

# CSCI4360/6360 Project 1 Report

## Ye Tian

This project features experiments of different feature selection algorithms with different regression models on different datasets.

### Feature Selection Algorithm to be used:

1. Forward Selection
2. Backward Elimination
3. Stepwise Regression
4. (Ridge Regression)
5. (Lasso Regression)

### Models to be used:

1. Multiple Linear Regression
2. Quadratic Regression
3. QuadraticX Regression
4. Cubic Regression
5. CubicX Regression

### Datasets to be tested on:


1. AutoMPG.csv
2. Concrete.csv
3. Winequality-red.csv
4. Forestfires.csv
5. AirQualityUCI.csv
6. PRSA\_data\_2010.1.1-2014.12.31.csv

Testing results:(The red lines in all the graphs are for R2 score, the green lines are for adjusted R2, while the blue lines are for cross validated R2)

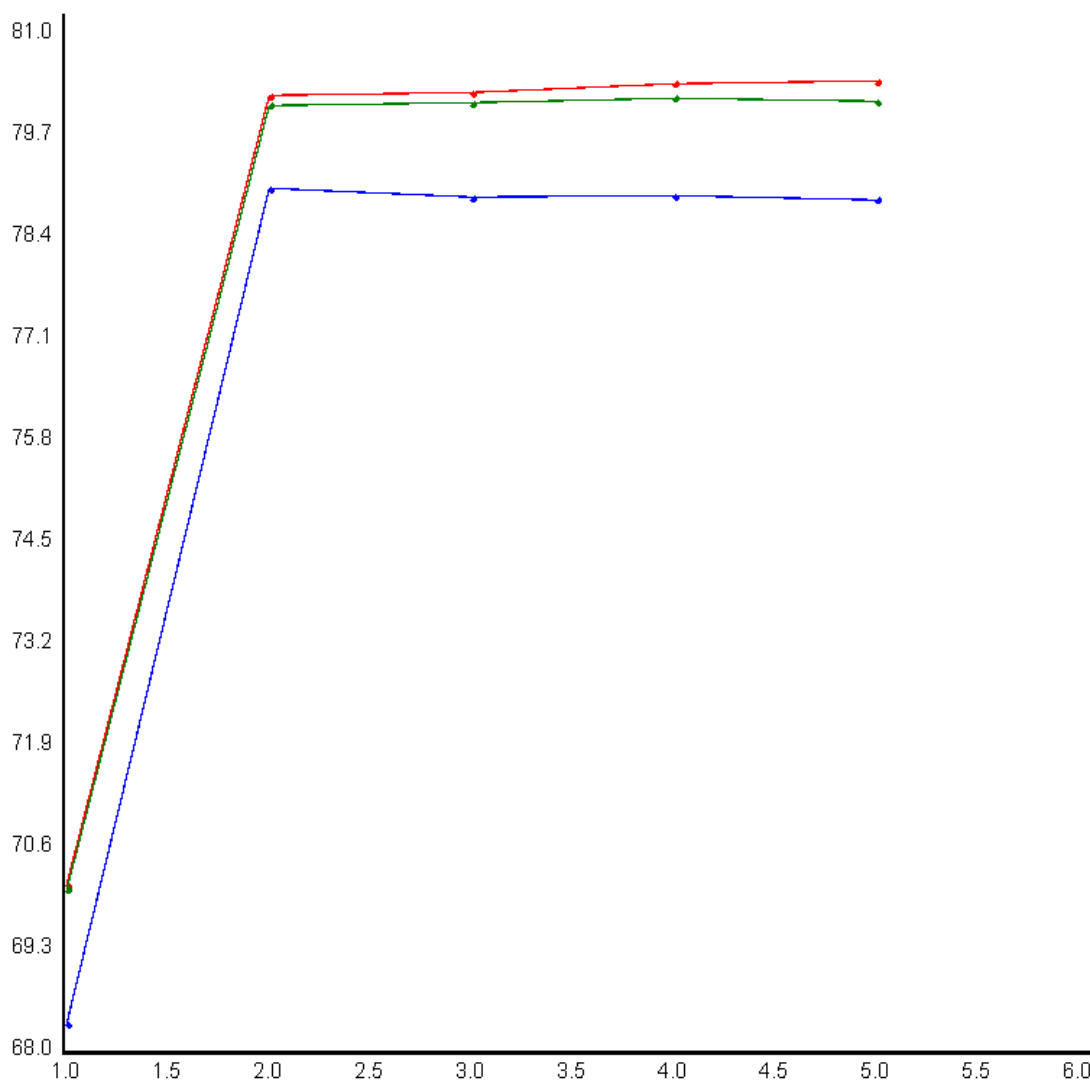
Here we exhibited the testing results with auto-mpg.csv datasets, where the input features we will use are mpg, cylinders, displacement, horsepower, weight and acceleration, to predict the response feature, which is the model year.


## Multiple Linear Regression

For multiple linear regression, the three feature selection algorithms all show a sudden jump of quality of fits as variables are added/removed. In forward selection and backward selection,  $R^2$  score and adjusted score behave similarly while there's a very noticeable gap from the cross-validated  $R^2$  score. With stepwise regression, all three qualities of fits ( $R^2$ ,  $R^2$  adjusted, and cross validated  $R^2$ ) behave pretty much similarly.

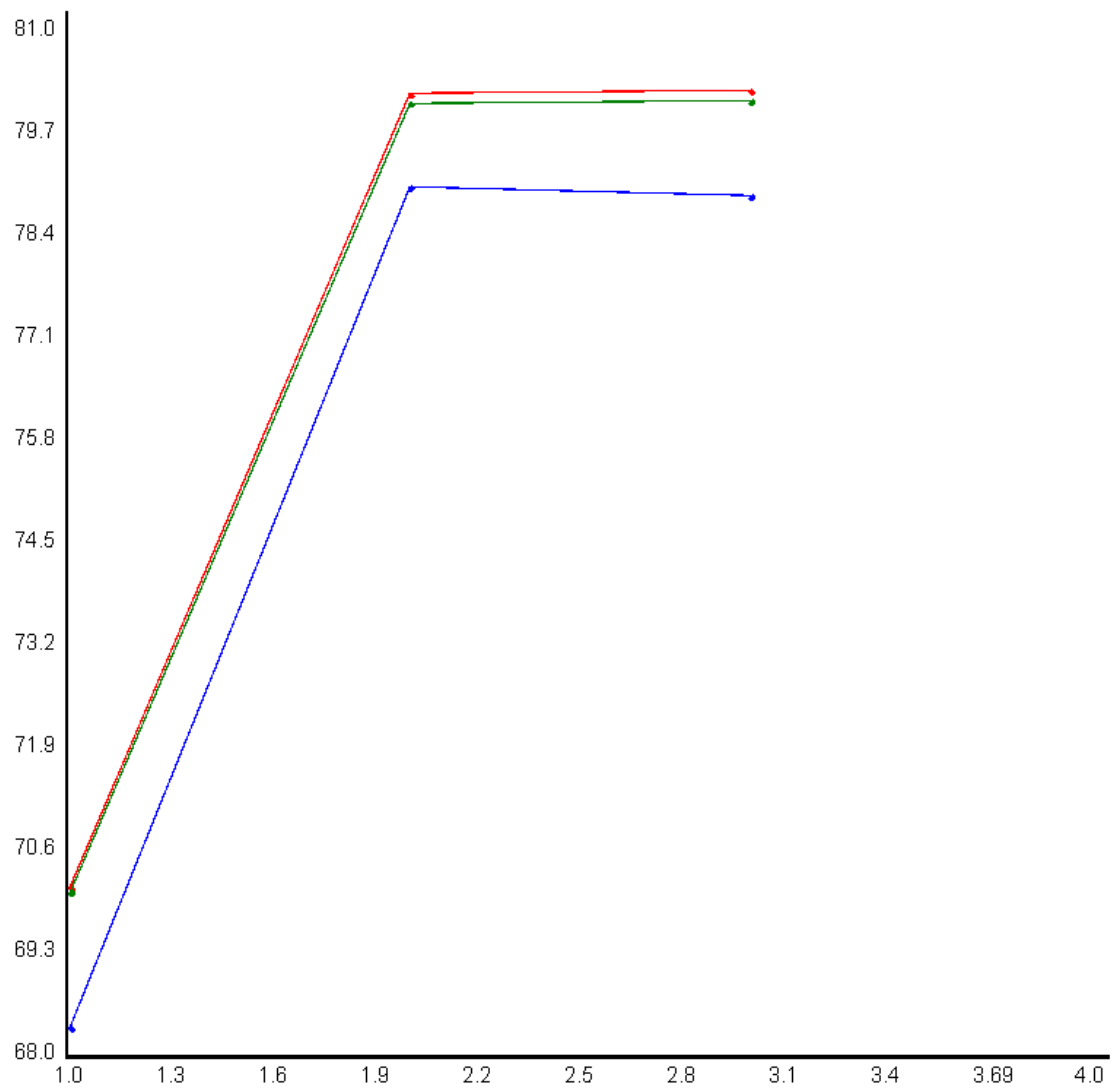
  $R^2$  vs  $n$  for Regression with forward selection

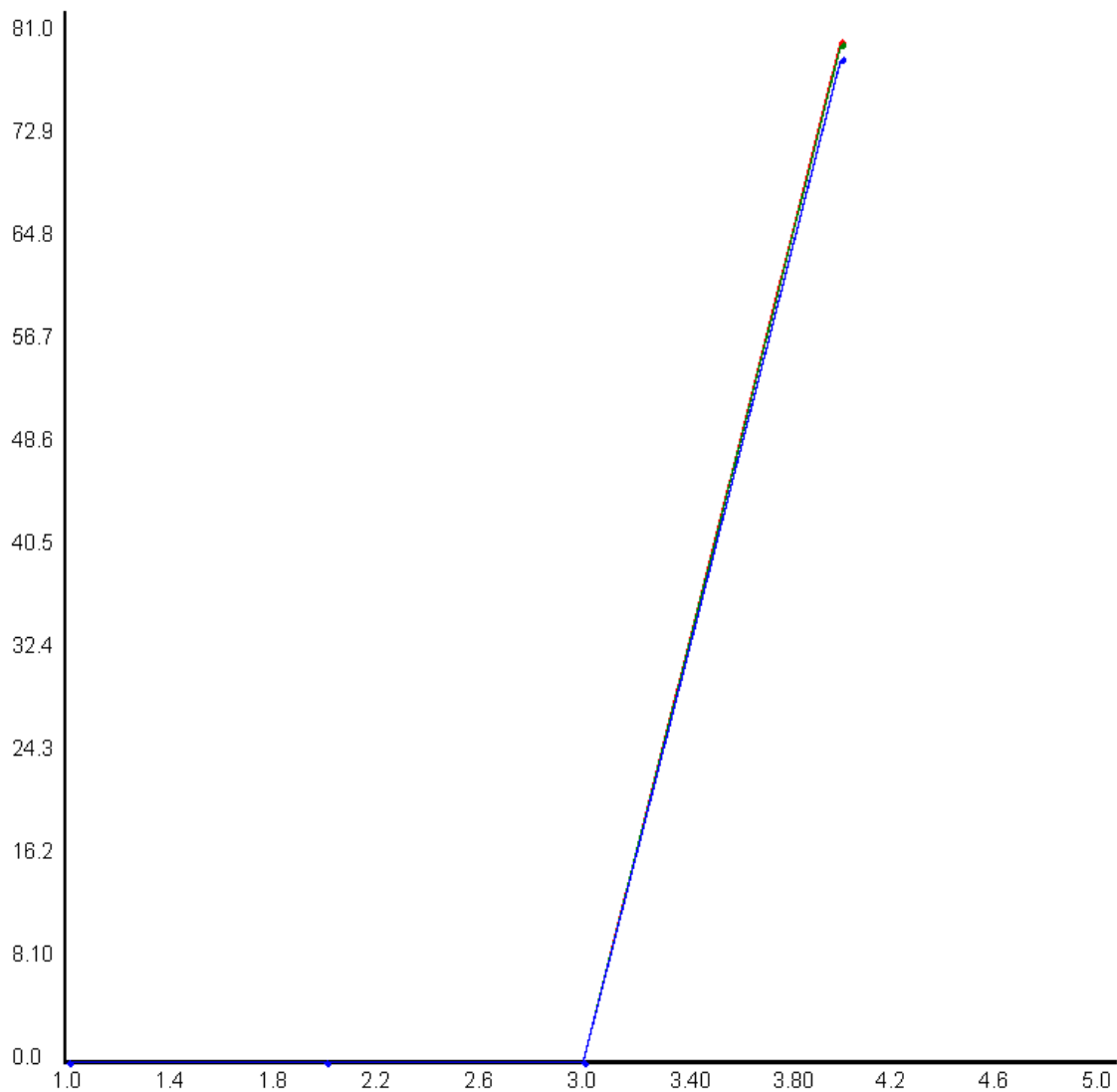
— □ ×



  $R^2$  vs  $n$  for Regression with backward elimination


— □ ×



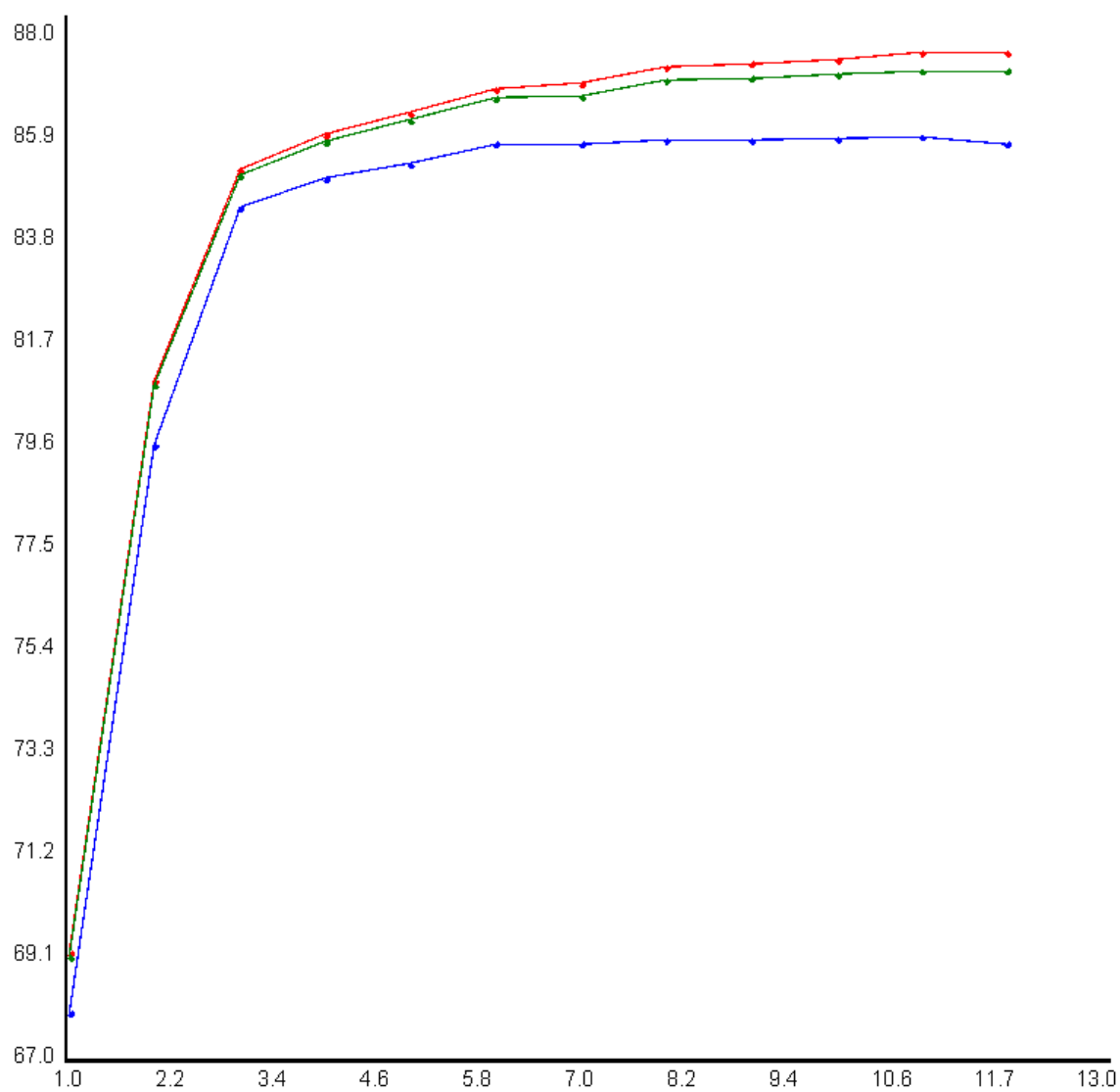



## Quadratic Regression:

For quadratic regression, all qualities of fits improve more gradually and smoothly with forward selection and backward elimination feature selection processes. As with multiple linear regression, R<sup>2</sup> score and R<sup>2</sup> adjusted score have more similar behavior while there's an obvious gap between cross validated R<sup>2</sup> and the other two. However, also as with multiple linear regression, all three qualities of fits behave similarly with stepwise regression, and there is a sudden jump as features selected go from 3 to 4, but then goes very smoothly.

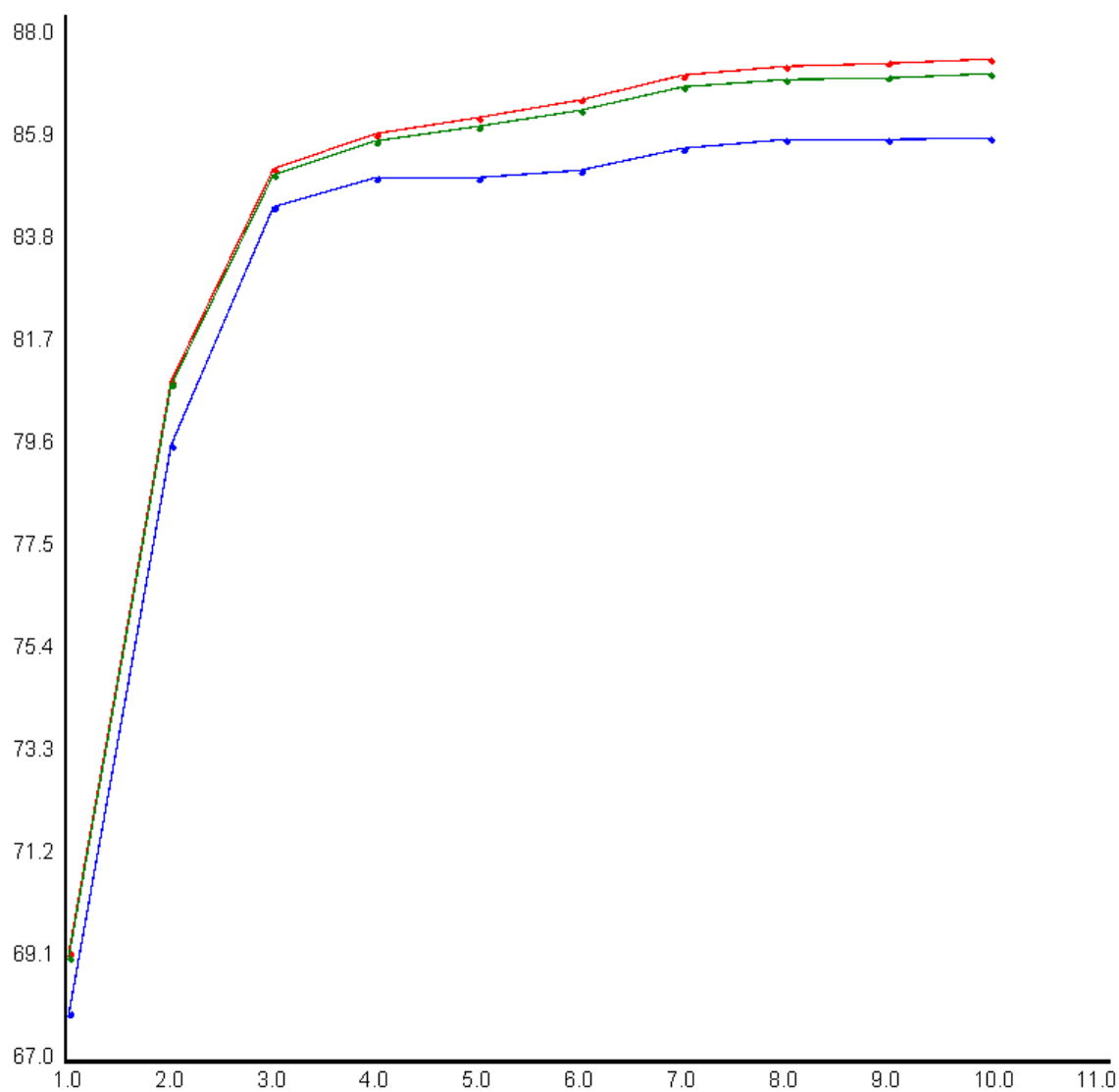
 R^2 vs n for QuadRegression with forward selection feature selection

— □ ×

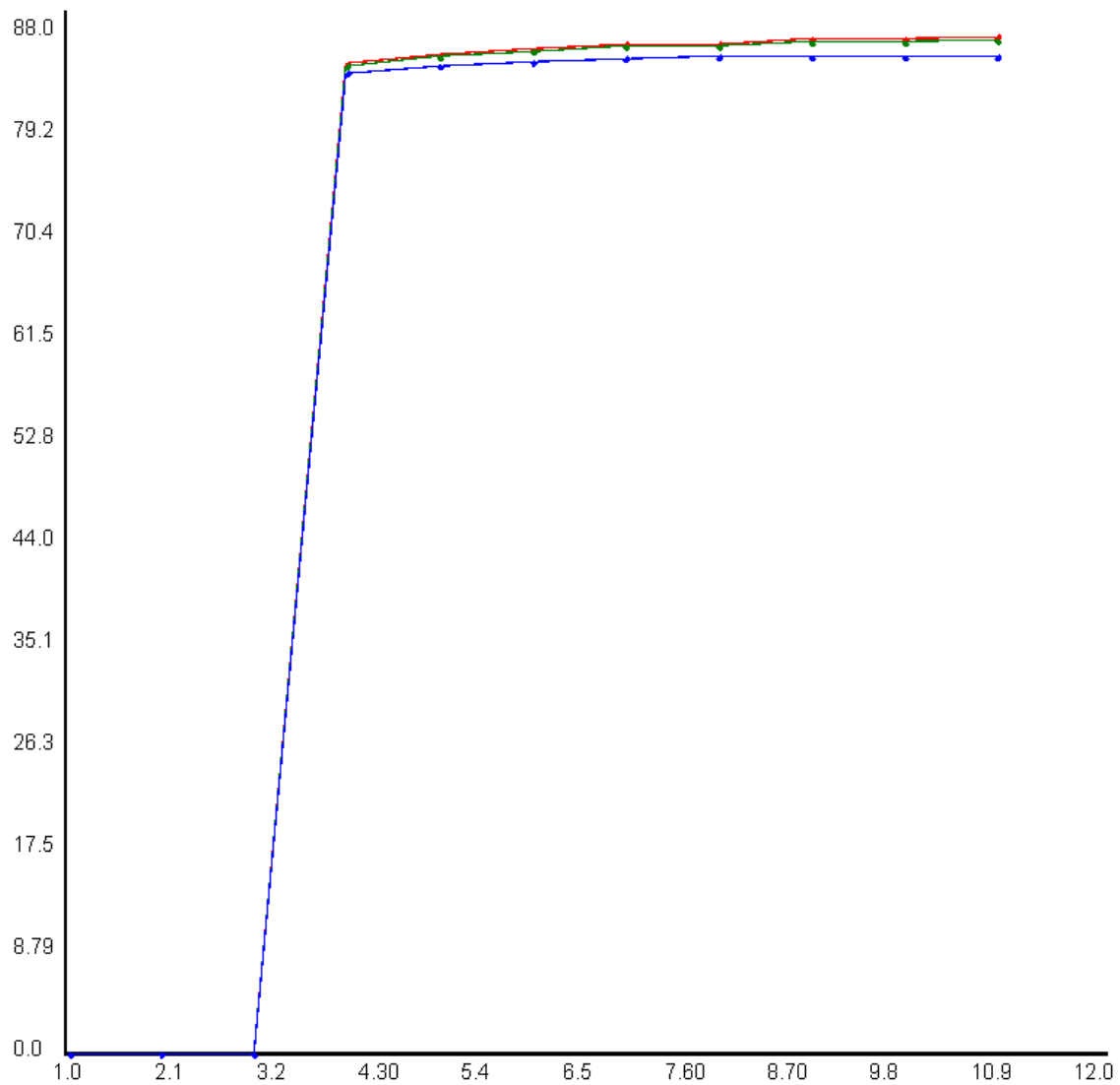


 R^2 vs n for QuadRegression with backward elimination feature selection

— □ ×




R<sup>2</sup> vs n for QuadRegression with stepwise regression feature selection

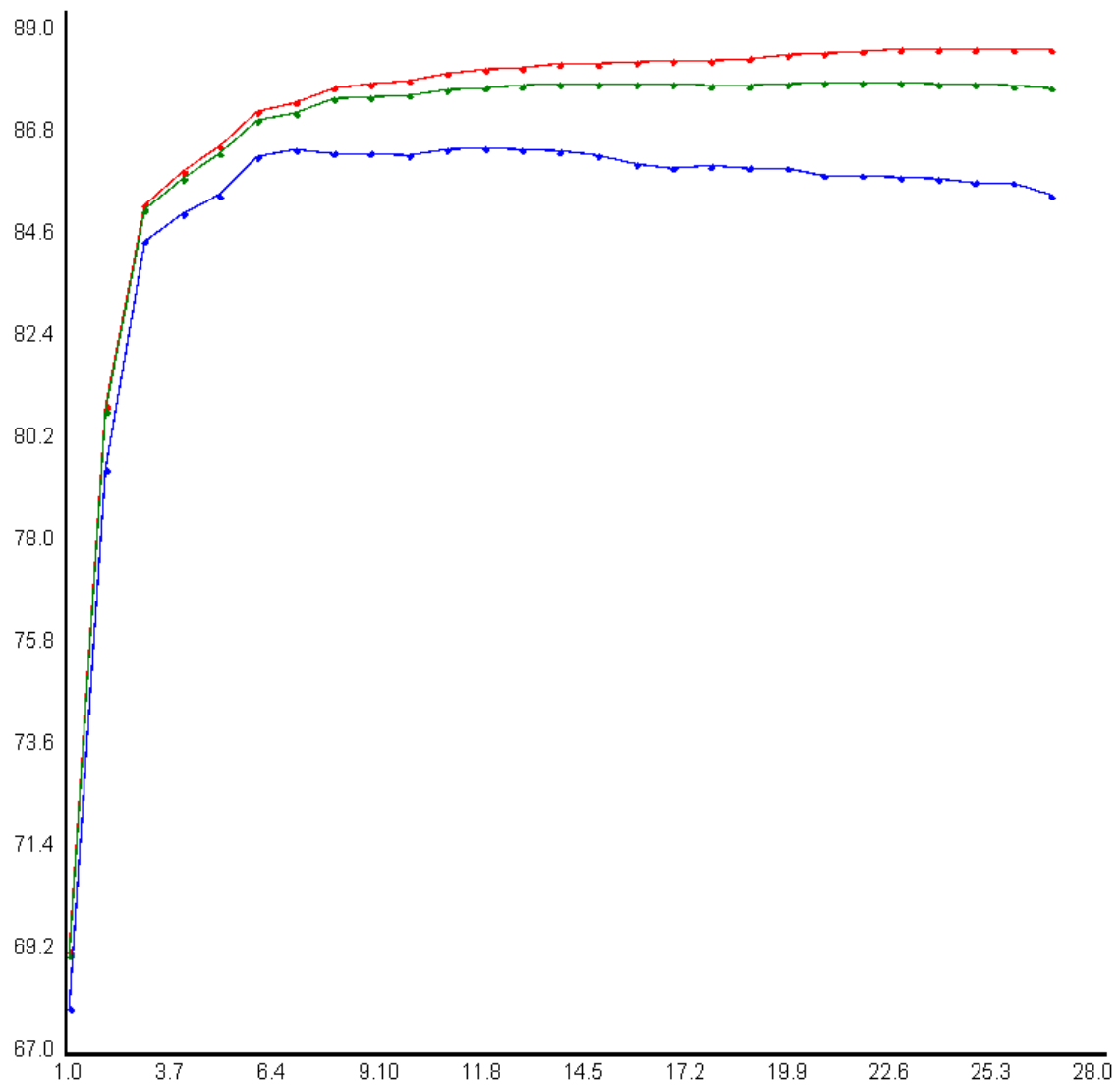


## QuadX Regression


The patterns for quadx regression are generally the same as those for quadratic regression. However, one noticeable difference is that, with forward selection and backward elimination, the cross validated R<sup>2</sup> scores slowly decreased after it reached its peak.

 R<sup>2</sup> vs n for QuadXRegression with forward selection

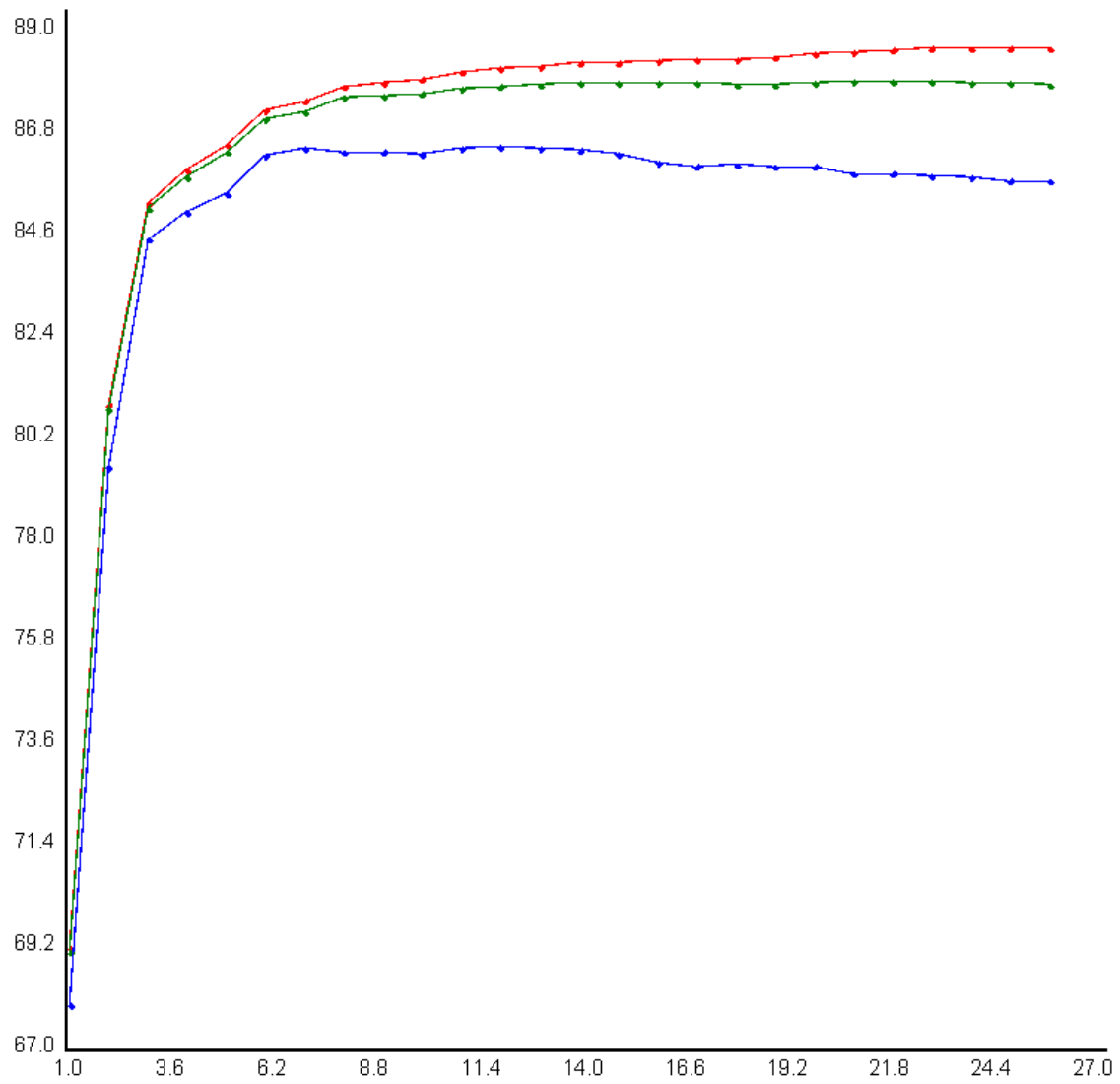
— □ ×



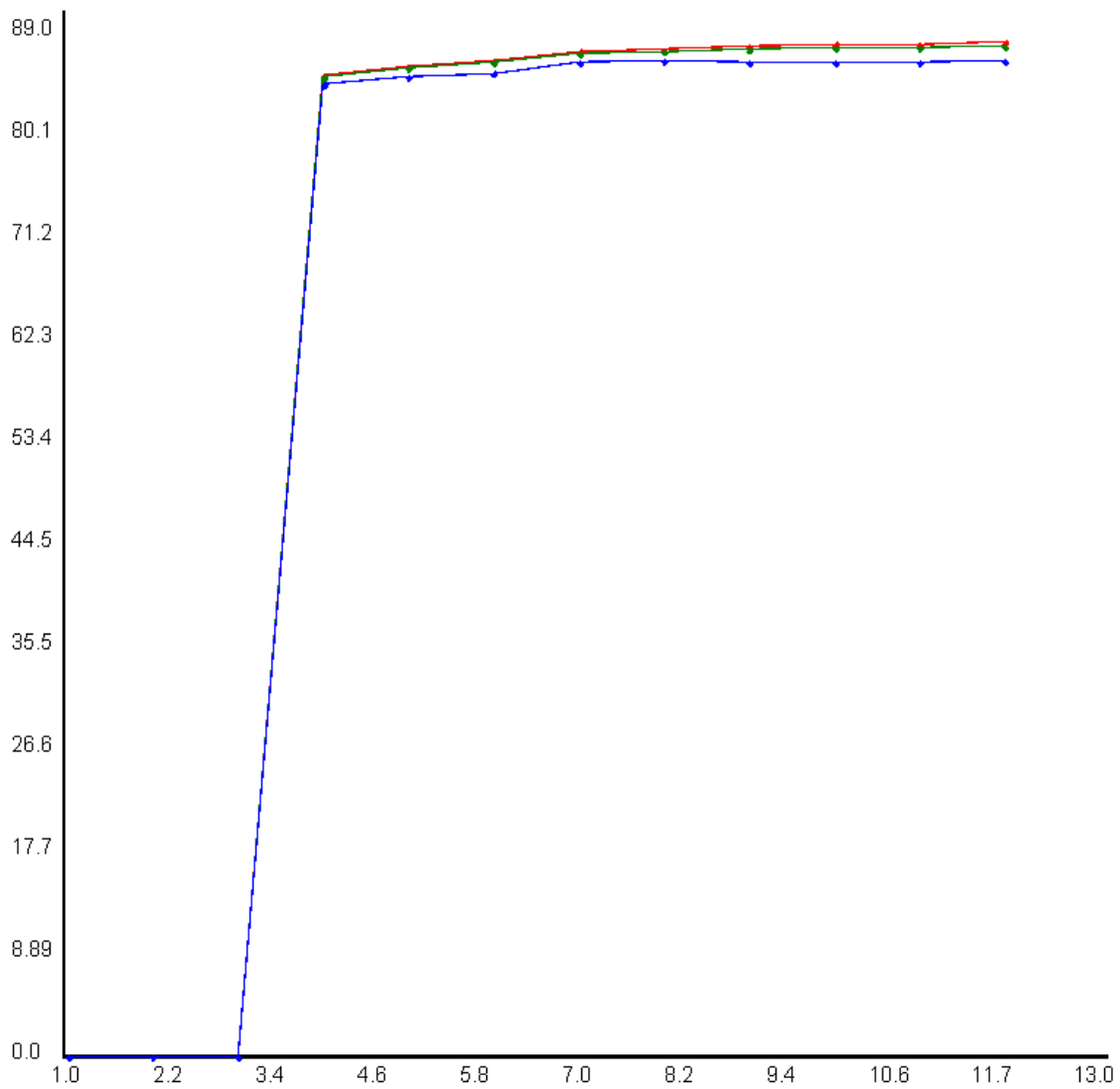


 R<sup>2</sup> vs n for QuadXRegression with backward elimination

— □ ×



R<sup>2</sup> vs n for QuadXRegression with stepwise regression

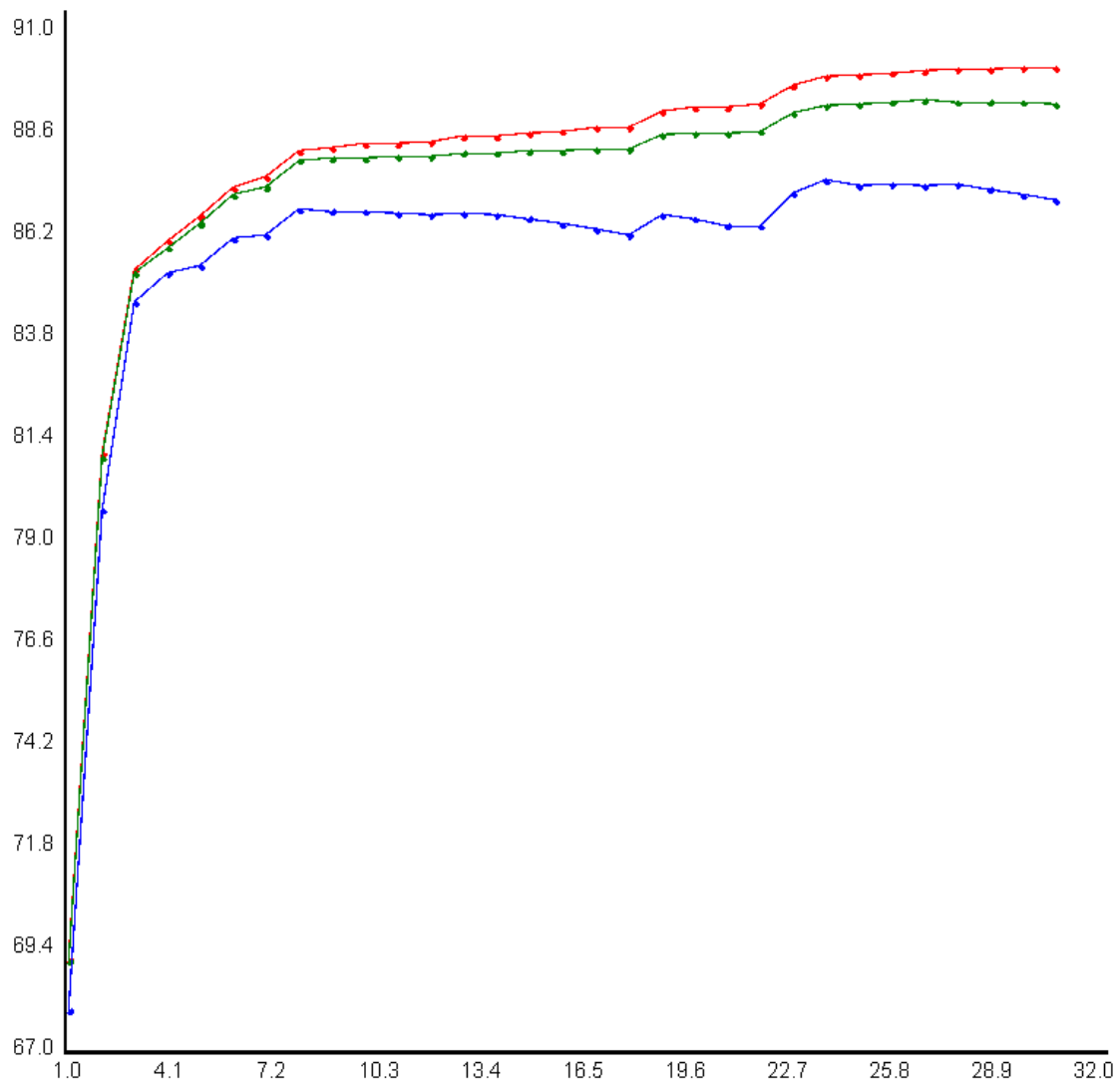


## Cubic Regression

One big feature where cubic regression differs from previous models is that, with forward selection and backward elimination, all three qualities of fits show “stepwise” jump behaviour, indicating that some features have a greater impact on quality of fit than others with cubic regression model.

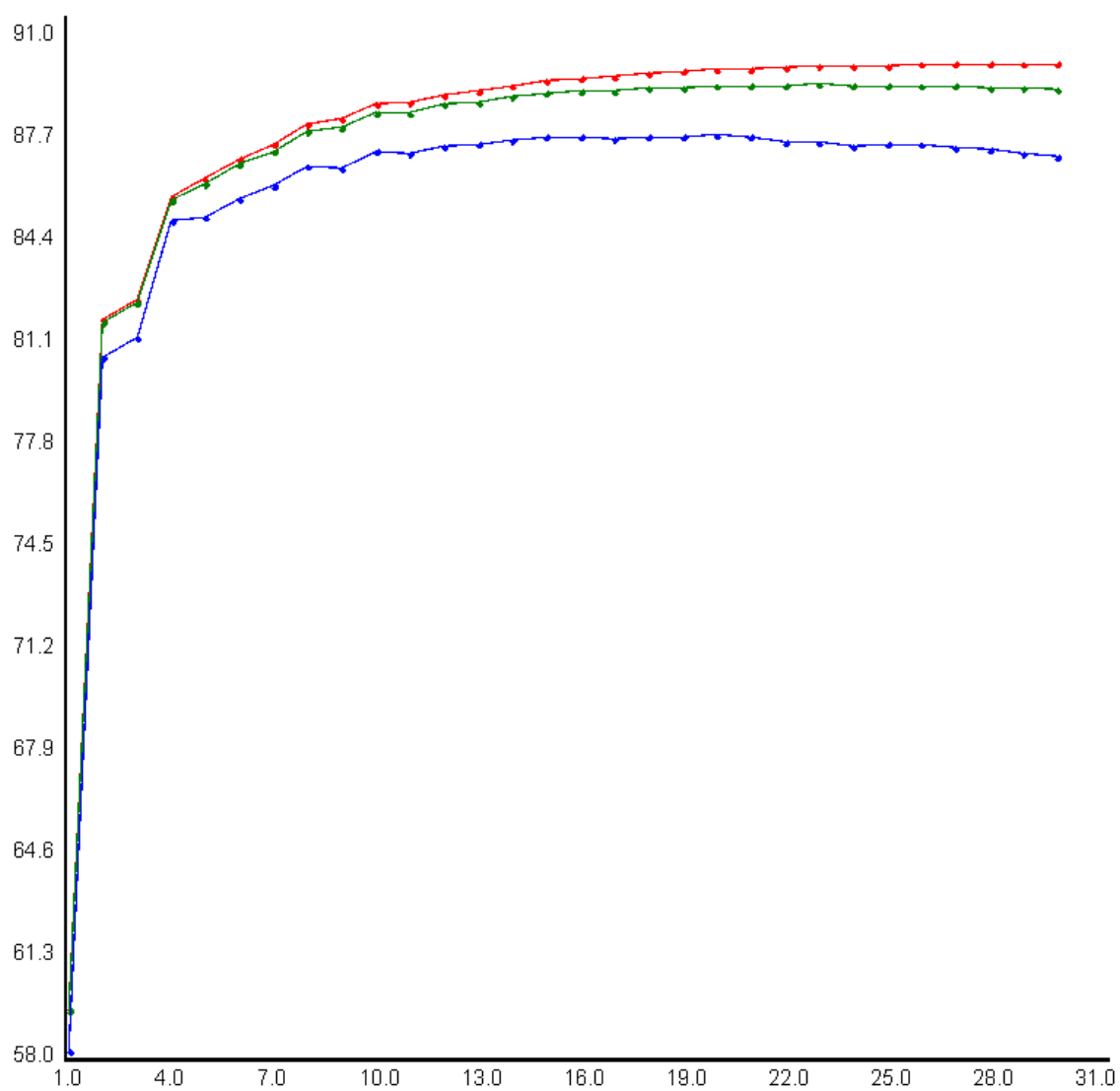
R<sup>2</sup> vs n for CubicRegression with forward selection

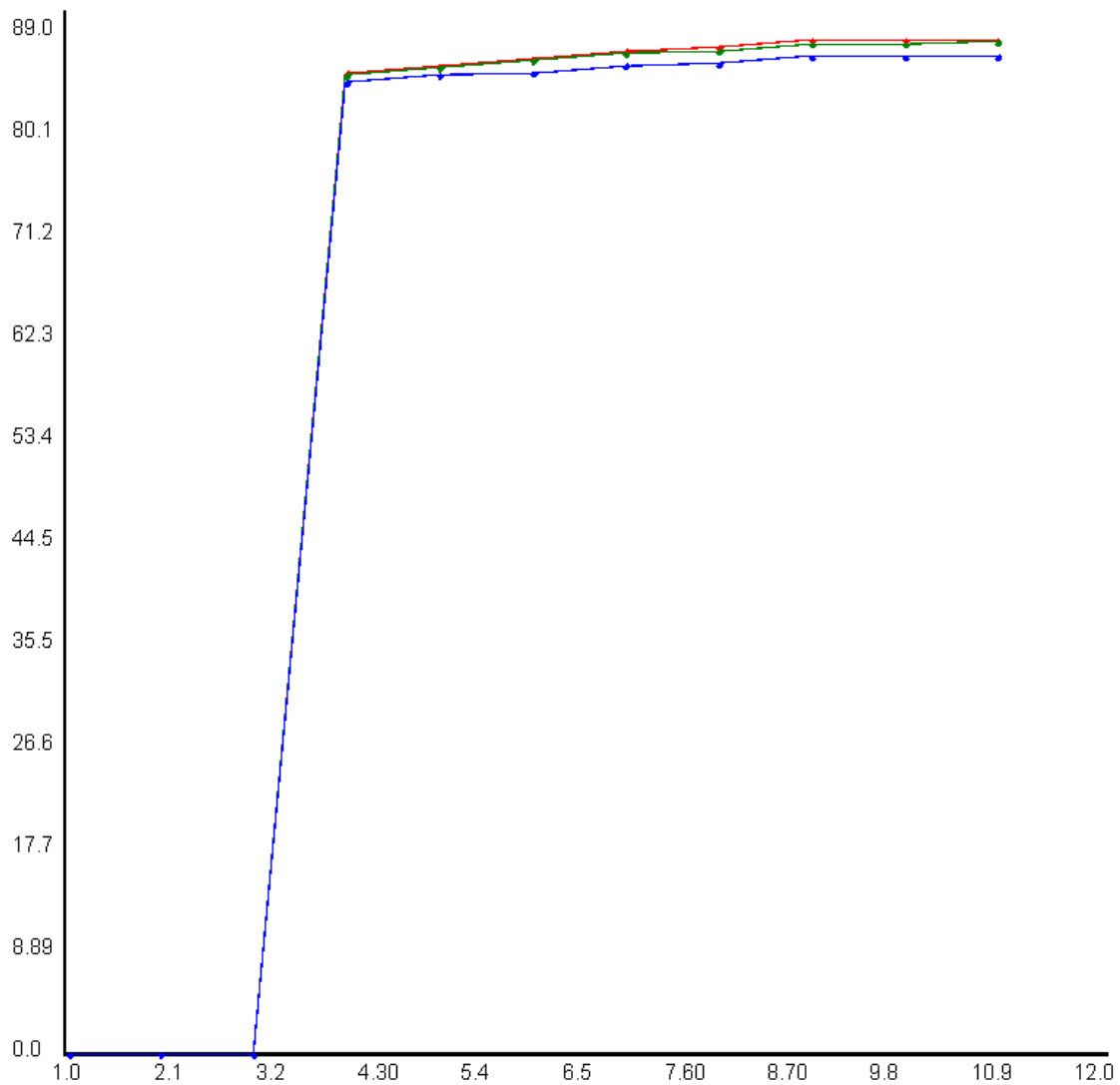
— □ ×



R<sup>2</sup> vs n for CubicRegression with backward elimination


— □ ×



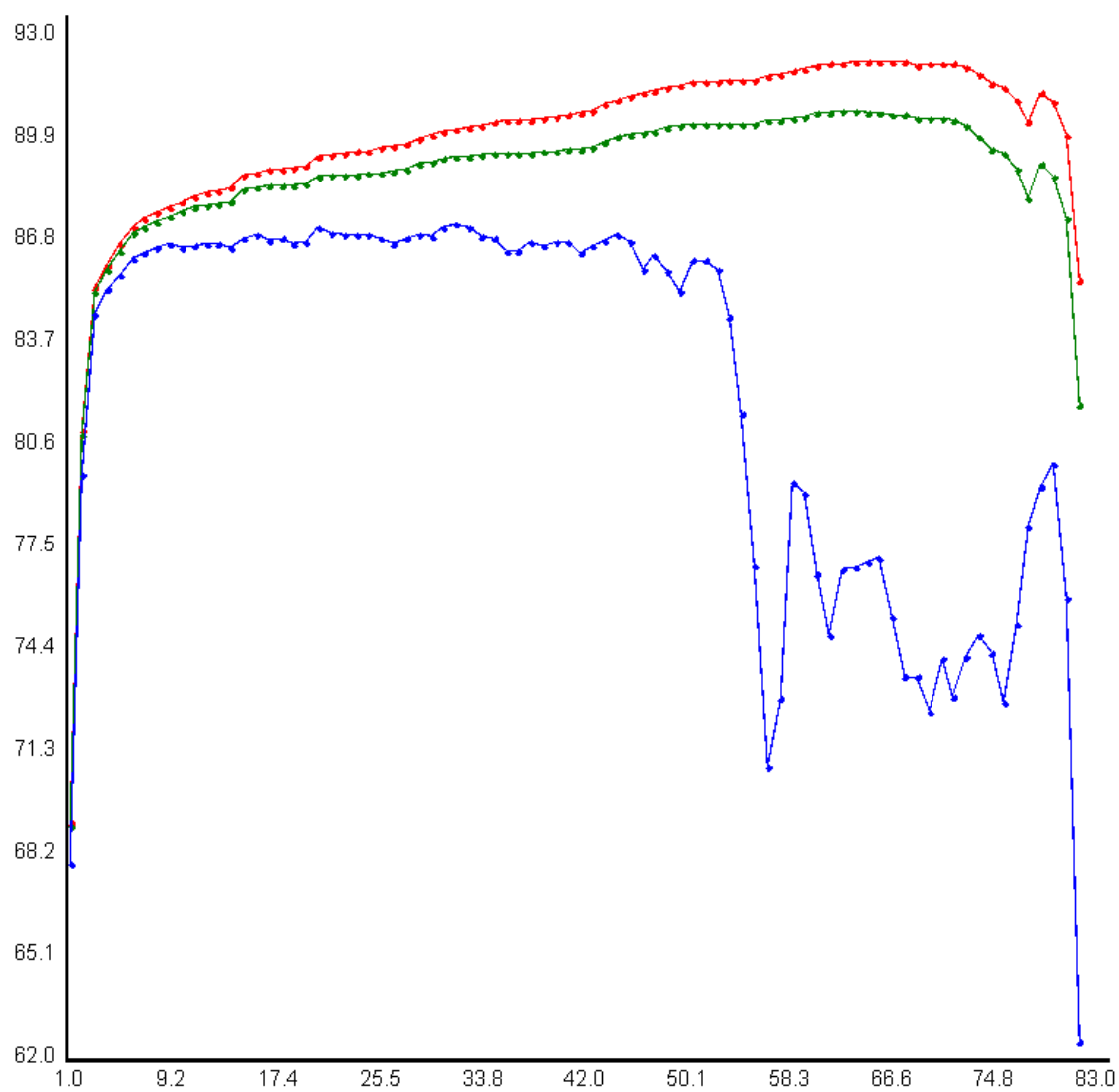



## CubicX Regression

With cubicX regression model, the forward selection process seems not working very well since it shows a sharp drop in all three qualities of fits towards the end, especially there's a jugged edge pattern for cross validated R<sup>2</sup> score.

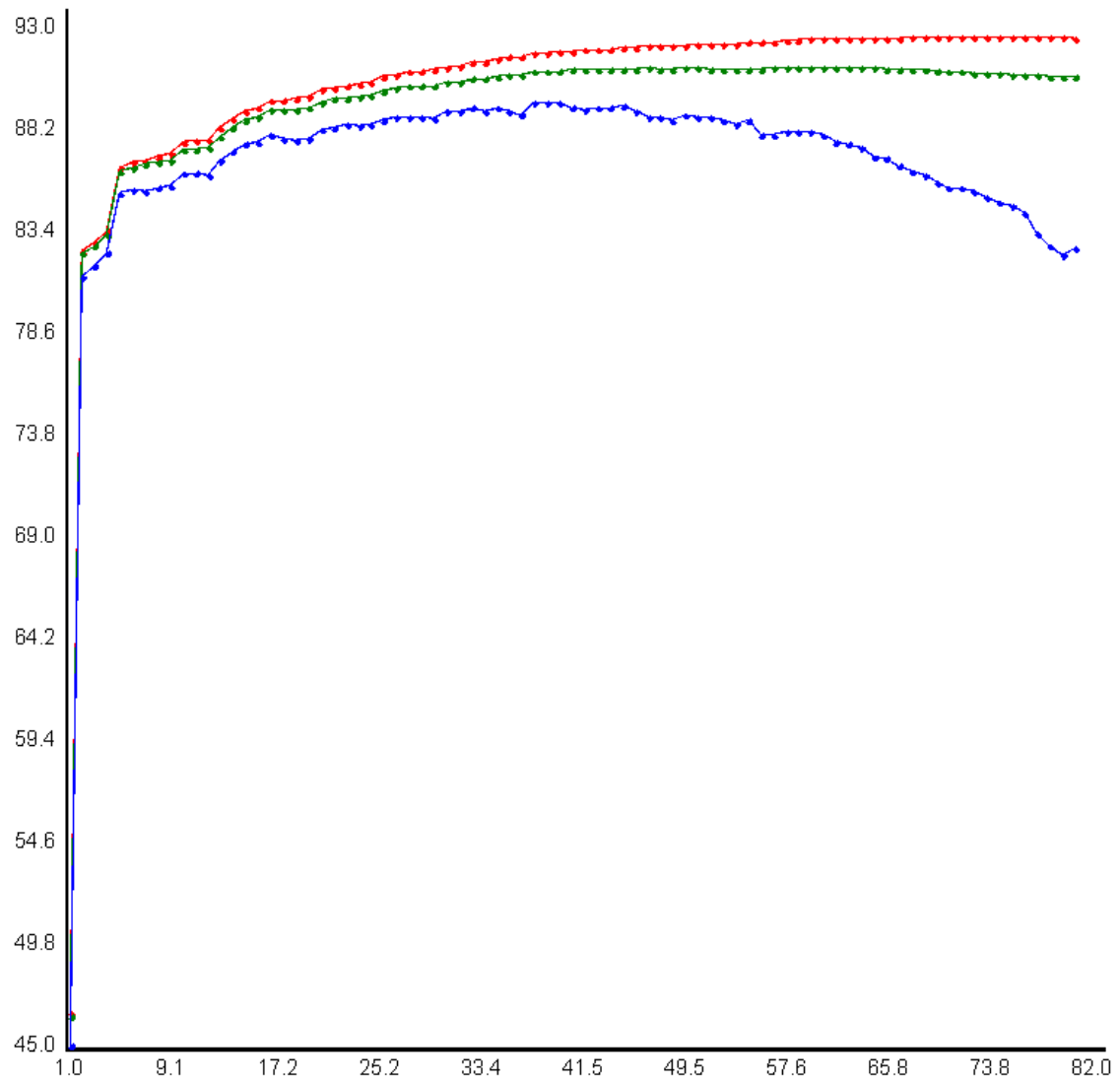
 R^2 vs n for CubicXRegression with forward selection


— □ ×



 R<sup>2</sup> vs n for CubicXRegression with backward elimination

— □ ×



 R^2 vs n for CubicXRegression with stepwise regression

— □ ×

