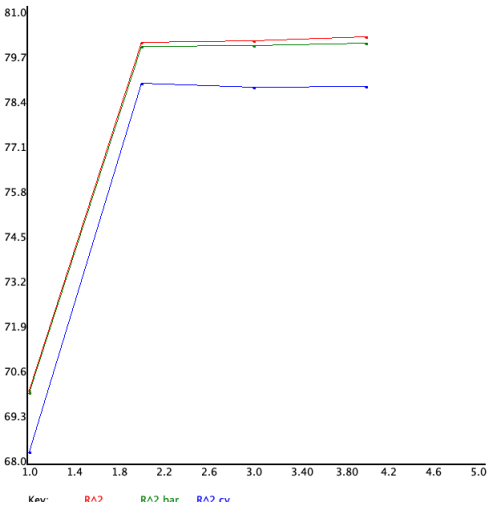
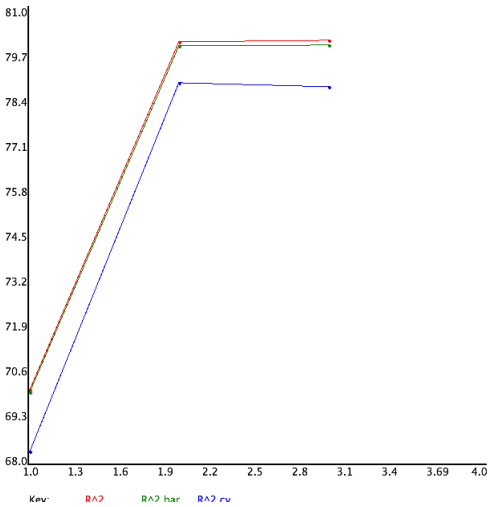
Project 1

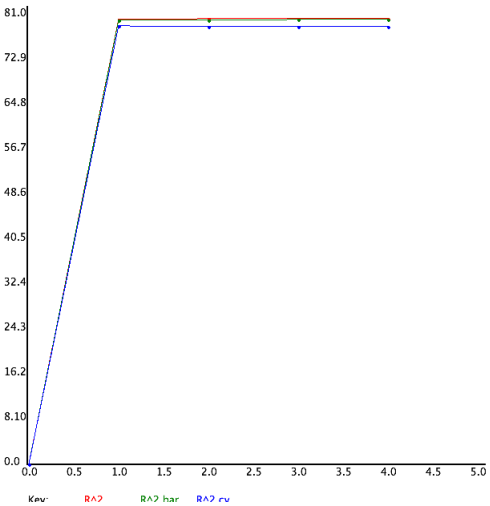
Mingyu Sun, Shubhangi Rai

1 Exercise 6 with Forward Selection, Backward Elimination, Stepwise Regression

We called rg.forwardSelAll () and rg.backwardElimAll () for Forward Selection and Backward Elimination, both of which yield the same results for AutoMPG dataset, as this data set is relatively small.





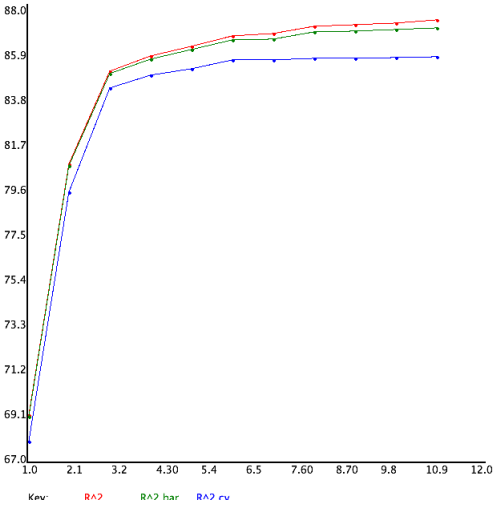


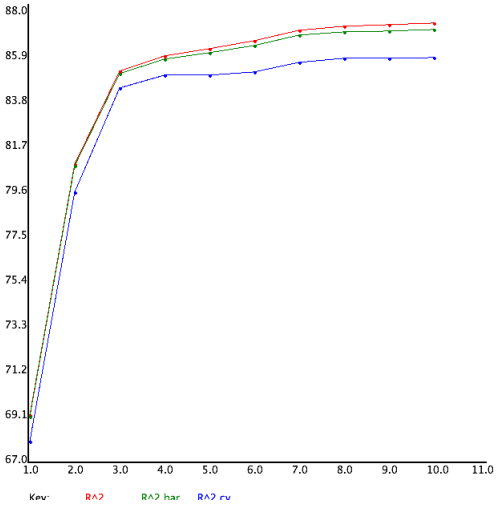
We called rg.stepwiseSelAll () for Stepwise Regression, the result is close to Forward Selection and Backward Elimination, apart from the first row, which corresponds to the first feature included. This is reasonable, as Stepwise Regression starts with no features, then adds the best features, like Forward Selection, but different from Forward Selection, after the second feature is added, we begin to check if we need to eliminate features, namely, do Backward Elimination. So with limited steps, Stepwise Regression should perform better than Forward Selection and Backward Elimination. We note that, even though the result of Stepwise Regression looks similar to Forward Selection and Backward Elimination, this is not always true. With data set of more features, or with Quad Regression, Cubic Regression etc., the difference with show up, as we explain in the following section.

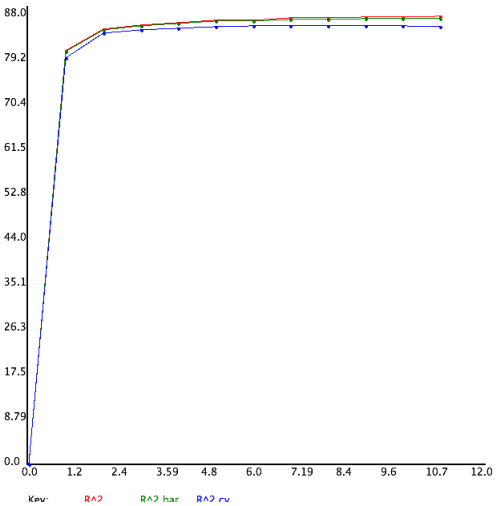
We used $\bar{R}^2$ as the selection criterion, we should not use R^2 as R^2 may always encourage us to select more features. However, within this example, the all the three criterion, $\bar{R}^2$, R^2, $\bar{R}^2_{cv}$, are close, $\bar{R}^2$, R^2, $\bar{R}^2_{cv}$ in particular, for data set with more features, and Quad Regression, Cubic Regression etc., this is not true.

2 Quad Regression, QuadX Regression with Forward Selection, Backward Elimination, Stepwise Regression

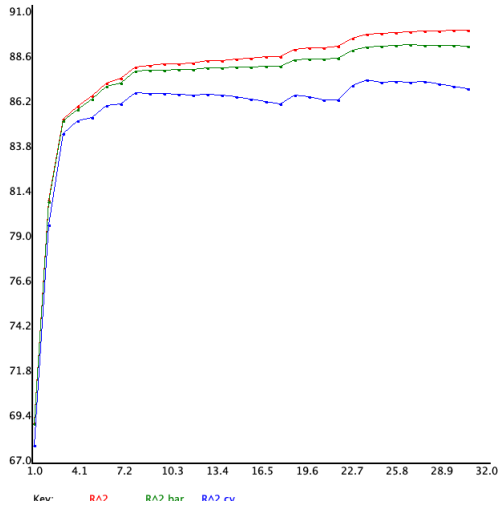
Different from the previous section with regression, the Forward Selection and Backward Elimination do not generate the same result for Quad Regression. This is more clearly demonstrated with $\bar{R}^2_{cv}$, as we can see from the plot (right), the $\bar{R}^2_{cv}$ occasionally go downward, meaning that eliminate certain features based on $\bar{R}^2$ hurt the model. Note that there are 13 terms for Quad Regression. Stepwise Regression also differ, and with Quad Regression, Stepwise Regression shows some advantages, it is more stable

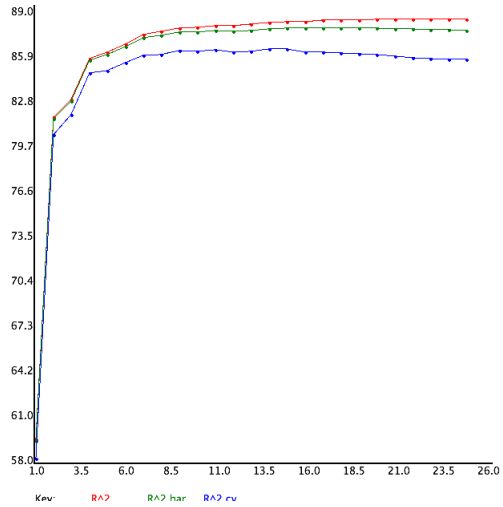


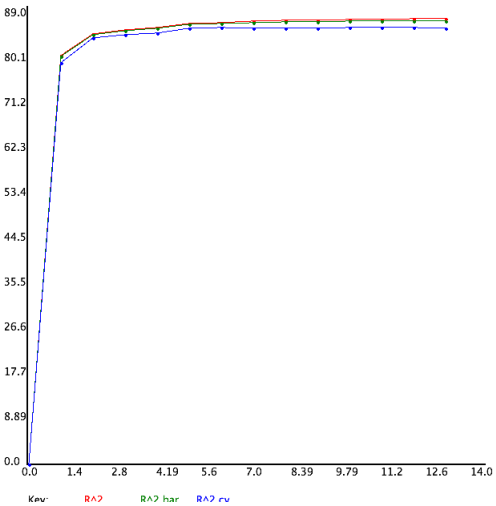




With QuadX Regression, the three Feature Selection methods vary more. There are 28 terms, some of them should correlate with each other. Neither Forward Selection nor Backward Elimination truly performs well, Stepwise Regression may be the best choice. This is evident if we compare the last plot (Stepwise Regression) with the previous two.

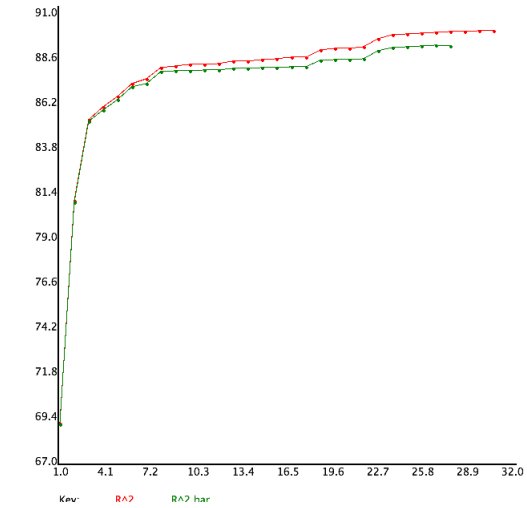


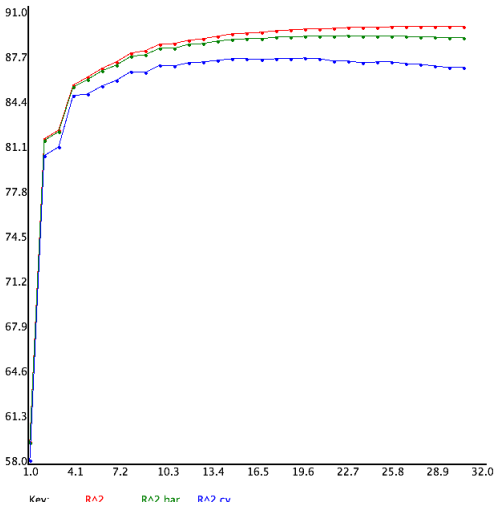


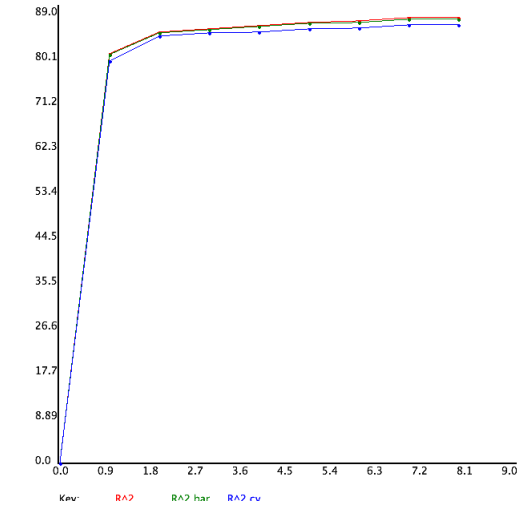


3 Cubic Regression, CubicX Regression with Forward Selection, Backward Elimination, Stepwise Regression

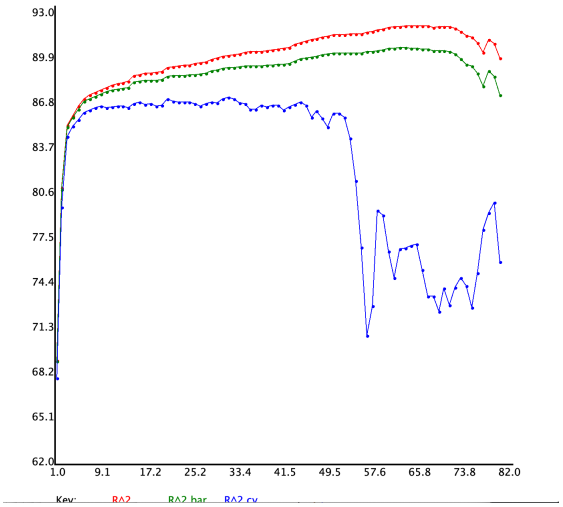
There are 34 terms for Cubic Regression, based on similar reasons, Stepwise Regression is a better choice than Forward Selection or Backward Elimination, this is even more clearly demonstrated in the following 3 plot (the last plot is Stepwise Regression)

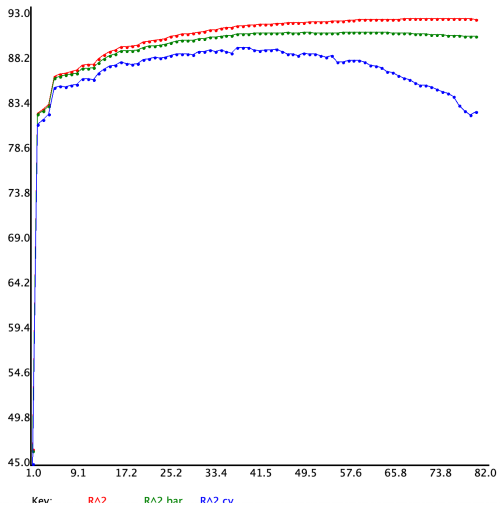


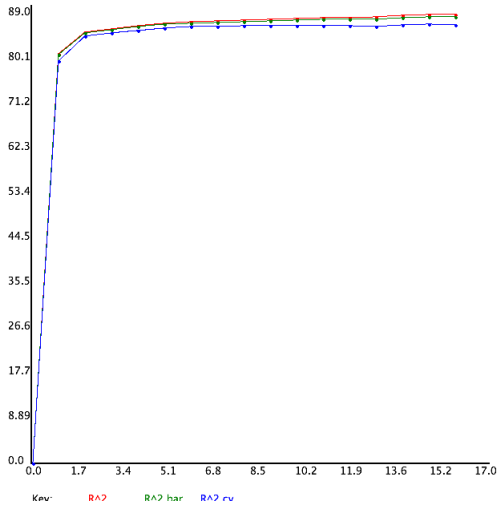




There are 84 terms for CubicX Regression, Forward Selection and Backward Elimination become even more unstable, while Stepwise Regression behaves well.

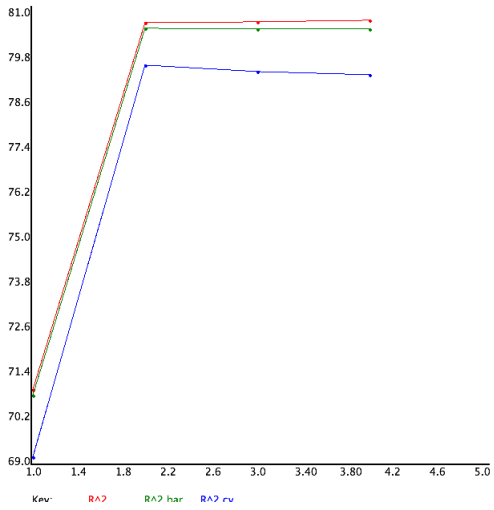

93.0
89.9
86.8
83.7
80.6
77.5
74.4
71.3
68.2
65.1
62.0
1.0
9.1
17.2
25.2
33.4
41.5
49.5
57.6
65.8
73.8
82.0
Key:
RA2
RA2 bar
RA2 cv

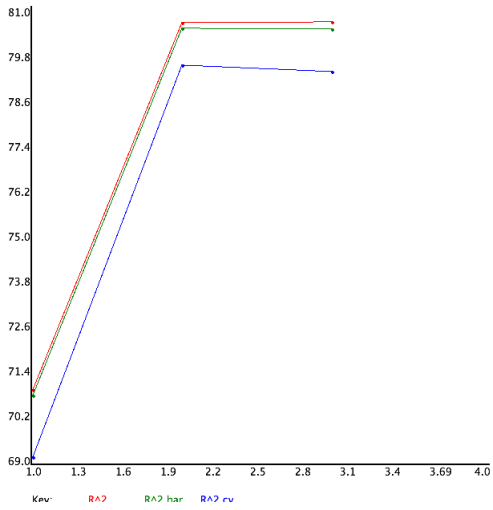

93.0
88.2
83.4
78.6
73.8
69.0
64.2
59.4
54.6
49.8
45.0
1.0
9.1
17.2
25.2
33.4
41.5
49.5
57.6
65.8
73.8
82.0
Key:
RA2
RA2 bar
RA2 cv

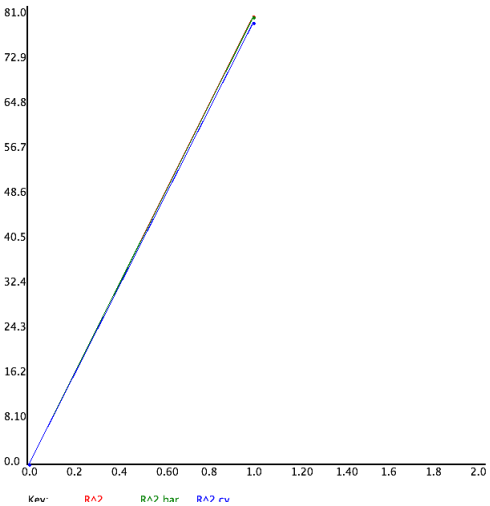

89.0
80.1
71.2
62.3
53.4
44.5
35.5
26.6
17.7
8.89
0.0
0.0
1.7
3.4
5.1
6.8
8.5
10.2
11.9
13.6
15.2
17.0
Key:
RA2
RA2 bar
RA2 cv

4 Ridge Regression, Lasso Regression with Forward Selection, Backward Elimination, Stepwise Regression

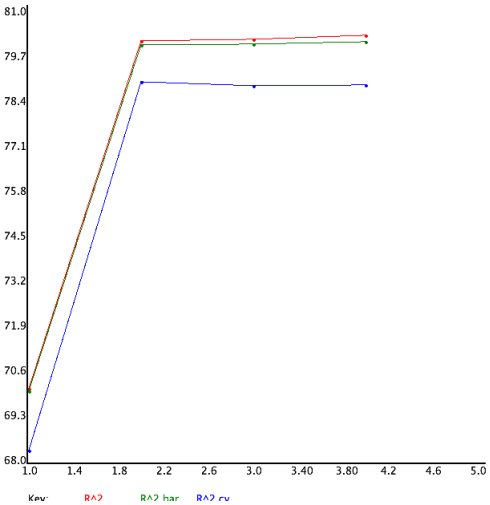
Though the Rubric asks for QR, QXR, CR, CXR, with Ridge Regression and Lasso Regression, based on 7.6 and 7.7 of our textbook, (the two equations) Ridge Regression and Lasso Regression do not involve Quad nor cubic terms, so I did Ridge Regression and Lasso Regression with Forward Selection, Backward Elimination, Stepwise Regression.
The AutoMPG dataset is relatively small, Ridge Regression and Lasso Regression do not truly show their advantages, however, they should be better for feature selection. The advantage is showed in RidgeRegressionstepwise, for this data set.
Results for Ridge Regression (in order of Forward Selection, Backward Elimination, Stepwise Regression)

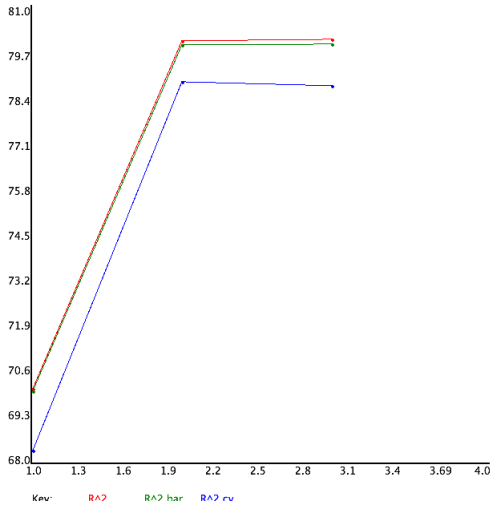


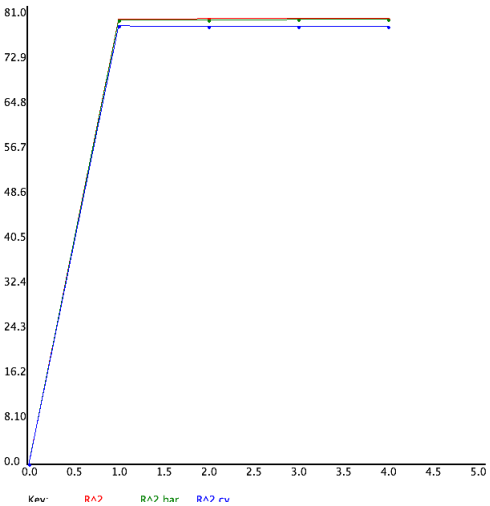




Results for Ridge Regression (in order of Forward Selection, Backward Elimination, Stepwise Regression)







5 Other data sets

We also use Concrete, diabetes, winequality-red, BPressure, basketball, and several other files from UCI Machine Learning Repository. For those data sets that are already included in scalation, we directly import that; for the rest, we add the csv files under the corresponding folder and import with $scalation.columnar_db.Relation$.
For small data sets (those with fewer features), the three feature selection methods do not differ that much. Forward Selection and Backward Elimination are alike. However, with more complicated features, Stepwise Regression has clear advantages Forward Selection or Backward Elimination. With whatever data set, Stepwise Regression also performs better at the beginning.

6 Python

In Python, feature selection can be done with $mlxtend.feature_selection$ (importing $SequentialFeatureSelector$), we could also write functions without $SequentialFeatureSelector$ for all three feature selection method. By adjusting $P - value$, we could get results agree with scalation.

7 Conclusion

We perform Forward Selection, Backward Elimination, Stepwise Regression, Stepwise Regression should be the most stable one, it also performs better at the beginning. The different criterion $\bar{R}^2$, R^2, $\bar{R}^2_{cv}$ do behave differently. Though we use $\bar{R}^2$ for this project, other criteria may yield different results, especially for Forward Selection and Backward Elimination.