Choosing the Optimal Model

All the selection methods discussed in Subset Selection results in a set of models for each k number of predictors p, and the last step of each of the algorithms were to select the **best** model among them.

In order to select the best model with respect to **test error**, we need to have an estimate of the test error using these two common approaches :

- Indirectly estimate test error by making an **adjustment** to the **training error** accounting for the bias due to Overfitting
- Estimating the test error directly, using either a **validation set** or Cross-Validation method

In this section we will discuss the following :

- Mallow's C_p
- AIC (Akaike information criterion)
- BIC (Bayesian information criterion)
- Adjusted R^2
- Validation set
- Cross-Validation

Indirect Estimate of Test Error Methods

It's known from previous chapters that's both R^2 and RSS are primarily used to fit a model to the training data, Since the least squares estimate the coefficients such that RSS is a small as possible

The following methods adjust the training error to indirectly estimate the test error :

Mallow's C_p

for a fitted least squares model containing d predictors , the C_p estimate of test error MSE is computed using :

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma^2}),$$

Where :

- $\hat{\sigma^2}$ is an estimate of the variance of the irreducible error ε for each Response
- $\hat{\sigma^2}$ is often estimated using the full model $\mathcal{M}_p$
- C_p adds penalty of $2d\hat{\sigma}^2$ to the training RSS to adjust for the fact that training error always underestimate the test error
- The penalty increases as the number of predictors d increase
- Since the RSS decrease the more predictors added
- $\hat{\sigma^2}$ is an unbiased estimate of σ^2 which makes C_p an unbiased estimate of the test error

Adjusted R^2

Recall in both Simple Linear Regression and Multiple Linear Regression we used the R^2 which is a measure for how much our model explained the data, the closer to one the better the goodness fit, in the R^2 the more predictors/variables added the more it increases **adjusted** R^2 fix that by :

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

- Adding d in the denominator penalize the increase of the number of predictors
- Adding predictors when they don't explain the variance in the data will decrease the values of the R^2 that's why it's called $\text{Adjusted } R^2$

AIC and BIC

While the adjusted R^2 and Mallow's C_p are exclusive for regression problems, AIC and BIC are more generalized for both **Classification and Regression** problems, they are based on the **log-likelihood** of the model ℓ, for both lower values indicates a model with low teste error

AIC

The AIC rewards models that fit the data well while penalizing unnecessary complexity, given by :

$$AIC = -2\log(\ell) + 2d$$

Where :

- ℓ is the likelihood function of the model in the case of

While the AIC provides useful measures to select a model, but it tend to favor complex models especially when dealing with smaller data sets, that's why a Adjusted AIC as devloped to account for potential bias :

$$AIC_c = AIC + \frac{2 * d * (d+1)}{n-k-1}$$

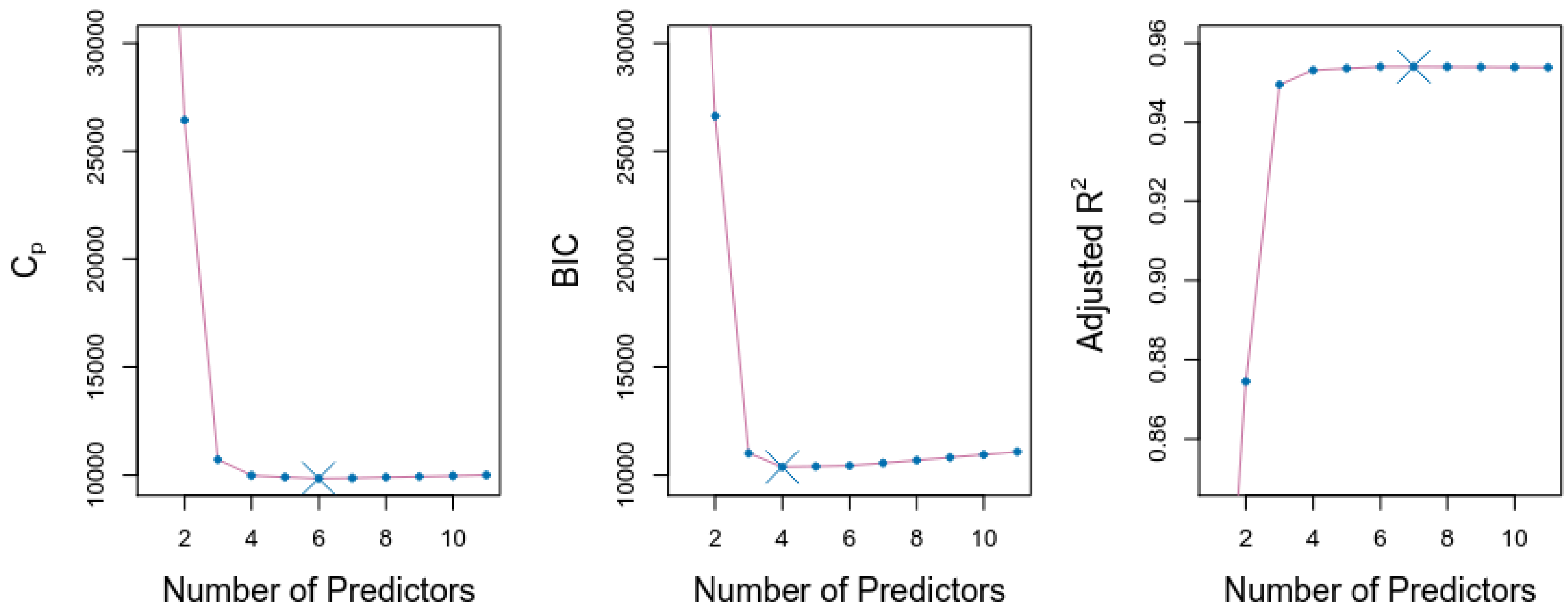
BIC

The BIC is stricter than AIC where it penalizes complexity, which favor simpler models since it aligns with Bayesian inference, given by :

$$BIC = -2\log(\ell) + d\log(n)$$

Where :

- n is the number of observations, Which make the **penalty term** increase with the size of the dataset



- Three plots showing the **Mallow's** C_p, BIC , $\text{Adjusted } R^2$

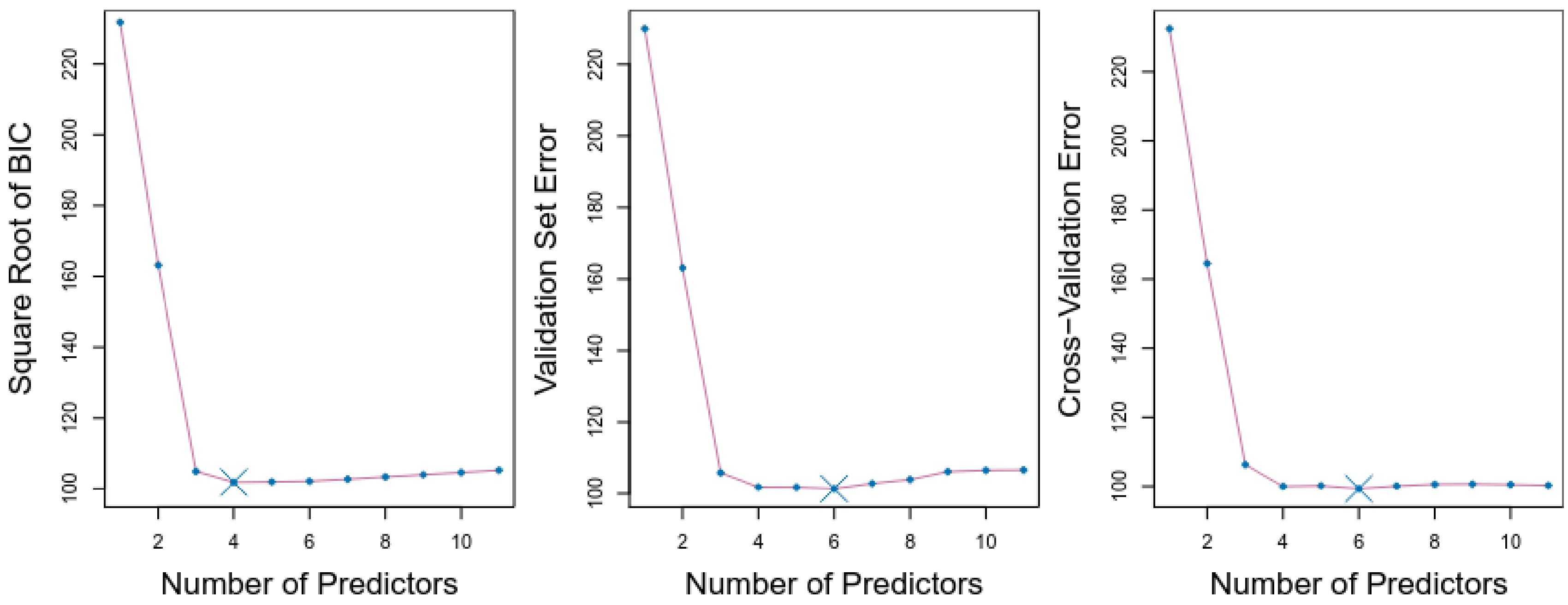
Direct Estimate of Test Error Methods

Instead of adjusting the training error for the test error the following methods discuss in this section estimate the test error directly :

- Validation Set
- Cross-Validation

The **Cross-Validation** focuses more into finding the best number of predictors k more than the best among each $\mathcal{M}_k$ (the exact subset)

- the error for each training fold
- the validation errors are averaged over all the folds for each model size k
- It calculate the test error for each model $\mathcal{M}_o$
- And then the best model size k is chosen on the full data


Square Root of BIC
Validation Set Error
Cross−Validation Error
Number of Predictors
Number of Predictors
Number of Predictors
100
120
140
160
180
200
220
2
4
6
8
10