MANCHESTER 1824

The University of Manchester

Model Assessment and Selection (II)

Ke Chen

Reading: Sects. 2.2, 5.2, 6.1 [Intro Stat Learn Python]

https://www.statlearning.com/

Lecture Goal

- Understanding motivations/ideas behind analytical methods
- Generic analytical methods: AIC, BIC and their implications
- Specific analytical methods used in regression
- Case study: feature (variable) selection in regression

Introduction

- To attain the ultimate goal of statistical learning, model assessment and selection rely on the performance of learning model(s) on unseen data but only a data set, i.e., a specific sample of the unknown population, is available during learning.

- Empirical methodology (held-out validation, cross-validation) addresses the issue by "simulating" a training-test scenario but suffers from a high computational burden.

- A general observation suggests that for a learning model, its training error ($\overline{\text{err}}$) is (nearly always) lower than its test one (Err) ; i.e., $\overline{\text{err}} < \text{Err}$.

- The bias-variance trade-off implies that test errors are determined by not only the properties of a given data set (e.g., sample size and "quality" on how informative it is) but also model complexity (flexibility or capacity of a learning model).

Analytical Methodology

- Analytical methods for model assessment and selection are yet another manner to measure test errors directly based on training errors estimated on a given data set.

- Motivated by the general observation, an analytical method has a generic form:

  test-error = training-error + penalty

- While "training-error" is estimated via a loss function on training data, "penalty" is a term reflecting the model complexity by applying the principle of Occam's razor: "prefer the simplest model that describes the data sufficiently well". Thus, the penalty term always penalises a model of higher complexity for the same/similar training error.

- There are miscellaneous "penalty" forms rooted from different theories/heuristics:
  - Akaike's information criterion (AIC) and Bayesian information criterion (BIC)
  - Mallow's C_p and adjusted R^2 also developed specifically for regression tasks

Akaike's Information Criterion (AIC)

- AIC is a generic analytical method to estimate test error based on training error, which was proposed by Hirotugu Akaike in 1973.
- The original problem setting was that finding the best one from a set of candidate probabilistic models to describe the given data subject to a unknow distribution.
- Akaike solved this problem by an approximation of Kullback-Liebler (KL) divergence between a candidate model and the "true" model, which takes into account both under-fitting and over-fitting risks in evaluation.
- As the original problem was set regarding probabilistic models and information theory, any learning models have to be (re)formulated from a probabilistic perspective so that AIC can be applied for model assessment and selection in machine learning.

Akaike's Information Criterion (AIC)

- The AIC score is in the following form:

$$\mathrm{AIC}[M(\Theta)] = -2\log(L[M(\hat{\Theta})]) + 2d(\Theta)$$

  - $M(\Theta)$: learning model with a collection of parameters, Θ
  - $L[M(\hat{\Theta})]$: likelihood function of $M(\Theta)$ by setting $\Theta = \hat{\Theta}$ (optimum)

    $\hat{\Theta}$ is achieved via the maximum likelihood estimate
  - $d(\Theta)$: the number of parameters in Θ
- AIC can directly approximate test error and prefers the model of the minimum AIC score in model selection.
- For AIC to be valid, there must be more training data than the number of parameters; i.e., $n > d(\Theta)$.

Bayesian Information Criterion (BIC)

- BIC is yet another generic analytical method to estimate test error based on training error, which was proposed by Gideon E. Schwarz in 1978.

- BIC provides an alternative yet more generic method from the Bayesian perspective to tackle the same problem encountered by AIC, which leads to a Bayesian method for model selection and model averaging (a soft version of model selection).

- Unlike AIC that approximates the KL divergence between a model and the "true" model, BIC uses the Laplace approximation to obtain an alternative solution.

- As the BIC was derived with the Bayesian treatment of probabilistic modelling, any learning models have to be (re)formulated from a probabilistic perspective so that BIC can be applied for model assessment and selection and even more.

Bayesian Information Criterion (BIC)

- The BIC score is in the following form:

$$\mathrm{BIC}[M(\Theta)] = -2\log(L[M(\widehat{\Theta})]) + d(\Theta)\log(n)$$

  - $M(\Theta)$: learning model with a collection of parameters, Θ
  - $L[M(\widehat{\Theta})]$: likelihood function of $M(\Theta)$ by setting $\Theta = \widehat{\Theta}$ (optimum)

    $\widehat{\Theta}$ is achieved via the maximum likelihood estimate
  - $d(\Theta)$: the number of parameters in Θ
  - n: the number of i. i. d. training data points

- The main difference between AIC and BIC is that BIC takes into account both model complexity and the number of training data in its penalty!

Bayesian Perspective of Model Selection

- Given a set of K candidate learning models $M_m = M(\Theta_m),\ m = 1, \cdots, K$
- Posterior probability of a model is defined with the Bayesian rule:

$$\Pr(M_m|\mathcal{Z}) = \frac{\Pr(M_m)\Pr(\mathcal{Z}|M_m)}{\Pr(\mathcal{Z})}, \qquad \text{where } \mathcal{Z} = \{X, Y\} \text{ is the training data set}$$

- To compare two models, form the posterior odd:

$$\frac{\Pr(M_m|\mathcal{Z})}{\Pr(M_i|\mathcal{Z})} = \frac{\Pr(M_m)}{\Pr(M_i)} \frac{\Pr(\mathcal{Z}|M_m)}{\Pr(\mathcal{Z}|M_i)}$$

If the odd $\frac{\Pr(M_m|\mathcal{Z})}{\Pr(M_i|\mathcal{Z})} > 1$, then choose M_m, where $\frac{\Pr(\mathcal{Z}|M_m)}{\Pr(\mathcal{Z}|M_i)}$ is Bayesian factor.

- As $\Pr(M_m)$ for $m = 1, \cdots, K$ is constant (under the assumption of uniform distribution), the odd is hence decided by Bayesian factor. Thus, $\Pr(\mathcal{Z}|M_m)$ has to be estimated.

Bayesian Perspective of Model Selection

- In Bayesian framework, the parameters are treated as random variables. Hence, there is always a prior distribution for the parameters of each model M_m; $\Pr(\Theta_m|M_m)$.
- By considering the prior of the parameters for model M_m,

$$\Pr(\mathcal{Z}|M_m) = \int \Pr(\mathcal{Z}, \Theta_m|M_m)d\Theta_m = \int \Pr(\mathcal{Z}|\Theta_m, M_m)\Pr(\Theta_m|M_m)d\Theta_m$$

- Applying the Laplace approximation to the integral along with other simplification:

$$\log[\Pr(\mathcal{Z}|M_m)] \approx \log\left(L\left[M\left(\widehat{\Theta}\right)\right]\right) - d(\Theta)\log(n)/2$$

$$\mathrm{BIC}[M(\Theta)] = -2\log[\Pr(\mathcal{Z}|M_m)] = -2\log\left(L\left[M\left(\widehat{\Theta}\right)\right]\right) + d(\Theta)\log(n)$$

- Hence, the ratio between two BIC estimates provides an approximation to Bayes factor and the posterior odd subsequently with considerably much less computational effort.

Bayesian Perspective of Model Selection

- Choosing the model with the minimum BIC is equivalent to choosing the model with the largest (approximate) log-likelihood and subsequently the posterior probability.
- Apart from model selection, the Bayesian framework can offer us more beyond the ranking; based on the BIC scores, BIC_m, estimated for $M_m = M(\Theta_m),\ m = 1, \cdots, K$, the posterior probability of each model can be estimated by

$$\widehat{\Pr}(M_m|\mathcal{Z}) = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_m)}{\sum_{k=1}^{K}\exp(-\frac{1}{2}\mathrm{BIC}_k)}, m = 1, \cdots, K.$$

- The posterior probabilities of candidate models, $\widehat{\Pr}(M_m|\mathcal{Z})$, $m = 1, \cdots, K$, may be the "proper" weights for model averaging (soft model selection manner using all models)

$$\hat{F}(\boldsymbol{x}) = \sum_{m=1}^{K} \widehat{\Pr}(M_m|\mathcal{Z}) \times \hat{f}\left(\boldsymbol{x}; \widehat{\Theta}_m\right).$$

Analytical Method for Regression

- Regression setting: $Y = f(X) + \varepsilon;\ E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$
- Learning an approximation of $f(X)$: $\hat{f}(X, \Theta)$ based on a training dataset $\mathcal{Z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$
- The loss function for learning: residual sum of squares (RSS):

$$\mathrm{RSS}(Y, \hat{f}(X, \Theta)) = \sum_{i=1}^n [y_i - \hat{f}(\boldsymbol{x}_i, \Theta)]^2$$

- From a probabilistic perspective, the regression setting amounts to assuming the system error is subject to Gaussian distribution: $N(0, \sigma_\epsilon^2)$. The log-likelihood function is

$$\log[L(\Theta|\mathcal{Z})] = -\frac{n}{2}\left[\log(2\pi) + \log\sigma_\epsilon^2\right] - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n [y_i - \hat{f}(\boldsymbol{x}_i, \Theta)]^2$$

- As σ_ϵ^2 is constant, minimising RSS is equivalent to maximising log-likelihood (via MLE).
- Thus, for a trained model, the RSS error on training set can be used to replace the negative log-likelihood in AIC and BIC; i.e., $\boldsymbol{\mathrm{RSS}(\hat{\Theta}) \approx -2\log\left(L\left[M(\hat{\Theta})\right]\right)}$.

Analytical Method for Regression

- For regression on a training data set of n points, AIC and BIC are

$$\mathrm{AIC}[M(\Theta)] = \frac{1}{n}[\mathrm{RSS} + 2d(\Theta)];\ \mathrm{BIC}[M(\Theta)] = \frac{1}{n}[\mathrm{RSS} + d(\Theta)\log(n)].$$

- Mallow's C_p

$$C_p = \frac{1}{n}[\mathrm{RSS} + 2d(\Theta)\hat{\sigma}_\varepsilon^2], \quad \hat{\sigma}_\varepsilon^2 = \mathrm{RSS}/(n-2).$$

- Adjusted R^2

$$\mathrm{AR}^2 = 1 - \frac{\mathrm{RSS}/(n-d(\Theta)+1)}{\mathrm{TSS}/(n-1)}, \quad \mathrm{TSS} = \sum_{i=1}^{n}[y_i - \bar{y}]^2.$$

Unlike others, adjusted R^2 prefers the model of the largest score in model selection.

Feature (Variable) Selection in Regression

- For regression, there are often a collection of natural features (variables/predictors) that might be associated with a response. However, it is possible that only a subset of features actually affect (are closely related to) the response in regression.

- In this case, the use of all features may lead to poor generalisation, while carefully selected feature subset can improve model accuracy for unseen (test) data.

- Feature selection is a task of identifying a subset of features from a collection of features so that a regression model based on the subset of selected features can achieve good generalisation as well as model interpretation (for linear regression)

- Methods: sequential best subset selection and forward/backward stepwise selection

Feature (Variable) Selection in Regression

- Best subset selection

For each possible combination of the p predictors (features):

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.

Feature (Variable) Selection in Regression

- Best subset selection
  - The best subset selection has to search through 2^p models. For computational reasons, it cannot be applied with very large p (applicable for $p < 40$ in practice).
  - An enormous search space could also lead to overfitting and high variance of the parameter (coefficient) estimates.
  - More attractive methods
    - Forward Stepwise Selection

      Begins with a null regression model containing no predictors (features), and then adds one predictor (feature) at a time that improves the model the most until no further improvement is possible.
    - Backward Stepwise Selection

      Begins with a full regression model containing all predictors (features), and then deletes one predictor (feature) at a time that improves the model the most until no further improvement is possible.

Feature (Variable) Selection in Regression

- 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.

Feature (Variable) Selection in Regression

- Backward stepwise selection

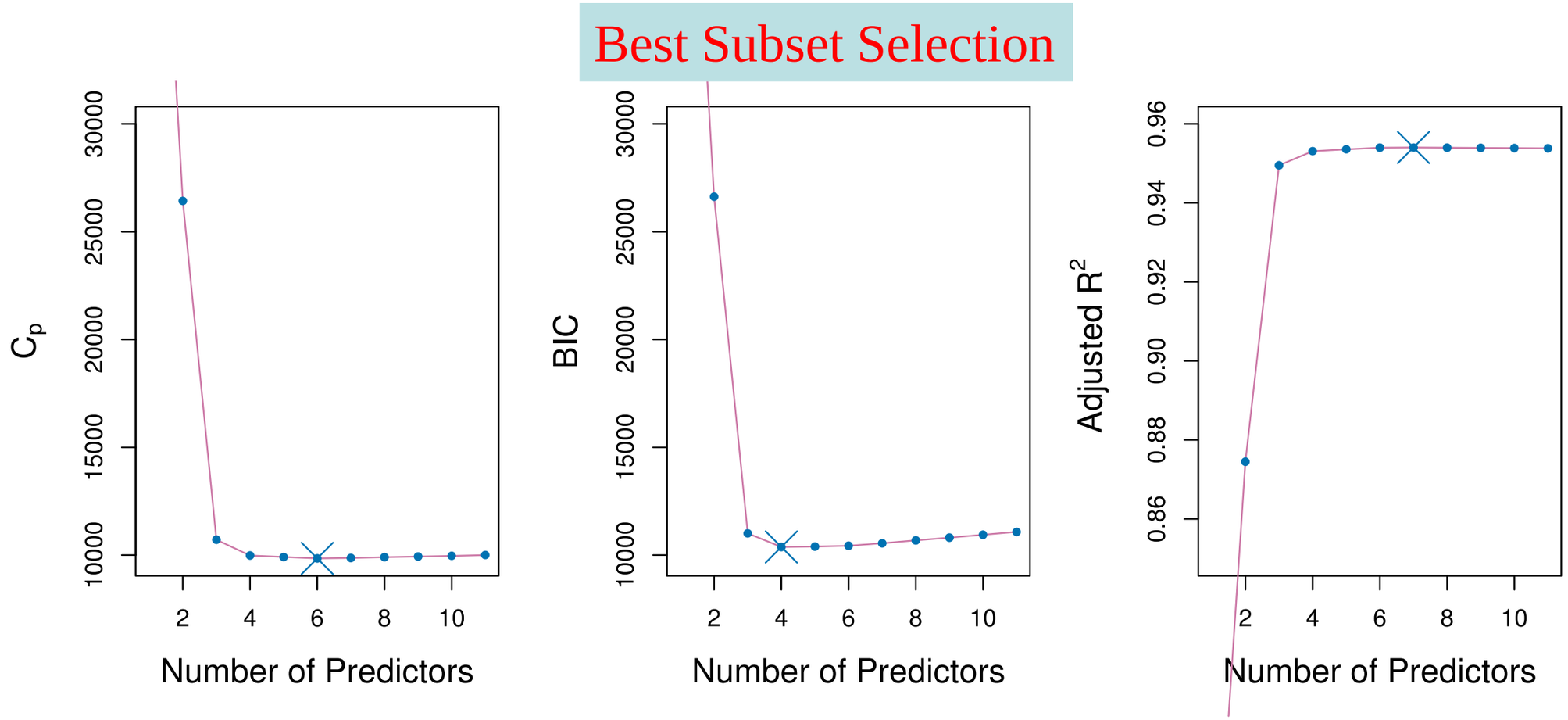
1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   - 2.1 Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   - 2.2 Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. 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.

Feature (Variable) Selection in Regression

- Stepwise selection (cont.)
  - Both forward and backward stepwise selection approaches search through only $1+p(p+1)/2$ models, so they can be applied in the feature selection problem where p is too large to apply best subset selection.
  - Both of these stepwise selection methods are not guaranteed to yield the best model containing a subset of the p predictors (features) due to the non-exhaustive search.
  - Forward stepwise selection can be used even when $n < p$, while backward stepwise selection requires that $n > p$, where n is the number of training examples.
  - There is a hybrid version of these two stepwise selection methods to overcome the weakness in forward and backward stepwise selection.

Feature (Variable) Selection in Regression

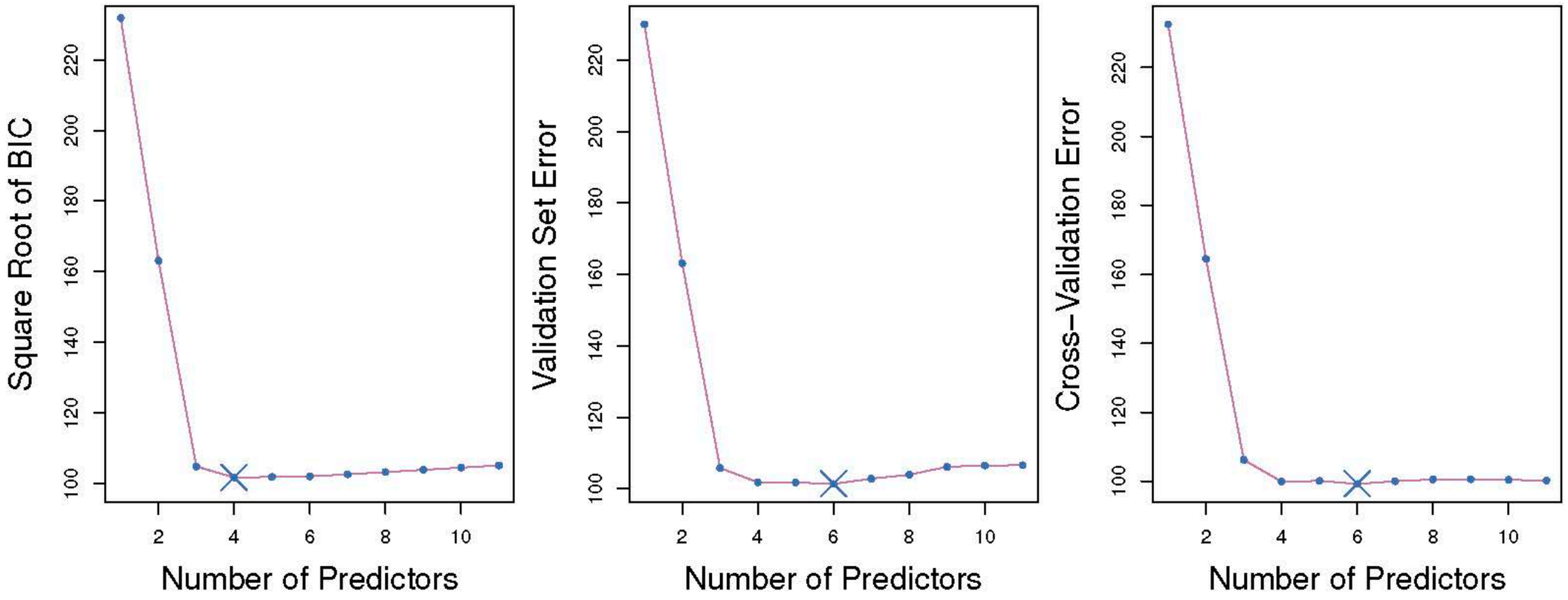
- Example: predicting credit based on 11 features, Sect. 3.3 in ISLR



Feature (Variable) Selection in Regression

- Example: predicting credit based on 11 features, Sect. 3.3 in ISLR

Best Subset Selection



Summary

- The analytical methods provide an alternative yet computationally efficient manner via the direct use of a "penalised" training error.
  - General methods: AIC and BIC
  - Regression: Mallow's C_p and adjusted R^2
- By means of BIC, model selection can be done from a Bayesian perspective.
- Robust estimate of test errors with analytical methods can be done via the bootstrap, yet another generic re-sampling technique.
- Feature (variable/predictor/subset) selection by means of analytical methods could improve generalisation as well as model interpretation in linear regression.
  - best subset selection: computationally expensive and only work for $p < 40$
  - stepwise selection: forward stepwise selection vs. backward stepwise selection; computationally tractable but no guarantee for finding out the "optimal" subset.