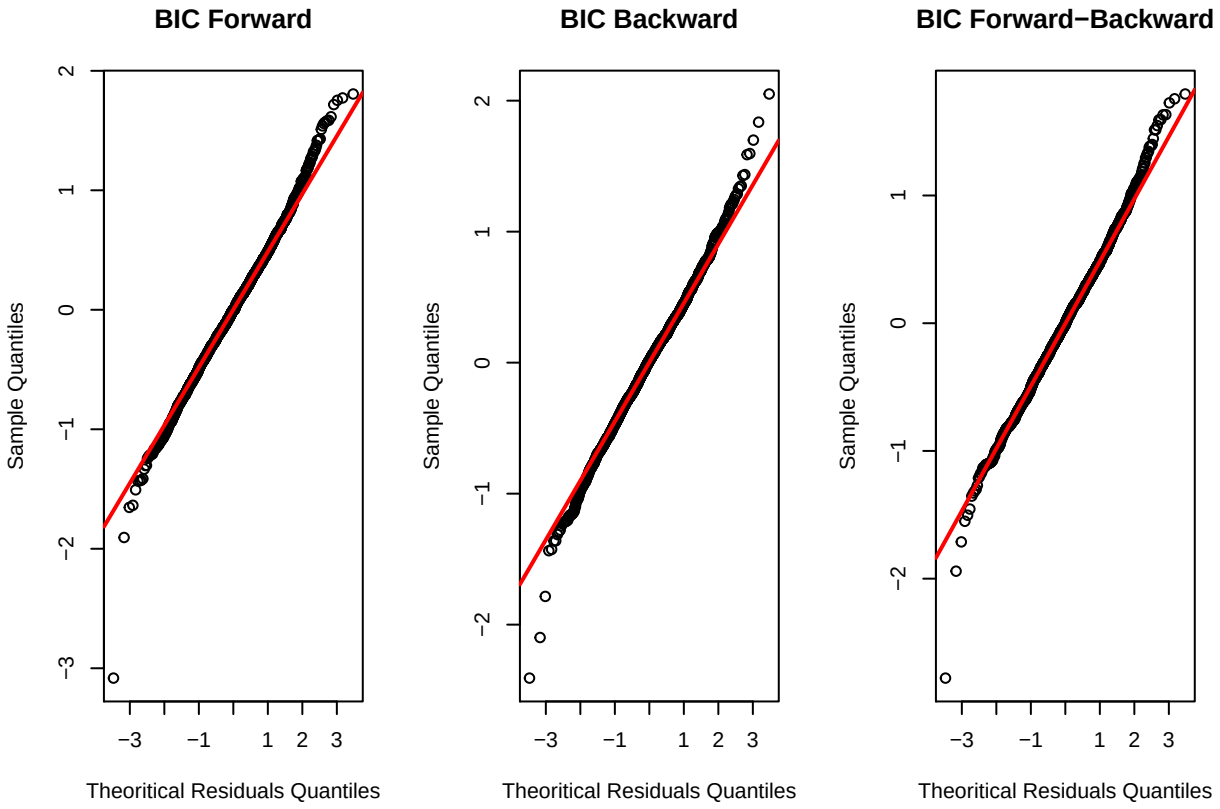
We can understand with such high R^2_{Adj} values that the models have a goodness of fit. However, that is not all, since now we will compare to check if the model assumptions are maintained.

-*Model Selection Criteria*

Going through the cross validation of 6 models, where it required more than 10 hours for one in the generic way, I decided to select model without cross validation, with the promise of training my data using $k = N$ cross validation once my model is selected. Below the R^2_{Adj} for all the six models are given (I tried using Kable but my Rstudio was crashing at the last minute. I tried updating according to piazza but it was too late to find a solution even after updating. Sorry for bad view):

```
## [1] "0.90353531946761"  "0.907930669198453" "0.905914393806386"
## [4] "0.871613272199826" "0.885560136497054" "0.880915366615913"
```

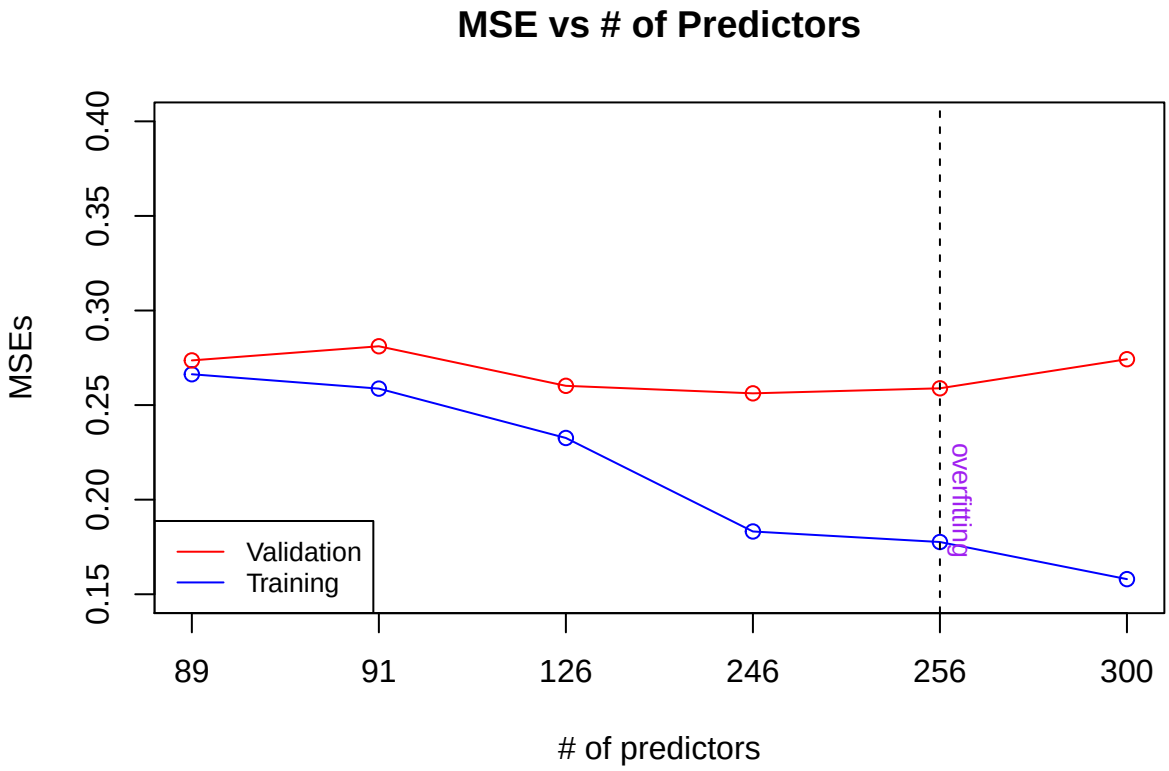
We noticed that all the 6 models have similar Residuals vs Fitted values, Residuals vs Indices, and the histograms maintain normality (showed in the appendix). However, all the 3 models from BIC, had a bit of high kurtosis.

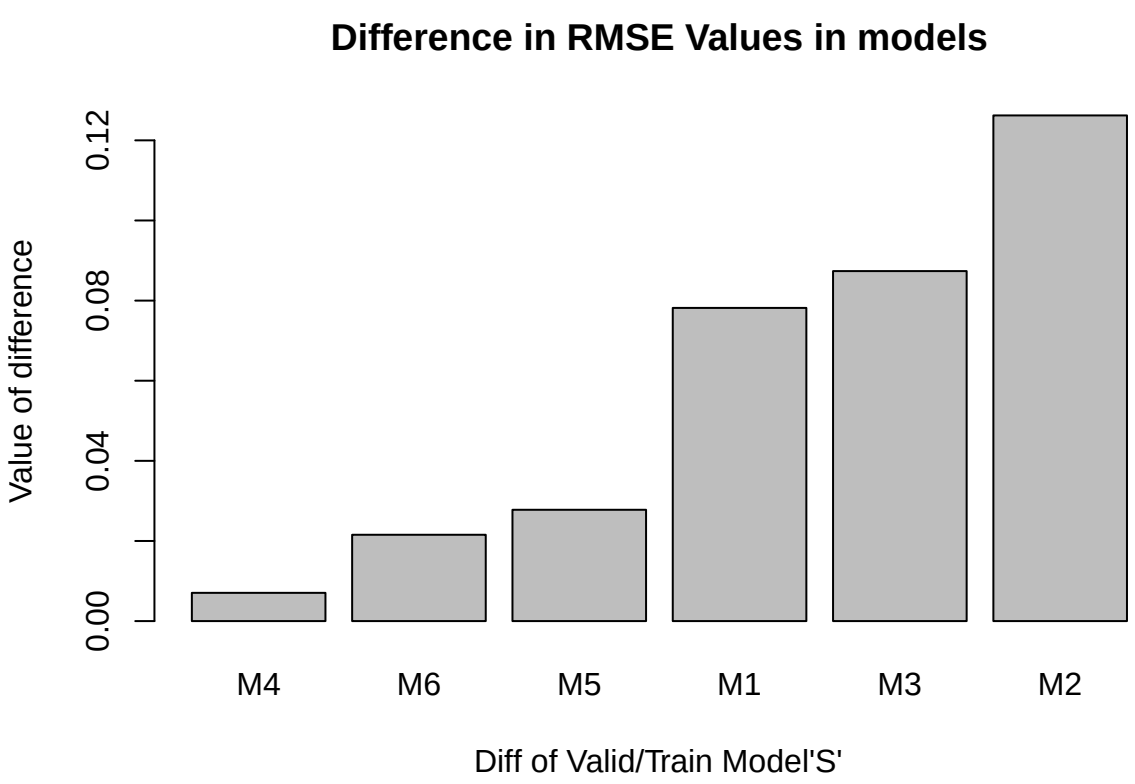


However, we will still continue using them, because transformation can become handy later on to fix them, and may be the best model is hiding here [spoiler alert: It is]

-*Getting my final model*

We will see below in the figure that of the 6 models, the 4^{th} model uses the least number of predictors and also has the least difference compared to it's validation set, after being run $k = N$ cross validation among the 6 models. A $8:2$ ratio was maintained between the training and validation set.


MSE vs # of Predictors
MSEs
0.15
0.20
0.25
0.30
0.35
0.40
Validation
Training
overfitting
89
91
126
246
256
300
# of predictors



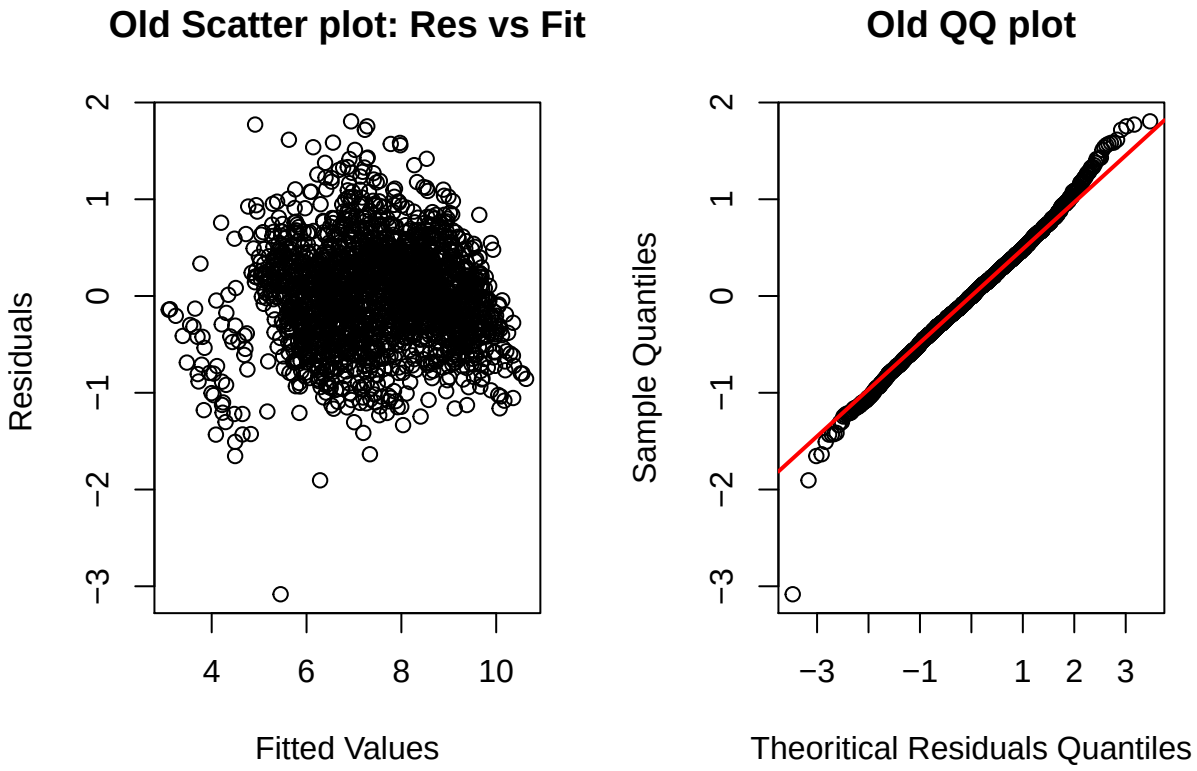
Among all the other models, the model 4 (BIC Forward) had the least difference as we can see in the barplot. In fact, choosing model 4 was wise since, if we follow the segmented line plots of MSE vs # of Predictors, we will notice that model 4 does not have much of an issue with overfitting.

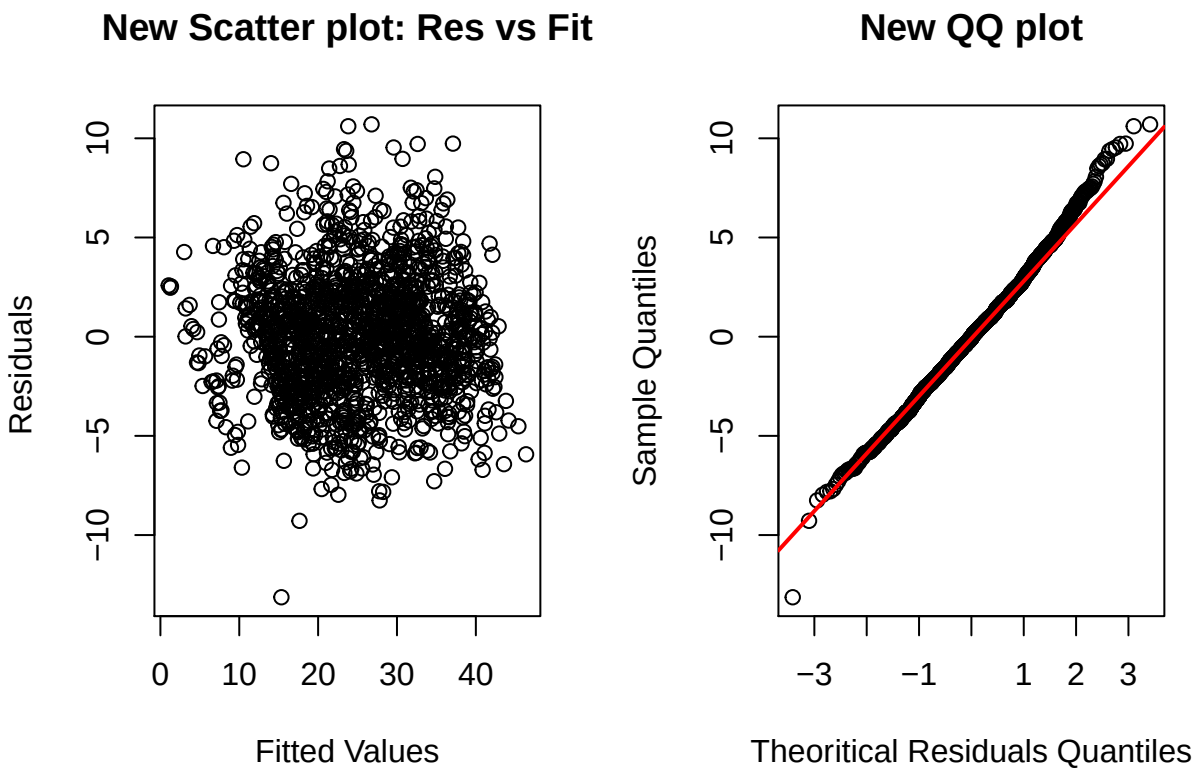
Also, having less number of predictors mean it is easier to interpret, therefore Model 4 was chosen.

-*Transformation:*

We see the the residuals vs fitted values does not really have a distribution which follows the MLR assumption. However, the QQ plot seems fine. So, we will go with the Box-Cox transformation and see that the new residuals vs fitted values distribution matches more like the MLR assumption (and then again look at QQ plot to see if it was made bettter). So we can say it helped stabilizing the non constant variance.

Along with that, the new QQ plot has a better matching with the MLR assumption by having all the plots lined with the line $y = x$


Old Scatter plot: Res vs Fit
Residuals
Fitted Values
Old QQ plot
Sample Quantiles
Theoritical Residuals Quantiles



Results and Discussion:
- While obtaining my final model, I started looking for outliers, however, I reached the conclusion that trying to remove some of the outliers after conducting a studentized residual test, the RMSE value only increases. RMSPE for model 4 before removing them was 0.5448438, while after removing them it increased to 0.5838417

- There were 6 models tested by my selection procedure. 3 from AIC (Forward, Backward, and Both) and 3 from BIC (Forward, Backward and Both)
- The figures on page 4 and 5 describe the differences in RMSE and MSE between validation and trainin sets.
- There are 89 variables in my model, some are positively, while some are negatively related to the respone variable (accuracy)
- All of them are uncorrelated variables, contributing to the prediction of the response variable.
- aliph1HC_scArgN_long = 0.563553314. With one unit increase in this variable *aliph1HC_scArgN_long*, the response variable increases by 0.563553314
- scArgN_bbO_medlong = 0.437025430. With one unit increase in this variable *scArgN_bbO_medlong*, the response variable increases by 0.437025430

- bbCA_bbO_vshor = -0.395812423 . With one unit increase in this variable *bbCA_bbO_vshor*, the response variable decreases by 0.395812423

Since, the differnce between the validation set and the training set for my model 4 was really small, and despite breaking some rules (since prediction was our main goal), out model fit the data good(except for the QQ plot which is still not bad) If we check page 6, the transformation has made the MLR assumption well maintained. The regression assumptions of normality, constant variance and independence among error terms were all maintained(check appendix)

I would have expected the MSPE to be 0.5448438 and I am confident about it since I have cross validated it N times, got rid of multicollinearity, made sure no overfitting(so no noise) remains, and checked for outliers. Hence, I could confidently say the MSPE should be around what I currently have

As seen previously, aliph1HC_scArgN_long = 0.563553314 makes a huge contribution to the prediction of the response variable. However, what I particularly found interesting was that the angles variable was not in my best mode (Model 4). The reason could be behind the fact that some variable distances can account for the ratio of the angle contributiong to the response variable(accuracy) [for instance, some distance A and B among atoms can form a ratio of the angle collectively.]