Linear model selection and regularization: Notes

Why do we need an alternative to least squares?

It helps increase prediction accuracy, especially in cases where there are more predictors than number of samples. It helps to control variance.
Model interpretability through feature selection. This can be done by reducing the coefficients of unrelated features to zero.

Three methods of feature selections:

Subset Selection:

Select a subset of predictors that are most related to the target variable. This can be done through forward, backward or an intelligent state space search.
We have a set of predictors but we want to select the best subset of those to fit to a model. The easiest way to do this is to select all possible subsets and then fit it to the model.

Best Subset Selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

5

Drawbacks of this approach:

1. If there are a lot of predictors then looking ad all the subsets is almost impossible because it's computationally expensive. ($\mathbf{2}^p$)
2. Another problem is that fitting all these models to so many subsets will just lead to overfitting so we probably won't choose the optimal model
3. So the alternative is stepwise methods

Stepwise:

Forward stepwise

1. You start with predictors and then incrementally add one after another
2. Do until you have all predictors

3. Whenever each additional predictor is added and it makes the predictions better it is important

Forward Stepwise Selection

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   2.1 Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   2.2 Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

So forward stepwise has a computational advantage over the previous but it does not always guarantee the best subset.
This happens due to correlated variables.

Backward stepwise
Exactly the opposite of backward stepwise. We take all the predictors and then remove one by one while fitting a model to each.

Both these methods go through only **1+p(p+1)/2** models
Optimal models have large R-squared(coefficient of determination) and small RSS.

Better ways to estimate test error

RSS is not always the best way to capture overfitting. **Cp, AIC, BIC and adjusted R squared** are a few other methods used to adjust training error for model size and different sizes of variables. We can also use these to decide how many predictors to use.
Mallows Cp:
Cp = (RSS + 2*predictors*variance^2) /n .....limits: N > Predictors

AIC : Akaike Information Criterion
AIC = -2*Log(likelihood) + 2* predictors (For linear models with gaussian distributions L=RSS)
Likelihood = RSS / variance^2

BIC : Bayesian information criterion
BIC = (RSS + log(n) * variance^2 * predictors)
BIC is imposing a higher penalty on models with many variables.

Adjusted R squared:

R-squared = 1 - ((RSS/n-d-1)/(TSS/n-1))
Larger R-squared value is better. Works when p > n also.

Cross Validation

Better approach than r-squared, AIC, BIC.
Don't have to calculate variance to find test error.

Some other fancy methods are,

Shrinkage

Penalize predictors that don't affect the target variable by reducing coefficient to zero. This is called regularization and also reduces the variance. Done using Ridge and Lasso methods.

Subset selection methods use least squares to select coefficients.

Ridge regression is basically = RSS + $\lambda \sum_{j=1}^{p} \beta_j^2$ -> minimize this value, λ is the tuning parameter (can be decided using cross validation).

$\lambda \sum_{j=1}^{p} \beta_j^2$ = shrinkage penalty , L2 penalty

Sum of least squares does not change if predictors are scaled but ridge regression is immediately affected as the penalty term changes with any multiplications to coefficients. Therefore apply ridge regression after standardising predictors. By using the formula below:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

Drawback: Uses all predictors in the final model, even though all may not be required (unlike subset selection)

Lasso Regression:

Tries to overcome the ridge regression drawback. Using absolute values of coefficients.

Lasso = RSS + $\lambda \sum_{j=1}^{p} |\beta_j|$ -> minimize

$\lambda \sum_{j=1}^{p} |\beta_j|$ = shrinkage penalty, L1 penalty

When lambda is large enough the coefficient and parameters overall effect shrinks to zero (similar to subset selection). This is also called variable selection and results in sparse models. Lambda is decided by cross validation.

Dimension Reduction

Transforming predictors to fit least squares model.

Choose a set of M predictors from the superset P.
Let's choose a subset z1,z2,z3 etc from P with x1,x2,x3...
Now new : $Zm = \sum_{j=1}^{p} \phi mj\ Xj$
Where ϕ is a constant that X is multiplied with to generate new parameters.
You basically try to transform previous parameters into new ones that are more relevant to the model, i=by doing so you reduce the unnecessary features.
Fit a new linear model using least squares:

$$yi = \theta 0 + \sum_{m=1}^{M} \theta m * Zim + \varepsilon i\ ;\ i = 1\ to\ n$$

If we look at it, it is like a recursive fit of linear models.

$\sum_{m=1}^{M} \theta m * Zim$ is actually $\sum_{p=1}^{P} \sum_{m=1}^{M} \theta p * \phi m * Xi$

Principal Component Analysis

The assumption is that a correlated linear combination of the predictors is also related to the response.
PLS is PCA in a supervised way.