Variable Selection

Thursday, October 10, 2024 10:37 AM

Variable Selection Introduction

We previously covered some factor-based models, like classification, clustering, and regression. In some cases, the model may perform better when using fewer factors.

There are two main reason why we would want to limit the number of factors in a model:

1. Overfitting: the model may fit too closely to random effects rather than the real effects
   a. Rule of thumb: when the number of factors is close to or larger than the number of data points, it is overfitted and likely to result in bad estimates
2. Simplicity: simple models are usually better than complex ones; less data is required, there is less chance of including insignificant factors, and the model is easier to interpret. Some factors may be illegal to use.

Models for Variable Selection

Forward Selection: a variable selection approach that starts with no variables in the model, then incrementally adds features to the model.

1. Start with a model with no factors
2. In each forward step, add the one factor that gives the single best improvement to the model
   a. Improvement could refer to improved accuracy, R-Squared, AIC, BIC, and other evaluation criteria.
   b. The definition of "good enough" is something you can set.
   c. A common rule of thumb is that a factor is "good enough" if its p-value is less than or equal to 0.15
3. Once the model has "enough" factors, fit the model with the current set of factors.
   a. Remove any factors that aren't good enough (those with a high p-value - p-value greater than 0.05)

Backwards Elimination: a variable selection approach that starts with all factors in the model, then incrementally removes the "worst" factors.

1. Start with a model with all factors
2. In each backwards step, remove the one factor that provides the worst improvement to the model; the factor that is "bad enough".
   a. The definition of "bad enough" is something you can set.
   b. A common rule of thumb is that a factor is "bad enough" if its p-value is greater than 0.15
3. Once the model has been reduced to enough factors, fit the model with the current set of factors
   a. Remove any factors that aren't good enough (those with a high p-value - p-value greater than 0.05)

Stepwise Regression: a combination of forward selection and backward elimination.

1. Either start with a model that has all factors or no factors.
2. Depending on the start, at each step you either remove a "bad enough" factor or add a "good enough" factor.\
3. After completing the incremental steps, remove any factors that aren't good enough anymore (those with a high p-value - p-value greater than 0.05)
4. Fit the model with a final set of factors

The three methods of variable selection above are Stepwise Selection methods: decisions are made step-by-step, taking the action that "looks best" at each step without considering future options. This is also known as a Greedy Algorithm. These are all classic variable selection methods.

More modern variable selection methods are listed below and they do take into account future options.

LASSO: a variable selection method that identifies that variables that are strongly associated with the response variable and uses them. It does this by adding a constraint to the standard regression equation such that the sum of the coefficients cannot be too large. Thus the "budget" for coefficients is mostly used on the important factors, leaving the other factors with coefficients at or close to zero.

$\min \sum_{i=1}^{m}(y_i - (a_0 + a_1x_1 + a_2x_2 + \cdots + a_jx_j))^2$ such that $\sum_{j=1}^{n}|a_j| \leq \tau$

When choosing the threshold τ, consider the number of variables and the quality of the model as more variables are added. The best approach generally is to use different values of τ with the LASSO approach, then see which gives the best tradeoff between the number of variables and model quality.

Note: whenever constraining the sum of coefficients, the data first be scaled, or else the units used to measure the variables will artificially impact how big the coefficients need to be.

Elastic Net: a variable selection method similar to the LASSO method, but it instead constrains the combination of absolute value of the coefficients and their squares.

$\min \sum_{i=1}^{m}(y_i - (a_0 + a_1x_1 + a_2x_2 + \cdots + a_jx_j))^2$ such that $\lambda \sum_{j=1}^{n}|a_j| + (1-\lambda)\sum_{j=1}^{n} a_j^2 \leq \tau$

Similarly to the LASSO method, you also need to choose values of τ and λ such that you get the best tradeoff between the number of variables and quality of the model. This method also requires that the data first be scaled.

LASSO vs Ridge Regression

If you take out the absolute value term from Elastic Net, you get a method known as Ridge Regression that may lead to better predictive models.

$\min \sum_{i=1}^{m}(y_i - (a_0 + a_1x_1 + a_2x_2 + \cdots + a_jx_j))^2$ such that $\sum_{j=1}^{n} a_j^2 \leq \tau$

Note: while Lasso Regression is used for Variable Selection, Ridge Regression is not.

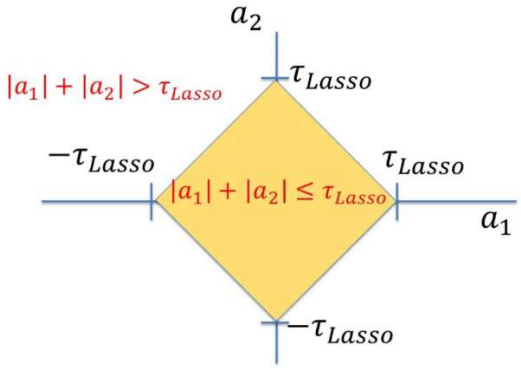
To explore how LASSO and Ridge Regression differ, let's consider two dimensional data:

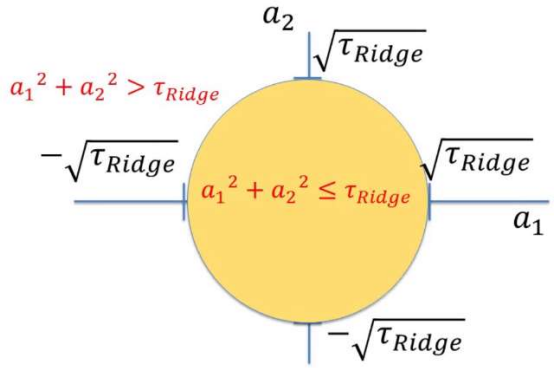
$i1$ $i2$ $LASSO$

$\min \sum_{i=1}^{m}(y_i - (a_1 x_{i1} + a_2 x_{i2}))^2$ such that $|a_1| + |a_2| \leq \tau_{LASSO}$

$\min \sum_{i=1}^{m}(y_i - (a_1 x_{i1} + a_2 x_{i2}))^2$ such that $a_1^2 + a_2^2 \leq \tau_{Ridge}$

Graphically, these restrictions on the coefficients look like the below:



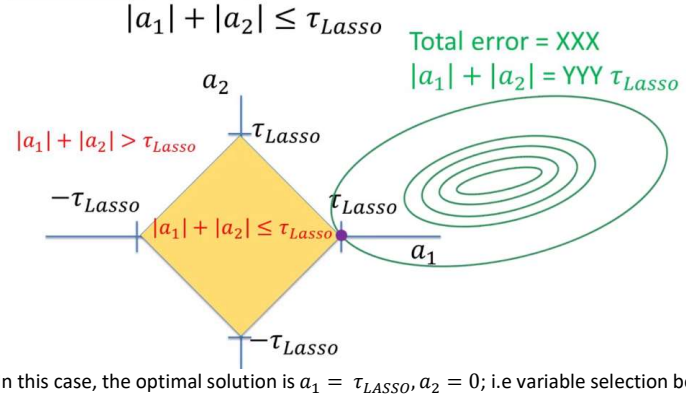


We choose coefficients a_1, a_2 to minimize the total error: $\min \sum_{i=1}^{m}(y_i - (a_1 x_{i1} + a_2 x_{i2}))^2$

If you rearrange the error equation, you can see that it is a quadratic function. Each point on each ellipse has the same total error. Depending on the choice of a_1, a_2, the value of the error can get higher or lower (changing the size of the ellipse). Since the objective is to minimize the total error, we want to find the smallest value of the error function that intersects with the shaded area of the restricted coefficients values.

Lasso regression

$|a_1| + |a_2| \leq \tau_{Lasso}$



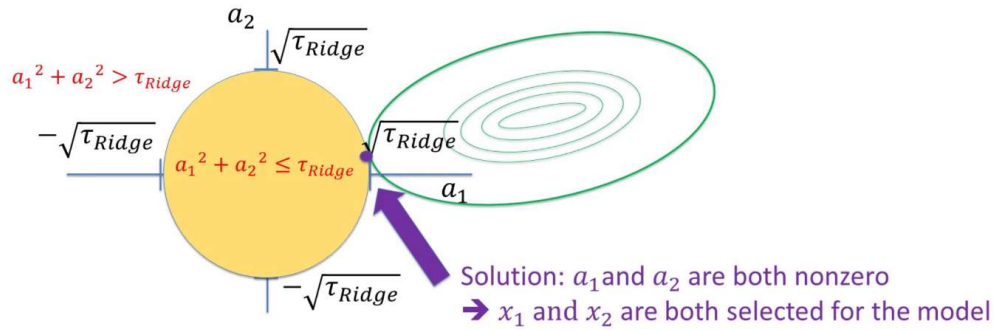
In this case, the optimal solution is $a_1 = \tau_{LASSO}, a_2 = 0$; i.e variable selection because x_2 won't be used in the model.

Note: it is possible to have a LASSO regression where both variables are selected, where the optimal minimized error is on the side of the diamond shape rather than a corner. But the more variables used and the more restrictive the threshold τ value, the less likely that becomes. Thus the LASSO method usually removes some variables and if it doesn't, you can reduce the value of τ until it does hit at a corner.

On the other hand, since the error and shaded restricted coefficient area of Ridge Regression are curved quadratic equations, it is very unlikely that they will first touch in an area where one of the values is zero. Thus no variables are removed with this method.

Ridge regression

$a_1^2 + a_2^2 \leq \tau_{Ridge}$



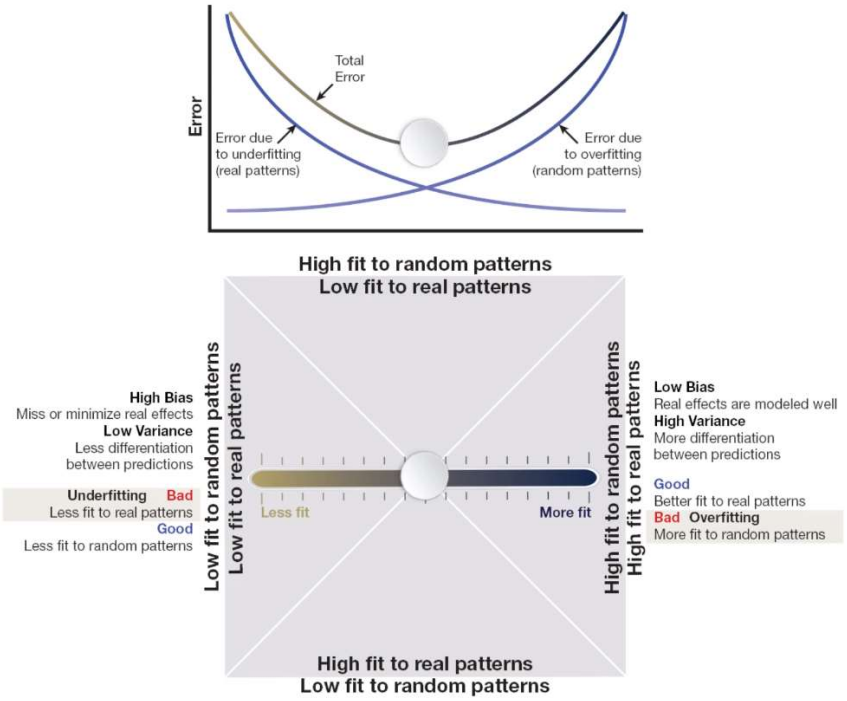
Bias-Variance Tradeoff

Recall that data has two kinds of patterns: real effects and random effects and we don't know which are which; all we can do is fit the model more to all patterns or less to

all patterns.

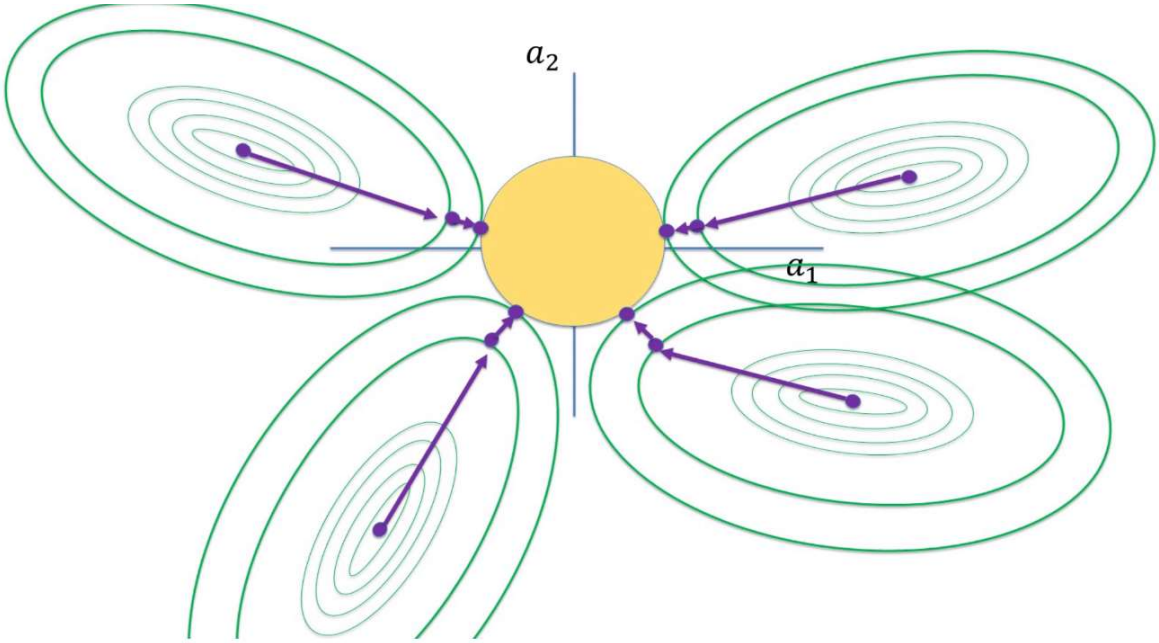
Consider a linear regression model as an example (though the same concepts apply to other model types):

- More fit to all patterns:
  - Better fit to real patterns, but also more fit to random patterns.
  - Danger of overfitting to random patterns
  - With a linear regression model, this means using more variables, which can include variance from random effects.
    - Low bias: real effects are modeled well
    - High variance: more differentiation between predictions
- Less fit to all patterns:
  - Less fit to real patterns, but also less fit to random patterns.
  - Danger of underfitting to real patterns
  - With a linear regression model, this means using fewer variables and likely missing out on important relationships between the predictors and response variable, so every prediction by the model gets closer to the constant a_0
    - High bias: missing or minimizing real effects
    - Low variance: less differentiation between predictions

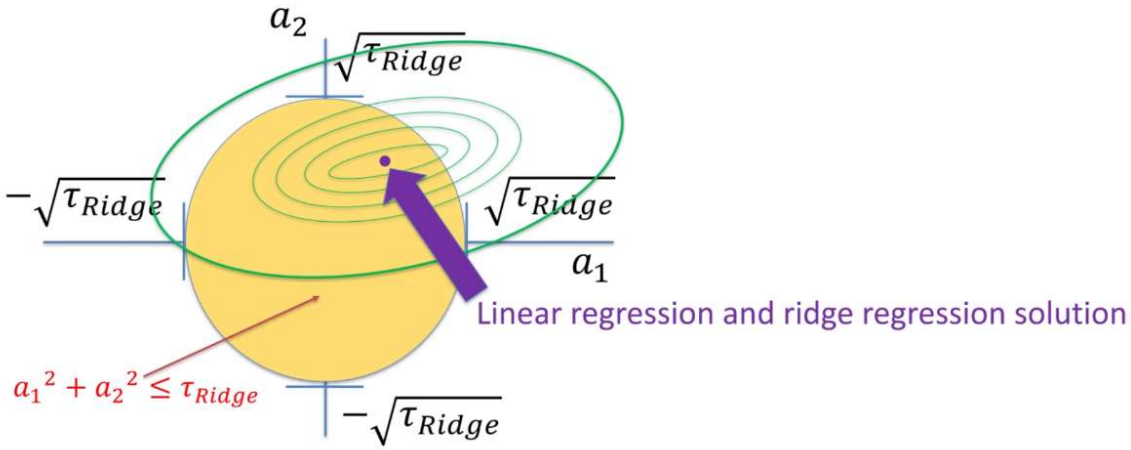


Ridge Regression Regularization

Recall that Ridge Regression restricts regression coefficient size, but does not reduce the number of variables. If we didn't care about the restriction on the coefficient size, then the ideal linear regression solution would be the center of the smallest error ellipses - where the error is an ellipse of size zero. Adding a ridge regression constraint moves the ideal solution inward toward the origin, with all coefficients getting smaller than those in the linear regression solution. The tighter the restraint gets, the smaller the magnitude of the coefficients.



However, if the linear regression solution is already inside the shaded area of the restricted coefficients, then adding the ridge regression constraint doesn't change the solution. To change the solution, you'd have to reduce the size of the circle until the linear regression solution is outside of the circle, and then the magnitude of the coefficients would decrease.



What does changing the magnitude of the coefficients do? Consider the effects/variables x_n and the magnitudes/coefficients a_n. Reducing the coefficients reduces the effects, lowering the fit of all patterns and reducing the variance - potentially leading to underfitting if you go too far.

Thus Ridge Regression is a regularization approach that can reduce overfitting not by variable selection, but by reducing the magnitude of effects.

Choosing a Variable Selection Model

How do you choose the right variable selection method?

Greedy Variable Selection methods are good for initial analysis, but often don't perform as well on other data. They perform quickly, but may result in a set of variables fit more to random effects and therefore they may appear to have a better fit than they actually do (like a higher r-squared value on the training data, but poorer r-squared values on previously-unseen data).

- Forward Selection
- Backward Elimination
- Stepwise Regression
  - Most common of the three greedy variable selection methods.

Global Optimization Variable Selection methods are slower to compute than greedy methods, but generally result in better predictions.

- LASSO
- Elastic Net
  - Advantages: Has the variable selection benefits of LASSO and the predictive benefits of Ridge Regression
  - Disadvantages: arbitrarily rules out some correlated variables like LASSO does and underestimates coefficients of very predictive variables like Ridge Regression

- If two highly predictive variables are very correlated, LASSO may just remove one. But since Ridge Regression reduces the size of the coefficients, that very predictive variable's effect has been reduced. And it's not always clear that LASSO removed the "worse" of the two correlated variables.
- You may have to check the correlation of the variables and manually remove one rather than letting LASSO make that choice.

Note that the Elastic Net threshold looks like a combination of LASSO and Ridge Regression's thresholds: $\lambda \sum_{j=1}^{n} |a_j| + (1-\lambda) \sum_{j=1}^{n} a_j^2 \leq \tau$. The Ridge component shrinks the size of the coefficients, reducing overfitting by adding some bias and reducing variation in predictions. Since Prediction Error is a function of bias an variance, trading off between the two can result in better predictions.

There isn't a good rule of thumb for when to use each approach, so try them all out, compare them, and choose the approach that appears to work the best.