MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 5 – Model Selection

Topic 2 – Information Criteria including AIC and BIC

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

1. Model Complexity Scores

2. The Akaike Information Criterion (AIC)

3. The Bayesian Information Criterion (BIC)

Model Complexity Scores

Model Complexity Scores

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that

Model Complexity Scores

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.

Model Complexity Scores

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that:

Model Complexity Scores

MANIPAL ACADEMY *of* HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data

Model Complexity Scores

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data (2) do not suffer from bias due to over-fitting on the training data.

Model Complexity Scores

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data (2) do not suffer from bias due to over-fitting on the training data.
- Model complexity scores typically take the form loss + penalty.

Model Complexity Scores

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data (2) do not suffer from bias due to over-fitting on the training data.
- Model complexity scores typically take the form loss + penalty.
- The loss term captures the goodness of fit of the model to the training data.

Model Complexity Scores

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data (2) do not suffer from bias due to over-fitting on the training data.
- Model complexity scores typically take the form loss + penalty.
- The loss term captures the goodness of fit of the model to the training data.
- The penalty term accounts for adding more predictors to the model which

Model Complexity Scores

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- A major drawback in using prediction error on test data estimated using cross validation for model selection is that it can be *computationally expensive*.
- An alternative is to use *model complexity scores* which are measures of performance that: (1) depend only on the training data (2) do not suffer from bias due to over-fitting on the training data.
- Model complexity scores typically take the form loss + penalty.
- The loss term captures the goodness of fit of the model to the training data.
- The penalty term accounts for adding more predictors to the model which balances the loss term which always decreases as more predictors are added to the model.

The Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and

The Akaike Information Criterion (AIC)

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.
- The AIC approach tends to work well for underlying population models that are actually complex.

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}} .$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.
- The AIC approach tends to work well for underlying population models that are actually complex.
- The resulting models typically

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.
- The AIC approach tends to work well for underlying population models that are actually complex.
- The resulting models typically (1) have a large number of predictors

The Akaike Information Criterion (AIC)

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)
INSPIRED BY LIFE

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}} .$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.
- The AIC approach tends to work well for underlying population models that are actually complex.
- The resulting models typically (1) have a large number of predictors (2) have good predictive ability

The Akaike Information Criterion (AIC)

- The AIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{2p}_{\text{penalty}}.$$

- The first term involves the RSS which is a measure of fit of the model and a constant term which plays no role in model selection.
- The second term is the penalty term for adding more predictors to the model.
- The AIC approach tends to work well for underlying population models that are actually complex.
- The resulting models typically (1) have a large number of predictors (2) have good predictive ability (3) have low interpretability.

The Bayesian Information Criterion (BIC)

MANIPAL
ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

The Bayesian Information Criterion (BIC)

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

- The only difference from AIC is in the last term which is $\log($number of samples$)$ here.

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

- The only difference from AIC is in the last term which is $\log($number of samples$)$ here.
- The BIC approach tends to work well for underlying population models that are actually simple.

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

- The only difference from AIC is in the last term which is $\log($number of samples$)$ here.
- The BIC approach tends to work well for underlying population models that are actually simple.
- The resulting models typically

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as
$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$
- The only difference from AIC is in the last term which is $\log(\text{number of samples})$ here.
- The BIC approach tends to work well for underlying population models that are actually simple.
- The resulting models typically (1) have a small number of predictors

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

- The only difference from AIC is in the last term which is $\log($number of samples$)$ here.
- The BIC approach tends to work well for underlying population models that are actually simple.
- The resulting models typically (1) have a small number of predictors (2) have optimal predictive ability

The Bayesian Information Criterion (BIC)

- The BIC for a model built using n samples and p predictors is defined as

$$\underbrace{n\log(RSS) - n\log(n)}_{\text{loss}} + \underbrace{\log(n)p}_{\text{penalty}}.$$

- The only difference from AIC is in the last term which is $\log($number of samples$)$ here.
- The BIC approach tends to work well for underlying population models that are actually simple.
- The resulting models typically (1) have a small number of predictors (2) have optimal predictive ability (3) have good interpretability.

Summary

- Necessity of model complexity scores
- Differences between AIC and BIC for model selection