STAT 331 – Final Report

Cameron Sequeira-Hogg, 20722977

STAT 331, Fall 2020

Summary

The objective of this project was to develop a multiple linear regression model to best predict the accuracy of computer-generated structures for the COVID-19 spike protein. To train the model, a training dataset with 1946 samples was provided. Alongside the response variable, accuracy, there were 685 explanatory variables. In initial analysis of the dataset, two of the explanatory variables were removed due to perfect multicollinearity. An initial test of the effect of removing highly multicollinear variables and observations considered outliers on predicted Root Mean Squared Error (RMSE) led to the removal of 115 more explanatory variables and 206 observations from the dataset. After this, nine models were tested against each other to gain a sense of what model selection techniques would produce the lowest RMSE. The three techniques with the lowest RMSE were tested against each other, trained on 80% of the initial dataset with outliers removed and tested on the remaining 20% of the data in the validation set. The model with the lowest RMSE of 0.571 in this test was selected by an ICM algorithm with a penalty of 2. This same selection technique was used on the entire dataset with the 115 highly multicollinear variables and 206 outliers removed, and the model generated was chosen as the final model used to predict the 1946 accuracy values in the test set.

Exploratory Analysis of Dataset

The training dataset consists of 1946 observations of 685 numerical explanatory variables. The response variable of interest is the accuracy of the computer-generated structures for the COVID-19 spike protein. The accuracy variable has a mean of 7.48813 and a standard deviation of 1.473993, with maximum and minimum values of 10.4823 and 2.3706, respectively.

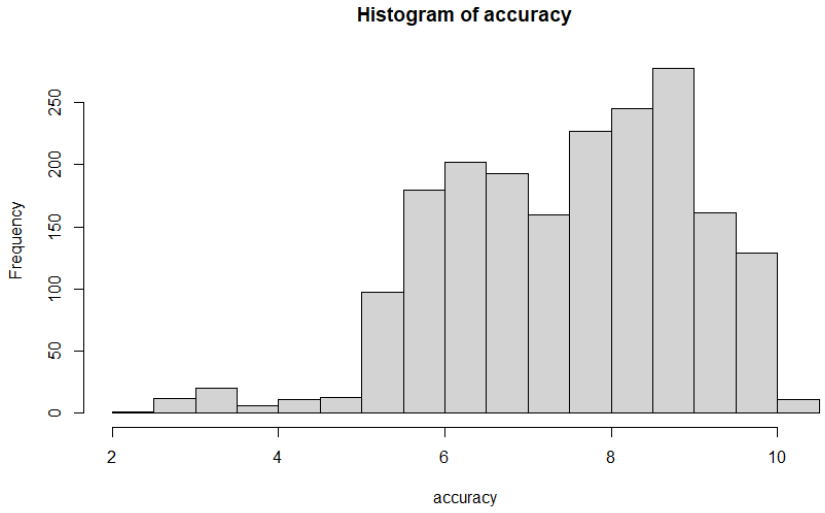

Figure 1: Histogram of the Response Variable, Accuracy.

A histogram of the response variable shows that most accuracy values are between 5.5 and 10. The histogram shows that the responses are left-skewed.

Methods

Multicollinearity Removal

Generating a summary of multicollinearity between variables showed perfect multicollinearity between one pair of explanatory variables: scArgN_bbC_short & scArgN_bbC_medshort. Perfect multicollinearity was also found between a third variable, scArgN_bbO_short, and the Intercept. Fitting a linear model with perfect multicollinearity is impossible, and as a result the variables scArgN_bbC_short and scArgN_bbO_short were removed. After the two explanatory variables were removed, a linear model could be fit using the dataset and the remaining 683 explanatory variables.

VIF Removal

A common preparation before model selection to reduce multicollinearity is to remove VIF values over a certain threshold, most commonly 10. This helps with mitigating the violations of model assumptions as well as tightening prediction intervals. However, minimizing VIF does not necessarily improve prediction overall, and the objective was to generate the lowest possible RMSE value for a separate set of test data.

Outlier Removal

Measuring Cook's distance is a common way of determining outliers. With a considerably sized dataset with over 1000 observations, outliers cannot be individually examined. As a result, a common practice is to remove observations with a Cook's Distance over a certain amount. The chosen threshold was 4/n (n being the number of observations in a dataset), a commonly used threshold to identify the most influential points in the final regression. Like removal of explanatory variables with a VIF of over 10, removing these outliers does not necessarily improve prediction overall.

Search Strategy

Earlier in the methods section two possible choices were presented: the removal of variables with a VIF of over 10, and the removal of outliers with a Cook's Distance of over 4/n. In addition to the choices regarding modifying the initial dataset, there was a choice of which model selection technique would be used.

There were four potential treatments for dataset modification: No VIF Removal & No Outlier Removal, No VIF Removal & Outlier Removal Present, VIF Removal Present & No Outlier Removal, and both VIF Removal & Outlier Removal. To compare the effects of the four potential dataset treatments, five models were run with an 80-20 training split under each

treatment to find their RMSE on the 20% validation data. A K-Fold analysis was considered, but with K=5 the already high computation time of the analysis would be multiplied by five.

The five models picked were forward/back BIC (penalty of log(n)), forward/back BIC with double penalty, forward/back BIC with quadruple penalty, ICM with penalty log(n), and the full model using all available variables.

Tables 1 & 2: Validation set RMSE values for the initial five models chosen.

VIF Removal	Outlier Removal	BIC with Penalty log(n)	BIC with Penalty 2log(n)	BIC with Penalty 4log(n)
No	No	0.651	0.643	0.743
No	Yes	0.495	0.555	0.610
Yes	No	0.622	0.678	0.755
Yes	Yes	0.45	0.548	0.584

VIF Removal	Outlier Removal	ICM with Penalty log(n)	Full Model (all variables)
No	No	0.681	0.698
No	Yes	0.466	0.427
Yes	No	0.615	0.658
Yes	Yes	0.447	0.440

The models were selected to show a variety of search strategies and penalties to determine which of the four treatment groups had the lowest RMSE on their validation data. Three different BIC models were chosen to see if both lower and higher penalty models performed better in one treatment group. The full model was mainly included to compare the VIF treatment effects. The treatment group with highly multicollinear explanatory variables and potential outliers removed performed the best, having the lowest RMSE for every type of model. As a result, this treatment group was selected for a more in-depth analysis and attempt at model selection.

Table 3: Validation and training set RMSE values for nine models chosen. The dataset used had highly multicollinear explanatory variables and potential outliers removed. The first six listed models were trained in K=5 Folds, and the RMSE listed is the average RMSE for predictions of the five folds. The last three listed models were run once with an 80-20 Training-Validation split due to the three models having a large computation time.

Model	RMSE (Validation)	RMSE (Train)
Forward/backwards BIC with penalty log(n)	0.501	0.150
Forward/backwards BIC with penalty 2log(n)	0.564	0.239

Forward/backwards BIC with penalty 4log(n)	0.642	0.354
Forward/backwards BIC with penalty log(n)/2	0.467	0.0977
ICM with penalty log(n)	0.488	0.375
Full Model (all available variables after highly multicollinear explanatory variables removed)	0.434	0.064
ICM with penalty log(n)/2	0.494	0.356
Forward/backwards AIC	0.457	0.073
ICM with penalty 2	0.461	0.271

The four stepwise forward/backward BIC models with penalties ranging from (1/2) log(n) to 4log(n) were added to show the effect of lower and higher penalties on their average validation RMSE and training RMSE. Similarly, the ICM models were added to compare their penalty effects on RMSE.

The three models with the lowest RMSE values from the last test were tested once again on an 80-20 Training-Validation split. This time, the outliers were removed after the data was split. This was to see if the apparent overfitting of the full model, which previously had the lowest RMSE, would affect its ability to predict the response of outliers.

Table 4: Validation and training set RMSE values for three models chosen. The models were run once with an 80-20 Training-Validation split due to the three models having a large computation time.

Model	**RMSE (Validation)**	**RMSE (Train)**
Full Model (all available variables after highly multicollinear explanatory variables removed)	0.591	0.137
Forward/backwards AIC	0.604	0.165
ICM with penalty 2	0.571	0.403

After the ICM model was shown to have the lowest validation RMSE, a new model was trained on the entire training dataset (with multicollinear variables and outliers removed) using the same selection technique.

Results and Discussion

The analysis discussed in the methods section revealed that removing variables with high multicollinearity and removing outliers with a Cook's Distance of over 4/n improved the RMSE in the tested models. As a result, the subsequent initial search for an ideal model removed 115

more explanatory variables and 206 observations from the dataset. Nine total models were tested using this treatment.

In the final trial between the three best performing models, the ICM model with a penalty of 2 performed the best with an RMSE value of 0.571 on the validation set. As a result, this model selection technique was used to train the final model, by applying it to the entire training set with the 115 highly multicollinear variables and 206 outliers removed. As the final model utilizes the same selection technique, the MSPE for the test data predictions would likely be around the 0.571.

The final model generated had 332 total coefficients. A complete list of coefficients is provided in the appendix. 275 of the coefficients are statistically significant at a 95% confidence level. The predicted intercept value is 3.211465, which is statistically significant at a 95% confidence level.

The plots of residuals over time and residuals versus fitted values show no apparent model violations.

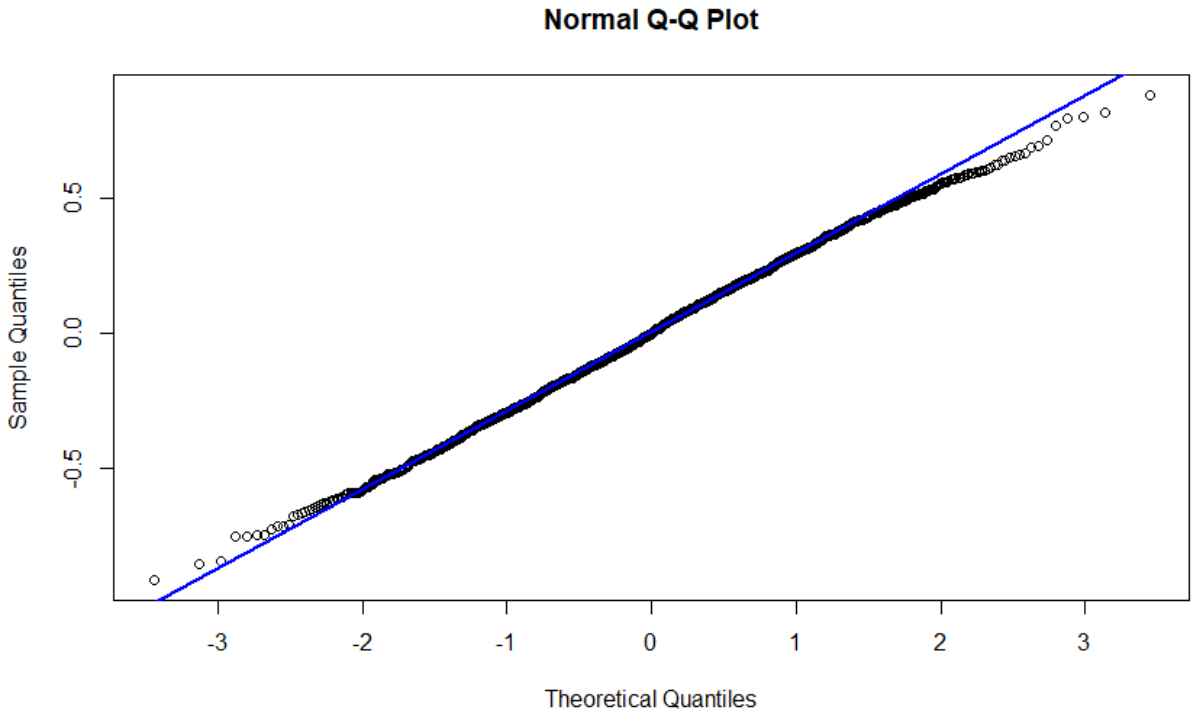

Figure 2: Normal Q-Q Plot for the Final Model

The Normal Q-Q plot shows that observations are slightly less extreme than they should be compared to a regular normal distribution. This is a model violation, and could be remedied by transforming the response, but the aim is to prioritize RMSE minimization over rectifying model assumptions.