1. Subset Selection
   1. With *k* predictors, all three models would be equivalent when *k* is either near 0 or near *p*. This is because there is little variation in the possible predictors chosen by any of the 3 models. However, with most values of *k*, best subset selection would be superior, as both forwards and backwards stepwise selection are highly likely to miss values due to the intrinsic design of their selection algorithms. In a scenario where *p=*3, if X2 and X3 formed the best two-variable model, for example, but X1 formed the best 1 one-variable model, then best subset selection would choose the correct model but forward selection must contain X1 and one other variable. A similar situation occurs with backwards selection when X1 and X2 form the best two-variable model, but X3 forms the best one-variable model.
   2. It is unclear as to which model would have the lowest test RSS. Training error is not an accurate representation of test error, and so we cannot be certain that the best subset selection k-variable model would perform the best.
   3. T/F
      1. True. Of the k+1 predictors in the model, included must be the k predictors from the k predictor model.
      2. True. Of the k predictors in the kth model, all of these predictors must be in the k+1th model, as the kth model is simply the k+1th model with the single different predictor excluded. As well, it is evident that the k-1th model is a subset of the kth model, so the same must be true for the kth model and the k+1th model.
      3. False. In general, the k+1 predictors chosen by forward selection will not match the k variables left in the backwards selection model, even if k+1>=p-k.
      4. False. In general, the k predictors chosen by the forward selection will not match the k+1 variables left in the backwards selection model.
      5. False. The k variables to include to make the best k-variable model may vary wildly from the k+1 variables to included to make the best k+1-variable model. This is due to effects such as correlation and/or collinearity, and synergy effects.
2. Comparisons with Least Squares
   1. The lasso, relative to least squares, is less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance. Unlike least squares, the lasso intrinsically performs variable selection, which leads to increased bias by assuming that some variables have coefficients of 0 and are completely unrelated to the response, but also lowers variance.
   2. Ridge regression, relative to least squares, is less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance. The shrinkage of coefficients performed by ridge regression reduces the impact of certain predictors and puts less weight on them, decreasing the variance of the model but increases bias by doing so.
   3. Non-linear methods, relative to least squares, are more flexible and hence will give improved prediction accuracy when their increase in variance are lower than their decrease in bias. Non-linear methods do not assume the shape of the model to be linear, which increases variance as there are many possible shapes of the relationship between x and y, with a significantly different one being optimal depending on what training set. However, there is a significant decrease in bias as well due to the lack of any premade assumptions about said relationship. If the true shape is predicted well by the non-linear model, then the variance increase will be minimal, and will be offset by the decrease in bias.
3. Minimization (Lasso)
   1. As we increase *s* from 0, the training RSS will decrease constantly. When *s* is 0, the Lasso simply returns a null model. As *s* increases, we are restricting the coefficients less and less, and the model fit the training set more and more accurately due to increased flexibility.
   2. Decreasing at first, then increasing in a u-shape. As *s* increases, up to a certain point, there will be a minimum where the RSS has the lowest possible sum of bias and variance. It begins increasing again after that as *s* -> infinity, the coefficients all approach 0, and the model approaches the null model. The null model consists of only an intercept, where the prediction for every future observation is simply the mean of the existing responses. Variance will steadily decrease as *s* increases from 0. This is because there is more and more being assumed about the model, as more and more variables will approach 0, leaving less and less possible variation in the predictors related to the response.
   3. Variance increases steadily as *s* increases due to the increased flexibility.
   4. Squared bias will steadily decrease as *s* increases from 0. This is because of the fact that, as *s* increases, more of the variables are allowed to be non-zero, and this increases flexibility and consequently decreases bias, as well as squared bias.
   5. The irreducible error is, as the name implies, irreducible. It is unaffected by any other factors in the model, and cannot be altered. Therefore, it will remain constant, regardless of what value *s* is.
4. Minimization (Ridge Regression)
   1. As the tuning parameter increases from 0, the training RSS increases steadily, as the model itself is becoming less flexible due to the increased restrictions on the coefficients imposed by a larger lambda value.
   2. The test RSS will, as the tuning parameter increases, decrease and then increase to form a u-shape. This is a result of the bias-variance balance, as the minimum test RSS occurs when the sum of the two is minimized, with all other values of the tuning parameter producing larger RSSes due to either large bias or large variance.
   3. Variance will steadily decrease as the tuning parameter increases, since the model is becoming more restricted and less flexible, thereby varying significantly less.
   4. Bias, as the counterpart of variance, will steadily increase as the tuning parameter increases, as the lack of flexibility means that the model is making stronger assumptions about the relationship between x and y.
   5. Irreducible error remains constant, as it is independent of everything else.
5. Suppose that , , , , , and . Let and .
   1. Ridge Regression:
   2. We partially differentiate the above function in order to find the ridge coefficients, as they are identical to the input values which minimize said function.

Taking , we obtain , as desired.

* 1. Lasso optimization is identical to ridge regression optimization except for the adjustment term used. Therefore, lasso optimization is as follows:
  2. Since the formulas for lasso and ridge regression are almost identical , we may follow the same procedure as in . We set the adjusting terms from the 2 partial derivatives equal to each other, and obtain

We now end up with two distinct cases.

**Case 1**: and have identical signs.

All values of this case allow the statement to hold true, so they work.

**Case 2**: and have opposite signs.

All values of this case lead the statement to be false, so they do not work.

Therefore, solutions to the lasso optimization problem in this example are pairs with either or holding true.

1. Exploring the simple case of ridge regression and lasso, .
   1. Plotted on R

To minimize, differentiate the equation and set it equal to 0. We obtain

* 1. Plotted on R

To minimize, differentiate the equation and set it equal to 0. We obtain

1. Bayesian Connection
   1. Let , with being independently and identically distributed for all , and .

Equivalently, we have , due to how we know that .

We obtain the prior probability function for :

This in turn allows us to obtain the probability density function, which is also the likelihood of the data:

* 1. Assume that are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter for all . In other words, assume that . We know that the overall prior, given all of the predictors , is . We write the posterior for as follows:
  2. Knowing the formula for the lasso, we may conclude that the most likely (the mode) is what results from the lasso estimates for the given posterior. In our case, the lasso would be

Through analyzing the 2 expressions, it is clear that maximizing the probability of the given posterior is analogous to minimizing the lasso expression. Note that as the lasso expression decreases, the value of the posterior must increase. As such, the and are what define the values of both aforementioned expressions. As both of these terms decrease, the lasso expression decreases, with the posterior increasing. Therefore, the lasso estimates, which minimize the above expression or decrease it as much as possible, will increase the posterior as much as possible and maximize it, thereby making the lasso estimates the most probable value for , or in other words, the mode for .

* 1. Assume that are independent and identically distributed according to a normal distribution with mean 0 and variance for all . We know the prior for our overall is . We write our posterior as follows:
  2. The ridge regression formula is

We may use a similar, analogous argument from to show that the mode for here is equivalent to the ridge coefficient estimates. Since follows a normal distribution in this case, the mean, mode, and median are all identical. Therefore, the ridge regression estimate in this case is simultaneously the mean and the mode under this posterior distribution, as desired.