Subset Selection

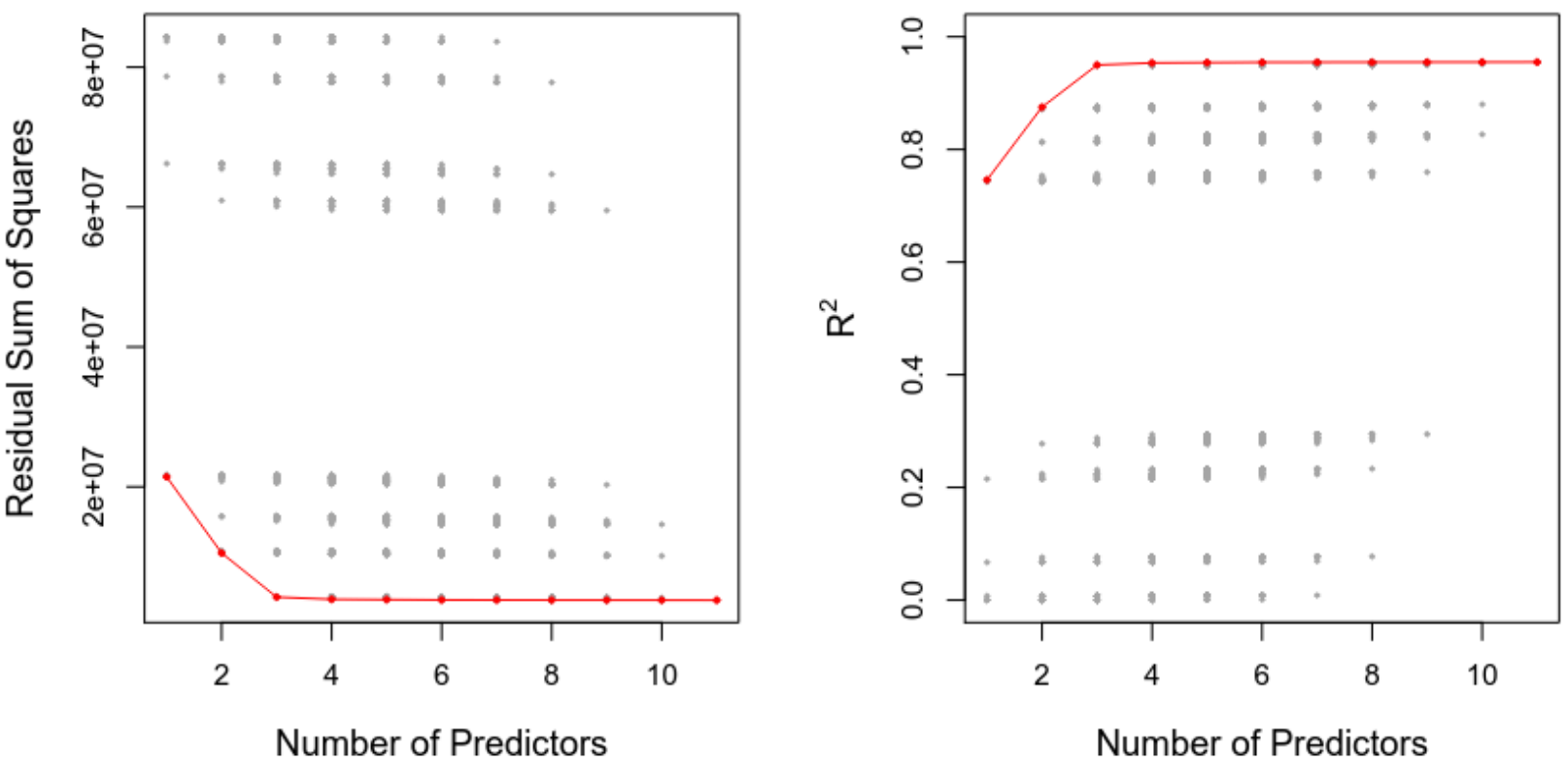
As a start the goal of **Subset Selection** is to identify a subset of the predictors p that we believe to be related to the Response Y and fit the model on the reduced set of variables

Best Subset Selection

Simply its **fitting** a separate least squares regression for each possible **combination** of p predictors, Algorithm steps

1. Let $\mathcal{M}_0$ be the **null model** which contain no predictors, the model predicts the sample mean for each observation
2. For $k = 1, 2, \dots p$:
    1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
    2. Pick the **best among** these $\binom{p}{k}$ models, call it $\mathcal{M}_k$ which have the smallest RSS or largest R^2
3. Select a single best model from among $\mathcal{M}_0 \dots \mathcal{M}_p$ using the prediction error on a **validation set** $C_p(AIC)$, BIC or adjusted R^2, or using Cross-Validation

- **Step** 2 in the algorithm reduces the possible models from 2^p which is the total amount of possible of models with the number of **predictors** we have to $p + 1$ by selecting the best model out of each class k based on the **training data**
- **Step** 3 among those $p + 1$ options by using a validation set or adjusted R^2, if Cross-Validation is used to select the best model then **Step** 2 is repeated on each training fold and the errors are averaged to select the best value



- This shows each model containing a subset of 10 predictors
- The red frontier tracks the best model among the k predictor class
- The graph goes to 11 since one of the variables are categorical and takes 3 values

Although in this example it's applied on the least squares the **Best Subset Selection** approach can be applied to other models such as Logistic Regression instead of the RSS we use deviance which is a measure the **deviance of the fitted logistic regression with respect to a perfect hypothetical model**

As someone can notice the **best subset selection** is simple and conceptually appealing which gives accurate results but same as **LOOCV** in the Cross-Validation is suffers from being computationally expansive and impossible in the model have 20 predictors then there is over **one million possibilities** and one it pass $p > 40$ it's impossible to compute even with new hardware

Stepwise Selection

As stated in the **Best Subset Selection** section that due to computational limitations it doesn't allow for large values of predictors , while also introducing a statistical problem where the larger the dimension search space the higher the chance of finding a model that perform good on the training data even tho they might not have any predictive abilities which also known as Overfitting and **high variance**

In this Section we discussing **Stepwise Selection** methods which are lighter yet powerful alternatives to the **Best Subset Selection**

Forward Stepwise Selection

Where **Best Subset Selection** consider all the possible 2^p models, the **Forward Stepwise Selection** consider much smaller set, Algorithm steps :

1. Let $\mathcal{M}_0$ denote as a **null model**, which contains no predictors
2. for $k = 0, \dots p-1$
    1. Consider all $p-k$ models, one of these predictors can be added to $\mathcal{M}_k$
    2. Choose the **best** among these $p-k$ models, and call it $\mathcal{M}_{k+1}$, in the least square the **best** is defined as having the smallest RSS or the highest R^2
3. Select as single best model among $\mathcal{M}_0, \dots \mathcal{M}_p$, using prediction error on a validation set AIC, BIC or adjusted R^2, Or Cross-Validation method

for example for $p = 20$ in the **Best Subset Selection** requires fitting $1,048,567$ and for **Forward Stepwise Selection** only 211 models, Sine

$$1 + \sum_{k=1}^{p-1} p - 1 = 1 + \frac{p(p+1)}{2}$$

Where Forward Stepwise Selection fail

It's clear that **Forward Stepwise Selection** has a computational advantage over the **Best Subset Selection**, but it's not guaranteed to find the best possible model out of all 2^p, for example :

- $p = 3$
- The **FSS** select X_1 as the best possible $\mathcal{M}_1$
- While the best model contain only X_2, X_3
- Since **FSS** selected X_1 as the best predictor in the previous iteration
- It will build upon it selection X_1, X_2 as the best possible model

Backward Step Selection

Like **Forward Stepwise Selection** the **Backward Stepwise Selection** offers an efficient alternative to the **Best Stepwise Selection**, from it's name it starts with the full model $\mathcal{M}_p$ containing all the predictors p and iteratively removes the predictor with the least contribution on at tie, Algorithm steps :

1. $\mathcal{M}_p$ denote the **full model** contains all the predictors p
2. for $k = p, p-1, \dots, 1$
    1. Consider all k models that contain all but one of the predictors in the previous model $\mathcal{M}_k$
    2. Choose the best among these k models can call it $\mathcal{M}_{k-1}$, in the least squares the best is defined by having the smallest RSS or the highest R^2
3. Select a single best model from each k iteration $\mathcal{M}_0, \dots \mathcal{M}_p$, using prediction error on a validation set AIC or BIC, or use Cross-Validation method

- Same as the **FSS** , the **BSS** approach searches through only $1 + \frac{p(p+1)}{2}$ and can be applied when p is too large
- The **BSS** can't be applied when the n is smaller than p is the initial full model cannot be fitted using least squares method
- **BSS** may not yield the best predictors selection same as **FSS** since it's forced to keep going backward and unable to re add predictors back

Hybrid Approaches

Generally the **best subset, forward stepwise, backward stepwise** approaches yields similar but not identical models, a **Hybrid approach** can be taken also called stepwise selection where predictors are added using the forward stepwise and after adding each new variable the backward stepwise is used so it removes and adds the right amount of predictors, this approach try to mimic the best subset selection with the computational advantages of both the forward and backward stepwise

WHY STEPWISE SELECTION IS FUCKING USELESS