Statistical Learning for Data Science

Lecture 18

唐晓颖
电子与电气工程系
南方科技大学

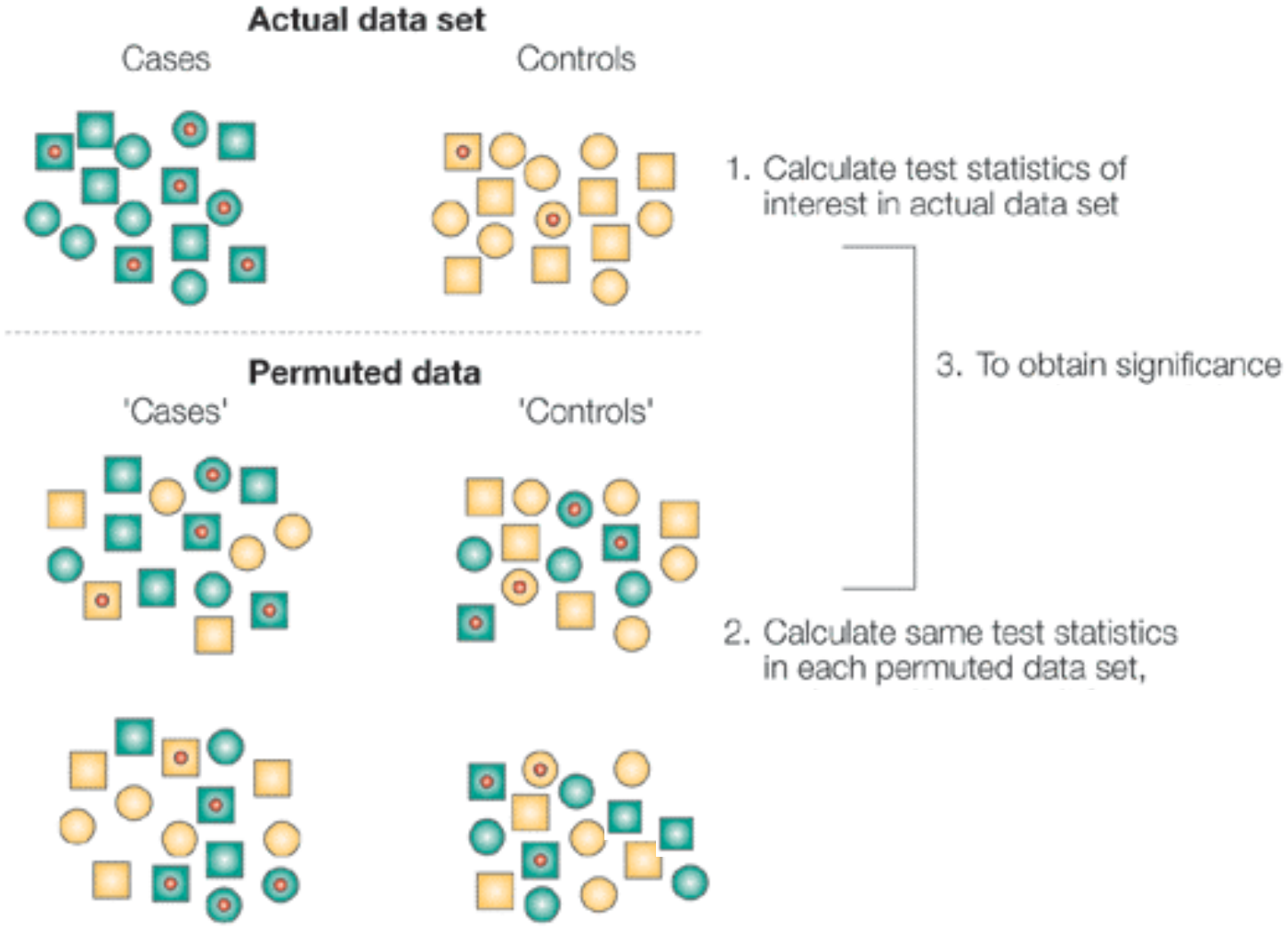
May 15, 2023

Permutation Test

A permutation test (also called a randomization test, re-randomization test, or an exact test) is a type of **statistical significance** test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

- Another form of resampling, but in this case it is done *without replacement*.
- They have been around since Fisher introduced them in the 1930's.
- Often used to conduct hypothesis testing.
- Rather than assume a distribution for the null hypothesis, we simulate what it would be by randomly reconfiguring our sample lots of times (e.g. 1000) in a way that “breaks” the relationship in our sample data.

Permutation Test



Permutation Test

- Suppose we have two groups A and B whose sample means are $\bar{x}_A$ and $\bar{x}_B$, and that we want to test, at a 5% significance level, whether they come from the same distribution. Let n_A and n_B be the sample size corresponding to each group.
- The permutation test is designed to determine whether the observed difference between the sample means is large enough to reject the null hypothesis H_0 that the two groups have identical probability distribution.

Permutation Test

1. The difference in means between the two samples is calculated: this is the observed value of the test statistic, $T(obs)$.

2. The observations of groups A and B are pooled.

3. The difference in sample means is calculated and recorded for every possible way of dividing these pooled values into two groups of size n_A and n_B (i.e., for every permutation of the group labels A and B). The set of these calculated differences is the exact distribution of possible differences under the null hypothesis that group label does not matter.

4. The one-sided p-value of the test is calculated as the proportion of sampled permutations where the difference in means was greater than or equal to $T(obs)$. The two-sided p-value of the test is calculated as the proportion of sampled permutations where the absolute difference was greater than or equal to $\left|T(obs)\right|$.

Permutation Test

- Monte Carlo testing
  - In most cases, there are too many possible orderings of the data to allow complete enumeration in a convenient manner.

    The total number of possible permutations by dividing the pooled values into two groups of size n_A and n_B is:

$$\frac{(n_A + n_B)!}{n_A! n_B!}$$

    Suppose we have a sample of $n_A = 14$ and $n_B = 12$. If so, then we have (14+12)!/14!12! = 9657700 possible permutations.

  - Thus, in most settings, we randomly generate some of the possible permutations (10000 or 40000) and call it good. This is usually called randomization test.

Permutation Test

- Monte Carlo testing
  - Thus, in most settings, we randomly generate, via Monte Carlo sampling, some of the possible permutations (10000 or 40000) and call it good.

1. Define a domain of possible inputs $(1, 2, 3, ..., n_A + n_B)$ → Index of subjects in the pooled group
2. Generate inputs randomly from a probability distribution (uniform distribution) over the domain $(1, 2, 3, ..., n_A + n_B)$.
3. Repeat step 2 for 10000 or 40000 times.

Linear Model Selection and Regularization

Goal of this chapter:

Discuss some ways in which the simple linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures.

Linear Model Selection and Regularization

- Why consider alternatives to least squares?
  - Prediction accuracy

    When $n \gg p$, the least squares estimates tend to perform well on test data.

    When $n < p$

    There can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training

    There is no longer a unique least squares coefficient estimate

Linear Model Selection and Regularization

- Why consider alternatives to least squares?
  - Model interpretability
    - Some or many of the variables used in a multiple regression model are in fact no associated with the response.
    - Including those *irrelevant* variables leads to unnecessary complexity in the resulting model.
    - By removing these variables—that is, by setting the corresponding coefficient estimates to zero– we can obtain a model that is more easily interpreted.
    - We will present some approaches for automatically performing *feature selection* (variable selection)—that is, for excluding irrelevant variables from a multiple regression model.

Linear Model Selection and Regularization

- Three classes of alternative methods
  - *Subset selection*: identify a subset of the p predictors that we believe to be related to the response, and then fit a model using least squares on the reduced set of variables.
  - *Shrinkage*: fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as **regularization**) has the effect of reducing variance and can also perform variable selection.
  - *dimension reduction*: project the p predictors into a M-dimensional subspace, where M<p. This is achieved by computing M different **linear combinations**, or **projections**, of the p variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Subset Selection

- Best subset selection

  We fit a separate least squares regression for each possible combination of the p predictors. That is, we fit all p models that contain exactly one predictor, all $\binom{p}{2} = \frac{p(p-1)}{2}$ models that contain exactly two predictors, and so forth.

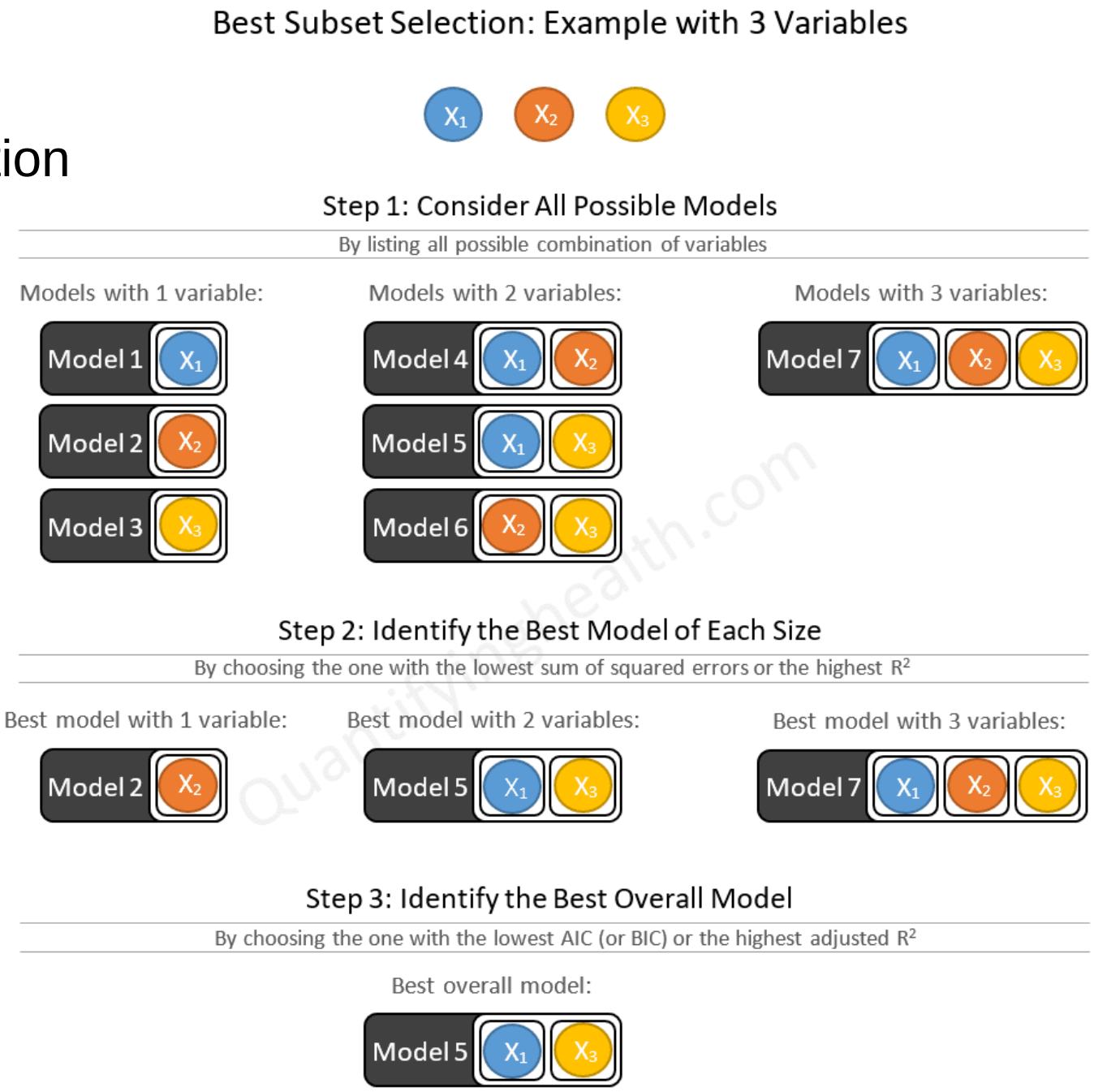
  We then look at all of the resulting models, with the goal of identifying the one that is best.

  Question: how many possible models we will need to test?

  Answer: 2^p

Subset Selection

- Best subset selection



Subset Selection

- Best subset selection

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.

Subset Selection

- Best subset selection

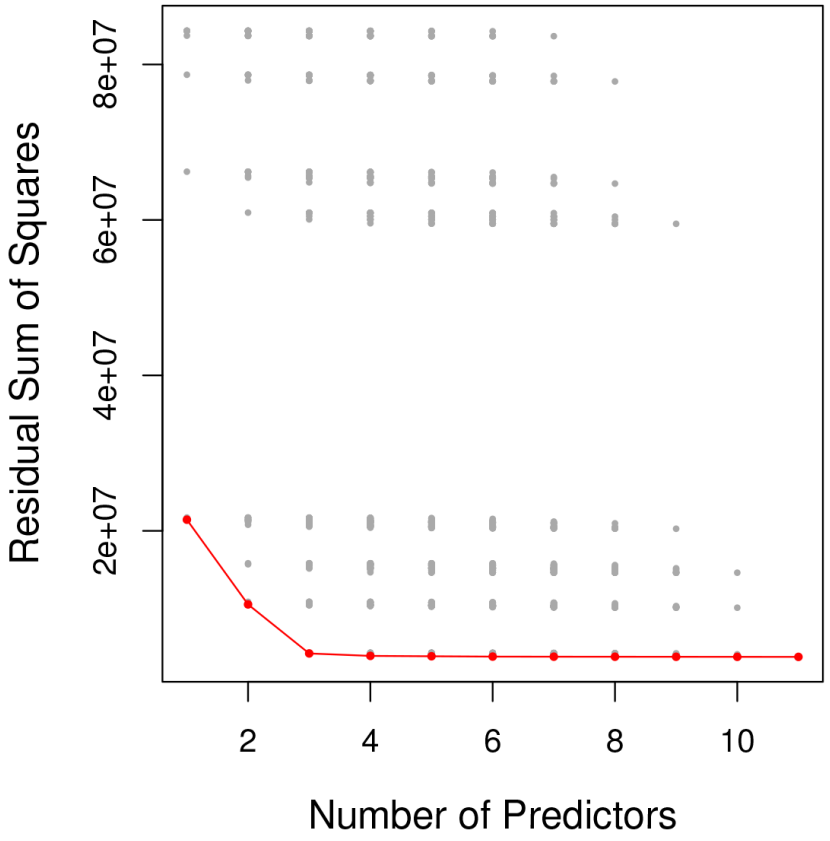
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

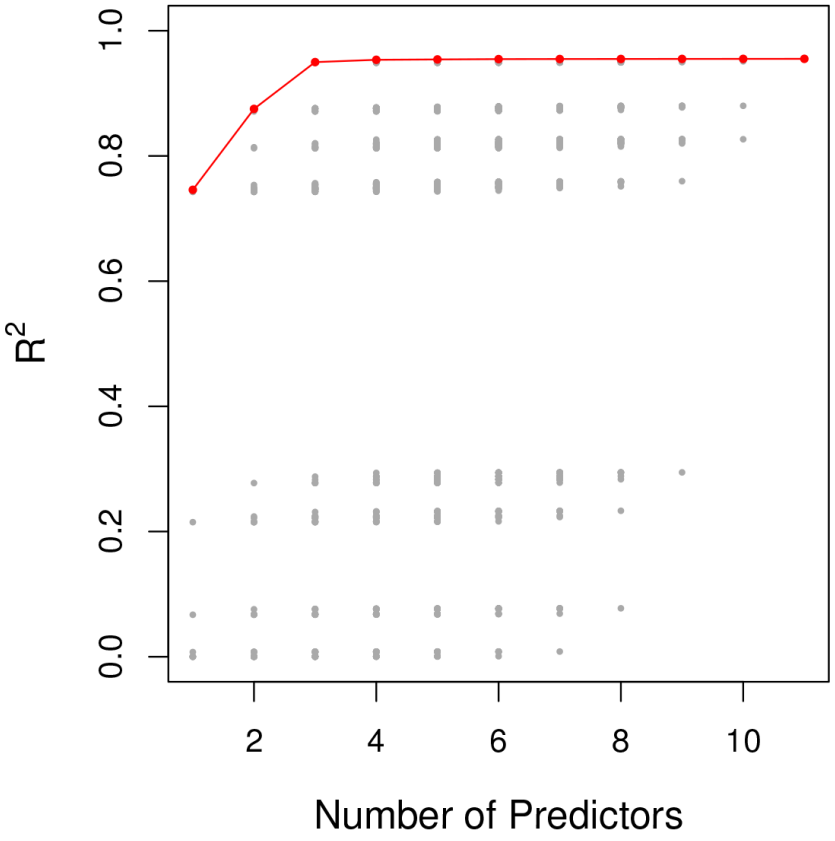
Step 2 identifies the best model (on the training data) for each subset size.

Subset Selection

- Best subset selection

Step2 identifies the best model (on the training data) for each subset size.





Subset Selection

- Best subset selection

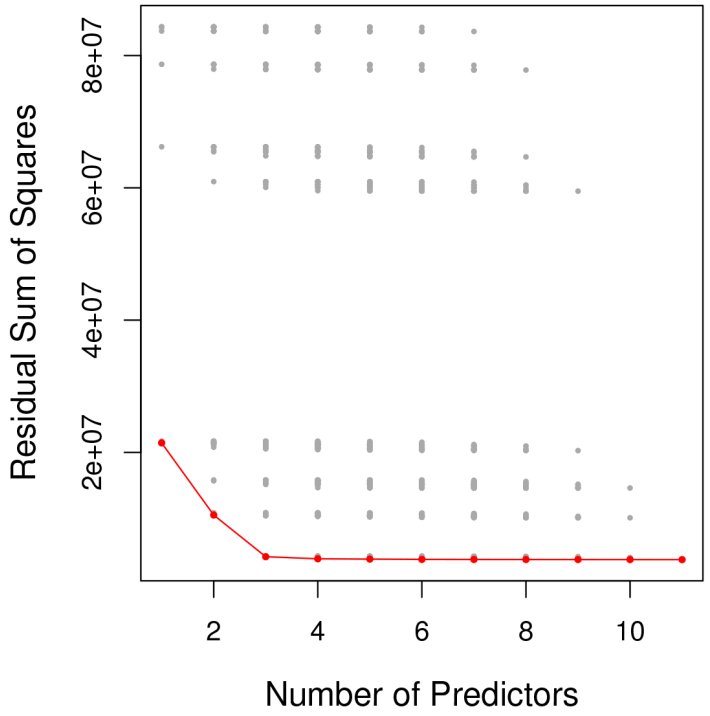
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

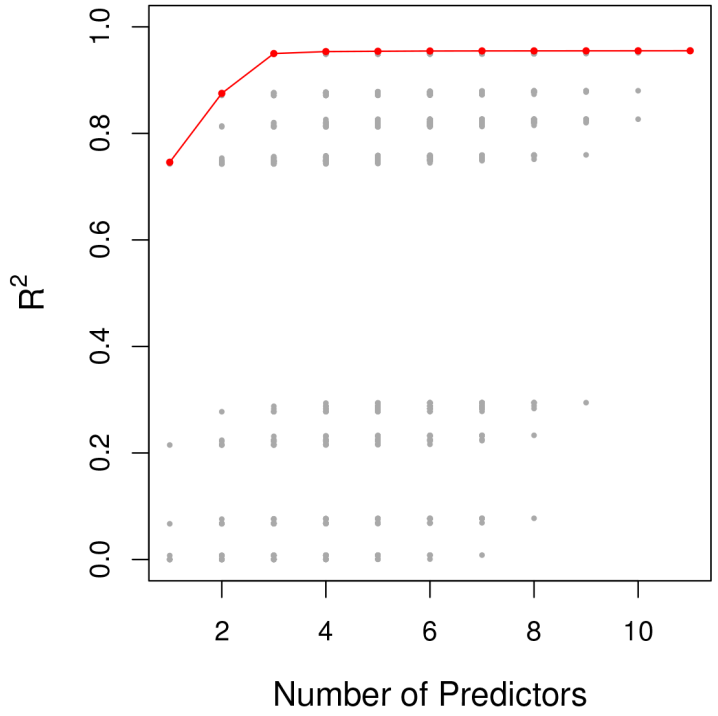
Why not select the single best model using RSS or R^2, as we did in step 2?

Subset Selection

- Best subset selection

Why not select the single best model using RSS or R^2, as we did in step 2?





We will always end up with a model involving all the variables.

A low RSS or a high R^2 indicates a model with a low *training error*, whereas we wish to choose a model that has a low *test error*.

Subset Selection

- Best subset selection

  In addition to least squares regression, the best subset selection applies to other types of models, such as logistic regression.

  In logistic regression, instead of ordering models by RSS in Step 2, we instead use the **deviance**; the deviance is negative two times the maximized log-likelihood. The smaller the deviance, the better the fit.

Subset Selection

- Best subset selection

Pro:

Simple and conceptually appealing.

Con:

When p is large, the best subset selection is computationally expensive.

When p is large, the best subset selection may also suffer from statistical problems. The larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data. Thus an enormous search space can lead to overfitting and high variance of the coefficient estimates.

Subset Selection

- Stepwise selection

  Given the various limitations of the best subset selection approach, especially when p is large, the **stepwise** methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Subset Selection

- Stepwise selection

  Given the various limitations of the best subset selection approach, especially when p is large, the **stepwise** methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

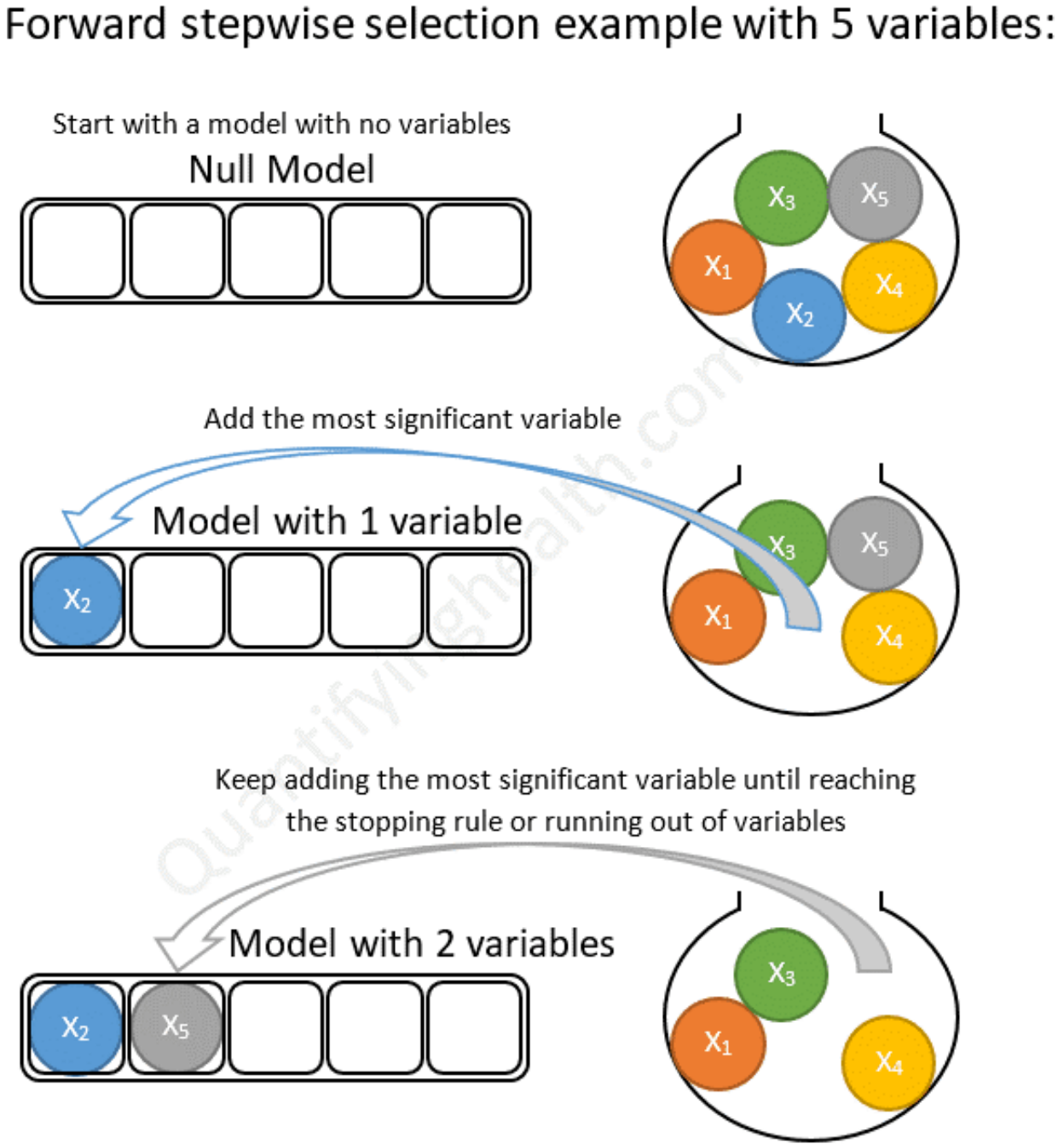
  - Forward stepwise selection
  - Backward stepwise selection
  - Bidirection stepwise selection

Subset Selection

- Forward stepwise selection
  - Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
  - In particular, at each step the variable that gives the greatest *additional* improvement to the fit is added to the model.

Subset Selection

- Forward stepwise selection



Subset Selection

- 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.

Subset Selection

- Forward stepwise selection

Computational advantage over best subset selection is clear.

Forward stepwise selection involves fitting one null model, along with p-k models in the kth iteration, for k=0,1,2,…,p-1. This amounts to a total of $1+\sum_{k=0}^{p-1}(p-k)=1+p(p+1)/2$ models.

The best subset selection involves fitting 2^p models.

This is a substantial difference: when p=20, best subset selection requires fitting 1,048,576 models, whereas forward stepwise selection requires fitting only 211 models.

Subset Selection

- Forward stepwise selection

  Though forward stepwise tends to do well in practice, it is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.

  Example: suppose that in a given data set with p=3 predictors, the best possible one-variable model contains X_1, and the best possible two-variable model instead contains X_2 and X_3 . Then forward stepwise selection will fail to select the best possible two-variable model. **Why?**

Subset Selection

- Forward stepwise selection

Credit data example

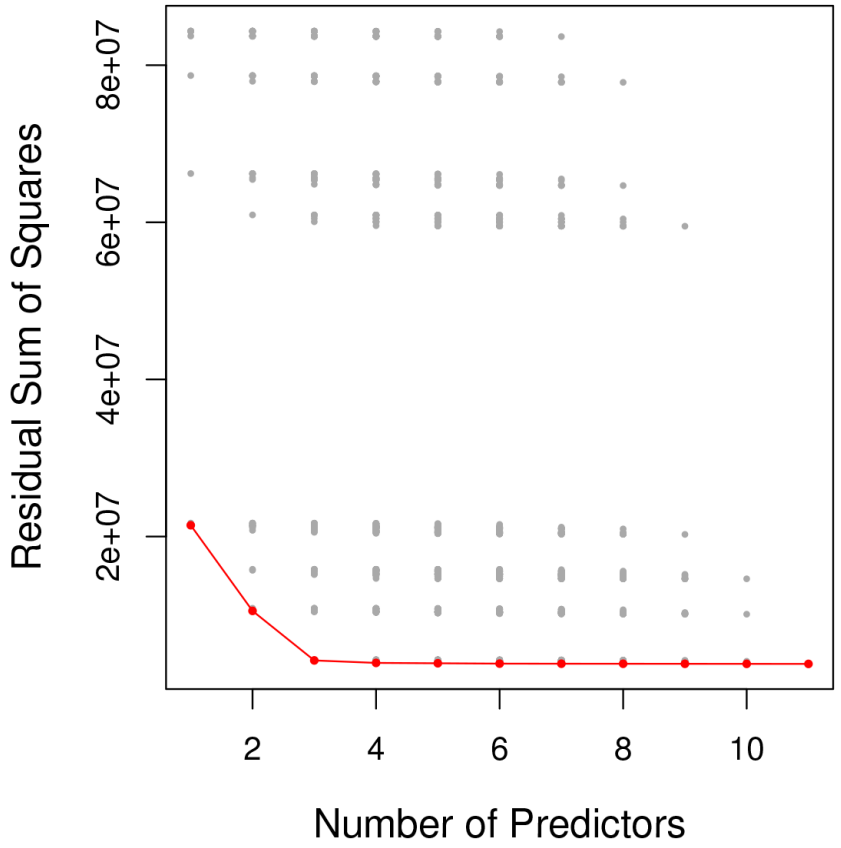
# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income	rating, income,
	student, limit	student, limit

Comment: The first four selected models for best subset selection and forward stepwise selection on the Credit data set. The first three models are identical but the fourth models differ.

Subset Selection

- Forward stepwise selection

Credit data example



There is not much difference between the three-variable and four-variable models in terms of RSS, so either of the four-variable models will likely be adequate.

Subset Selection

- Forward stepwise selection

  Comment: Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$; however, in this case, it is possible to construct submodels $M_0, ..., M_{n-1}$ only, since each submodel is fit using least squares, which will not yield a unique solution if $p \geq n$.

Subset Selection

- Backward stepwise selection
  - Like forward stepwise selection, backward stepwise selection provides an efficient alternative to best subset selection.
  - However, unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

Subset Selection

- 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.

Subset Selection

- Backward stepwise selection

  - Like forward stepwise selection, the backward selection approach searches through only $1 + p(p + 1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection.

  - Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.

  - Backward selection requires that *the number of samples n is larger than the number of variables p* (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.