Subset Selection

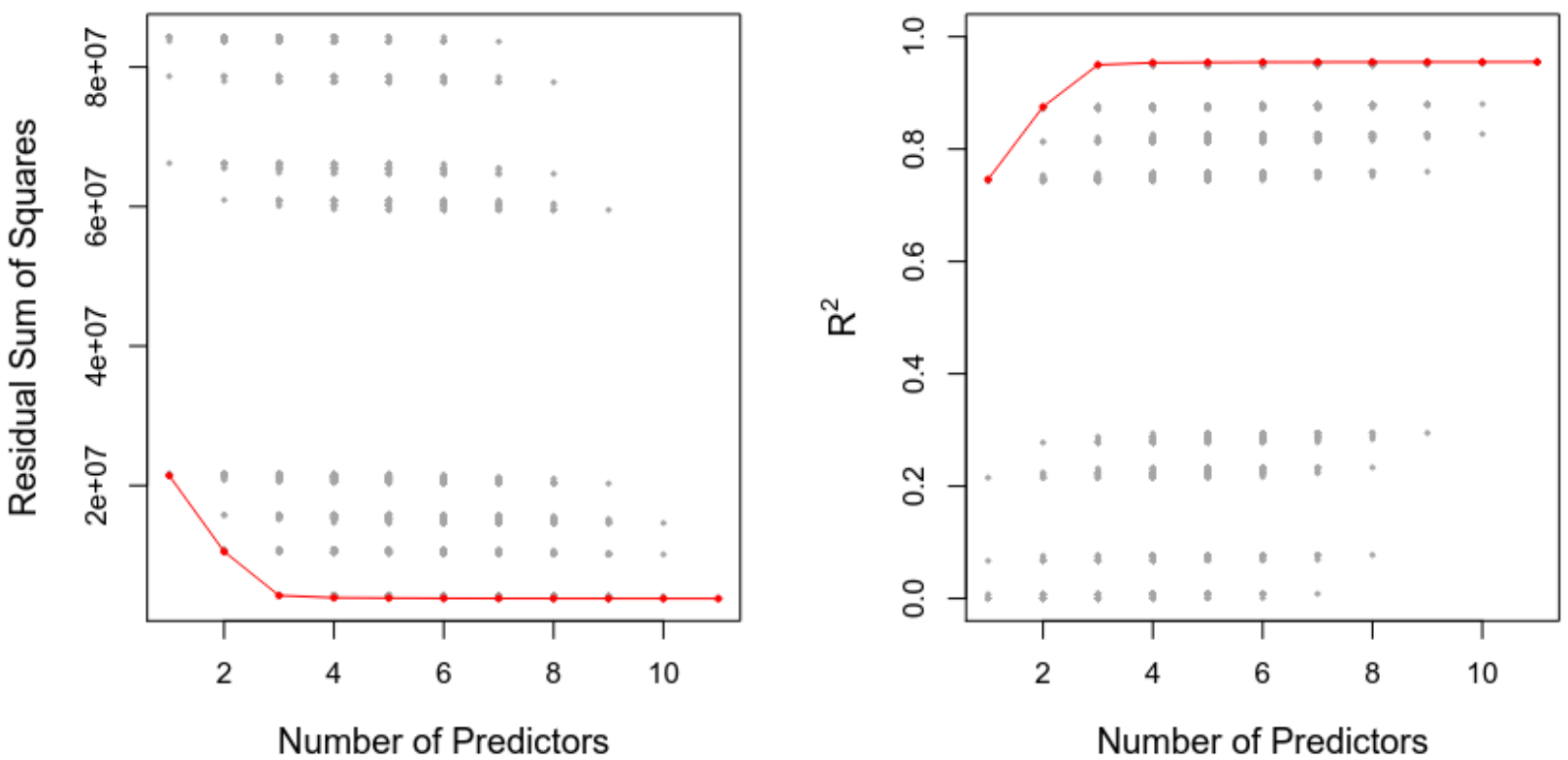
As a start the goal of **Subset Selection** is to identify a subset of the predictors p that we believe to be related to the Response Y and fit the model on the reduced set of variables

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