Best subset selection – Start with null model, create model for each predictor, then create a model for each predictor combination (p1+p2) then p3 until pn. Select best model based on the Adj R squared and ect. ALL COMBINATIONS are calculated- therefore unfeasible for large nr of p. 2p nr of models.

STEPWISE Selection – Forward and backward

Forward – start with null model, then create model for p. Then choose the best model out of all for p=1, then KEEP IT, CANNOT be replaced. Then in the next step, choose from p1+p2, again best model, KEEP IT, and so on.

Backward – start with full model (all p) and reduce by removing the least significant variables.

Like forward stepwise selection, the backward selection approach searches through only 1+*p*(*p*+1)*/*2 models, and so can be applied in settings where *p* is too large to apply best subset selection.2 Also like forward stepwise selection, backward stepwise selection is not guaranteed to yield the *best* model containing a subset of the *p* predictors. Backward selection requires that the number of samples *n* is larger than the number of variables *p* (so that the full model can be fit). In contrast, forward stepwise can be used even when *n < p*, and so is the only viable subset method when *p* is very large.

Cp – Unbiased estimate of the MSE. The smaller the value, the better.

AIC – Same as Cp – lower the better.

Under certain circumstances, AIC and Cross Validation essentially do the same thing, but there are important cases where CV is more flexible.

BIC – same as AIC – smaller the better

Adjusted R2 - The intuition behind the adjusted R2 is that once all of the correct variables have been included in the model, adding additional noise variables will lead to only a very small decrease in RSS. Since adding noise variables leads to an increase in d, such variables will lead to an increase in RSS /d.and consequently a decrease in the adjusted R2

Cross Validation has an advantage over R2 and all the aforementioned statistics, because it makes less assumptions about the true underlying model of the relationship.

One standard error rule – pg 229

If a set of models appear to be more or less equally good. Then we might as well choose the simplest model – that is the model with the smallest number of predictors!

**RIDGE REGRESSION**

Shrinkage method – using lambda to shrink the coefficients b1,b2,…bp. The closer the parameter(bi) is to 0, the smaller the shrinkage penalty – lambda. The bigger- bigger the lambda. If lambda= 0 => same output as Least Squares (no shrinkage penalty).

Consider SCALING

Advantage over Least squares – bias – variance tradeoff – decreases variance, because of lambda

If p>n – prefer to use Ridge regression over Least Squares, because it has unique solution(the LS doesn’t have unique)

Advantages over best subset selection method – computational, because for every defined lambda only one solution is calculated.

ONE DISADVANTAGE – ALWAYS WILL INCLUDE ALL OF THE PREDICTORS IN THE FINAL MODEL. They may be shrunk into values close to 0, but will be there and therefore hard to interpret.

Outperforms Lasso, if ALL or close to all variables are significant and none of the coefficients=0

**LASSO**

Like Ridge Regr., the lasso shrinks the parameter coefficients towards zero, but in this case it HAS THE OPPORTUNITY TO MAKE THEM EXACTLY 0, thus eliminating them from the model and in a way making a variable selection.

Produces simpler and more interpretable models than RR, Performs Variable Selection

Outperforms RR, if number of components are = 0

To determine the number of predictors that are related to the response – cross validation.

**SUMMARY - RR + LASSO**

In the case of a more general data matrix X, the story is a little more complicated than what is depicted in Figure 6.10, but the main ideas still hold approximately: ridge regression more or less shrinks every dimension of the data by the same proportion, whereas the lasso more or less shrinks all coefficients toward zero by a similar amount, and sufficiently small coefficients are shrunken all the way to zero.

**How to Select Lambda**

Cross-validation provides a simple way to tackle this problem. We choose a grid of . values, and compute

the cross-validation error for each value of ., as described in Chapter 5. We then select the tuning parameter value for which the cross-validation error is smallest. Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

**Dimension Reduction**

CONSIDER THIS WHEN p>n or p~n