CSCI 4360 - Project 1 Analysis

Elise Karinshak, Raj Patel, Lauren Wilkes

Conceptual Analysis

This report observes the relationship between changes in the number of features throughout features selection processes and quality of fit evaluation metrics.

The selection processes we compare are: forward selection, backwards elimination, and stepwise selection. Variable selection is an essential process, as the goal for regression is to effectively predict the response variable while minimizing avoiding variables in the models. These processes assist in identifying variables to include in optimal models. Forward selection begins with a null model and adds features while the metric of interest (in our case, adjusted R^2) approaches its optimal value. In contrast, backwards elimination begins with a full model and eliminates variables until the metric of interest achieves an optimal value. Stepwise selection incorporates both of these approaches by adding or removing a feature at each step as the metric of interest increases. While these methods can produce the same outcomes, this is not a guarantee, and the different methods sometimes identify different optimal parameters.

The metrics we assess are R^2 (the proportion of total error explained by the model), Adjusted R^2 (an R^2 value with a penalty for additional features), Cross validated R^2 (the value of R^2 when incorporating testing through cross-validation methods), and AIC (Akaike Information Criterion, another predictor of error). When assessing these metrics, the goal is to maximize Adjusted R^2 and Cross validated R^2 measures while minimizing AIC; R^2 is helpful for model quality of fit analysis, but less essential when comparing features (as R^2 always increases as terms are added regardless of whether these additional predictors are useful).

The expected relationship between the R^2 metrics is as follows: R^2, since there is no penalty for additional terms, is expected to be larger than adjusted R^2, which accounts for additional parameters. These are both expected to be larger than R^2 cross validated, as this incorporates random testing, and the model should perform worse with the testing data than the training data. It is also worth noting that the adjusted R^2 and cross validated R^2 measures get worse after reaching the optimal value. This occurs for adjusted R^2 because the penalty for the additional terms overpowers the minimal gain in R^2 (indicating the additional predictors is not truly valuable to the model), and cross validated R^2 decreases as the model becomes increasingly overfit (resulting in poor testing results).

In addition to these metrics, lasso and ridge regression techniques supplement this variable selection analysis. Lasso is based on the L-1 norm and forces coefficients to zero. Ridge regression acts similarly, but is based on the L-2 norm and shrinks coefficients (although they rarely equal zero). These metrics can similarly help indicate the most important predictors and help improve variable selection.

We conducted this analysis with the following 6 datasets:
Dataset 1: AutoMPG (https://archive.ics.uci.edu/ml/datasets/Auto+MPG)
Dataset 2: Concrete (https://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test)
Dataset 3: White Wine Quality (https://archive.ics.uci.edu/ml/datasets/Wine+Quality)
Dataset 4: Parkinsons (https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring)
Dataset 5: Absenteeism at Work
(https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work)
Dataset 6: Obesity
(https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+)

For each of these datasets, we analyzed 5 models of varying complexity (multiple linear, quadratic, quadratic with interaction, cubic, and cubic with interaction) to observe the relationship between these quality of fit metrics and their trends, and we used this information to identify the optimal number of parameters for each model.

Dataset Discussion

This discussion is to highlight key insight for each dataset, as well as note observations and applications of the concepts discussed in the introduction (for models of varying complexity); for the full output and all plots for each dataset, refer to our code.

Dataset 1: AutoMPG

Multiple Linear Ridge Regression:

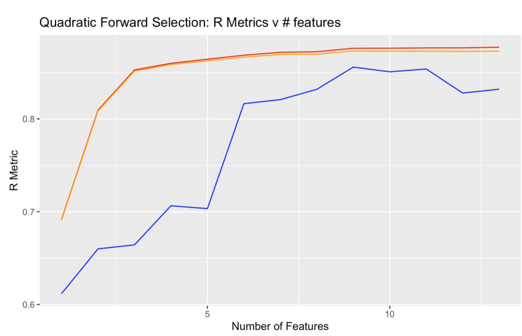
```
SUMMARY
    Parameters/Coefficients:
    Var       Estimate        Std. Error        t value          Pr(>|t|)        VIF
---------------------------------------------------------------------------------
    x0    -0.32983        0.33210          -0.99315         0.32064         NA
    x1    0.00768         0.00736          1.04352          0.29671         19.64168
    x2    -0.00039        0.01384          -0.02829         0.97743         9.39804
    x3    -0.00679        0.00067          -10.14088        0.00000         10.73168
    x4    0.08527         0.10204          0.83572          0.40332         2.62558
    x5    0.75337         0.05262          14.31760         0.00000         1.24483


    Residual standard error: 3.43524 on 385.0 degrees of freedom
    Multiple R-squared:  0.80926,       Adjusted R-squared:  0.80628
    F-statistic: 272.2340982172825 on 6.0 and 385.0 DF,  p-value: 0.0
---------------------------------------------------------------------------------

Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 0, lambda = optimal_ridge_lambda)

  Df  %Dev Lambda
1  7 81.03 0.6493
```

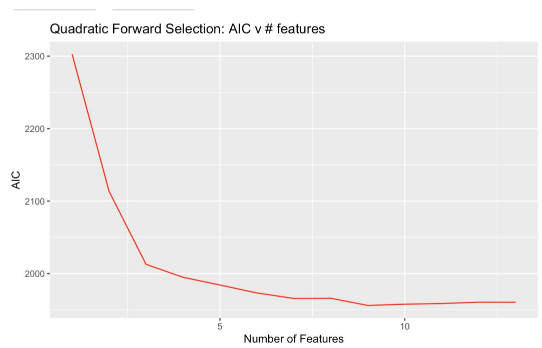
Multiple Linear Lasso Regression:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate     Std. Error       t value          Pr(>|t|)        VIF
----------------------------------------------------------------------------------
    x0   -0.52252         0.32953          -1.58563         0.11282         NA
    x1   0.01022          0.00739          1.38316          0.16662         19.38972
    x2   -0.02087         0.01223          -1.70683         0.08785         7.18602
    x3   -0.00639         0.00066          -9.62893         0.00000         10.32071
    x4   -0.05201         0.09256          -0.56196         0.57414         2.11495
    x5   0.61025          0.02410          25.31786         0.00000         0.25572


    Residual standard error: 3.47202 on 386.0 degrees of freedom
    Multiple R-squared:  0.80464,       Adjusted R-squared:  0.80211
    F-statistic: 317.97399938752665 on 5.0 and 386.0 DF,  p-value: 0.0
----------------------------------------------------------------------------------

 Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 1, lambda = optimal_lasso_lambda)

   Df  %Dev    Lambda
 1  7 82.09 0.005027
```

Both lasso and ridge regression resulted in very similar models. Both of these shrunk the coefficients of less meaningful parameters; ridge accomplishes this using the L2 norm and lasso accomplishes this using the L1 norm. In the ridge model, x1, x2, and x3 are multiplied by very small coefficients (making the the terms insignificant); similarly, in the lasso model, x1, x2, and x3 are also insignificant. Thus, both methods resulted in consistent recommendations regarding which features are meaningful predictors.

Quadratic





This output suggests about 10 features is optimal, as this is where the minimum AIC and maximum adjusted r^2 values occur.

Quadratic Ridge:

```
fname = Array(x0, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11)
SUMMARY
    Parameters/Coefficients:
    Var      Estimate    Std. Error      t value         Pr(>|t|)        VIF
----------------------------------------------------------------------------------
    x0   -5304.78884    38.15531        -139.03147      0.00000         NA
    x1   -36.72436      6.98182         -5.26000        0.00000         294.77161
    x2   2.55936        0.57022         4.48839         0.00001         270.82059
    x3   0.68156        0.09813         6.94557         0.00000         215.53010
    x4   -0.00126       0.00018         -6.99214        0.00000         158.76922
    x5   -0.49256       0.20651         -2.38511        0.01707         130.49091
    x6   0.00180        0.00070         2.55640         0.01058         97.18286
    x7   -0.05657       0.01316         -4.29900        0.00002         261.35563
    x8   0.00001        0.00000         4.29132         0.00002         201.27701
    x9   -21.50572      2.72981         -7.87812        0.00000         123.79695
    x10  0.60855        0.08077         7.53477         0.00000         105.72141
    x11  151.29797      0.21783         694.57015       0.00000         2.71450

    Residual standard error: 13.88251 on 379.0 degrees of freedom
    Multiple R-squared:  0.99953,       Adjusted R-squared:  0.99951
    F-statistic: 67069.8555500694 on 12.0 and 379.0 DF,  p-value: 0.0
----------------------------------------------------------------------------------

-------------------------
| Quad Ridge Regression |
-------------------------

 Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 0, lambda = optimal_ridge_lambda)

   Df  %Dev Lambda
 1 13 82.66 0.6493
```

Quadratic Lasso:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate    Std. Error      t value         Pr(>|t|)        VIF
----------------------------------------------------------------------------------
    x0   -5705.20167    29.89321        -190.85276      0.00000         NA
    x1   -6.86276       5.87248         -1.16863        0.24255         294.77161
    x2   0.32933        0.47962         0.68665         0.49231         270.82059
    x3   0.20627        0.08254         2.49904         0.01245         215.53010
    x4   -0.00029       0.00015         -1.94872        0.05133         158.76922
    x5   -0.04557       0.17370         -0.26236        0.79304         130.49091
    x6   0.00077        0.00059         1.30410         0.19220         97.18286
    x7   -0.01766       0.01107         -1.59513        0.11068         261.35563
    x8   0.00000        0.00000         0.75680         0.44917         201.27701
    x9   -4.82337       2.29607         -2.10071        0.03567         123.79695
    x10  0.14779        0.06793         2.17553         0.02959         105.72141
    x11  152.24350      0.18322         830.93815       0.00000         2.71450

    Residual standard error: 11.67673 on 380.0 degrees of freedom
    Multiple R-squared:  0.99967,       Adjusted R-squared:  0.99966
    F-statistic: 103435.40619088199 on 11.0 and 380.0 DF,  p-value: 0.0
----------------------------------------------------------------------------------

-------------------------
| Quad Lasso Regression |
-------------------------

 Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 1, lambda = optimal_lasso_lambda)

   Df  %Dev    Lambda
 1 13 87.67 0.0006493
```

Again, the results for lasso and ridge for the quadratic model were fairly consistent with each other, as well as with the output of forward / backward selection. Both most dramatically shrunk x4, x6, and x8. We can also observe that Lasso does in fact shrink the parameters more. We also observe how adding the quadratic terms increased the R squared.

QuadX Ridge:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate    Std. Error      t value        Pr(>|t|)       VIF
--------------------------------------------------------------------------------
    x0   -563.91576     84.55809        -6.66897        0.00000        NA
    x1   -114.15343     21.03480        -5.42688        0.00000        12095.66551
    x2   4.43508        0.44060         10.06592        0.00000        19643.11095
    x3   7.26882        0.68412         10.62502        0.00000        6473.76716
    x4   -0.53312       0.03971         -13.42623       0.00000        10757.74455
    x5   70.76766       3.07749         22.99526        0.00000        711.28545
    x6   8.11845        0.92122         8.81274         0.00000        219.47454
    x7   -1.82907       1.04615         -1.74839        0.08040        4120.84699
    x8   0.12871        0.03767         3.41694         0.00063        13012.57500
    x9   -0.00213       0.05892         -0.03621        0.97111        5635.44017
    x10  -0.00526       0.00272         -1.93327        0.05320        7128.04114
    x11  0.21681        0.42071         0.51536         0.60630        805.59947
    x12  1.59634        0.23362         6.83313         0.00000        7746.67640
    x13  -0.00150       0.00043         -3.46737        0.00053        4183.73357
    x14  -0.00156       0.00093         -1.67764        0.09342        3721.81982
    x15  0.00013        0.00005         2.34210         0.01918        7249.17408
    x16  -0.00784       0.00895         -0.87647        0.38078        1230.43568
    x17  -0.06111       0.00545         -11.21077       0.00000        15762.42641
    x18  -0.00327       0.00137         -2.37642        0.01748        1673.74482
    x19  0.00034        0.00010         3.44009         0.00058        4338.24194
    x20  -0.01562       0.01841         -0.84894        0.39591        502.99348
    x21  -0.09338       0.00744         -12.54759       0.00000        3820.96353
    x22  -0.00001       0.00000         -2.08647        0.03694        1916.10591
    x23  0.00110        0.00081         1.35261         0.17618        912.72501
    x24  0.00687        0.00047         14.60768        0.00000        7818.42521
    x25  0.11839        0.08366         1.41523         0.15700        512.76681


    Residual standard error: 6.52929 on 365.0 degrees of freedom
    Multiple R-squared:  0.99930,       Adjusted R-squared:  0.99925
    F-statistic: 19945.428571830387 on 26.0 and 365.0 DF,  p-value: 0.0
--------------------------------------------------------------------------------

-------------------------
| QuadX Ridge Regression |
-------------------------
```

```
Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 0, lambda = optimal_ridge_lambda)

  Df  %Dev Lambda
1 34 86.04 0.6493
```

QuadX Lasso:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate    Std. Error      t value        Pr(>|t|)       VIF
--------------------------------------------------------------------------------
    x0   -6139.25579    198.52566       -30.92424       0.00000        NA
    x1   22.27774       33.02121        0.67465         0.49990        12120.91764
    x2   1.24300        0.69513         1.78817         0.07375        19880.95960
    x3   5.57365        1.44944         3.84537         0.00012        11816.49593
    x4   -0.23604       0.06970         -3.38666        0.00071        13477.94627
    x5   5.88126        9.78221         0.60122         0.54769        2922.27535
    x6   157.68690      2.25753         69.84930        0.00000        535.94944
    x7   1.08528        1.65019         0.65767         0.51075        4169.35201
    x8   -0.05719       0.05982         -0.95608        0.33903        13343.01681
    x9   0.04052        0.09336         0.43400         0.66429        5752.03787
    x10  0.00074        0.00427         0.17252         0.86303        7132.24500
    x11  0.47316        0.66142         0.71537         0.47438        809.66763
    x12  -0.51640       0.36667         -1.40833        0.15903        7760.01693
    x13  0.00075        0.00068         1.10091         0.27093        4228.92792
    x14  -0.00362       0.00146         -2.48948        0.01279        3721.86785
    x15  0.00001        0.00008         0.09797         0.92196        7255.37556
    x16  -0.01765       0.01403         -1.25767        0.20851        1230.57107
    x17  -0.00663       0.00868         -0.76473        0.44443        16233.31696
    x18  0.00018        0.00222         0.08284         0.93397        1779.57741
    x19  0.00016        0.00016         0.99452         0.31997        4567.06405
    x20  0.02928        0.02990         0.97919         0.32749        539.86776
    x21  -0.07818       0.01551         -5.04230        0.00000        6745.02793
    x22  -0.00000       0.00000         -0.28893        0.77264        1942.99294
    x23  -0.00071       0.00129         -0.55371        0.57977        935.96921
    x24  0.00290        0.00082         3.52450         0.00042        9728.40265
    x25  0.12751        0.13210         0.96526         0.33442        519.87754
    x26  -0.12886       0.09986         -1.29041        0.19691        2104.23787


    Residual standard error: 10.23924 on 365.0 degrees of freedom
    Multiple R-squared:  0.99975,       Adjusted R-squared:  0.99974
    F-statistic: 56915.892727037935 on 26.0 and 365.0 DF,  p-value: 0.0
--------------------------------------------------------------------------------

-------------------------
| QuadX Lasso Regression |
-------------------------
```

```
Call:  glmnet(x = train_matrix, y = unlist(response), alpha = 1, lambda = optimal_lasso_lambda)

  Df  %Dev   Lambda
1 30 88.82 0.001807
```

As the number of variables increases, ridge and lasso are shrinking more and more features (minimizing the effects of insignificant predictors).

To show how these relationships between shrunken variables continue with ridge and lasso regressions for other complex models, we have also included our output for Cubic and CubicXR. The effectiveness of these methods removing ineffective parameters is especially apparent for the CubicXR model, as this model includes hundreds of parameters prior to applying shrinking effects.

Cubic Ridge:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate      Std. Error      t value       Pr(>|t|)      VIF
--------------------------------------------------------------------------------
    x0   26856.74173    21788.25617     1.23262       0.21772       NA
    x1   13284.02087    3106.10737      4.27674       0.00002       36877.29356
    x2   -50.77135      39.20036        -1.29518      0.19526       21740.43093
    x3   876.00464      89.14788        9.82642       0.00000       15370.35376
    x4   -11.58869      4.22197         -2.74485      0.00605       17005.43519
    x5   7246.26737     900.00636       8.05135       0.00000       8505.79829
    x6   -9149.27502    468.44635       -19.53111     0.00000       7935.11072
    x7   -1222.49887    457.13557       -2.67426      0.00749       110018.48189
    x8   1.70436        4.23673         0.40228       0.68748       23017.46513
    x9   -5.01473       5.91178         -0.84826      0.39629       7931.51699
    x10  0.26831        0.24431         1.09823       0.27211       8043.44975
    x11  6.62341        37.68247        0.17577       0.86048       903.66758
    x12  -83.42790      20.09424        -4.15183      0.00003       8013.47789
    x13  0.09668        0.09763         0.99031       0.32202       29653.62268
    x14  -0.22295       0.09031         -2.46877      0.01356       4925.62895
    x15  -0.00704       0.00480         -1.46711      0.14235       8131.47815
    x16  -1.29522       0.82902         -1.56235      0.11821       1477.23393
    x17  1.00811        0.47908         2.10427       0.03536       17023.82692
    x18  -1.03566       0.35334         -2.93104      0.00338       15469.55870
    x19  0.03872        0.00944         4.09938       0.00004       5577.93639
    x20  -9.40318       1.84563         -5.09484      0.00000       707.19815
    x21  -7.93263       0.88255         -8.98833      0.00000       7513.71129
    x22  -0.00242       0.00084         -2.89582      0.00378       27122.99489
    x23  0.25361        0.08011         3.16580       0.00155       1242.80293
    x24  0.14848        0.04625         3.21029       0.00133       10583.88748
    x25  -160.89771     37.83484        -4.25263      0.00002       14664.80856
    x26  -49.17991      5.54247         -8.87329      0.00000       2228.95812
    x27  183.59345      2.86533         64.07413      0.00000       4178.68334
    x28  62.63004       26.48551        2.36469       0.01805       31776.12359
    x29  -0.00005       0.00009         -0.54815      0.58359       4514.32603
    x30  0.00140        0.00069         2.02254       0.04312       2704.05114
    x31  0.00000        0.00000         2.91758       0.00353       5592.38883
    x32  2.63137        0.68050         3.86682       0.00011       3007.41967

    Residual standard error: 552.17825 on 358.0 degrees of freedom
    Multiple R-squared:  0.99994,       Adjusted R-squared:  0.99993
    F-statistic: 177336.41299897514 on 33.0 and 358.0 DF,  p-value: 0.0
--------------------------------------------------------------------------------

--------------------------
| Cubic Ridge Regression |
--------------------------
```

Cubic Lasso:

```
SUMMARY
    Parameters/Coefficients:
    Var      Estimate      Std. Error      t value       Pr(>|t|)      VIF
--------------------------------------------------------------------------------
    x0   433567.31363   1477.03365      293.53923     0.00000       NA
    x1   345.42721      210.63178       1.63996       0.10101       36877.29356
    x2   -9.71883       2.65826         -3.65609      0.00026       21740.43093
    x3   6.09993        6.04531         1.00904       0.31296       15370.35376
    x4   0.94813        0.28630         3.31165       0.00093       17005.43518
    x5   140.70283      61.03135        2.30542       0.02114       8505.79829
    x6   -17274.95286   31.76635        -543.81299    0.00000       7935.11072
    x7   -8.60440       30.99934        -0.27757      0.78134       110018.48189
    x8   -0.42419       0.28730         -1.47647      0.13982       23017.46513
    x9   0.12087        0.40089         0.30151       0.76302       7931.51699
    x10  0.02495        0.01657         1.50578       0.13212       8043.44975
    x11  -1.88478       2.55533         -0.73759      0.46076       903.66758
    x12  -3.96777       1.36263         -2.91184      0.00359       8013.47789
    x13  0.01170        0.00662         1.76800       0.07706       29653.62268
    x14  0.00719        0.00612         1.17377       0.24049       4925.62895
    x15  -0.00053       0.00033         -1.62287      0.10462       8131.47815
    x16  0.08699        0.05622         1.54740       0.12177       1477.23393
    x17  0.10771        0.03249         3.31558       0.00091       17023.82325
    x18  -0.06486       0.02396         -2.70672      0.00680       15469.55813
    x19  0.00004        0.00064         0.05526       0.95593       5577.93539
    x20  -0.26923       0.12516         -2.15113      0.03147       707.19810
    x21  0.07631        0.05985         1.27514       0.20226       7513.71128
    x22  -0.00003       0.00006         -0.60875      0.54269       27122.99328
    x23  -0.00275       0.00543         -0.50706      0.61212       1242.80289
    x24  -0.01130       0.00314         -3.60190      0.00032       10583.88662
    x25  -4.78012       2.56566         -1.86311      0.06245       14664.80855
    x26  -0.26779       0.37585         -0.71250      0.47616       2228.95803
    x27  228.00659      0.19430         1173.45252    0.00000       4178.68327
    x28  0.69613        1.79604         0.38759       0.69832       31776.12359
    x29  -0.00001       0.00001         -1.36680      0.17169       4514.31884
    x30  0.00011        0.00005         2.30511       0.02116       2704.05105
    x31  0.00000        0.00000         0.86558       0.38672       5592.38883
    x32  0.08053        0.04615         1.74512       0.08096       3007.41967

    Residual standard error: 37.44439 on 359.0 degrees of freedom
    Multiple R-squared:  1.00000,       Adjusted R-squared:  1.00000
    F-statistic: 3.9771637354515254E7 on 32.0 and 359.0 DF,  p-value: 0.0
--------------------------------------------------------------------------------

--------------------------
| Cubic Lasso Regression |
--------------------------
```

CubicX Ridge:

SUMMARY

Parameters/Coefficients:

Var	Estimate	Std. Error	t value	Pr(>|t|)	VIF
x0	257.43171	52085.88513	0.00494	0.99606	NA
x1	754.46294	9070.40059	0.08318	0.93371	15333646.26662
x2	−1091.43573	196.43065	−5.55634	0.00000	26617820.20461
x3	959.78325	412.10277	2.32899	0.01986	16015446.19206
x4	149.04590	18.48639	8.06247	0.00000	15897459.70804
x5	4726.40466	2682.46587	1.76196	0.07808	3684330.50052
x6	−5309.48651	1196.97660	−4.43575	0.00001	2526218.28560
x7	1255.45194	587.74947	2.13603	0.03268	8868013.40768
x8	−49.83231	20.92541	−2.38143	0.01725	27378548.79746
x9	66.75606	36.29819	1.83910	0.06590	14579930.98385
x10	−0.86985	1.61673	−0.53803	0.59056	17175413.92445
x11	864.15023	262.80441	3.28819	0.00101	2143194.72104
x12	−300.66122	198.91746	−1.51149	0.13066	38290401.51115
x13	0.23528	0.23423	1.00449	0.31514	8322752.57070
x14	0.03366	0.67986	0.04951	0.96052	13611088.99275
x15	0.07798	0.04117	1.89404	0.05822	29156398.15652
x16	−12.75864	5.59749	−2.27935	0.02265	3283779.72700
x17	30.73161	4.55589	6.74547	0.00000	75068181.85144
x18	0.75111	0.95170	0.78923	0.42997	5472068.63598
x19	−0.25489	0.07901	−3.22595	0.00126	19034257.49003
x20	38.97229	13.26532	2.93791	0.00330	1781368.98877
x21	−29.68779	9.09077	−3.26571	0.00109	38872824.77485
x22	−0.00264	0.00242	−1.08754	0.27680	11089703.02440
x23	−1.31162	0.53306	−2.46057	0.01387	2682771.66834
x24	−3.27413	0.41908	−7.81261	0.00000	42369341.72177
x25	210.51663	47.27086	4.45341	0.00001	1116209.97987
x26	−239.17234	54.89311	−4.35706	0.00001	10660983.87770
x27	148.87143	7.18009	20.73392	0.00000	1279430.21590
x28	5.03860	26.97772	0.18677	0.85184	1607537.69654
x29	0.94378	1.48566	0.63526	0.52526	10404492.43678
x30	0.52762	1.68315	0.31347	0.75392	2632536.35994
x31	−0.17992	0.07942	−2.26533	0.02349	3660680.86585
x32	17.53628	11.91443	1.47185	0.14106	594095.76789
x33	−17.10762	6.57568	−2.60165	0.00928	5950372.03724
x34	−0.03839	0.03072	−1.24947	0.21149	10020134.99163
x35	0.00895	0.05042	0.17750	0.85912	5539051.43212
x36	0.00731	0.00303	2.41190	0.01587	11978237.75990
x37	−0.19724	0.46078	−0.42805	0.66862	2062354.81972
x38	0.43293	0.22304	1.94104	0.05225	16582743.83675
x39	−0.01423	0.05438	−0.26172	0.79354	1391473.66905
x40	−0.00266	0.00450	−0.59075	0.55469	5207045.74275
x41	−1.34733	0.77608	−1.73607	0.08255	903064.00582
x41	−1.34733	0.77608	−1.73607	0.08255	903064.00582
x42	−0.59013	0.35353	−1.66927	0.09506	7198361.06074
x43	−0.00015	0.00014	−1.08721	0.27694	3072253.69548
x44	−0.01365	0.03322	−0.41092	0.68113	1211398.44545
x45	0.04053	0.01844	2.19862	0.02790	11861659.01908
x46	−7.98077	3.16963	−2.51789	0.01181	130785.46221
x47	−7.76768	2.64598	−2.93565	0.00333	1236531.99336
x48	2.98002	1.22358	2.43550	0.01487	7908409.47642
x49	0.00027	0.00023	1.14988	0.25019	1392442.24932
x50	0.00023	0.00051	0.45629	0.64818	1420247.63172
x51	−0.00007	0.00003	−2.10908	0.03494	3611870.20740
x52	0.00550	0.00602	0.91343	0.36101	810912.83663
x53	−0.00040	0.00278	−0.14320	0.88614	6224848.55527
x54	−0.00013	0.00101	−0.13286	0.89430	1201882.00470
x55	−0.00006	0.00008	−0.69011	0.49013	4001557.71049
x56	0.02415	0.01348	1.79184	0.07316	717380.55739
x57	−0.00420	0.00736	−0.57053	0.56832	8319716.37297
x58	0.00000	0.00000	1.74361	0.08123	2709680.39002
x59	−0.00148	0.00084	−1.75680	0.07895	1918324.09830
x60	−0.00130	0.00052	−2.48336	0.01301	25105652.62836
x61	0.25021	0.08408	2.97602	0.00292	185635.05050
x62	0.07984	0.06387	1.24998	0.21131	2340300.50647
x63	−0.19552	0.02839	−6.88711	0.00000	15571207.69559
x64	−0.00241	0.00134	−1.79490	0.07267	495873.41346
x65	0.00031	0.00014	2.20815	0.02723	2673216.15563
x66	−0.09064	0.02710	−3.34438	0.00082	556837.91311
x67	0.00590	0.00977	0.60312	0.54643	2960789.74028
x68	−0.00000	0.00000	−0.36129	0.71788	1864112.01765
x69	0.00479	0.00179	2.66745	0.00764	1381834.06835
x70	0.00209	0.00088	2.36913	0.01783	12305107.92876
x71	−0.86688	0.20423	−4.24462	0.00002	116018.57227
x72	−0.07476	0.12307	−0.60748	0.54353	787789.25934
x73	0.17925	0.05476	3.27352	0.00106	7277135.37557
x74	−0.00000	0.00000	−0.61198	0.54055	614461.77732
x75	0.00007	0.00004	1.65289	0.09835	623215.56397
x76	0.00003	0.00003	1.15905	0.24644	8449366.66816
x77	0.01253	0.00755	1.66014	0.09689	257733.65104
x78	0.00472	0.00611	0.77176	0.44026	2039319.96891
x79	0.01854	0.00256	7.23635	0.00000	8702059.66176
x80	−1.83380	0.53436	−3.43180	0.00060	90419.74105
x81	−0.99794	0.42460	−2.35033	0.01876	563873.46535
x82	1.88975	0.31573	5.98531	0.00000	2615109.14534

Residual standard error: 79.07627 on 308.0 degrees of freedom
Multiple R-squared: 1.00000, Adjusted R-squared: 1.00000
F-statistic: 3438160.0865125037 on 83.0 and 308.0 DF, p-value: 0.0

| CubicX Ridge Regression |

CubicX Lasso:

x81, x82)
SUMMARY

Parameters/Coefficients:

Var	Estimate	Std. Error	t value	Pr(>|t|)	VIF
x0	10661.32315	35067.47293	0.30402	0.76111	NA
x1	20159.02832	6106.76633	3.30110	0.00096	15333646.26662
x2	−685.94020	132.24951	−5.18671	0.00000	26617820.20461
x3	2294.39890	277.45361	8.26949	0.00000	16015443.05964
x4	11.43689	12.44620	0.91891	0.35814	15897459.25911
x5	20087.92375	1806.00499	11.12285	0.00000	3684330.50052
x6	−7831.50756	805.88038	−9.71795	0.00000	2526218.28560
x7	−286.10676	395.70997	−0.72302	0.46967	8868013.40768
x8	0.85227	14.08830	0.06050	0.95176	27378548.79746
x9	−79.10110	24.43823	−3.23678	0.00121	14579930.98385
x10	2.18685	1.08849	2.00907	0.04453	17175413.92445
x11	−649.76126	176.93650	−3.67228	0.00024	2143194.72104
x12	−332.01686	133.92378	−2.47915	0.01317	38290401.51115
x13	0.06178	0.15770	0.39178	0.69522	8322752.57070
x14	0.91245	0.45772	1.99345	0.04621	13611088.99275
x15	0.00328	0.02772	0.11833	0.90581	29156398.15652
x16	10.80793	3.76858	2.86790	0.00413	3283779.72700
x17	14.08679	3.06731	4.59255	0.00000	75068166.83617
x18	−3.06971	0.64074	−4.79085	0.00000	5472025.24625
x19	0.08423	0.05320	1.58341	0.11333	19034116.62765
x20	−50.28228	8.93105	−5.63005	0.00000	1781361.68447
x21	−41.59441	6.12048	−6.79594	0.00000	38872803.30089
x22	−0.00403	0.00163	−2.47108	0.01347	11089699.01021
x23	0.63575	0.35889	1.77145	0.07649	2682766.47208
x24	−0.40195	0.28215	−1.42459	0.15427	42369334.94547
x25	−142.89552	31.82572	−4.48994	0.00001	1116209.18201
x26	−407.06033	36.95750	−11.01428	0.00000	10660983.60010
x27	175.03441	4.83409	36.20834	0.00000	1279430.07160
x28	8.70930	18.16310	0.47951	0.63158	1607537.69654
x29	−0.71089	1.00024	−0.71072	0.47726	10400944.83379
x30	1.82810	1.13320	1.61322	0.10670	2632181.47801
x31	−0.02370	0.05347	−0.44323	0.65760	3660638.80725
x32	4.78941	8.02154	0.59707	0.55046	594089.20720
x33	1.00367	4.42716	0.22671	0.82065	5950370.14645
x34	0.00841	0.02068	0.40677	0.68417	10017872.09655
x35	−0.03859	0.03395	−1.13666	0.25568	5538610.38055
x36	0.00169	0.00204	0.82695	0.40827	11978221.89436
x37	0.11073	0.31023	0.35693	0.72115	2062343.99810
x38	0.00851	0.15017	0.05667	0.95481	16582703.14071
x39	0.09651	0.03661	2.63617	0.00838	1391473.65830
x40	−0.00504	0.00303	−1.66389	0.09613	5206991.16267
x41	0.95736	0.52251	1.83223	0.06692	903060.99443
x42	0.59871	0.23802	2.51540	0.01189	7198321.38982
x43	−0.00003	0.00009	−0.31045	0.75622	3072201.21804
x44	−0.04168	0.02237	−1.86334	0.06241	1211385.36493
x45	−0.01042	0.01241	−0.83923	0.40134	11861506.68776
x46	3.76645	2.13399	1.76497	0.07757	130785.21571
x47	6.24697	1.78144	3.50669	0.00045	1236530.79642
x48	1.23012	0.82379	1.49325	0.13537	7908400.72741
x47	6.24697	1.78144	3.50669	0.00045	1236530.79642
x48	1.23012	0.82379	1.49325	0.13537	7908400.72741
x49	0.00002	0.00016	0.09683	0.92286	1392278.86157
x50	−0.00046	0.00035	−1.31616	0.18812	1420196.92870
x51	−0.00000	0.00002	−0.18299	0.85481	3611870.20451
x52	−0.00669	0.00405	−1.64992	0.09896	810910.53132
x53	0.00082	0.00187	0.43635	0.66258	6224812.10056
x54	0.00072	0.00068	1.06127	0.28857	1201880.91448
x55	0.00005	0.00005	0.98661	0.32383	4001554.21367
x56	0.00280	0.00907	0.30889	0.75740	717380.48517
x57	−0.01151	0.00496	−2.32257	0.02020	8319653.14368
x58	−0.00000	0.00000	−0.57915	0.56249	2709669.94207
x59	0.00102	0.00057	1.80771	0.07065	1918323.45278
x60	−0.00039	0.00035	−1.10069	0.27103	25105512.46587
x61	−0.06611	0.05661	−1.16784	0.24287	185635.05043
x62	−0.13270	0.04300	−3.08587	0.00203	2340296.81367
x63	−0.06526	0.01911	−3.41447	0.00064	15571201.42352
x64	0.00005	0.00090	0.05525	0.95594	495872.24863
x65	−0.00013	0.00009	−1.39147	0.16408	2673202.69213
x66	0.01257	0.01825	0.68868	0.49102	556836.72113
x67	0.03366	0.00658	5.11405	0.00000	2960765.24373
x68	0.00001	0.00000	1.86532	0.06214	1864111.98293
x69	−0.00167	0.00121	−1.38397	0.16637	1381833.54239
x70	−0.00063	0.00059	−1.06137	0.28852	12305004.52813
x71	0.11271	0.13750	0.81974	0.41236	116018.55161
x72	0.56737	0.08286	6.84731	0.00000	787783.36946
x73	0.17482	0.03687	4.74215	0.00000	7277134.11151
x74	−0.00000	0.00000	−0.01012	0.99192	614461.77732
x75	0.00003	0.00003	0.89528	0.37064	623215.50887
x76	0.00005	0.00002	2.30435	0.02120	8449355.19121
x77	0.00185	0.00508	0.36457	0.71544	257733.65103
x78	−0.00827	0.00411	−2.01098	0.04433	2039310.61399
x79	0.00297	0.00172	1.72067	0.08531	8702053.97005
x80	0.07208	0.35976	0.20036	0.84120	90419.67886
x81	1.50791	0.28587	5.27490	0.00000	563873.00726
x82	2.08210	0.21257	9.79491	0.00000	2615108.81582

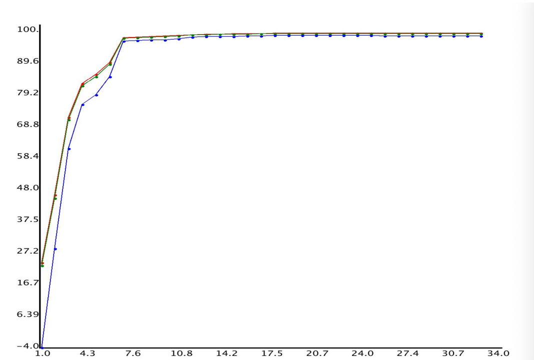
Residual standard error: 53.23913 on 309.0 degrees of freedom
Multiple R-squared: 1.00000, Adjusted R-squared: 1.00000
F-statistic: 7677521.963732964 on 82.0 and 309.0 DF, p-value: 0.0

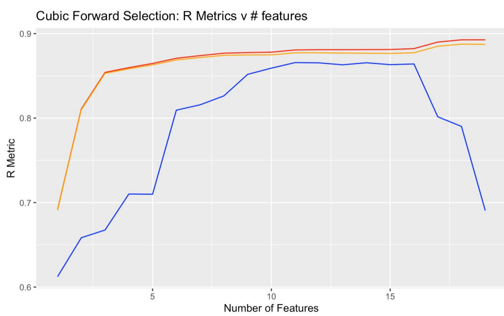
| CubicX Lasso Regression |

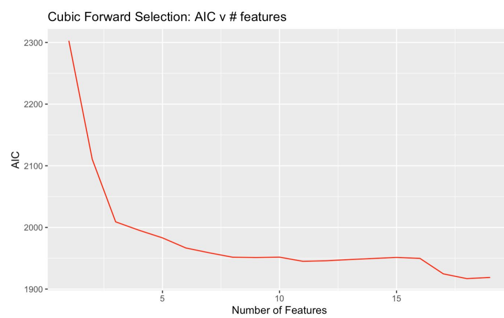
From the cubic and cubicXR lasso and ridge regressions, we see again parameter shrinkage. Additionally, although the R squared is extremely high and indicates a great model fit in the cubic models, there are so many parameters that our analysis becomes a little too complex to make meaningful inferences about the best parameters, and it is a harder model to interpret; these selection methods prevent overfitting and simplify the model. Although the R squared is high, we also see many parameters that are essentially zero, and we can recognize that it might be better to use a smaller model instead.

Dataset 2: Concrete

Cubic Forward:

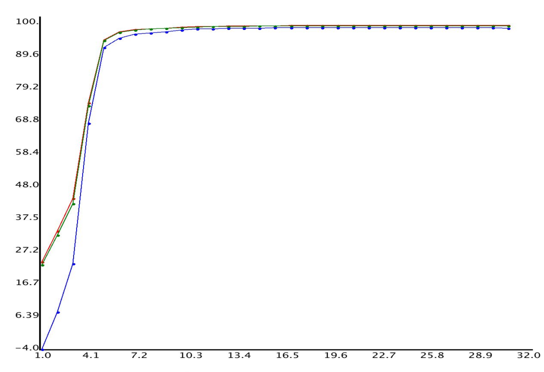


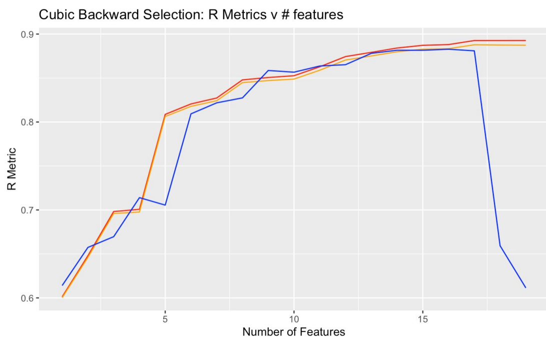


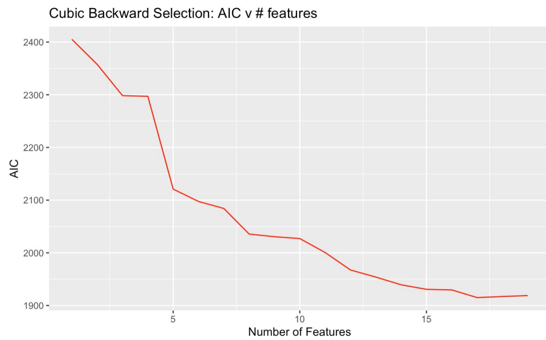


From forward selection, we see as expected that the R squared values approach 1 as more variables are added. However, it is especially notable in the second plot that although the R squared values are increasing, the R squared cv suggests that overfitting happens after around 15 variables. This is something we have seen consistently, especially dealing with cubic models that have so many variables.

Cubic Backward:





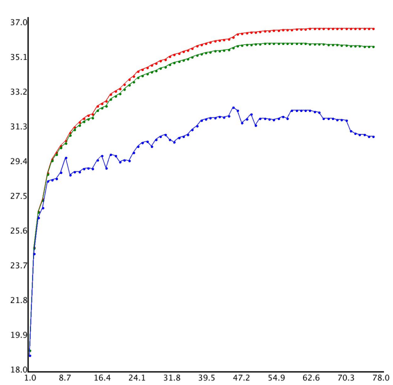


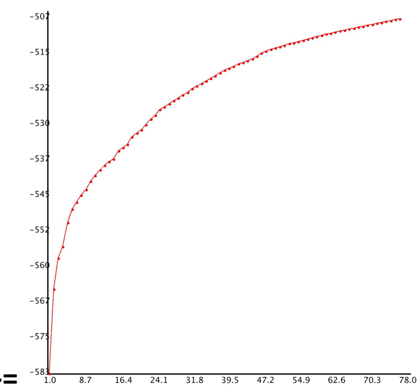
We included the output from forward and backward to comment on the consistency in the output for both selection processes in both R and Scalation; while concurrence between the forward and backward processes is not guaranteed, in this case, both methods suggest that after about 15 parameters, the model does not really improve. Additionally, as apparent in the forward selection output as well, there is a steep drop in the cross validated R^2 values, because as the model becomes overfit with unnecessary parameters, it becomes more ineffective with the testing data (as opposed to the training data).

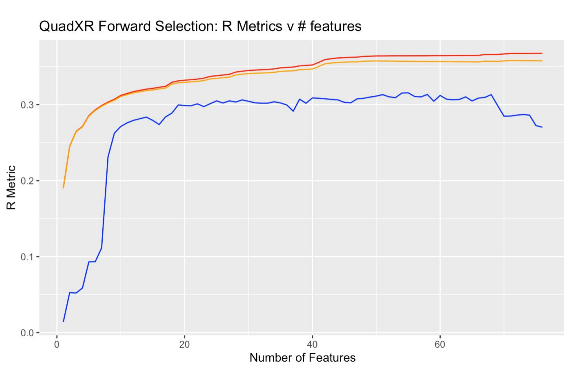
Dataset 3: Wine Quality

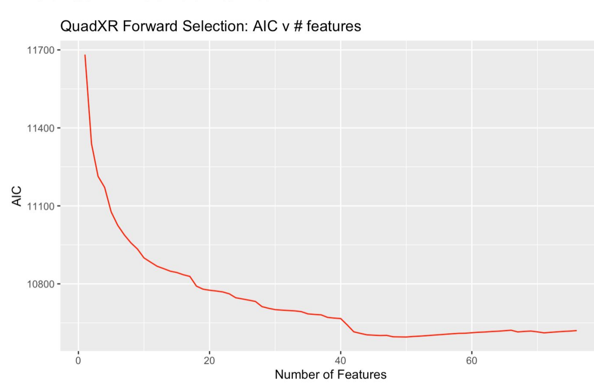
For this dataset, we are more closely analyzing the more complex QuadXR model, which incorporates all variables, their quadratics terms, and their interaction terms.

QuadXR Forward:

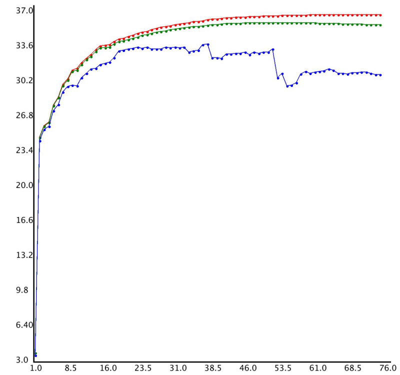


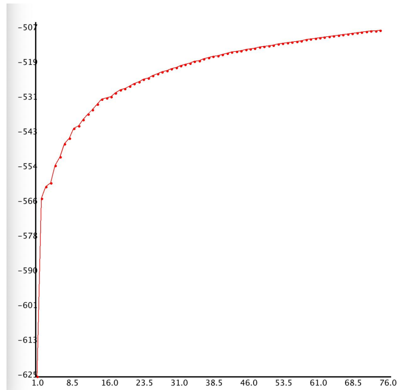




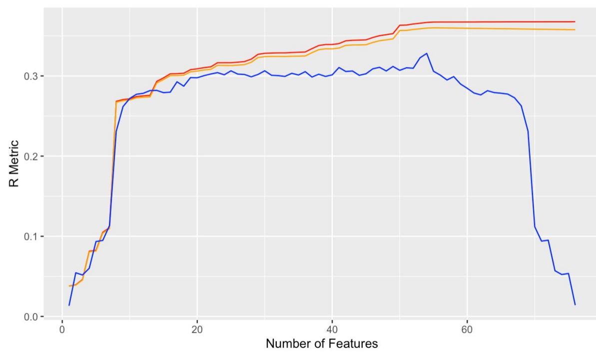


QuadXR Backward:

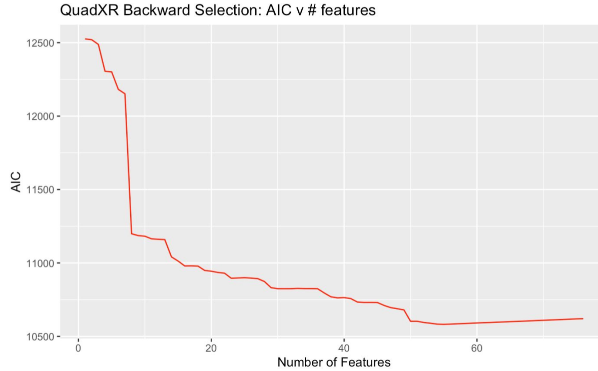




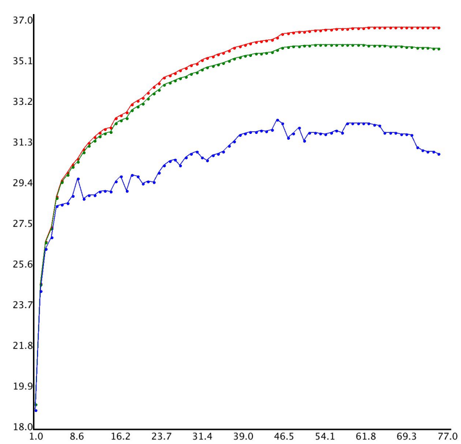
QuadXR Backward Selection: R Metrics v # features

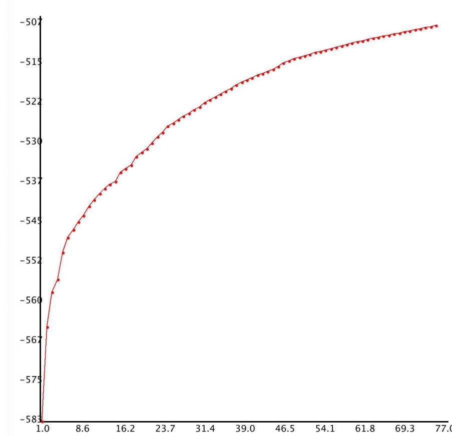


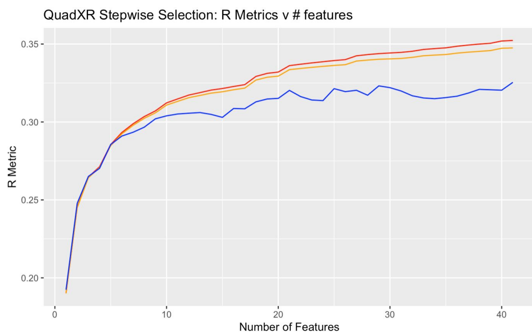
QuadXR Backward Selection: AIC v # features

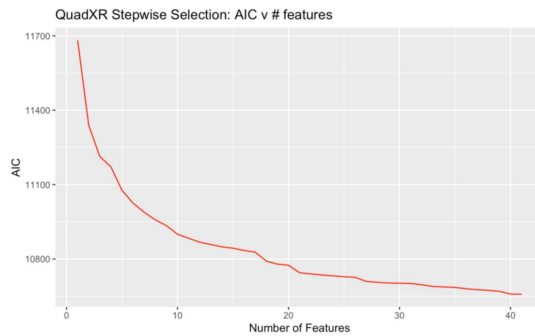


QuadXR Stepwise:







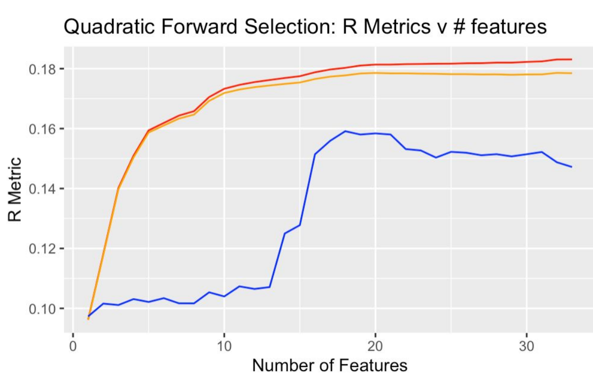


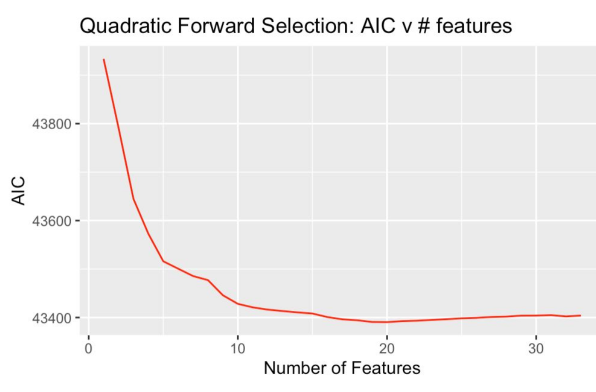
For the QuadXR model, the relationship between the metrics is consistent across selection methods. The drop off in the effectiveness of the model for R^2 cross validated is also significant and adjusted R^2 begins to decrease (even though R^2 continues to increase). Based on these metrics, forward selection and backwards elimination suggest about 40 variables as optimal, while stepwise suggests closer to 50. Considering the total number of features in the full model, the output of all three of these methods is generally consistent for this complex model.

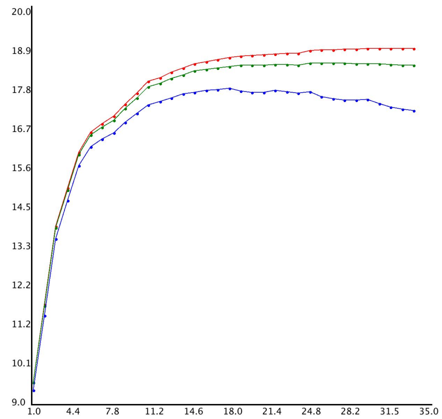
Dataset 4: Parkinsons

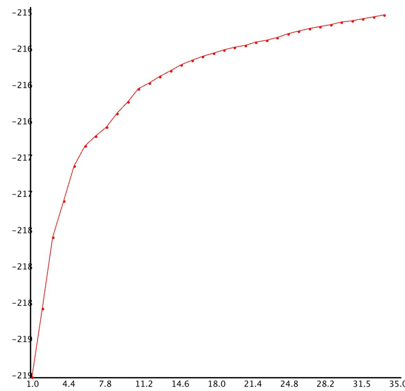
For this data, we used the quadratic model. Forward/backwards and stepwise selection about 35 features are optimal from the plots. The R^2 crossed validation is lowest at about 35 features. Also from the plots we can see that the adjusted R^2 is at a lower point then R^2 when we have 35 features. The AIC for all the selection is at it lowest when there are 35 features. The quadratic model is fairly consistent for all the models.

Quadratic Forward:

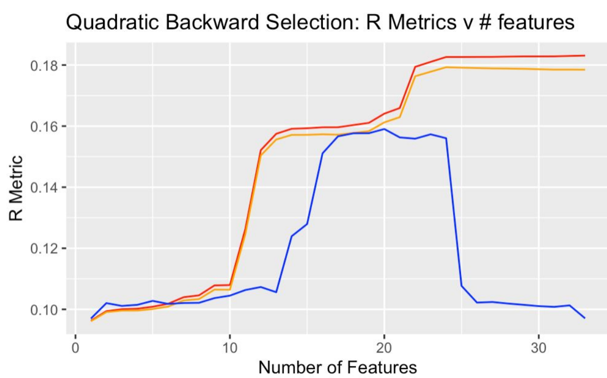


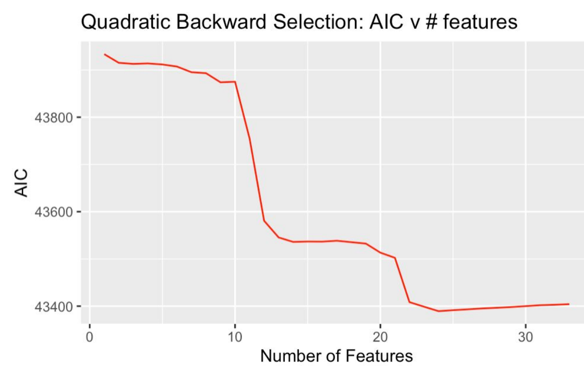


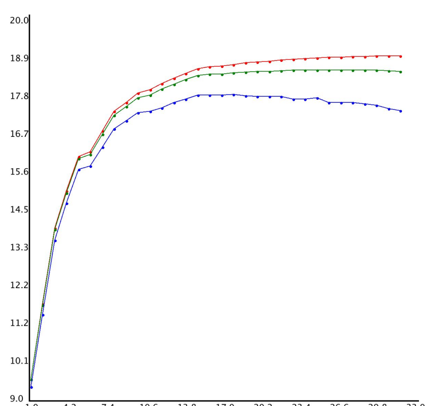


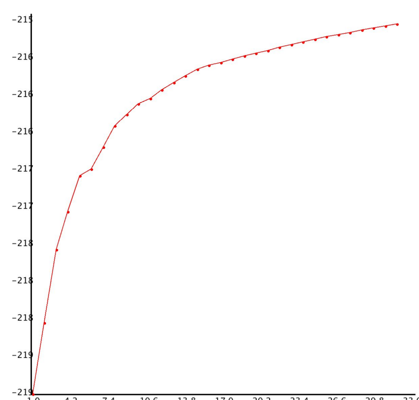


Quadratic Backward:

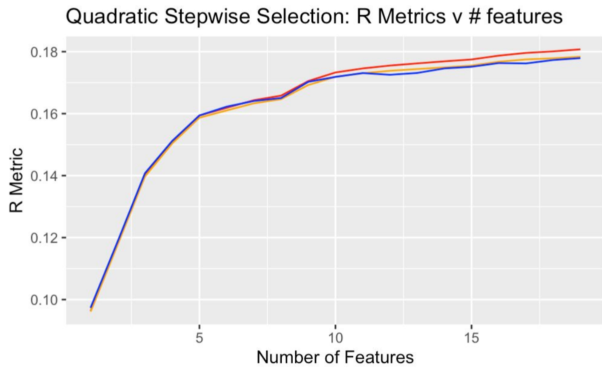


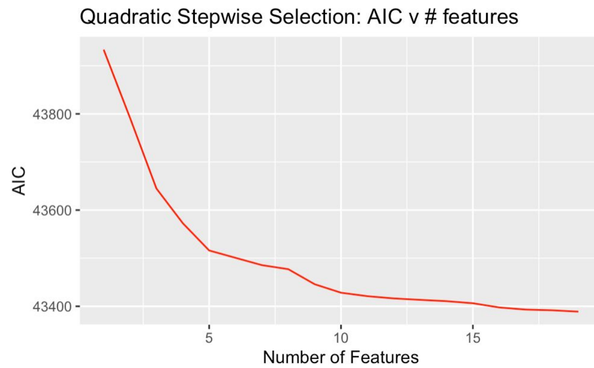


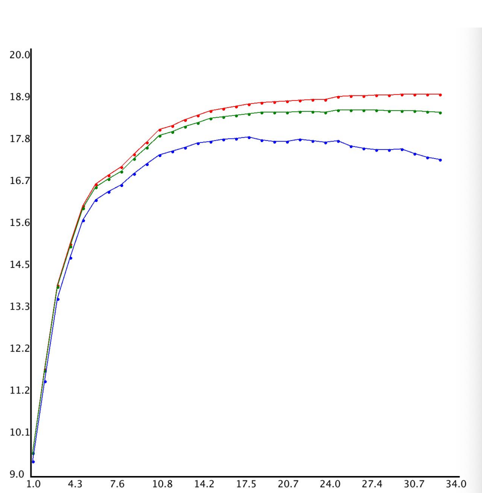


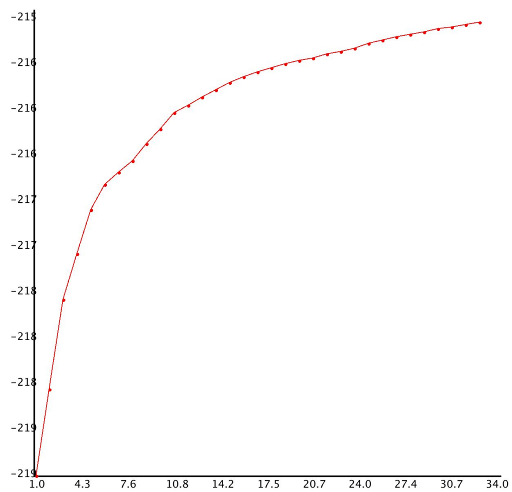


Quadratic Stepwise:





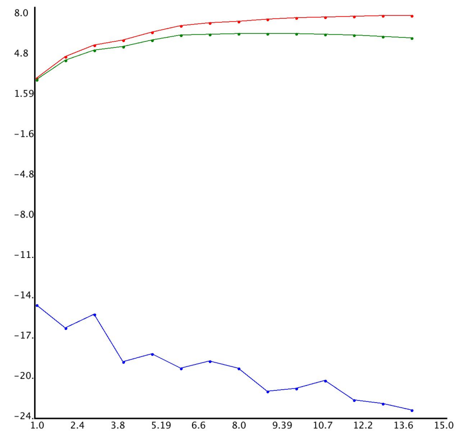


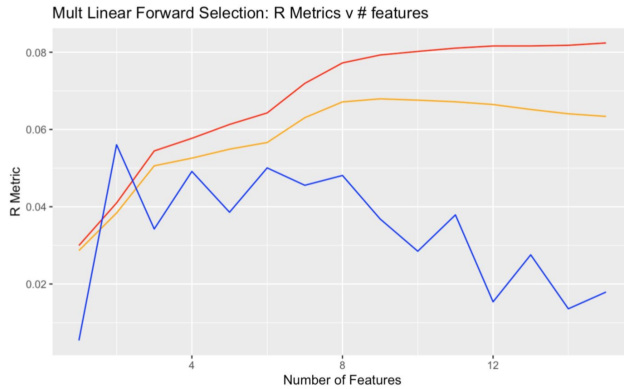


Dataset 5: Absenteeism

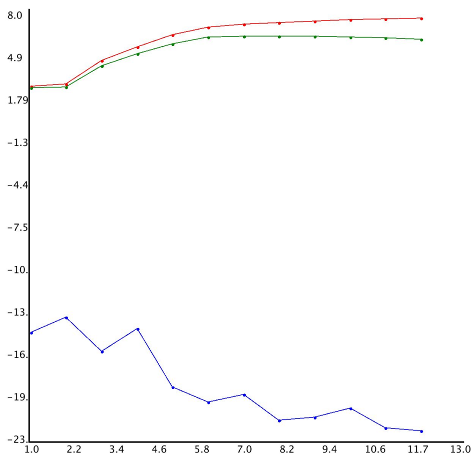
This dataset was noteworthy for the general ineffectiveness of the model in predicting the response. The low R^2 values indicate that even as features are added, the model still performs very poorly. The negative cross validated values computed in scalation suggest that the model is arbitrarily worse than the null model (which would predict the mean of y observations and result in an R^2 value equal to 0).

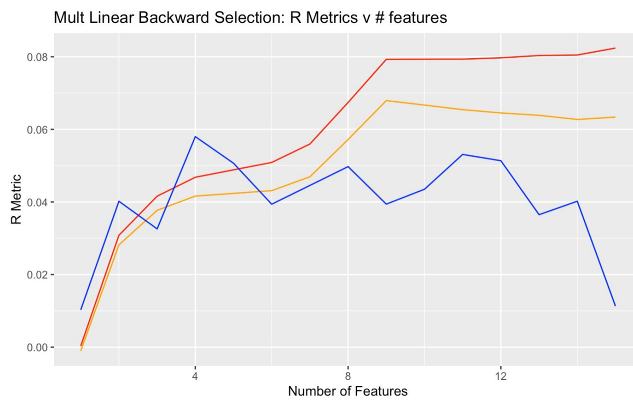
Multiple Linear Forward:



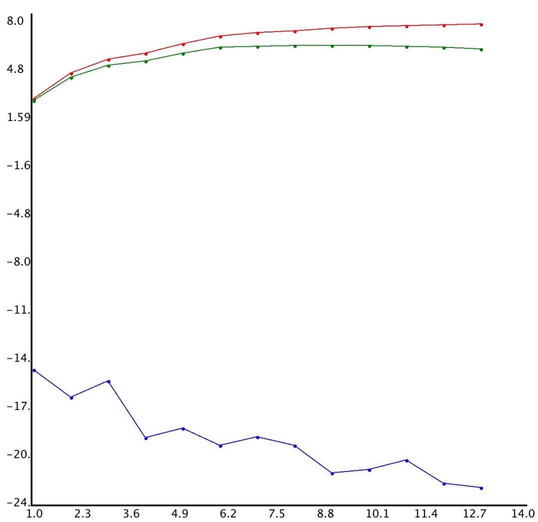


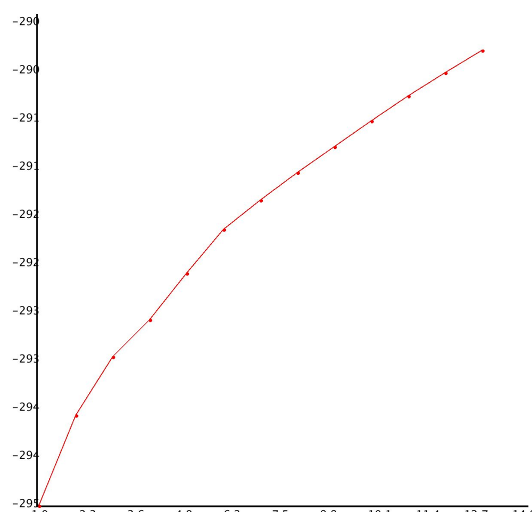
Multiple Linear Backward:





Multiple Linear Stepwise:

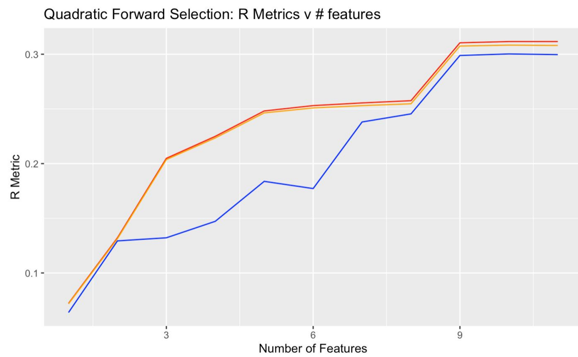


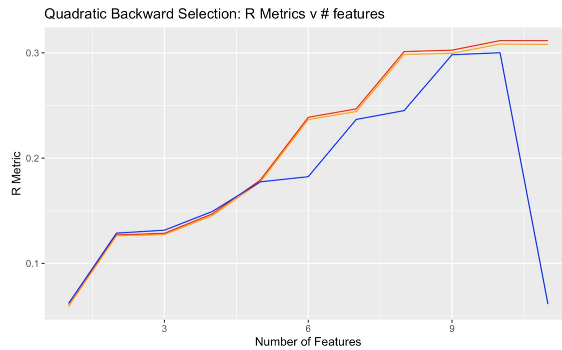


Dataset 6: Obesity

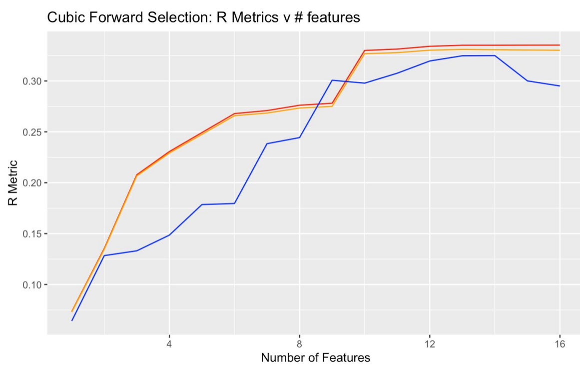
For this dataset, we are comparing the output for quadratic, cubic, and cubicXR. The quadratic model includes the variables and their quadratic terms, the cubic model includes the variables and their quadratic and cubic terms, and the cubicXR also includes interaction terms. As terms are added, the R^2 values slightly increase; however, they stabilize at larger values as the gain from additional terms becomes more incremental.

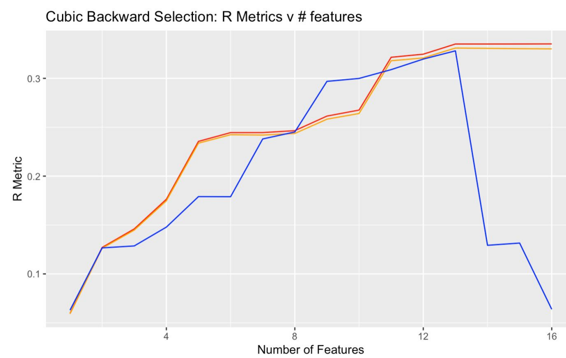
Quadratic:



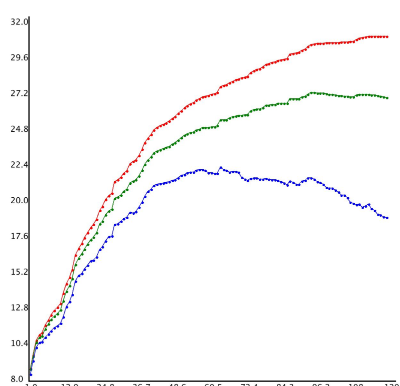


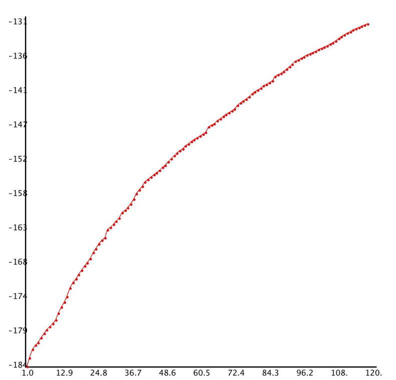
Cubic:



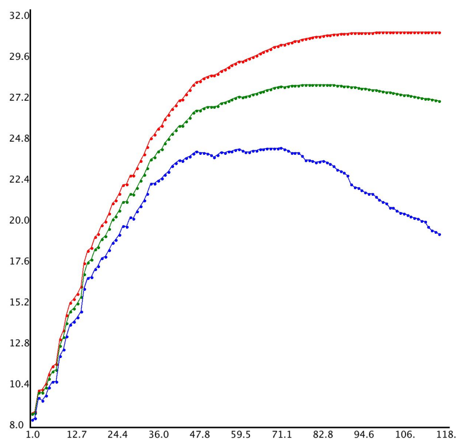


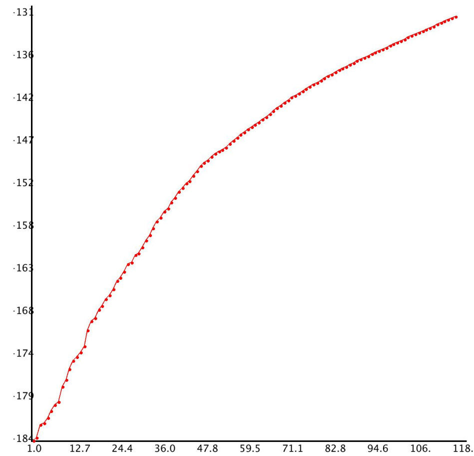
CubicXR Forward:





CubicXR Backward:





CubicXR Step: