Homework5

Name: Chenqing Ji
Student ID: 11911303
Statistical Learning for Data Science

Due time: June 7, 2023 (Wednesday) 16:00pm

1 Proof

Choosing the Optimal Model

- AIC

In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

Prove this.（HW）

Proof:

To prove that in the case of the linear model with Gaussian errors, the maximum likelihood (ML) and least squares(LS) are the same things and that C_p and AIC are equivalent, let's start with the linear model formulation:

$$y = X\beta + \varepsilon \tag{1}$$

In equation (1), where y is the response variable, X is the predictor matrix, β is the vector of coefficients, and ε is the vector of Gaussian errors.

The ML estimation seeks to find the values of β that maximize the likelihood function based on the Gaussian error assumption. Since the errors ε are Gaussian, the likelihood function can be written as:

$$L(\beta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right) \tag{2}$$

where σ^2 is the variance of the Gaussian errors, and n is the number of observations.

To maximize the likelihood function, we can take the logarithm of the likelihood (log-likelihood) and find the β values that minimize the negative log-likelihood (which is equivalent to maximizing the likelihood). The negative log-likelihood can be written as:

$$-2\log(L(\beta)) = n\log\left(2\pi\sigma^2\right) + (y - X\beta)^\top (y - X\beta) \tag{3}$$

We can see that minimizing the negative log-likelihood is equivalent to minimizing the residual sum of squares (RSS), which is given by:

$$RSS = (y - X\beta)^\top (y - X\beta) \tag{4}$$

This is precisely the objective of the LS estimation, where we aim to minimize the sum of squared residuals. Hence, the maximum likelihood (ML) and least squares(LS) estimation are equivalent in the case of the linear model with Gaussian errors.

Then, the C_p statistic is used for model selection and is defined as:

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right) \tag{5}$$

where d is the number of predictors in the model, n is the number of observations and σ^2 is an estimate of the variance of the error associated with each response measurement.

On the other hand, the Akaike Information Criterion (AIC) is another model selection criterion, given by:

$$AIC = -2\log(L(\beta)) + 2d \tag{6}$$

Comparing the C_p and AIC expressions in equations (5) and (6), we can see that they have a similar structure. The first term in both criteria is related to the residual sum of squares (RSS) or negative log-likelihood, and the second term penalizes the number of predictors (d). The only difference is the constant term ($2d\hat{\sigma}^2$ in C_p and $2d$ in AIC).

However, if we assume a linear model with Gaussian errors, the maximum likelihood estimation can be shown to be equivalent to the LS estimation. Consequently, the negative log-likelihood ($-2\log(L(\beta))$) is proportional to the RSS, and we can conclude that C_p and AIC are equivalent since the constant term ($2d\hat{\sigma}^2$ in C_p and $2d$ in AIC) does not affect the model selection decision.

In summary, in the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

2

We now review k-fold cross-validation.

(a) Explain how k-fold cross-validation is implemented.

(b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i The validation set approach?

ii LOOCV?

Solution:

(a) The k-fold cross-validation method is an alternative to the leave-one-out cross-validation method (LOOCV). This method randomly divides the set of observations into k groups, or folds, of roughly the same size. The first fold serves as the verification set, and then the model is fitted on the remaining k-1 folds. Repeat this step k times, each time using a different observation group as the verification set. The whole process will give the estimate of k test errors, and the error estimates of k-fold cross validation will be calculated by averaging these k test errors.

(b)

(i) Compared with the verification set approach, k-fold cross-validation has the following advantages and disadvantages:

Advantages:

(1) Better model assessment: k-fold cross-validation allows us to make the most out of your limited data by utilizing it for both training and validation. It partitions the data into k subsets or folds, and each fold is used as both a training and validation set. This provides a more reliable estimate of the model's performance than using a single validation set. Based

on that, k-fold cross-validation can be more accurate to estimate the test error rate obtained by fitting the model on the whole data set.

(2) Better usage of data: k-fold cross-validation divides the dataset into multiple subsets or folds, ensuring that every data point is used for both training and validation at some point. This leads to more efficient use of the available data compared to the validation set approach, where a portion of the data is completely held out for validation and not used for training.

(3) Reduced bias: k-fold cross-validation provides a more reliable estimate of a model's performance because it averages the results over multiple iterations. The validation set approach may suffer from bias if the split between training and validation data is not representative of the overall dataset. k-fold cross-validation helps mitigate this bias by providing an average performance estimate across different data partitions.

Disadvantages:

(1) More complex computations are required: Performing k-fold cross-validation requires training and evaluating the model k times, which can be computationally expensive, especially if the model is complex or if the dataset is large. The computational cost is higher compared to the validation set approach, which involves training the model only once on a single validation set.

(2) Variability in performance estimates: Although k-fold cross-validation provides a more robust performance estimate, it can also introduce higher variability in the results due to different random partitions of the data. This variability may make it harder to compare the performance of different models or make definitive conclusions about the model's performance.

(ii) Compared with the LOOCV, k-fold cross-validation has the following advantages and disadvantages:

Advantages:

(1) Reduced computational cost: LOOCV is computationally expensive since it requires training and evaluating the model as many times as there are data points in the dataset. In contrast, k-fold cross-validation performs the training and evaluation k times, where k is typically much smaller than the total number of data points, resulting in significantly lower computational cost.

(2) Lower variance: LOOCV tends to have higher variance compared to k-fold cross-validation. In LOOCV, each model is trained on almost identical datasets, where only one data point is left out each time. This can lead to overfitting and higher variability in the performance estimates. k-fold cross-validation, by using different subsets for training and validation, provides a more stable estimate of the model's performance.

(3) More representative model assessment: With k-fold cross-validation, the model is evaluated on multiple independent subsets of data, providing a more representative assessment of its generalization performance. It allows for a more comprehensive evaluation of the model's ability to generalize to unseen data by simulating different training and validation set combinations. LOOCV, on the other hand, may be prone to overfitting since each model is trained on almost the same dataset, leaving out only one data point at a time.

Disadvantages:

(1) Less exhaustive exploration of hyperparameters: LOOCV provides a more exhaustive exploration of hyperparameters since each model is trained on almost the entire dataset. In contrast, k-fold cross-validation may not explore the hyperparameter space as thoroughly since each model is trained on a smaller subset of data.

(2) Potential information leakage: Similar to the validation set approach, there is a risk of information leakage between the training and validation sets in k-fold cross-validation when performing certain steps such as feature selection, hyperparameter tuning, or model selection within each fold. This can lead to over-optimistic performance estimates if not handled properly.

In a conclusion, k-fold cross-validation offers advantages such as reduced computational cost, lower variance, and a more representative assessment of model performance compared to LOOCV. However, it has the disadvantage of potential information leakage and may explore the hyperparameter space less exhaustively. The choice between the two techniques mainly depends on the balance between bias and variance.

3

Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

Solution:

When we use statistical learning methods to predict the response Y of predictor X, we can employ resampling techniques such as cross-validation or bootstrap. These techniques help estimate the variability and generalization performance of the model's predictions. Here are some common estimation methods:

(1) Cross-validation method: Use techniques such as k-fold cross-validation or leave-one-out cross-validation (LOOCV). In k-fold cross-validation, divide the training set into k equally sized subsets or "folds." Then, iteratively use k-1 folds for training and the remaining fold for validation/testing. Repeat this process k times, rotating the fold used for validation each time. Calculate the standard deviation of the predictions obtained from each iteration.

(2) Bootstrap method: Bootstrap resampling involves creating multiple bootstrap samples by randomly sampling the original training set with replacement. Each bootstrap sample has the same size as the original training set but may contain duplicate instances. Train the model on each bootstrap sample and calculate the predictions for the validation/test set. Repeat this process a sufficient number of times to obtain a collection of predictions and then calculate the standard deviation of these predictions.

(3) Construct the confidence interval: Construct the confidence interval is to estimate the standard deviation of a prediction based on sample data and confidence level. Specifically, we can use the observations and confidence levels of the sample data to construct confidence intervals, and then use the formula for the confidence interval to estimate the standard deviation of the prediction.

(4) Variance-covariance matrix estimation: Variance-covariance matrix estimation is a method to estimate the standard deviation of a prediction using variances and covariance matrices. Specifically, we can use the variance and covariance matrix of the sample data to estimate the variance and covariance matrix of the population data, and then use the variance and covariance matrix of the population data to estimate the predicted standard deviation. Variance-covariance matrix estimation is usually applicable to nonlinear models and complex models, such as neural networks and decision trees.

In a conclusion, it should be noted that the above methods are not independent of each other and can be used in combination to better estimate the predicted standard deviation.

4

We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. Explain your answers:

(a) Which of the three models with k predictors has the smallest training RSS?

(b) Which of the three models with k predictors has the smallest test RSS?

(c) True or False:

i The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.

ii The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$ variable model identified by backward stepwise selection.

iii The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$ variable model identified by forward stepwise selection.

iv The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

v The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

Solution:

(a) To determine which model with k predictors has the smallest training RSS for each approach, we need to understand how these methods select the models.

Best Subset Selection: In best subset selection, all possible combinations of predictors are evaluated to determine the best subset of predictors for each k. The training RSS is calculated for each model, and the model with the smallest training RSS is chosen as the best model for that particular k. Therefore, among the p + 1 models with different numbers of predictors, the best subset selection will provide the model with k predictors that has the smallest training RSS.

Forward Stepwise Selection: In forward stepwise selection, the algorithm starts with an empty model and iteratively adds predictors one at a time based on their individual performance in reducing the RSS. The training RSS is calculated after each predictor is added. The algorithm continues until the desired number of predictors is reached. Since forward stepwise selection adds predictors incrementally, the model with k predictors obtained from this approach is the one that includes the k best predictors selected during the forward stepwise process. However, there is no guarantee that the model with k predictors from forward stepwise selection will have the smallest training RSS among the p + 1 models.

Backward Stepwise Selection: In backward stepwise selection, the algorithm starts with a model containing all predictors and iteratively removes one predictor at a time based on their individual performance in reducing the RSS. The training RSS is calculated after each predictor is removed. The algorithm continues until the desired number of predictors (k) is reached. The model with k predictors obtained from the backward stepwise selection is the one that includes the k predictors remaining after the backward stepwise process. Similar to the forward stepwise selection, the model with k predictors from backward stepwise selection may not necessarily have the smallest training RSS among the p + 1 models.

In summary, among the three approaches, the best subset selection is the one that guarantees to select the model with k predictors that has the smallest training RSS. Forward stepwise and backward stepwise selection methods may not always identify the model with the smallest training RSS for a specific k. The performance of forward and backward stepwise selection depends on the order of predictor selection and removal, which may not necessarily lead to the best overall model for a given k.

(b) To determine which of the three models with k predictors has the smallest test RSS, we compare the models obtained from each approach directly:

Best Subset Selection: For each value of k, we obtain a model with k predictors. Among these models, we compare the test RSS values and select the one with the smallest test RSS as the best model with k predictors.

Forward Stepwise Selection: For each value of k, we obtain a model by adding predictors sequentially. However, the model with k predictors in forward stepwise selection might not be the same as the best subset selection model with k predictors. Hence, we need to compare the test RSS

of the forward stepwise selection model with the test RSS of the best subset selection model with k predictors.

Backward Stepwise Selection: For each value of k, we obtain a model by removing predictors sequentially. Similarly to forward stepwise selection, the model with k predictors in backward stepwise selection might not be the same as the best subset selection model with k predictors. Hence, we need to compare the test RSS of the backward stepwise selection model with the test RSS of the best subset selection model with k predictors.

In a conclusion, we need to compare the test RSS values of the models obtained through best subset selection, forward stepwise selection, and backward stepwise selection for each value of k. The model with the smallest test RSS among the three approaches for a specific k will have the lowest prediction error and is considered the best model with k predictors.

(c)

i **True:** In forward stepwise selection, predictors are added to the model one at a time based on a specified criterion, such as RSS or R^2. When we move from the k-variable model to the (k + 1)-variable model, we select an additional predictor that improves the model fit the most. Since we are adding predictors, the k-variable model will be a subset of the (k + 1)-variable model.

ii **True:** In backward stepwise selection, predictors are removed from the model one at a time based on a specified criterion, such as RSS or R^2. When we move from the k-variable model to the (k + 1)-variable model, we reintroduce a previously removed predictor that improves the model fit the most. Since we are adding predictors back into the model, the k-variable model will be a subset of the (k + 1)-variable model.

iii **False:** The predictors in the k-variable model identified by backward stepwise selection are not necessarily a subset of the predictors in the (k + 1)-variable model identified by forward stepwise selection. Backward stepwise selection removes predictors, while forward stepwise selection adds predictors. The criteria and order of selection differ between the two approaches, so there is no guarantee that the predictors selected in the k-variable model by backward stepwise will be a subset of the predictors selected in the (k + 1)-variable model by forward stepwise.

iv **False:** The predictors in the k-variable model identified by forward stepwise selection are not necessarily a subset of the predictors in the (k + 1)-variable model identified by backward stepwise selection. Forward stepwise selection adds predictors, while backward stepwise selection removes predictors. The criteria and order of selection differ between the two approaches, so there is no guarantee that the predictors selected in the k-variable model by forward stepwise will be a subset of the predictors selected in the (k + 1)-variable model by backward stepwise.

v **True:** In best subset selection, all possible combinations of predictors are considered, and the model with the best subset of predictors is chosen based on a specified criterion. When we move from the k-variable model to the (k + 1)-variable model, we consider all possible subsets that include the additional predictor. Since we are considering all subsets, the k-variable model will always be a subset of the (k + 1)-variable model. Therefore, the predictors in the k-variable model identified by best subset selection will be a subset of the predictors in the (k + 1)-variable model identified by best subset selection.