5 Resampling

2 common methods. Cross-Validation and Bootstrap

5.1 Cross Validation

Validation Set approach: Split samples into 2 parts. Use one as train set, another as validation set. Evaluate the test accuracy.

LOOCV : Leaving One out Cross Validation approach: Leave one sample out, and use the rest for training. Use that one sample for testing. Average the accuracy score. This is lot of correlated training sets. With ordinary least squares linear regression, this simplifies to,

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

Otherwise it is a computationally intensive approach

K-fold CV: Split into K groups. Train on K-1 and test on the one fold. Repeat K times. Average the error to find a score.

LOOCV is special case with K=1.

K-fold is in between Cross Validation approach and LOOCV. CV is biased, LOOCV has too much variance.

K-fold is bias variance trade off. Can be used effectively to do hyperparameter tuning. It is normally effective in spotting the point where both training error and test error are low.

K-fold underestimates test error rates (why?), however effective in spotting optimum point

Note that in regression problem, MSE is used to measure CV error, and in classification problems, misclassification count is used.

5.2 Bootstrap: Already studied in Inferential stats

6 Linear Model and Regularization

The linear model has distinct advantages in terms of inference and, on real-world problems, is often surprisingly competitive in relation to non-linear methods. Hence, before moving to the non-linear world, we discuss in this chapter some ways in which the simple linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures.

Why do we want to do this? It can result in better prediction accuracy and better interpretability

1. Better prediction accuracy: Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimates will have low bias. If n >> p, then the usual least squares tends to have low variance, and tends to do well. If n is not so larger than p, then the least squares tends to overfit. And if p > n, then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all. By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training.
2. Model Interpretability: Some predictors are not related to target at all. Eliminating irrelevant features is important. Least squares is unlikely to make coefficients for such predictors zero.

Many methods:

1. Subset selection: Select a subset of features and fit least squares. To do it exhaustively, it will take a long time as there are 2^p possibilities, for p features. So, this is how it is done.

1.1 Best subset selection

- A model with no predictors, is called null model Mo. Simply predicts sample mean of each.
- For k = 1 to p, fit all possible pCk models and call it Mk, best as per small RSS or large R^2 - This selects best within each subset size.
- Pick the best among M0,M1….Mp based on Cross Validated score, AIC, BIC or adjusted R^2 - This selects best inter subset size

Low RSS or high R^2 indicates low training error and not low test error. Hence in the third step CV or AIC are used.

For logistic regression, we use “deviance” instead of RSS. -2*max likelihood is deviance. Smaller deviance, better fit. It is used in broader range of models.

1.2 Forward Stepwise selection:

- Let Mo be null model
- For k = 0 to p - 1, consider p - k models that augment Mk with one additional predictors and select the best predictors (that give highest R^2 or small RSS)
- Among Mo to Mp, choose the best based on CV score, AIC, BIC or adjusted R^2

Unlike first one, this has 1+ p(p+1)/2 models only. Kth iteration has p-k models to fit.

2. Shrinkage: Also called regularization. Depending on what type is applied, some coefficients can turn out to be zero.
3. Dimensionality Reduction: This approach involves projecting the p predictors into a M-dimensional subspace, where M <p. This is achieved by computing M different linear

combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Choosing the optimal model:

C_p, AIC, BIC, and Adjusted R^2

AIC: Akaike information criterion**, BIC:** Bayesian Information Criterion

Recall MSE = RSS/n

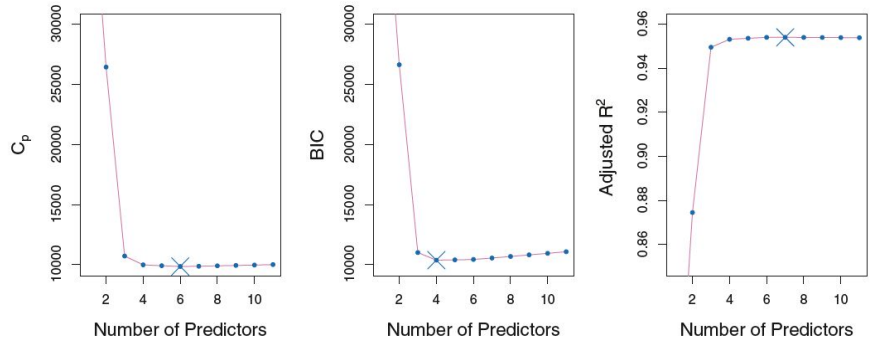
For a fitted least squares model containing d predictors, the C_p estimate of test MSE is computed using the equation

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right), \tag{6.2}$$

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right), \qquad \text{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right).$$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

Note all the differences in equations. For least squares, Cp and AIC are proportional. Since log n > 2 for any n > 7, BIC tends to penalize models with more variables heavily and results in model with lower number of variables than that chosen with Cp



Shown for the Best models of each size of dataset

Olden days Cp, AIC, BIC and adjusted R2 were metrics to evaluate the model. But nowadays with fast computers, cross validations are more widely used.

6.2 Shrinkage methods

Fit all the p predictors, but shrink some to zero. Ridge and Lasso regression

Ridge regression: Least squares minimizes RSS, sum (y - yhat)^2. Ridge regression minimizes RSS + lambda * sum (beta^2) for non intercept coeffs. Recall Andrew Ng lectures and this expression. Why not intercept while shrinking? Below:

> each variable with the response; however, we do not want to shrink the intercept, which is simply a measure of the mean value of the response when $x_{i1} = x_{i2} = \ldots = x_{ip} = 0$. If we assume that the variables—that is, the columns of the data matrix $\mathbf{X}$—have been centered to have mean zero before ridge regression is performed, then the estimated intercept will take the form $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i/n$.

Some more notes, best refer the book.

But note that ordinary least square does not need scaling. But ridge regression needs it.

Advantages are rooted in bias variance tradeoff.

> In general, in situations where the relationship between the response and the predictors is close to linear, the least squares estimates will have low bias but may have high variance. This means that a small change in the training data can cause a large change in the least squares coefficient estimates. In particular, when the number of variables p is almost as large as the number of observations n, as in the example in Figure 6.5, the least squares estimates will be extremely variable. And if $p > n$, then the least squares estimates do not even have a unique solution, whereas ridge regression can still perform well by trading off a small increase in bias for a large decrease in variance. Hence, ridge regression works best in situations where the least squares estimates have high variance.

Important from above para that when data is approx linear, coeffs can vary by great amount with small change in training data.

Lasso is another alternative, which can even set coefficients of unimportant predictors to zero. Ridge does not do that. The difference in expression is, instead of beta^2 , we have absolute of beta.

> The *lasso* is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients, $\hat{\beta}_\lambda^L$, minimize the quantity lasso

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|. \qquad (6.7)$$

Lasso does variable selection and produces sparse models. Depending on the value of lambda, lasso can have different number of variables in the model.! Wow.

Another Formulation for Ridge Regression and the Lasso

One can show that the lasso and ridge regression coefficient estimates solve the problems

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{6.9}$$

for every value of λ, there is some s such that the Equations (6.7) and (6.8) will give the same lasso coefficient estimates, and similar statement for ridge.

When $p = 2$, then (6.8) indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$. Similarly, the ridge regression estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq s$.

s can be seen as budget for either l1 or l2 norm, corresponding to Lasso and Ridge. When s is large, then budget is not restrictive. When it is extremely large, then least squares fits within that budget.

The formulations (6.8) and (6.9) reveal a close connection between the lasso, ridge regression, and best subset selection. Consider the problem

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

(6.10) ==> No more than s predictors can be non zero and RSS should be minimized. Selection problem