Regression Model Building for Military Coups and Politics in Sub-Saharan Africa

Table Of Contents:

Abstract:

In this report I utilize the 'africa' dataset in the faraway library to explore the bias-variance tradeoff. The africa dataset is a subset of a larger study on factors affecting regime stability in a specific region of the African continent. The study consists of a dataframe with 47 observations on 9 variables. My research began by generating an OLS regression followed by three different stepwise regressions and an all-possible-regression to determine the best predictor variables for miltcoup based on R^2, variance, and MSE(bias) to minimize total error. After selecting the best model from these procedures, I perform a k-fold cross validation test to measure the performance of the model on the new test data sets. I also employ the same exact technique on a lesser quality model to display the differences in model building. Based on the bias-variance tradeoff, my results show that the best fit model- with a balance between low MSE and low variance- should use the regressors oligarchy, pollib, parties, popn, pctvote, size.

Introduction:

Using the Africa dataset I wanted to create a model that balanced bias and variance which are inversely related. Since bias is an error due to simplifying real world problems and trying to accommodate this with more flexible methods leads to higher variance, there is a common question as to which metric to forego. More flexible models also tend to accommodate noise and this can lead to higher total error. The goal in variable selection is to build a model with low prediction error while maintaining this tradeoff. Through running multiple regressions I sought to find the best balance between bias and variance. The first model I created is an ordinary least squares with all the variables included. The next step was to conduct 3 stepwise model building techniques; forward selection, backward elimination, and both (bidirectional). The forward selection technique begins with just the intercept and then sequentially adds the variable that most improves the fit. Backward elimination begins with the full least squares model and removes the least useful predictor until a specified threshold is reached. Stepwise regression is a combination of both previous procedures, and it is able to both add and remove variables as it iterates through the process. To validate my results and show reliability I conducted k-fold cross validation. Cross-validation is a resampling procedure used to evaluate the performance of a given model. This was all analyzed through the lens of the Africa dataset, which I outline below.

Materials and Methods:

In the faraway package there is the africa dataset which is the subset of a larger study on factors affecting regime stability in Sub-Saharan Africa. There are 47 observations for each of the 9 variables. Some values are not available, and therefore will be replaced with 0's in my program.

Miltcoup: (Response variable Y) number of successful military coups from independence to 1989

Oligarchy: (X1) number years country ruled by military oligarchy from independence to 1989

Pollib: (X2) Political liberalization - 0 = no civil rights for political expression, 1 = limited civil rights for expression but right to form political parties, 2 = full civil rights

Parties: (X3) Number of legal political parties in 1993

Pctvote: (X4) Percent voting in last election

Popn: (X5) Population in millions in 1989

Size: (X6) Area in 1000 square km

Numelec: (X7) Total number of legislative and presidential elections

Numregim: (X8) Number of regime types

Dataset Overview:

Bias: Distance between the predicted value generated by the model and the real value. With lower algorithm complexity, bias is low which means the distances between predicted and actual values are low but this results in the data being under-fitted and not well represented. This oversimplifies the real model and results in poor predictive accuracy.

Variance: Illustrates the spread of the data which is useful when fitting training data but is harmful when trying to predict data that the model has not been exposed to yet. This metric looks to overfit the noise to incorporate all independent variables by being flexible. With higher complexity, the variance is lowered but this is also problematic because if it is too rigid then data will not be represented correctly

Trade-off: There is correlation between complexity and bias and complexity and variance. High complexity means high bias and low variance, but low complexity means low bias and high variance. Since the two are negatively correlated, this is an investigation into how to find the middle ground between the two metrics.

OLS Regression: Input miltcoup as the dependent variable and everything else as independent variables. This determines the least squares line which minimizes the vertical distance from the data points to the regression line. This type of regression exhibits high bias and low variance compared to nonlinear models.

Forward Selection: This procedure aims to add in independent variables based on if they have the next largest correlation with the dependent variable, miltcoup.

It begins with the simplest model (no independent variables)
$Y = \beta 0 + \epsilon$

Then it starts to consider all single-independent variable models and determine which ones have the largest correlation to Y based on best fit and RMSE
$Y = \beta 0 + \beta 1X1 + \epsilon$
$Y = \beta 0 + \beta 2X2 + \epsilon$
$Y = \beta 0 + \beta 3X3 + \epsilon$
$Y = \beta 0 + \beta 4X4 + \epsilon$
$Y = \beta 0 + \beta 5X5 + \epsilon$
In my study, the most influential predictor variable is oligarchy.

It repeats this process until the remaining predictor variables no longer provide much value. This process was performed in R with the OSLRR package using the ols_step_forward_p

Backwards Elimination: This procedure starts with the full model and eliminates the least useful variable, one at a time.

First it starts with all independent variables in the model
$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \epsilon$

This process is repeated until it is no longer significant to remove variables. This process was performed in R with the OSLRR package using the ols_step_backward_p

Both Directions: This procedure involves techniques from both forward and backward selection. It starts out with the null model and adds in variables to test at each step. It continuously aims to maximize R^2 and checks all t-values for significance.

The starting model is the same as it was in forward selection
$Y = \beta 0 + \epsilon$

Then the program proceeds to add in variables and test for significance/fit
$Y = \beta 0 + \beta 1X1 + \epsilon$
$Y = \beta 0 + \beta 2X2 + \epsilon$
$Y = \beta 0 + \beta 3X3 + \epsilon$
$Y = \beta 0 + \beta 4X4 + \epsilon$
$Y = \beta 0 + \beta 5X5 + \epsilon$

Then test the t-values for each coefficient in the model and continue adding and subtracting until all significant independent variables have been included. This process was performed in R with the OSLRR package using the ols_step_both_p function

All Possible Regressions: This regression presents all possible combinations of the model that can exist. There are 8 independent variables in my dataset therefore there are 2^8 possible subsets to be tested. This technique is extremely powerful and will always contain the best possible models, but it is also not always feasible given a large number of predictors. ols_step_all_possible was used to perform this function.

Determining Best Subset: As part of my project I wanted to discuss the importance of the bias-variance tradeoff. Viewing all possible regressions is perhaps one of the best ways to discuss this topic. A common metric of assessing the utility of models is R^2, but

even this metric is susceptible to artificial inflation as a result of including extraneous variables. Therefore, since the goal of balancing variance and bias is to get the lowest possible total error I chose RMSE and Mallow's CP as 2 metrics I would like to carefully consider.

RMSE: The standard deviation of the residuals. As already mentioned, this is the definition of bias- the distance between the predicted values and the true values. Therefore it is logical to assume that the RMSE decreases with an increase in variables because complexity is increasing. However, this will be accompanied by an increase in variance which is undesired.

In a simple manner the RMSE can be found with

Squaring the points of the residuals.
Finding the average of the residuals.
Taking the square root of the result.

Mallow's C(p): This calculation determines the fit of a model based on OLS by stating how much of the error is left unexplained by the partial model. This is useful when viewed with all possible subsets because you can see where C(p) is the lowest while maintaining a high R2 and a low RMSE.

$C(p) = RSS(p) /\sigma 2 + 2p - N$

R^2: Measures the proportion of variation in Y that is explained by the predictor variables. This value is how well the regression line "fits" the data and is one of many measures in deciding which model is ideal. An R^2 value of .65 means that 65% of the variability of the response variable is explained by all the X variables. The goal is to maximize R^2 while limiting the increase in error.
The Formula for R-Squared Is
R2= Explained Variation / Total Variation

Model Validation:

1.K-fold cross validation

Cross-validation refers to a technique for measuring the performance of a given predictive model on new test data sets.

The k-fold cross-validation method evaluates the model performance on different subsets of the training data and then calculates the average prediction error rate. The algorithm is as follow:

1. Split the data into K subsets of equal size
2. For each fold estimate a model on all the subsets except one
3. Use the subset fold to test the model by calculating a CV metric of choice
4. Average the CV metric across subsets to get the CV error

The obvious advantage is that it uses all the data to estimate the model. However, finding the appropriate K value is not always simple. A common starting value for K is either 5 or 10. As k increases, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias becomes smaller. Although smaller k values tend to have more bias, the choice of 5 in this dataset works well due to the large sample size.

2.K-fold repeated cross validation

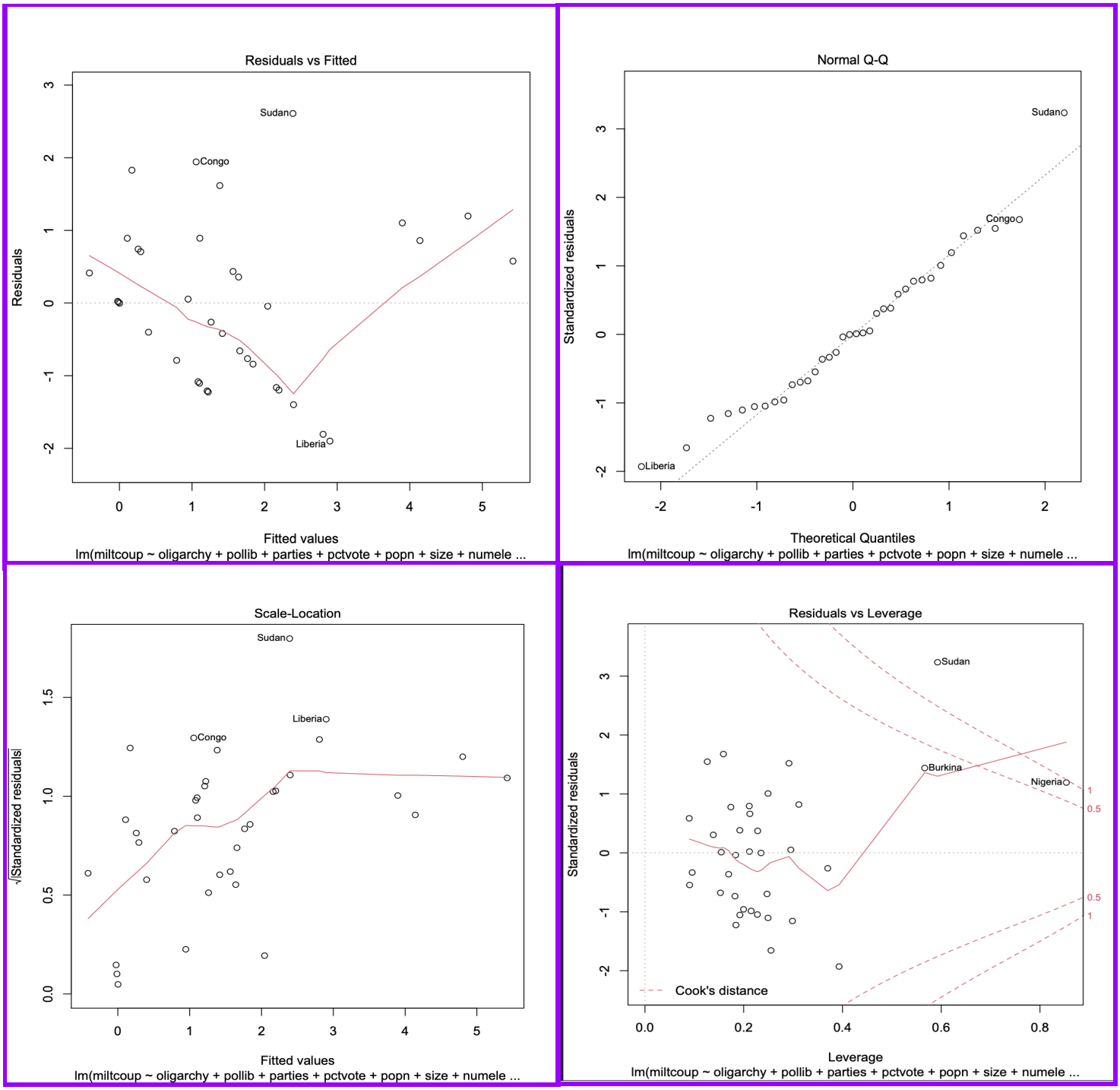
Repeat the process of splitting the data into K-folds multiple times. The mean error from the repeated process gives you the final model error. In this report I use 10-fold cross validation with 6 repeats for the data.

Results:

OLS Regression: The ordinary least squares line with the full model. Certain variables are clearly very significant.

```
> lmod = lm(miltcoup ~ oligarchy + pollib + parties + pctvote + popn + size + numelec + numregim, africa)
> summary(lmod)

Call:
lm(formula = miltcoup ~ oligarchy + pollib + parties + pctvote +
    popn + size + numelec + numregim, data = africa)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89967 -0.90106  0.00466  0.77182  2.60931

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8360534  1.0887267   1.686  0.10324
oligarchy    0.1325624  0.0578618   2.291  0.02999 *
pollib      -1.2275514  0.4191202  -2.929  0.00684 **
parties      0.0627987  0.0234284   2.680  0.01238 *
pctvote      0.0205813  0.0144650   1.423  0.16624
popn         0.0240960  0.0137765   1.749  0.09164 .
size        -0.0005773  0.0004559  -1.266  0.21631
numelec     -0.0502585  0.0878190  -0.572  0.57186
numregim    -0.1193285  0.3054271  -0.391  0.69909
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.264 on 27 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.6036,  Adjusted R-squared:  0.4861
F-statistic: 5.138 on 8 and 27 DF,  p-value: 0.0005869
```

The plot function gives a visual representation of the residuals and normality below.



Forward Selection:

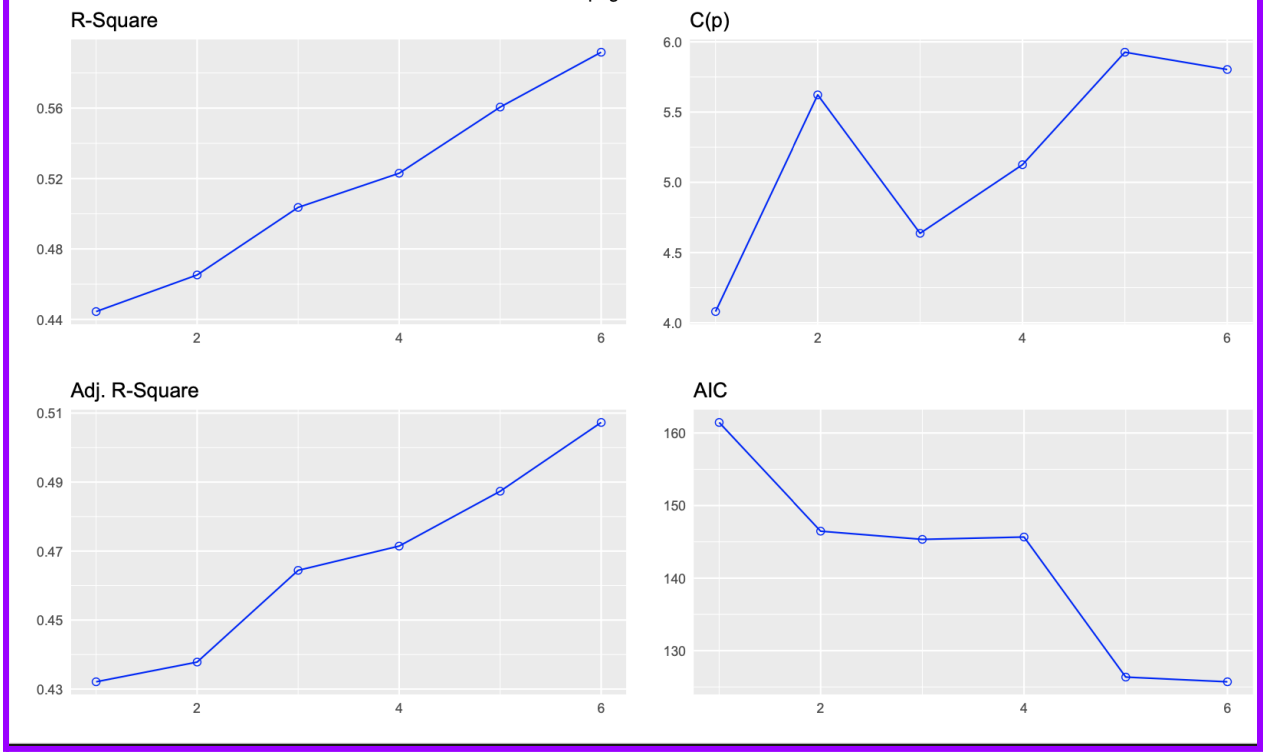
2 Variables not included in the final model: numregim and numelec.
The graphs below show different metrics of model validity.
Low RMSE, higher C(p) value.

```
> ForwardModel = ols_step_forward_p(lmod)
> ols_step_forward_p(lmod)
```

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	oligarchy	0.4444	0.4321	4.0798	161.4563	1.2925
2	pollib	0.4652	0.4378	5.6227	146.4681	1.3054
3	parties	0.5036	0.4644	4.6369	145.3417	1.2742
4	popn	0.5230	0.4715	5.1255	145.6657	1.2658
5	pctvote	0.5606	0.4873	5.9264	126.3598	1.2621
6	size	0.5918	0.5073	5.8033	125.7107	1.2373

page 1 of 2

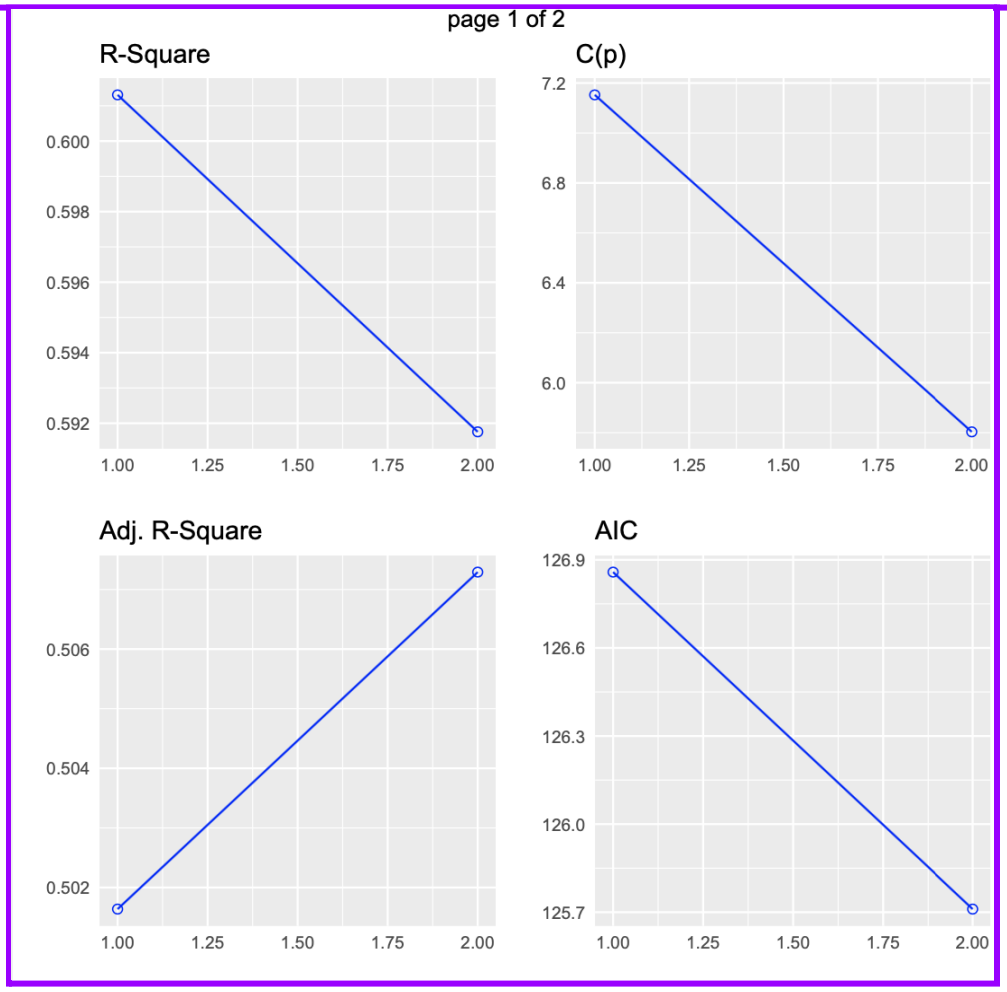


Backward Elimination:

2 Variables removed: numregim and numelec.

This results in the same model as forward selection.

```
> BackwardModel = ols_step_backward_p(lmod)
> BackwardModel


                        Elimination Summary
-------------------------------------------------------------------------
        Variable                    Adj.
Step    Removed     R-Square    R-Square      C(p)       AIC         RMSE
-------------------------------------------------------------------------
   1    numregim      0.6013      0.5016    7.1526    126.8582    1.2444
   2    numelec       0.5918      0.5073    5.8033    125.7107    1.2373
-------------------------------------------------------------------------
```

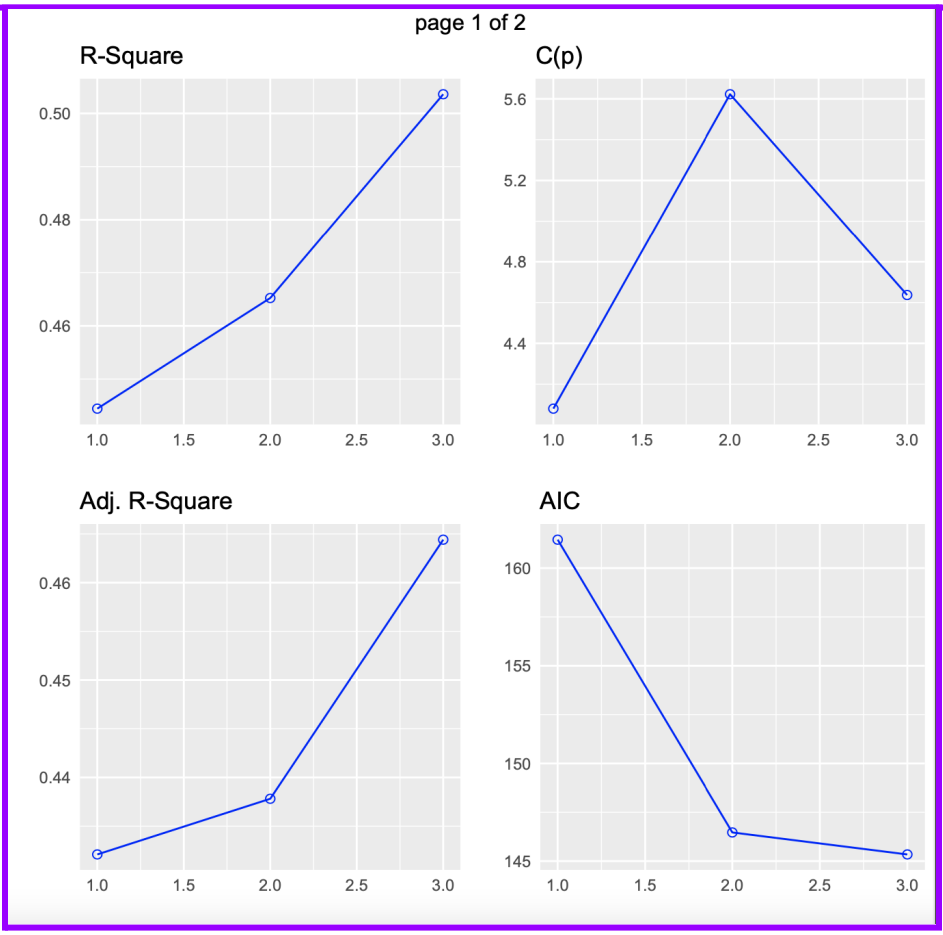


Both Directions:

3 variables added to the model: oligarchy, pollib, and parties.

Noteworthy point: results in a much lower C(p) value than the prior 2 models.

```
> Both = ols_step_both_p(lmod)
> Both

                        Stepwise Selection Summary
--------------------------------------------------------------------------------
                     Added/                   Adj.
Step    Variable     Removed     R-Square    R-Square     C(p)       AIC        RMSE
--------------------------------------------------------------------------------
   1    oligarchy    addition       0.444       0.432    4.0800    161.4563    1.2925
   2     pollib      addition       0.465       0.438    5.6230    146.4681    1.3054
   3     parties     addition       0.504       0.464    4.6370    145.3417    1.2742
--------------------------------------------------------------------------------
```

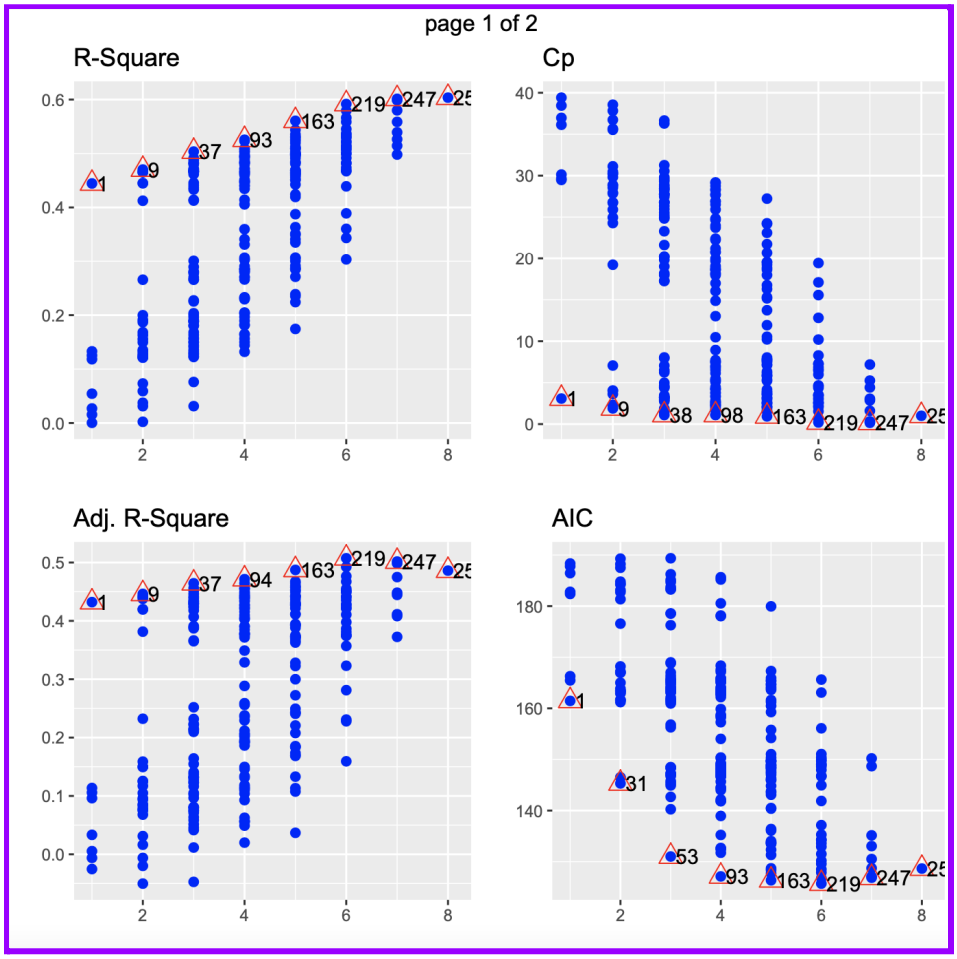


All Possible Regressions: 2^8 possible combinations results in 256 different models.

```
> All_Possible = ols_step_all_possible(lmod)
All
> All_Possible
    Index N                                        Predictors     R-Square Adj. R-Square Mallow's Cp
1       1 1                                        oligarchy 0.4444436324   0.432097935    4.079845
8       2 1                                         numregim 0.1328580617   0.113588241   30.484727
5       3 1                                             popn 0.1252599463   0.105821278   31.128619
2       4 1                                           pollib 0.1182758621   0.096232759   30.628038
3       5 1                                         parties 0.0543733640   0.033359439   37.135803
6       6 1                                             size 0.0270043617   0.005382236   39.455151
7       7 1                                          numelec 0.0157320249  -0.006140597   40.410409
4       8 1                                          pctvote 0.0003696485  -0.025261899   37.969516
14      9 2                                 oligarchy numelec 0.4702892409   0.446211479    3.889595
10     10 2                                 oligarchy parties 0.4663562493   0.442099715    4.222891
12     11 2                                   oligarchy popn 0.4653243838   0.441020947    4.310335
9      12 2                                 oligarchy pollib 0.4652369439   0.437813197    5.622700
15     13 2                               oligarchy numregim 0.4451297745   0.419908401    6.021699
13     14 2                                   oligarchy size 0.4448007463   0.419564417    6.049582
11     15 2                                 oligarchy pctvote 0.4124578488   0.381534578    9.064039
33     16 2                                    popn numregim 0.2656688820   0.232290195   21.229861
18     17 2                                    pollib popn 0.1998658762   0.158833357   26.277568
16     18 2                                  pollib parties 0.1911539742   0.149674691   26.955649
23     19 2                                    parties popn 0.1869339613   0.149976414   27.902141
21     20 2                                 pollib numregim 0.1682856444   0.125633626   28.735581
26     21 2                               parties numregim 0.1628853285   0.124834662   29.940109
35     22 2                                   size numregim 0.1555967727   0.117214808   30.557767
27     23 2                                    pctvote popn 0.1496965708   0.104943759   28.770409
32     24 2                                   popn numelec 0.1354949658   0.096199282   32.261267
36     25 2                                numelec numregim 0.1330244306   0.093616450   32.470629
19     26 2                                    pollib size 0.1286648842   0.083981032   31.819420
31     27 2                                     popn size 0.1253322726   0.085574649   33.122490
20     28 2                                  pollib numelec 0.1237446307   0.078808458   32.202382
22     29 2                                 parties pctvote 0.1222829775   0.076087345   30.826353
30     30 2                               pctvote numregim 0.1215270478   0.075291629   30.883046
17     31 2                                  pollib pctvote 0.1212290048   0.067970157   29.848410
24     32 2                                   parties size 0.0732921160   0.031169030   37.532560
25     33 2                               parties numelec 0.0591271536   0.016360206   38.732950
34     34 2                                   size numelec 0.0375643594  -0.006182715   40.560259
28     35 2                                  pctvote size 0.0312020774  -0.019787287   37.657169
29     36 2                                pctvote numelec 0.0022890323  -0.050222071   39.825568
37     37 3                       oligarchy pollib parties 0.5035984748   0.464408881    4.636872
44     38 3                         oligarchy parties popn 0.4914732701   0.455994661    4.094385
53     39 3                           oligarchy popn numelec 0.4872766454   0.451505249    4.450022
41     40 3                        oligarchy pollib numelec 0.4824842229   0.441627714    6.280277
46     41 3                      oligarchy parties numelec 0.4823586616   0.446244150    4.866789
39     42 3                         oligarchy pollib popn 0.4800501239   0.439001450    6.469732
57     43 3                     oligarchy numelec numregim 0.4719184331   0.435075533    5.751532
55     44 3                        oligarchy size numelec 0.4703373035   0.433384092    5.885522
54     45 3                          oligarchy popn numregim 0.4692796613   0.432252661    5.975151
42     46 3                      oligarchy pollib numregim 0.4672596920   0.425201247    7.465262
52     47 3                           oligarchy popn size 0.4671664790   0.429992047    6.154229
47     48 3                     oligarchy parties numregim 0.4664205058   0.429194030    6.217446
45     49 3                        oligarchy parties size 0.4663816790   0.429152494    6.220736
40     50 3                         oligarchy pollib size 0.4652411334   0.423023328    7.622374
43     51 3                      oligarchy parties pctvote 0.4616056804   0.417952087    7.378087
56     52 3                        oligarchy size numregim 0.4455844578   0.406904304    7.983168
38     53 3                       oligarchy pollib pctvote 0.4411940409   0.388805982   10.057296
48     54 3                         oligarchy pctvote popn 0.4367763829   0.391109603    9.240216
50     55 3                      oligarchy pctvote numelec 0.4332248182   0.387270074    9.506574
49     56 3                         oligarchy pctvote size 0.4139957880   0.366481933   10.948698
51     57 3                     oligarchy pctvote numregim 0.4127331838   0.365116956   11.043390
79     58 3                          parties popn numregim 0.3007264142   0.251939885   20.258959
73     59 3                           parties pctvote popn 0.2894237290   0.231809437   20.291258
85     60 3                         pctvote popn numregim 0.2803487625   0.221998662   20.971856
59     61 3                           pollib parties popn 0.2801932341   0.223366384   22.025376
58     62 3                         pollib parties pctvote 0.2775764144   0.209849203   21.200421
69     63 3                           pollib popn numregim 0.2699311849   0.212294173   22.824111
91     64 3                         popn numelec numregim 0.2659611906   0.214749181   23.205089
90     65 3                              popn size numregim 0.2658232807   0.214601649   23.216776
63     66 3                           pollib pctvote popn 0.2272886660   0.154846978   24.625252
62     67 3                       pollib parties numregim 0.2254269052   0.164276398   26.288053
68     68 3                           pollib popn numelec 0.2035485034   0.140670754   27.990935
67     69 3                             pollib popn size 0.2003781370   0.137250095   28.237697
60     70 3                          pollib parties size 0.1964301352   0.132990409   28.544985
61     71 3                        pollib parties numelec 0.1912177818   0.127366554   28.950683
76     72 3                     parties pctvote numregim 0.1888866932   0.123120749   27.831258
78     73 3                         parties popn numelec 0.1884558924   0.131836536   29.773167
77     74 3                           parties popn size 0.1878363837   0.131173806   29.825667
```

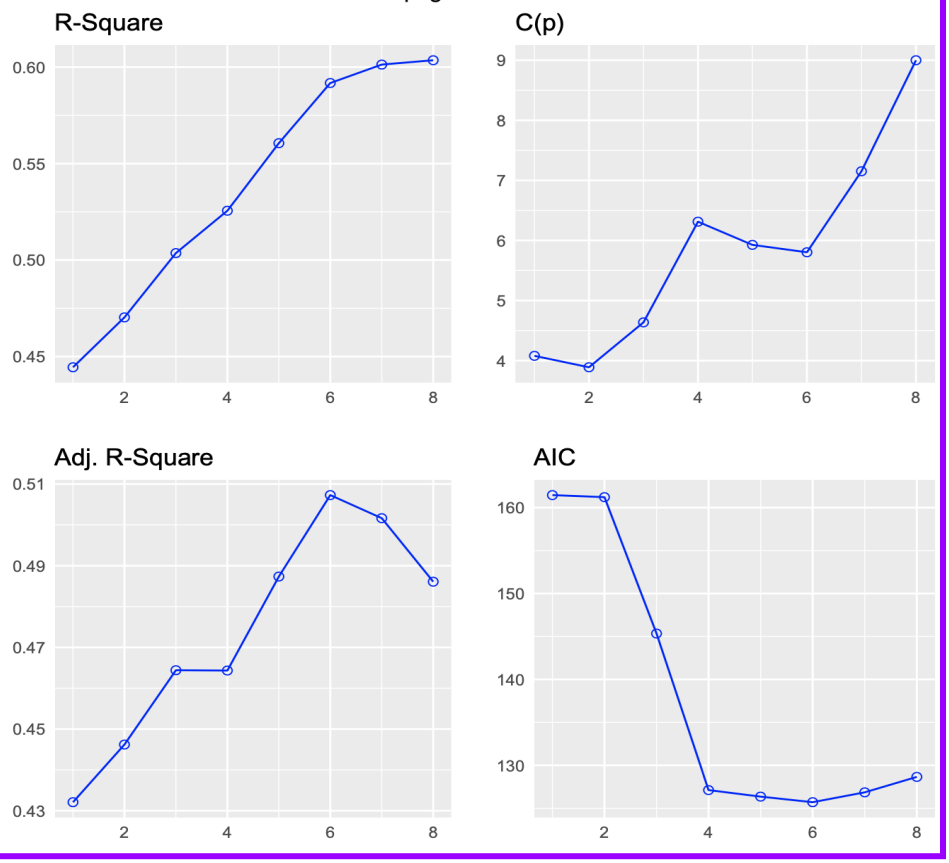
```
166    164 5                  oligarchy pollib parties pctvote numregim 0.5436525008   0.467354421    7.079466
164    165 5                      oligarchy pollib parties pctvote size 0.5354130889   0.457981937    7.640539
165    166 5                    oligarchy pollib parties pctvote numelec 0.5310072439   0.452841785    7.940598
167    167 5                         oligarchy pollib parties popn size 0.5301634515   0.464908375    6.569216
168    168 5                       oligarchy pollib parties popn numelec 0.5263177291   0.460528525    6.868544
169    169 5                     oligarchy pollib parties popn numregim 0.5236395718   0.457478401    7.076995
172    170 5                   oligarchy pollib parties numelec numregim 0.5194151925   0.452667303    7.405795
183    171 5                        oligarchy parties pctvote popn size 0.5125290405   0.442890332    7.558976
170    172 5                      oligarchy pollib parties size numelec 0.5098787826   0.441806391    8.148051
189    173 5                       oligarchy parties popn size numelec 0.5092950591   0.449452993    6.584102
171    174 5                     oligarchy pollib parties size numregim 0.5088549480   0.440640357    8.227740
191    175 5                   oligarchy parties popn numelec numregim 0.5032057509   0.442621086    7.100132
179    176 5                       oligarchy pollib popn size numelec 0.5007588907   0.431419848    8.857888
181    177 5                     oligarchy pollib popn numelec numregim 0.4999557914   0.430505207    8.920396
185    178 5                   oligarchy parties pctvote popn numregim 0.4989935507   0.427421201    8.574100
184    179 5                    oligarchy parties pctvote popn numelec 0.4988747767   0.427285459    8.583008
190    180 5                      oligarchy parties popn size numregim 0.4980098482   0.436791537    7.540451
182    181 5                     oligarchy pollib size numregim 0.4967642466   0.426870392    9.168807
199    182 5                      pollib parties pctvote popn numelec 0.4955682811   0.411496328   10.354157
197    183 5                       oligarchy popn size numelec numregim 0.4913338499   0.429301393    8.106200
192    184 5                   oligarchy parties size numelec numregim 0.4846197343   0.421768482    8.675178
180    185 5                     oligarchy pollib popn size numregim 0.4832118981   0.411435773   10.223639
173    186 5                      oligarchy pollib pctvote popn size 0.4811642054   0.394691573   11.335142
186    187 5                   oligarchy parties pctvote size numelec 0.4694422383   0.393648272   10.790367
193    188 5                    oligarchy pctvote popn size numelec 0.4684000580   0.392457209   10.868528
188    189 5                oligarchy parties pctvote numelec numregim 0.4682826161   0.392322990   10.877336
187    190 5                  oligarchy parties pctvote size numregim 0.4647503458   0.388286110   11.142246
174    191 5                    oligarchy pollib pctvote popn numelec 0.4643865500   0.375117642   12.477778
176    192 5                    oligarchy pollib pctvote size numelec 0.4627842651   0.373248309   12.586901
175    193 5                  oligarchy pollib pctvote popn numregim 0.4609388793   0.371095359   12.712581
178    194 5                oligarchy pollib pctvote numelec numregim 0.4593950737   0.369294253   12.817721
177    195 5                  oligarchy pollib pctvote size numregim 0.4562043125   0.365571698   13.035027
195    196 5                oligarchy pctvote popn numelec numregim 0.4516235534   0.373284061   12.126719
194    197 5                  oligarchy pctvote popn size numregim 0.4510940308   0.372678892   12.166432
196    198 5                oligarchy pctvote size numelec numregim 0.4424957338   0.362852267   12.811281
200    199 5                  pollib parties pctvote popn numregim 0.4241422268   0.328165931   15.218605
198    200 5                    pollib parties pctvote popn size 0.4194144024   0.322650136   15.540592
215    201 5                 parties pctvote popn numelec numregim 0.3875035031   0.300004004   16.935546
214    202 5                   parties pctvote popn size numregim 0.3634726983   0.272540227   18.737790
203    203 5                pollib parties pctvote numelec numregim 0.3514261456   0.243330503   20.170910
201    204 5                  pollib parties pctvote size numelec 0.3492997241   0.240849678   20.315729
206    205 5                  pollib parties popn numelec numregim 0.3409665508   0.249434127   21.295151
205    206 5                   pollib parties popn size numregim 0.3349731081   0.242608262   21.761644
213    207 5                   parties pctvote popn size numelec 0.3068697021   0.207851088   22.982859
217    208 5                  parties popn size numelec numregim 0.3054464571   0.220744806   23.858966
210    209 5                  pollib pctvote popn numelec numregim 0.3004876363   0.183902242   23.640060
209    210 5                   pollib pctvote popn size numregim 0.2915981952   0.173531228   24.245472
202    211 5                  pollib parties pctvote size numregim 0.2873776559   0.168607265   24.532911
218    212 5                  pctvote popn size numelec numregim 0.2860091658   0.184010475   24.547341
204    213 5                    pollib parties popn size numelec 0.2846932717   0.185345115   25.675120
212    214 5                    pollib popn size numelec numregim 0.2711091857   0.169874350   26.732423
207    215 5                  pollib parties size numelec numregim 0.2387800736   0.133055084   29.248724
208    216 5                    pollib pctvote popn size numelec 0.2350135139   0.107515766   28.099154
216    217 5                 parties pctvote size numelec numregim 0.2241568854   0.113322155   29.186091
211    218 5                  pollib pctvote size numelec numregim 0.1744202010   0.036823568   32.225841
219    219 6              oligarchy pollib parties pctvote popn size 0.5917561274   0.507291878    5.803315
220    220 6            oligarchy pollib parties pctvote popn numelec 0.5794536383   0.492444046    6.641172
221    221 6          oligarchy pollib parties pctvote popn numregim 0.5662832409   0.476548739    7.538138
223    222 6          oligarchy pollib parties pctvote size numregim 0.5586113475   0.467289557    8.060629
224    223 6        oligarchy pollib parties pctvote numelec numregim 0.5454117950   0.451359063    8.959581
222    224 6         oligarchy pollib parties pctvote size numelec 0.5373708517   0.441654476    9.507206
225    225 6           oligarchy pollib parties popn size numelec 0.5348397433   0.455097985    8.205242
226    226 6         oligarchy pollib parties popn size numregim 0.5309546555   0.450546882    8.507633
227    227 6       oligarchy pollib parties popn numelec numregim 0.5295562194   0.448908714    8.616479
228    228 6       oligarchy pollib parties size numelec numregim 0.5227667124   0.440955292    9.144933
242    229 6        pollib parties pctvote popn numelec numregim 0.5203708078   0.421137182   10.664989
234    230 6         oligarchy parties pctvote popn size numelec 0.5143955198   0.428700612    9.418995
235    231 6       oligarchy parties pctvote popn size numregim 0.5129486870   0.426998455    9.527504
238    232 6       oligarchy parties popn size numelec numregim 0.5093069918   0.435703041    8.583091
233    233 6        oligarchy pollib popn size numelec numregim 0.5070626866   0.422559147   10.367239
240    234 6          pollib parties pctvote popn size numelec 0.5010569024   0.397827296   11.980356
236    235 6        oligarchy parties pctvote popn numelec numregim 0.4990831548   0.410686065   10.567380
229    236 6          oligarchy pollib pctvote popn size numelec 0.4917804012   0.386631519   12.612129
232    237 6        oligarchy pollib pctvote size numelec numregim 0.4824337310   0.375351055   13.248681
230    238 6          oligarchy pollib pctvote popn size numregim 0.4818849980   0.374688791   13.286052
237    239 6       oligarchy parties pctvote size numelec numregim 0.4737749904   0.380911753   12.465423
239    240 6         oligarchy pctvote popn size numelec numregim 0.4684234718   0.374615849   12.866772
231    241 6        oligarchy pollib pctvote popn numelec numregim 0.4672379180   0.357011280   14.283587
241    242 6          pollib parties pctvote popn size numregim 0.4391137179   0.323068280   16.198976
246    243 6        parties pctvote popn size numelec numregim 0.3890591939   0.281246110   18.818874
243    244 6         pollib parties pctvote size numelec numregim 0.3602891892   0.227935228   21.567295
244    245 6           pollib parties popn size numelec numregim 0.3434387466   0.230885389   23.102730
245    246 6         pollib pctvote popn size numelec numregim 0.3034607232   0.159349149   25.437579
247    247 7      oligarchy pollib parties pctvote popn size numelec 0.6013101465   0.501637683    7.152642
248    248 7    oligarchy pollib parties pctvote popn size numregim 0.5987423166   0.498427896    7.327523
249    249 7  oligarchy pollib parties pctvote popn numelec numregim 0.5800155986   0.475019498    8.602900
250    250 7  oligarchy pollib parties pctvote size numelec numregim 0.5586322786   0.448290348   10.059204
251    251 7    oligarchy pollib parties popn size numelec numregim 0.5393172812   0.444470839    9.856738
254    252 7      pollib parties pctvote popn size numelec numregim 0.5264823329   0.408102916   12.248765
253    253 7    oligarchy parties pctvote popn size numelec numregim 0.5144114377   0.411407803   11.417801
252    254 7    oligarchy pollib pctvote popn size numelec numregim 0.4980550241   0.372568780   14.184798
255    255 8 oligarchy pollib parties pctvote popn size numelec numregim 0.6035514267   0.486085183    9.000000
```

Plots of all possible regression models.



Best Subsets: The best performing models for each number of predictors.
Both forward selection (Model 6) and bidirectional selection (Model 3) are listed below.
Plots are also displayed to see how the different models compare in terms of accuracy and error.

```
> Best_Subset = ols_step_best_subset(lmod)
> Best_Subset
                        Best Subsets Regression
--------------------------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------------------------
     1         oligarchy
     2         oligarchy numelec
     3         oligarchy pollib parties
     4         oligarchy pollib parties pctvote
     5         oligarchy pollib parties pctvote popn
     6         oligarchy pollib parties pctvote popn size
     7         oligarchy pollib parties pctvote popn size numelec
     8         oligarchy pollib parties pctvote popn size numelec numregim
--------------------------------------------------------------------------

                                                Subsets Regression Summary
-----------------------------------------------------------------------------------------------------------------------------------------
                        Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)       AIC        SBIC       SBC        MSEP       FPE       HSP       APC
-----------------------------------------------------------------------------------------------------------------------------------------
  1        0.4444      0.4321      0.3773   4.0798   161.4563   28.0693   167.0067   78.5219   1.7417   0.0380   0.6049
  2        0.4703      0.4462        0.39   3.8896   161.2172   28.1147   168.6178   76.6100   1.7331   0.0379   0.6019
  3        0.5036      0.4644      0.3088   4.6369   145.3417   26.8320   154.0300   68.2860   1.7782   0.0439   0.6009
  4        0.5256      0.4643      0.2602   6.3121   127.1210   26.0727   136.6221   60.1000   1.8956   0.0555   0.6275
  5        0.5606      0.4873      0.0837   5.9264   126.3598   26.5493   137.4444   57.5821   1.8584   0.0549   0.6152
  6        0.5918      0.5073     -0.1085   5.8033   125.7107   27.5006   138.3788   55.4076   1.8286   0.0547   0.6053
  7        0.6013      0.5016     -0.1123   7.1526   126.8582   29.6957   141.1099   56.1150   1.8926   0.0574   0.6265
  8        0.6036      0.4861     -0.1414   9.0000   128.6552   32.2694   144.4904   57.9457   1.9960   0.0614   0.6607
-----------------------------------------------------------------------------------------------------------------------------------------
```

K-fold cross validation for model chosen by forward selection: This function gives the error associated with each model.

```
> print(Forward_Model)
Linear Regression

47 samples
 6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 42, 42, 42, 43, 43, 42, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.442329  0.3663575  1.181811
```

K-fold cross validation for model chosen by both stepwise selection: This function gives the error associated with each model.

```
> print(Both_Model)
Linear Regression

47 samples
 3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 42, 43, 43, 42, 43, 43, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.298264  0.3834142  1.090605
```

K-fold repeated cross validation for model chosen by forward selection: Another look at the error.

```
> print(Forward_Model_Repeat)
Linear Regression

47 samples
 6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 6 times)
Summary of sample sizes: 43, 43, 42, 43, 41, 42, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.465511  0.3459891  1.211749
```

K-fold repeated Cross Validation for Model chosen by both stepwise selection:

```
> print(Both_Model_Repeat)
Linear Regression

47 samples
 3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 6 times)
Summary of sample sizes: 41, 42, 41, 43, 42, 43, ...
Resampling results:

  RMSE      Rsquared   MAE
  1.333669  0.4780234  1.122983
```

Another look at the error.

Discussion:

The ordinary least squares model, using all 8 predictors, resulted in the largest R^2 value of any model- 0.6036. This was expected as it includes more predictor variables than any other model. However, when adjusted for the number of predictors in the model, OLS drastically drops to an adjusted R^2 value of 0.4861. This means that just under 50% of the variation in the Y variable is explained by the 8 predictor variables. Though this seems like a significant value, it does not provide the level of explanation a researcher may want.

Forward selection and backward elimination both led to the creation of the same model. The variables numregim and numelec were left out in both instances, thereby suggesting neither of these variables are very useful in predicting the number of successful military coups from independence to 1989. Though I am not surprised these variables were left out, I am surprised to see that the variable size was useful enough to be left in the model. This model has a C(p) of 5.8033 and an RMSE of 1.2373.

The both directions regression model included just 3 variables: oligarchy, pollib, and parties. This was not expected, especially when there were 8 variables to choose from. However, upon inspecting the C(p) and RMSE values, I understood why this was the case. As a result of there being just 3 variables, this model has a C(p) of 4.6370 and an RMSE of 1.2742. This means that adding just 3 variables to get from the bidirectional model to the forward model results in over 1.15 increase in C(p), while RMSE remains nearly the same.

I decided the 2 models I wanted to compare using cross validation were the forward model (6 variables) and the bidirectional model (3 variables). Both of these models were listed under ‘best subset’, therefore I thought it would be useful to contrast them.
The forward model had a significantly higher C(p) of 5.8033. However the RMSE of the forward model, 1.2373, is slightly below that of the bidirectional model. Additionally, the 6 variable model has the largest adjusted R^2 out of any model, with a value of 0.5073.

Despite the large increase in C(p), I believe the 6 variable forward selection model is the best one to use. It has high predictive power and a reasonable error rate given the number of predictor variables.
Both cross validation techniques would seemingly confirm the previous statement, as it gives similar values for RMSE and MAE with both models that were tested.

Sources:

Adapted from "Data : A Collection of Problems from Many Fields for the Student and Research Worker" by D. Andrews and A. Herzberg published by Springer-Verlag, in 1985

Bratton, Michael, and Nicholas Van De Walle. 1997. “Political Regimes and Regime Transitions in Africa, 1910-1994.” *Study Number I06996.* Ann Arbor: Inter-University Consortium for Political and Social Research.

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2017.

Linear Models with R, Julian J. Faraway, (Taylor, 2nd, 2014)

CrossValidationOverview.pdf, Lecture 10 (February 24 2021)

R Code:

 # Attaching the dataset and building OLS model

```
> library(faraway)
> data(africa)
> str(africa)

> lmod = lm(miltcoup ~ oligarchy + pollib + parties + pctvote + popn + size + numelec +
numregim, africa)
> summary(lmod)

> library(MASS)
> library(olsrr)
```

 # The next few paragraphs are for different types of model building

```
> ForwardModel = ols_step_forward_p(lmod)
> ols_step_forward_p(lmod)
> plot(ForwardModel)

> BackwardModel = ols_step_backward_p(lmod)
> BackwardModel
> plot(BackwardModel)

> Both = ols_step_both_p(lmod)
> Both
> plot(Both)

> All_Possible = ols_step_all_possible(lmod)
> All_Possible
> plot(All_Possible)

> Best_Subset = ols_step_best_subset(lmod)
> Best_Subset
> plot(Best_Subset)
```

 # Cross Validation begins here

```
> library(caret)
> set.seed(1234)
```

```
> train.control = trainControl(method = "cv", number = 10)
> library(imputeTS)
> africa2 = na.replace(africa, 0)

> Forward_Model = train(miltcoup ~ oligarchy + pollib + parties + popn + pctvote + size,
africa2, method = "lm", trControl = train.control)
> print(Forward_Model)


> Both_Model = train(miltcoup ~ oligarchy + pollib + parties, africa2, method = "lm",
trControl = train.control)
> print(Both_Model)

> summary(Forward_Model)
> summary(Both_Model)
```

 # Repeated Cross Validation

```
> set.seed(1543)
> train.control = trainControl(method = "repeatedcv", number = 10, repeats = 6)
> Forward_Model_Repeat = train(miltcoup ~ oligarchy + pollib + parties + popn +
pctvote + size, africa2, method = "lm", trControl = train.control)
> print(Forward_Model_Repeat)

> Both_Model_Repeat = train(miltcoup ~ oligarchy + pollib + parties, africa2, method =
"lm", trControl = train.control)
> print(Both_Model_Repeat)
```
