Multiple Linear Regression Variable Selection Methods

Module 5

DATA 5600
Introduction to Regression and Machine Learning for Analytics
Marc Dotson

Module Overview

- Methods to select a subset of predictor variables
  - Best subsets
  - Forward selection
  - Backward selection
  - Stepwise/sequential replacement

- Shrinkage methods
  - Ridge regression
  - LASSO regression
  - Elastic net regression

Environmental Impact Data Set

Which variable is the response?

- To what extent do environmental conditions affect human mortality?

We only have enough observations to support including ~6 variables in the model

- $n = 60$

Different cities in the US

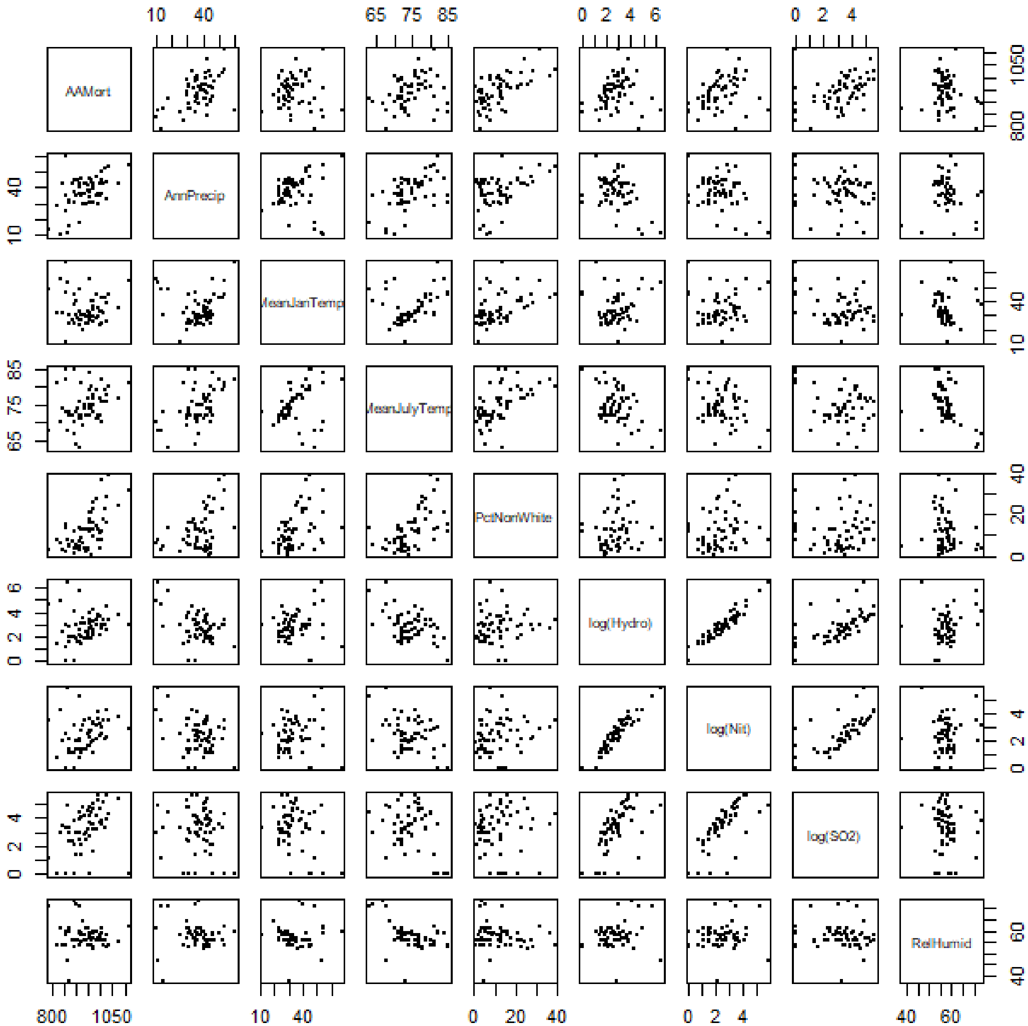
- Do you anticipate problems with the multicollinearity assumption?

These three are what the researchers are interested in

Variable	Description
AnnPrecip	Mean annual precipitation
MeanJanTemp	Average January temperature (in degrees Fahrenheit) hypothermia
MeanJulyTemp	Average July temperature (in degrees Fahrenheit) hyperthermia
PctGT65	Percent of population greater than 65 years old
PopPerHouse	Population per household
School	Median school years completed
PctSound	Percent of housing units that are "sound"
PopPerSqMile	Population per square mile
PctNonWhite	Percent of population that is nonwhite
PctWhiteCollar	Percent of employment in white-collar jobs
PctU20000	Percent of families with income under $20,000
log(Hydrocarbons)	Relative pollution potential of hydrocarbons
log(Nitrogen)	Relative pollution potential of oxides in nitrogen
log(SO2)	Relative pollution potential of oxides in sulfur dioxide
RelHumid	Annual average relative humidity
AAMort	Age-adjusted mortality

Environmental Impact Data Set

- Is the "no multicollinearity" assumption met?

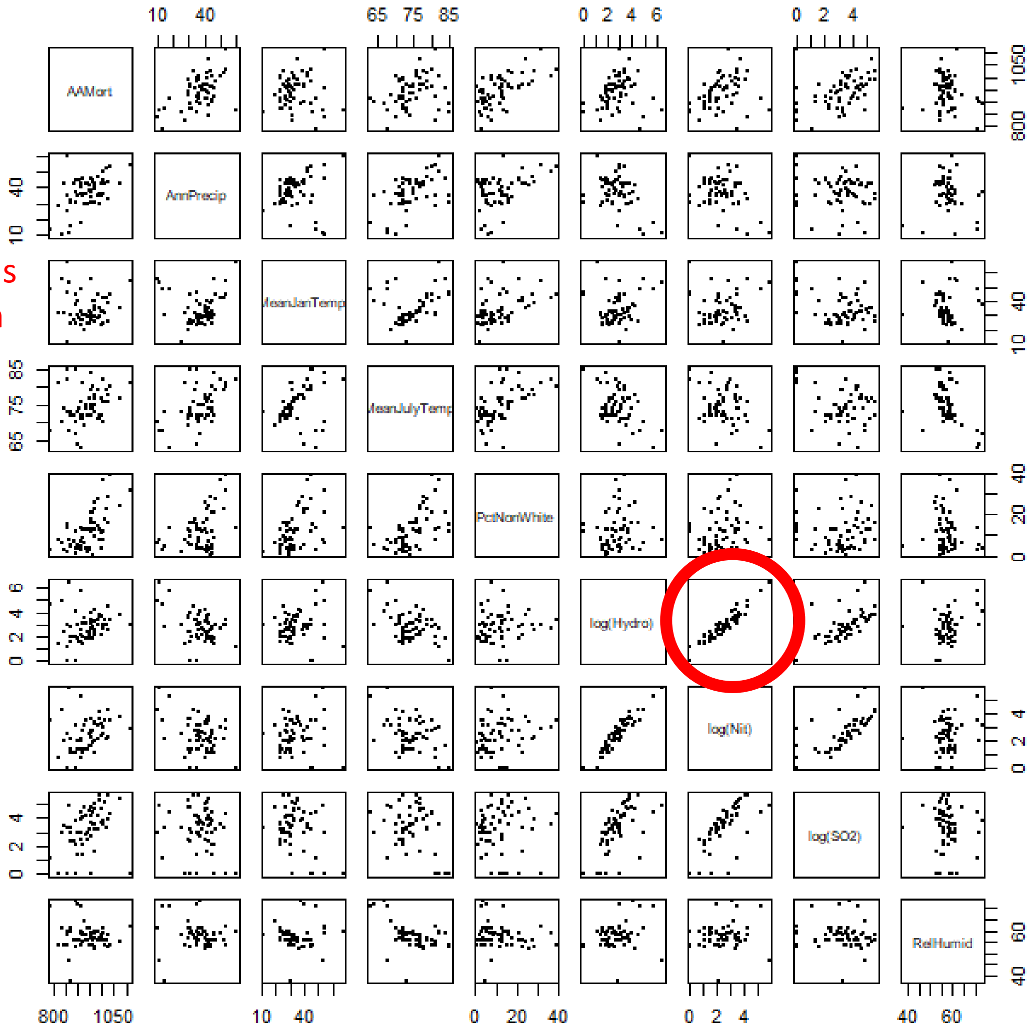


Variable	VIF
AnnPrecip	4.1
MeanJanTemp	6.3
MeanJulyTemp	5.4
PctGT65	7.3
PopPerHouse	4.5
School	4.3
PctSound	3.8
PopPerSqMile	1.8
PctNonWhite	7.3
PctWhiteCollar	2.6
PctU20000	8.4
log(Hydrocarbons)	18.1
log(Nitrogen)	18.1
log(SO2)	6.2
RelHumid	1.7

Environmental Impact Data Set

- Is the "no multicollinearity" assumption met?

Do not just throw Hydrocarbons and Nitrogen out of the model!



Variable	VIF
AnnPrecip	4.1
MeanJanTemp	6.3
MeanJulyTemp	5.4
PctGT65	7.3
PopPerHouse	4.5
School	4.3
PctSound	3.8
PopPerSqMile	1.8
PctNonWhite	7.3
PctWhiteCollar	2.6
PctU20000	8.4
log(Hydrocarbons)	18.1
log(Nitrogen)	18.1
log(SO2)	6.2
RelHumid	1.7

Recall from Module 4:

How we can deal with multicollinearity:

1. Remove some variables from the model – choose a subset of predictor variables

2. Apply a shrinkage method (ridge regression, lasso, elastic net, etc.)

3. Combine the correlated variables somehow, like with principal component regression

Ultimately, we want model *parsimony*: we want the simplest model that will get the job done

Remove Variables From Model

- Given 3 predictor variables, how do we choose which of these models is the "best?"

$$\hat{y}_i = \hat{\beta}_0$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i2}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i3}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i3}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i2} + \hat{\beta}_2 x_{i3}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$

- **It is best to eliminate variables based on contextual grounds**
- There are automatic statistical procedures that can help, but there is no guarantee that you get a "right" subset.
- If we use automatic statistical procedures, <u>we will need **metrics** to summarize how "good" a model is so we can directly compare subsetted models</u>.

Metrics to Summarize Model Performance

Metrics to Summarize Model Performance

Suggestion: AIC for prediction (keeps more variables than BIC), BIC for Inference (want something actionable)

- R^2: but which model will always have the highest R^2?

- **Adjusted $\boldsymbol{R^2}$:** only increases if covariates improve the model, higher value = better model

When n is small, high probability that AIC will select too many parameters (overfit).

- **AIC:** "Akaike's Information Criteria", penalizes the inclusion of additional variables, lower value = better model

AICc is AIC with extra penalty term for the number of parameters. A $n \to \infty$, the extra penalty term converges to 0, and thus AICc converges to AIC

$$\text{AIC} = n \times \log\left(\sum (y_i - \hat{y}_i)^2/n\right) + 2p$$

Penalties for including more covariates

- **BIC:** "Bayesian Information Criteria", variant of AIC with an even stronger penalty for the inclusion of additional variables, lower value = better model

BIC has larger penalty than AIC (log(n)>2) when n>8ish

$$\text{BIC} = n \times \log\left(\sum (y_i - \hat{y}_i)^2/n\right) + p \times \log(\text{n})$$

AIC vs BIC Example (Supervisor Data Set from Module 4)

- Suppose we only have the Learn and Raises variables to predict Rating

```
sm.OLS(super['Rating’],
sm.add_constant(super[['Learn’,
                       'Raises']]))
```

$$\boxed{\begin{aligned} AIC &= 130.96 + 6 \\ &= 136.96 \end{aligned}}$$

$$\begin{aligned} BIC &= 130.96 + 10.20 \\ &= 141.16 \end{aligned}$$

- Within each column, BIC is always larger
- Across AICs (and across the BICs), both parts of the computation differ because the model changes
- Example of choosing different models based on AIC/BIC

```
sm.OLS(super['Rating’],
sm.add_constant(super[['Learn']]))
```

$$\begin{aligned} AIC &= 134.16 + 4 \\ &= 138.16 \end{aligned}$$

$$\boxed{\begin{aligned} BIC &= 134.16 + 6.80 \\ &= 140.96 \end{aligned}}$$

```
sm.OLS(super['Rating’],
sm.add_constant(super[[‘Raises']]))
```

$$\begin{aligned} AIC &= 136.09 + 4 \\ &= 140.09 \end{aligned}$$

$$\begin{aligned} BIC &= 136.09 + 6.80 \\ &= 142.89 \end{aligned}$$

Metrics to Summarize Model Performance

- The previous metrics work well for summarizing how well a model explains the data.

- But, what if we want to understand how well our model predicts for new data?

  - Answer: apply cross-validation

k-Fold Cross-Validation

1. Randomly split your data into k groups (or folds) of roughly equal size

2. Remove the first fold from the data set
   - called the test set, validation set, or hold-out set

3. Fit the regression model on the remaining $k - 1$ folds
   - called a training set

4. Use the fitted regression model (obtained from the training set) to make predictions on the test set and compute the MSE

- Repeat steps 2-4 k times where a different fold is used each time for the test set
  - Obtain k estimates of the test error (MSE_1, MSE_2, …, MSE_k)

k-Fold Cross-Validation

- **PMSE:** "Predictive Mean Square Error" (or just MSE, for short)
  - How far off predictions are (in the test set) on average
  - lower value = better model

$$\text{PMSE} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

How to choose the value of k

- Commonly used values for k
  - $k = n$ (called leave one out cross-validation, LOOCV)
    - Results in fitting the model n times
  - $k = 10$
    - Results in fitting the model 10 times
    - Often preferred over LOOCV (bias-variance trade-off)
  - $k = 5$
    - Results in fitting the model 5 times
    - Often preferred over LOOCV (bias-variance trade-off)

- Also need to consider the sample size n and think about how many data points will be in each fold

Next Steps

Now, that we have metrics to summarize model performance, we can learn more about these first two remedies for multicollinearity:

1. Remove some variables from the model – choose a subset of predictor variables

2. Apply a shrinkage method (ridge regression, lasso, elastic net, etc.)

1. Choose a Subset of Predictor Variables

Two Main Classes of Variable Selection Procedures

a. Check all possible subsets ("best subset" method)
   - Fit all possible regression models
   - Choose the "best" model based on some metric (e.g. AIC, BIC, PMSE)

b. Stepwise methods
   - Take a "structured" approach to building a good subset of predictors
   - This results in not all possible models being considered
   - Choose the "best" model based on some metric (e.g. AIC, BIC, PMSE)

a. Check All Possible Subsets (Best Subset Method)

All Possible Subsets

- We are working with $p-1$ predictors

$$X_1, X_2, \dots, X_{p-1}$$

- Recall: p represents the number of βs in the model (including the intercept)

- Let $p-1$ represent the total number of possible predictors
- Let $s-1$ represent the number of predictors selected for the model (subset)

- Total number of possible subsets: $\sum_{s-1=0}^{p-1} \binom{p-1}{s-1} = 2^{p-1}$

All Possible Subsets: Notes

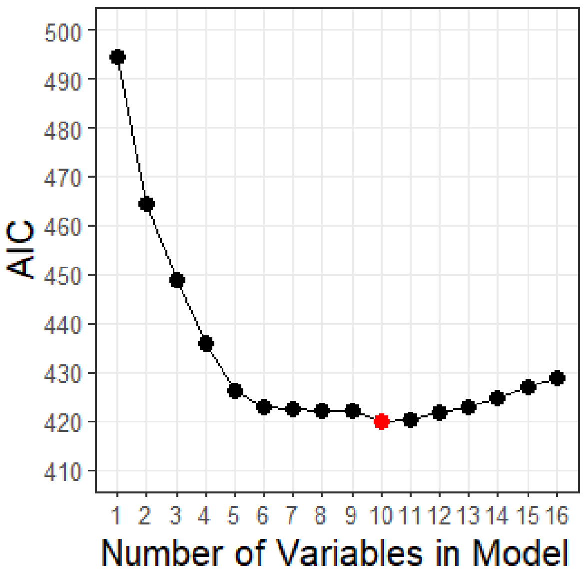
- We need the sample size n to be greater than the maximum number of parameters in the model
  - For sound results, we usually need n to be quite substantially larger than the number of parameters in the model (often 6-10 times larger)

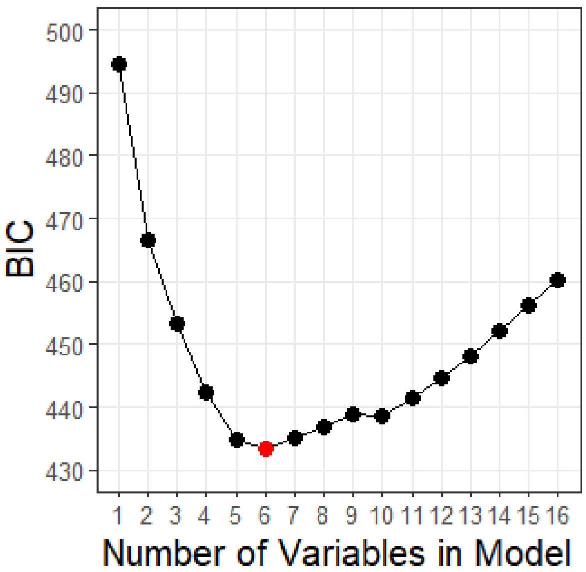
- Can only feasibly do this if we have a relatively small number of predictors
  - That being said, some smart CS people created a “leaps and bounds” algorithm to do this that can handle around 30-40 predictors

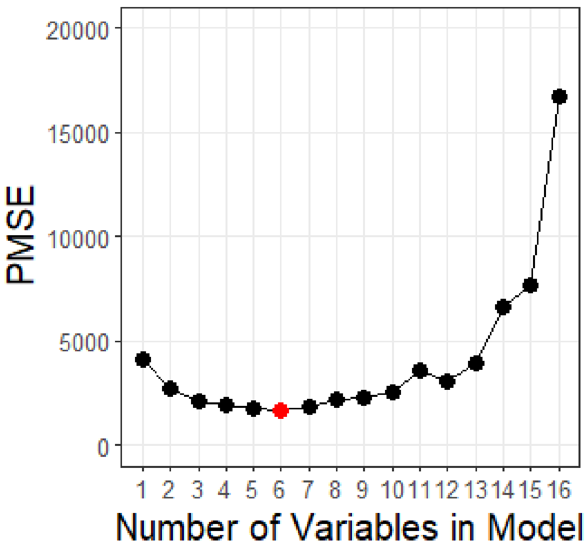
- Good to do if it is feasible

Environmental Impact Data Set: Best Subsets Selection

Subset







Environmental Impact Data Set: Best Subsets Model Comparison

Subset

Best AIC Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1954.2498   352.0943   5.550 1.08e-06 ***
AnnPrecip       2.6827     0.7582   3.538 0.000881 ***
MeanJanTemp    -2.5929     0.5893  -4.400 5.68e-05 ***
MeanJulyTemp   -3.1548     1.6648  -1.895 0.063891 .  
PctGT65       -13.7653     6.7280  -2.046 0.046037 *  
PopPerHouse  -148.8139    57.4636  -2.590 0.012550 *  
School        -20.4745     6.7398  -3.038 0.003781 ** 
PctNonWhite     4.1544     0.9917   4.189 0.000114 ***
log.Hydro     -33.9540    14.4758  -2.346 0.023006 *  
log.Nit        45.3215    12.6971   3.569 0.000801 ***
```

Best BIC Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1002.2638    87.7046  11.428 4.93e-16 ***
AnnPrecip      2.2616     0.6236   3.626 0.000637 ***
MeanJanTemp   -2.0340     0.4873  -4.174 0.000110 ***
School       -13.9333     6.0968  -2.285 0.026244 *  
PctNonWhite    3.7287     0.6380   5.844 3.02e-07 ***
log.Nit       19.4866     4.3556   4.474 4.00e-05 ***
```

Best PMSE Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1002.2638    87.7046  11.428 4.93e-16 ***
AnnPrecip      2.2616     0.6236   3.626 0.000637 ***
MeanJanTemp   -2.0340     0.4873  -4.174 0.000110 ***
School       -13.9333     6.0968  -2.285 0.026244 *  
PctNonWhite    3.7287     0.6380   5.844 3.02e-07 ***
log.Nit       19.4866     4.3556   4.474 4.00e-05 ***
```

Code!

b. Stepwise Methods

Stepwise Methods

- Automatically select a model based on some metric (e.g. AIC, BIC, PMSE) (very convenient)
- There is no guarantee the "right" model has been chosen
- These methods are best used as "confirmatory" approaches
- Three main stepwise methods:
  - Forward (terrible) Please, <u>never</u> do this!
  - Backward (okay)
  - Stepwise/Sequential Replacement (hybrid)

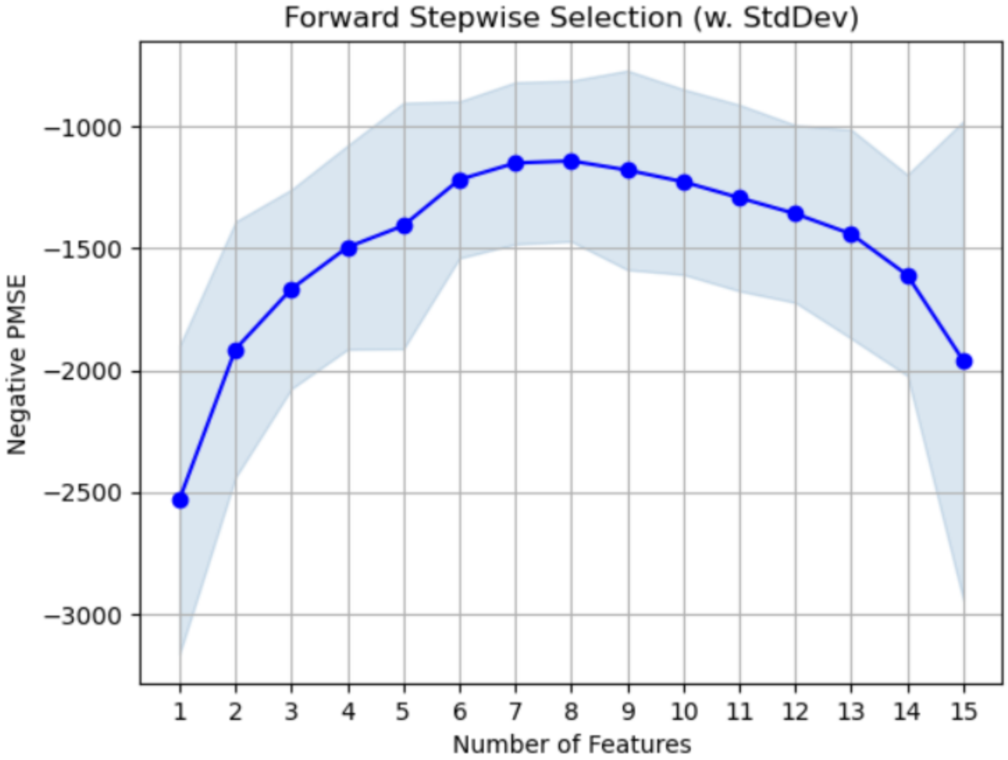
Forward Selection

1. Find the predictor that is most correlated with the response
   a. Fit a regression model with that one predictor
   b. Leave the predictor in the model if a chosen metric (e.g. AIC, BIC, PMSE) improves (compared to the intercept-only model)
2. Given the previously entered predictor, find the predictor with the highest partial correlation with the response
   a. Add this predictor to the model created in Step 1
   b. Leave the predictor in the model if the chosen metric is improved (compared to the model in Step 1)
3. Continue this process of adding predictors one at a time until adding predictors no longer improves the chosen metric

Problems with this method?

Environmental Impact Data Set: Forward Selection

Given we only have n=60, we can’t really afford to choose more than 6 variables



Only the PMSE is programmed in Python (not AIC or BIC)

Best model: AnnPrecip, MeanJanTemp, MeanJulyTemp, PopPerHouse, School, PctNonWhite, log.Hydro, log.Nit

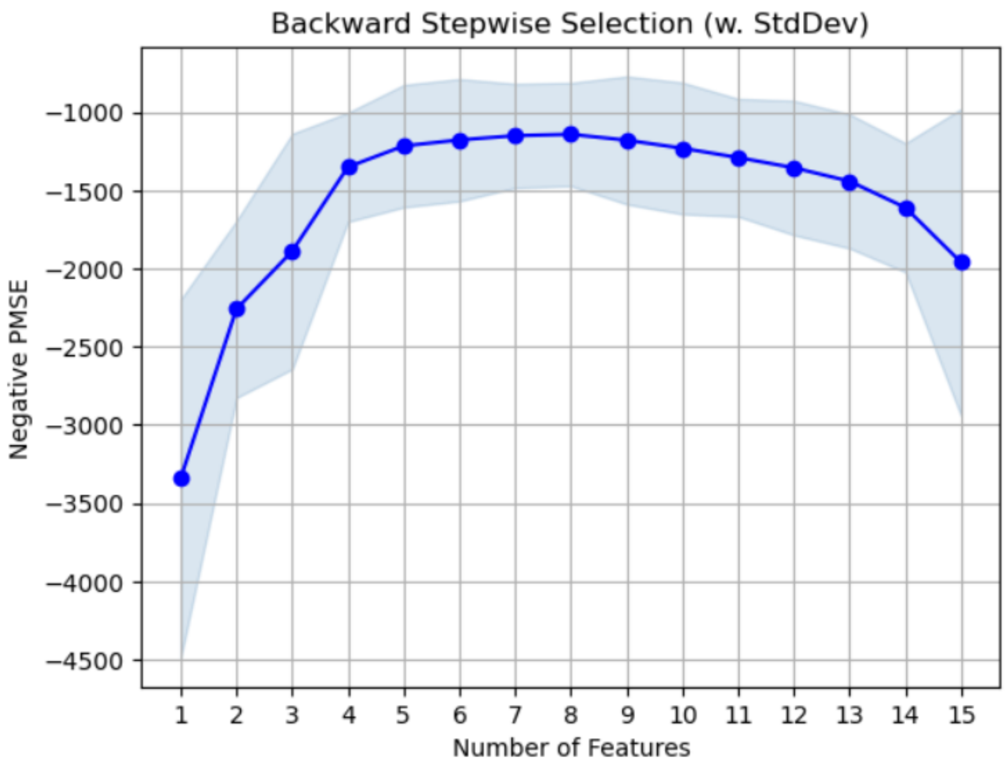
Best model with 5 predictors: MeanJanTemp, PopPerHouse, School, PctNonWhite, log.Nit

Backward Selection

1. Fit a model with all $p - 1$ predictors
   a. Obtain the model metric (e.g. AIC, BIC, PMSE)
   b. Consider removing a predictor from the model and compare that subsetted model's metric to the previous model's metric: remove the predictor if the metric with the variable omitted improves.
2. Repeat with the remaining $p - 2$ predictors
3. Continue this process of removing predictors one at a time until removing predictors no longer improves the chosen metric

Environmental Impact Data Set: Backward Selection

Subset



Best model: AnnPrecip, MeanJanTemp, MeanJulyTemp, PopPerHouse, School, PctNonWhite, log.Hydro, log.Nit

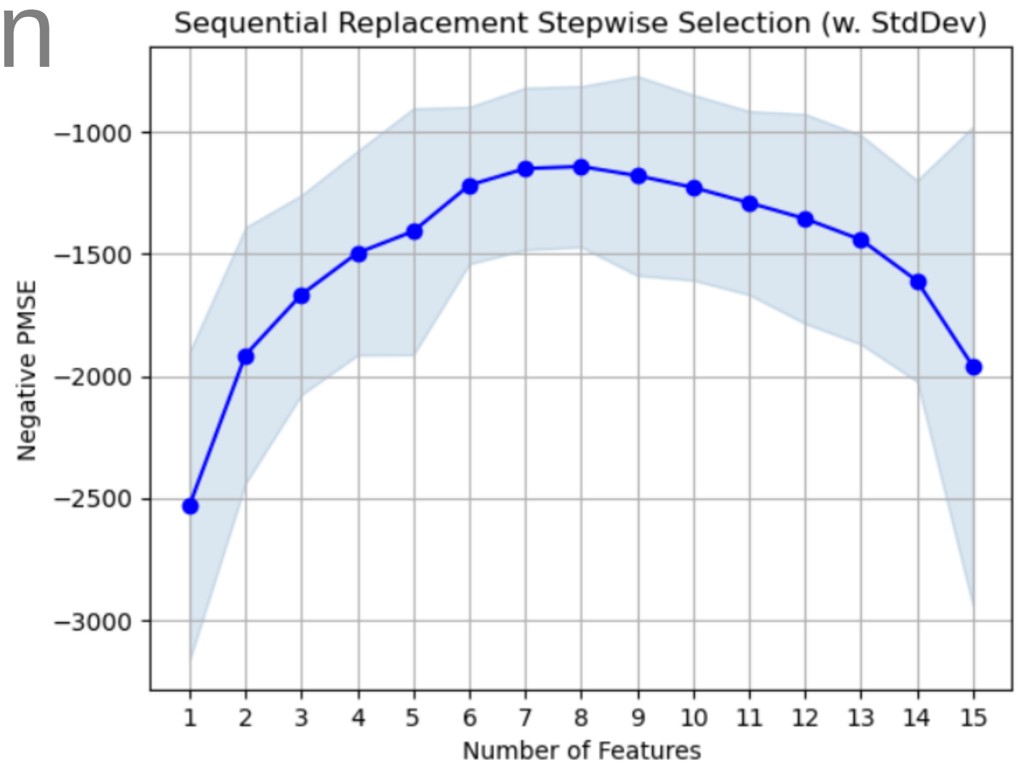
Best model with 4 predictors: AnnPrecip, MeanJanTemp, PctNonWhite, log.Nit

Stepwise/Sequential Replacement Subset Selection

1. Start with an intercept-only model
2. Take a “forward” step: add a predictor to the model if the predictor improves some chosen metric (e.g. AIC, BIC, PMSE)
3. Take a “backward” step: evaluate the predictor(s) in the model and drop a predictor if that would improve the chosen metric
4. Iterate the “forward” and “backward” steps until the model stays the same (i.e. adding or removing variables results in a worse metric value)

Environmental Impact Data Set: Stepwise/Sequential Replacement Selection

Subset



Best model: AnnPrecip, MeanJanTemp, MeanJulyTemp, PopPerHouse, School, PctNonWhite, log.Hydro, log.Nit

Best model with 5 predictors: MeanJanTemp, PopPerHouse, School, PctNonWhite, log.Nit

Things to Keep in Mind

- Different methods can give you different results (e.g. backward vs. best subsets)
- Different metrics can give you different results (e.g. PMSE vs. BIC)
- There is no “correct” model

- *Once you choose a model, you will need to re-check the multiple linear regression assumptions*

Code!

2. Apply a Shrinkage Method

Overview

- Shrinkage/penalty methods shrink the coefficients in the model toward zero.

- We will discuss three different shrinkage methods:
  - Ridge regression (coefficients never reach exactly zero)
  - **LASSO** ("least absolute shrinkage and selection operator", coefficients can be exactly zero)
  - **Elastic net** (coefficients can be exactly zero)

Note that the variables should all be standardized before fitting a shrinkage method.

- Note that LASSO and elastic net are variable selection methods since they can result in variables being dropped from the model

Recall OLS Properties

- Recall that ordinary least squares regression finds the $\hat{\beta}$s that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- These $\hat{\beta}$s are **unbiased** and have the **lowest variance** *if model assumptions are met*

- What is one side-effect of multicollinearity?

- Shrinkage methods introduce a small level of bias but reduce the variance of the estimates

Recall OLS Properties

- Since one of the symptoms of multicollinearity is inflated variance of the β_k estimates, can we (greatly) reduce the variance of the estimates by biasing them (slightly)?

- Think of two sampling distributions (draw)

  1. Unbiased estimate (with inflated variance)

  2. Biased estimate (with reduced variance)

- Allowing estimates (the β_ks) to be slightly biased could greatly reduce their variance, and we could avoid a major problem of multicollinearity.

Shrinkage Methods

- Ridge regression finds the $\hat{\beta}$s that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{k=0}^{p-1} \beta_k^2$$

- LASSO finds the $\hat{\beta}$s that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{k=1}^{p-1} |\beta_k|$$

- Elastic net finds the $\hat{\beta}$s that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{k=0}^{p-1} \beta_k^2 + \lambda_2 \sum_{k=1}^{p-1} |\beta_k|$$

Variance-Bias Tradeoff Summary

When OLS assumptions are met
(no multicollinearity)

Linear Regression
Unbiased/Accurate
Lowest Variance/Most Precise

When OLS assumptions are
not all met
(multicollinearity)

Linear Regression
Unbiased/Accurate
Inflated Variance/Not the Most Precise

Ridge Regression/LASSO/Elastic Net
Biased/Not Accurate
Lower Variance/More Precise

Shrinkage Methods

- We will not go into the details of these methods during class, but if you are interested, these slides provide a pretty good overview: https://ww2.amstat.org/meetings/csp/2014/onlineprogram/handouts/T3-Handouts.pdf

- We will discuss the main ideas of these methods, however.

- One of the important thing for you to know is that you have to select the value for the λ parameter (called "alpha" in Python) for these methods.

- Another important thing is that your predictors must be standardized!

- Note a major limitation to these shrinkage methods: traditional inference is not directly applicable to the estimates (be cautions)
  - Need "bootstrapping" to evaluate precision of the estimates (computationally intensive and beyond the scope of this course)

Ridge Regression

Shrinkage Methods: Ridge Regression

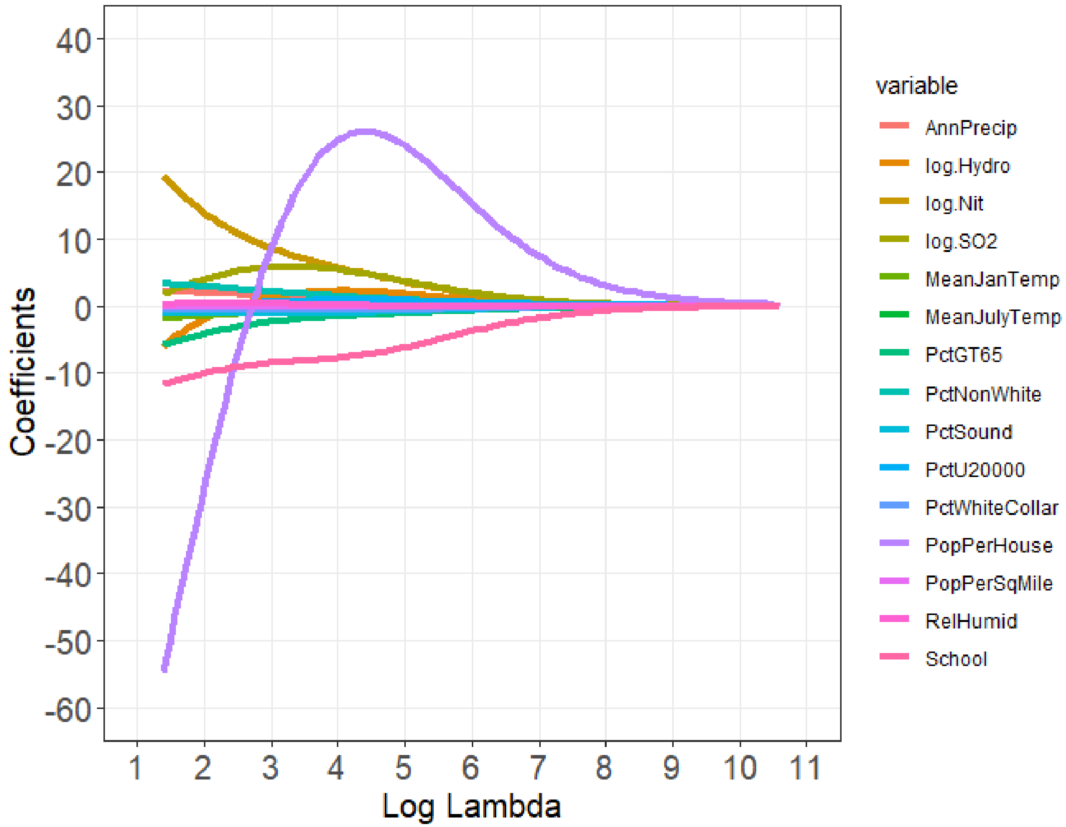
- Ridge regression tends to produce more precise predicted values ($\hat{y}_i$) than OLS *when predictors are highly related* (multicollinearity)

- Ridge regression keeps all variables in the model

- Why do ridge regression?
  - Sometimes we may want to adjust for multicollinearity but still keep certain predictors in the model (for mechanistic theory) with “correct” signs of coefficients

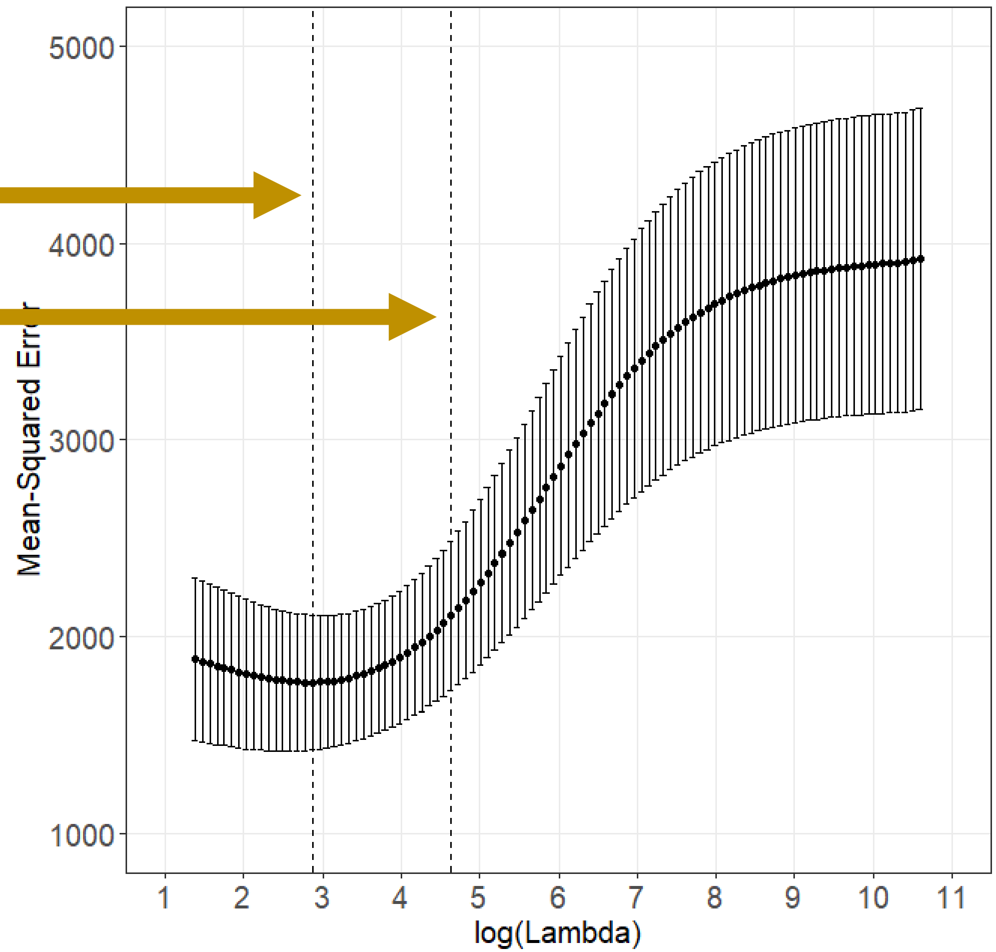
Environmental Impact Data Set: Shrinkage
Ridge Regression

- When (log) lambda is small, coefficients from ridge regression will be very similar to coefficients from OLS
- As (log) lambda increases, the coefficients are pulled toward zero (but never reach zero exactly)



Environmental Impact Data Set: Shrinkage Ridge Regression

- We want to pick the (log) lambda
  a. with the smallest MSE, or
  b. within one standard error of the smallest MSE



Generally, option (b) is preferable

Environmental Impact Data Set: Shrinkage Ridge Regression

- Coefficients using the lambda resulting in the smallest MSE

(Intercept)	990.296967184
AnnPrecip	1.642746339
MeanJanTemp	-0.961958428
MeanJulyTemp	0.379522490
PctGT65	-2.510240042
PopPerHouse	4.979715585
School	-8.624800121
PctSound	-1.058803626
PopPerSqMile	0.004242604
PctNonWhite	2.285625849
PctWhiteCollar	-0.576831752
PctU20000	0.551365261
log.Hydro	1.126212268
log.Nit	9.204845836
log.SO2	5.741615716
RelHumid	0.348072968

- Coefficients using the lambda resulting in the MSE being within one standard error of the smallest MSE

(Intercept)	917.632541713
AnnPrecip	0.822908119
MeanJanTemp	-0.261968437
MeanJulyTemp	0.533913942
PctGT65	-1.224043164
PopPerHouse	25.707749860
School	-6.857118151
PctSound	-0.901340368
PopPerSqMile	0.002998716
PctNonWhite	1.110880572
PctWhiteCollar	-0.573216186
PctU20000	0.876282363
log.Hydro	2.088061878
log.Nit	4.379673674
log.SO2	4.424749922
RelHumid	0.033115976

LASSO

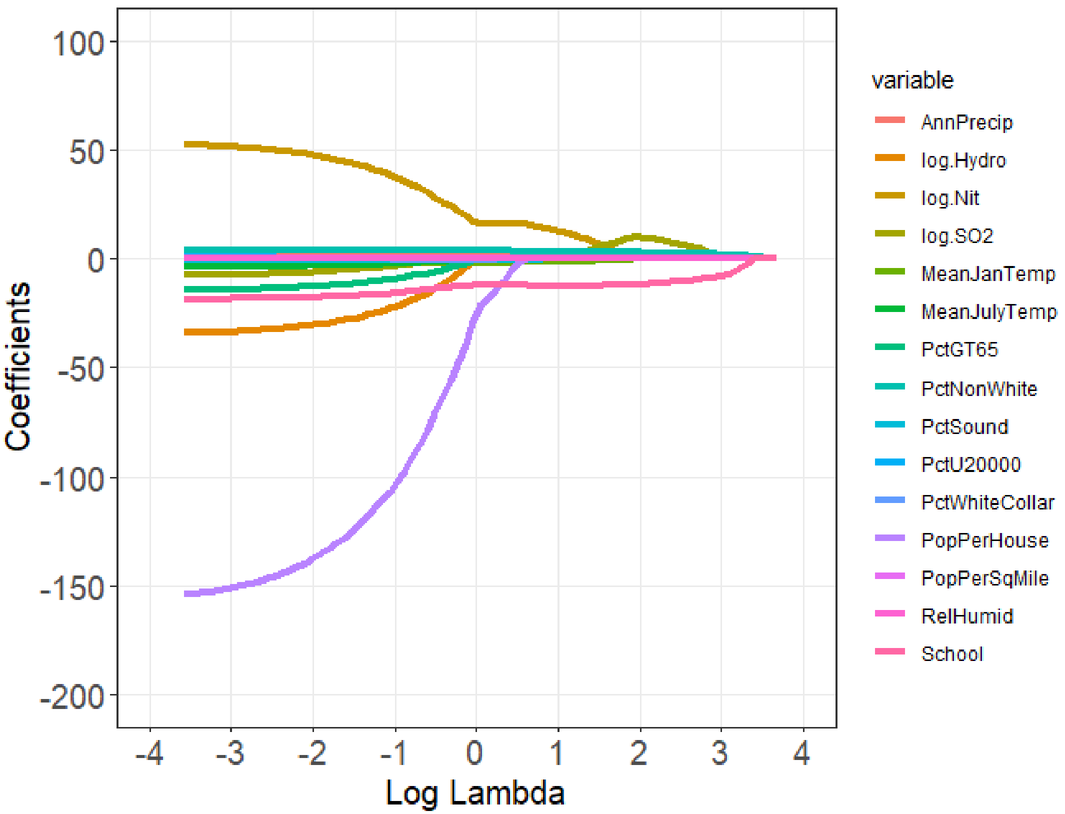
Shrinkage Methods: LASSO

- Allows the estimates to shrink all the way to zero – essentially a variable selection method

- Some limitations:
  - LASSO is known to give more biased estimates for nonzero coefficients (solution is to use “Adaptive LASSO”)
  - When there is high multicollinearity, LASSO tends to select only one variable from the group of correlated predictors (could miss out on some variable effects)
  - When there is high multicollinearity, LASSO is out-performed (prediction-wise) by ridge regression

Environmental Impact Data Set: Shrinkage LASSO

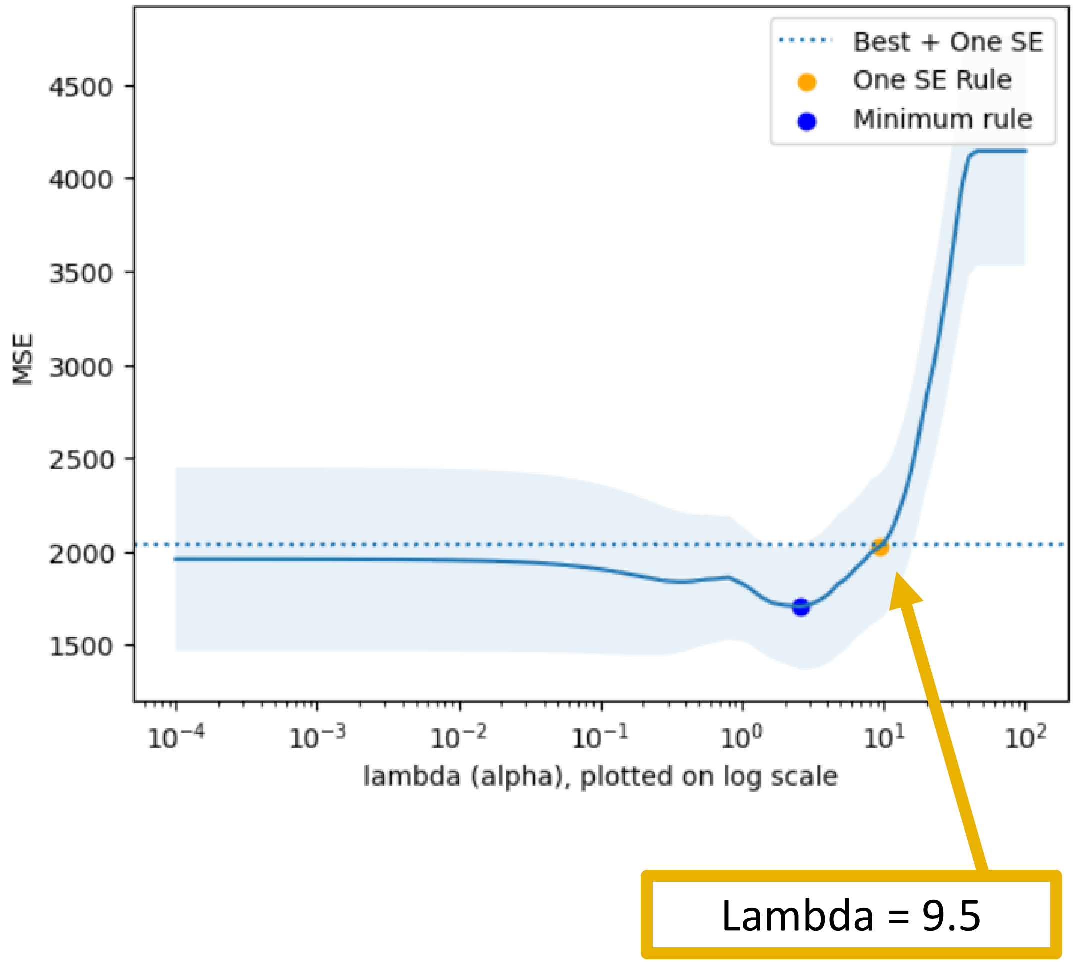
- When (log) lambda is small, coefficients from LASSO will be very similar to coefficients from OLS
- As (log) lambda increases, the coefficients are pulled toward zero, and some coefficients equal zero, causing them to drop from the model



Environmental Impact Data Set: Shrinkage LASSO

- We want to pick the (log) lambda
    a. with the smallest MSE, or
    b. within one standard error of the smallest MSE

Generally, option (b) is preferable



Environmental Impact Data Set: Shrinkage LASSO

- (standardized) coefficients using the lambda resulting in the smallest MSE

AnnPrecip	18.1
MeanJanTemp	-14.8
MeanJulyTemp	-0.0
PctGT65	-0.0
PopPerHouse	-0.0
School	-10.5
PctSound	-0.7
PopPerSqMile	3.7
PctNonWhite	30.8
PctWhiteCollar	-0.0
PctU20000	0.0
log(Hydrocarbons)	0.0
log(Nitrogen)	15.9
log(SO2)	1.8
RelHumid	0.0

- (standardized) coefficients using the lambda resulting in the MSE being within one standard error of the smallest MSE

AnnPrecip	9.1
MeanJanTemp	-0.0
MeanJulyTemp	0.0
PctGT65	-0.0
PopPerHouse	0.0
School	-9.3
PctSound	-0.0
PopPerSqMile	0.0
PctNonWhite	23.8
PctWhiteCollar	-0.0
PctU20000	0.0
log(Hydrocarbons)	0.0
log(Nitrogen)	0.0
log(SO2)	12.8
RelHumid	0.0

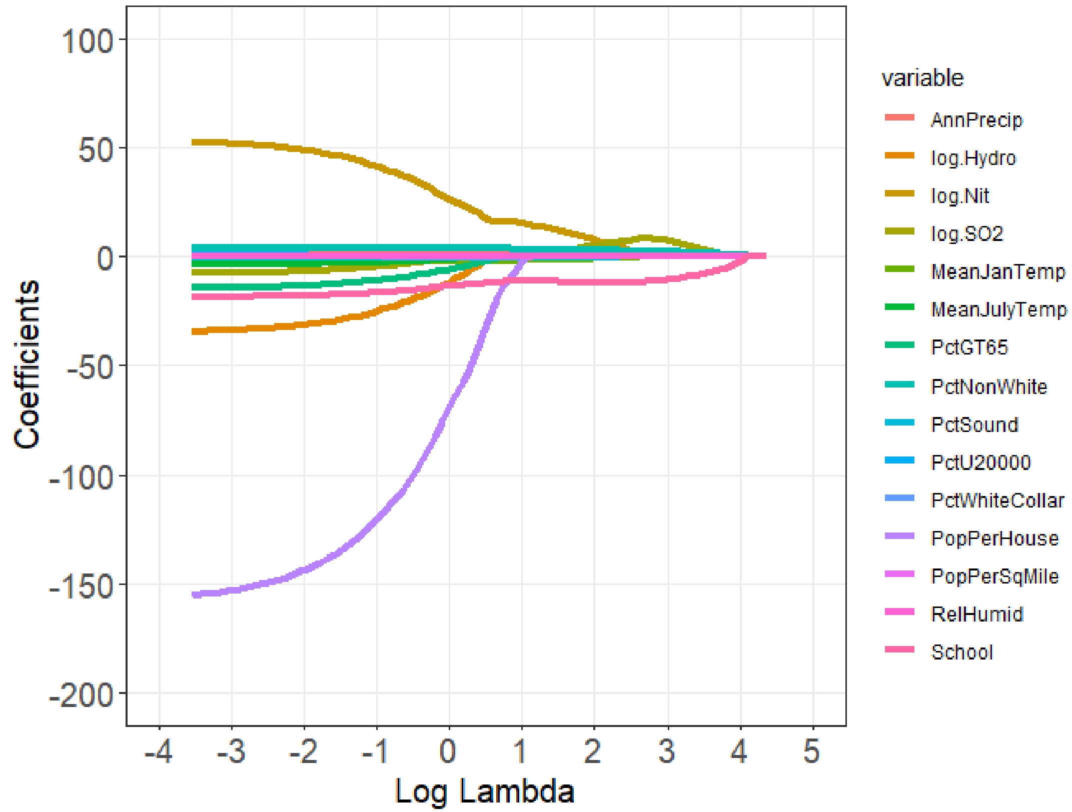
Elastic Net

Shrinkage Methods: Elastic Net

- Overcomes the limitations of LASSO:
  - When there is high multicollinearity, elastic net can select more than one variable from the group of correlated predictors
  - When there is high multicollinearity, elastic net can achieve better predictive performance (in general, elastic net should not do worse than ridge or LASSO)

Environmental Impact Data Set: Shrinkage Elastic Net

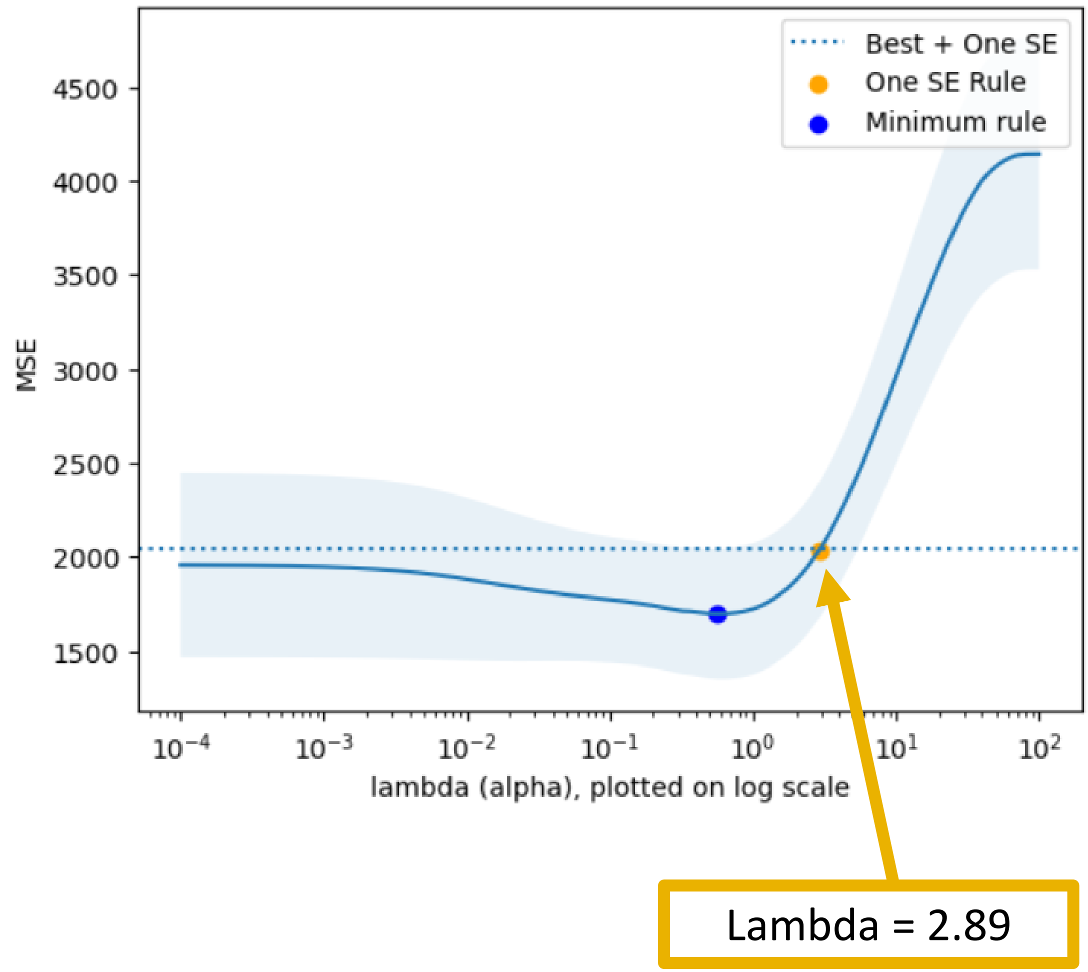
- When (log) lambda is small, coefficients from elastic net will be very similar to coefficients from OLS
- As (log) lambda increases, the coefficients are pulled toward zero, and some coefficients equal zero, causing them to drop from the model



Environmental Impact Data Set: Shrinkage Elastic Net

- We want to pick the (log) lambda
  a. with the smallest MSE, or
  b. within one standard error of the smallest MSE

Generally, option (b) is preferable



Environmental Impact Data Set: Shrinkage Elastic Net

- (standardized) coefficients using the lambda resulting in the smallest MSE

AnnPrecip	16.1
MeanJanTemp	-9.2
MeanJulyTemp	1.3
PctGT65	-3.3
PopPerHouse	0.7
School	-7.3
PctSound	-5.3
PopPerSqMile	5.9
PctNonWhite	20.5
PctWhiteCollar	-2.5
PctU20000	2.0
log(Hydrocarbons)	0.9
log(Nitrogen)	10.7
log(SO2)	8.5
RelHumid	1.4

- (standardized) coefficients using the lambda resulting in the MSE being within one standard error of the smallest MSE

AnnPrecip	8.3
MeanJanTemp	-2.2
MeanJulyTemp	2.0
PctGT65	-1.3
PopPerHouse	3.3
School	-5.8
PctSound	-4.5
PopPerSqMile	4.1
PctNonWhite	10.4
PctWhiteCollar	-2.3
PctU20000	3.2
log(Hydrocarbons)	1.9
log(Nitrogen)	5.1
log(SO2)	6.7
RelHumid	0.0

Environmental Impact Data Set Summary (PMSE/CV Method)

Variable	All Possible/ Best Subset	Forward	Backward	Stepwise/Sequential Replacement	LASSO (1se)	Elastic Net (1se)
AnnPrecip	X		X		X	X
MeanJanTemp	X	X	X	X		X
MeanJulyTemp						X
PctGT65						X
PopPerHouse		X		X		X
School	X	X		X	X	X
PctSound						X
PopPerSqMile						X
PctNonWhite	X	X	X	X	X	X
PctWhiteCollar						X
PctU20000						X
log.Hydro						X
log.Nit	X	X	X			X
log.SO2				X	X	X
RelHumid						

Summary

Variable Selection Methods

- **All Possible/Best Subsets**
- Forward Selection
- **Backward Selection**
- **Stepwise/Sequential Replacement**
- **LASSO Regression**
- **Elastic Net Regression**

Variable Selection Metrics

- Adjusted R^2
- **AIC**
- **BIC**
- **PMSE (aka MSE, based on cross-validation)**