Subset Selection in Linear Models

(ISLR 6.1, 6.4)

Yingbo Li

Southern Methodist University

STAT 4399

Outline

High dimensional data

New technology permits the collection of many variables. For example,

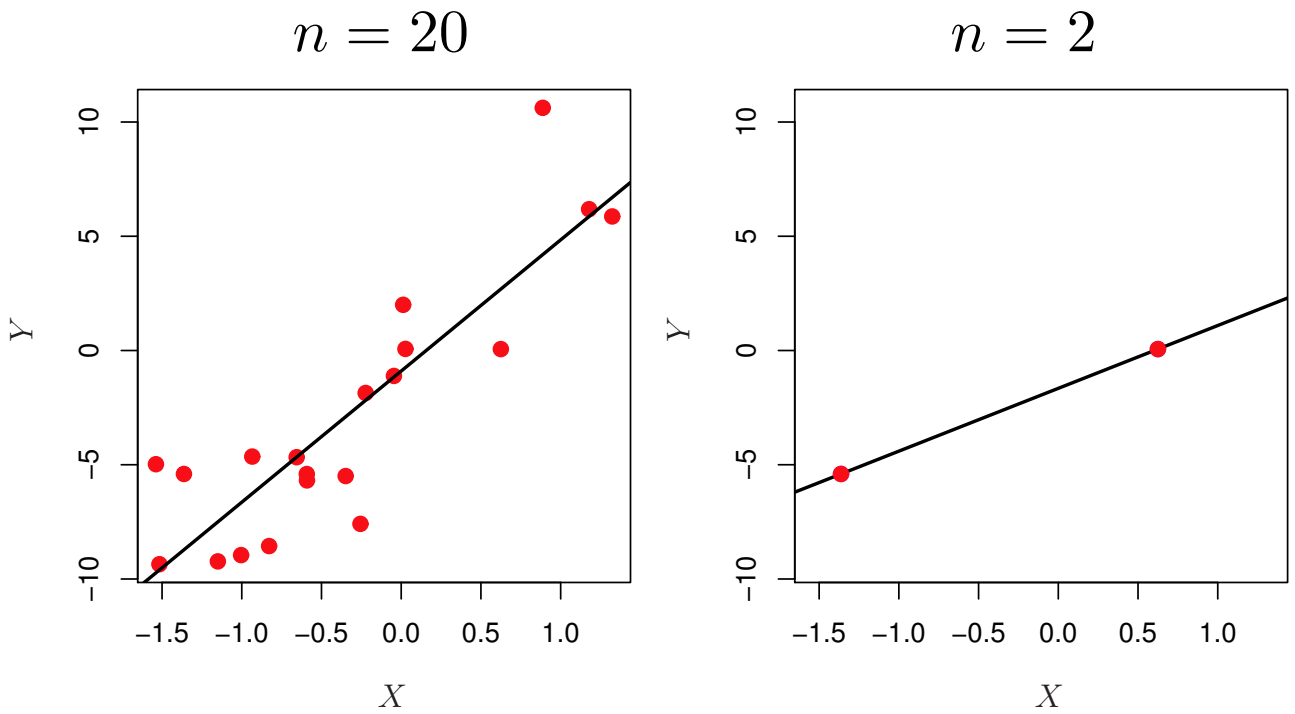
- Predict blood pressure, using
  - usual measurements: age, gender, BMI
  - genetic biomarkers: single nucleotide polymorphisms (SNPs)
- Half million SNP's can be obtained for each patient: $p \approx 500,000$
- Due to the cost of genetic tests, number of patients $n \approx 200$

We call a dataset *high dimensional*, if $p > n$.

- Many classical statistics approaches for low dimensional data ($n \gg p$) are not applicable to high dimensional data.
- For datasets that $p < n$ but n and p are close, classical approaches still have problems.

A simple illustration

Suppose we have $p = 1$ predictor, and fit linear regression using OLS.



When $n = 2$

- Number of observations equals number of regression coefficients
- A perfect fit to the data: residuals are zero. RSS $= 0$.

Overfitting

- In general, when $n \leq p+1$, all OLS residuals are zero.
  A least squares regression is too flexible and hence overfits the data.
- Recall that the design matrix $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ (including the intercept, a column of all one), the OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  - $\text{rank}(\mathbf{X}) = \text{rank}(\mathbf{X}^T\mathbf{X}) = \min(n, p+1)$
  - $\mathbf{X}^T\mathbf{X}$ is a $(p+1) \times (p+1)$ square matrix
  - When $p+1 > n$, $\mathbf{X}^T\mathbf{X}$ is not of full rank, so its inverse does not exist.
- Even when $p < n$ but if $n \approx p$, there can be *multicollinearity* issues.

Improving on the least squares regression estimates

There are 2 reasons we might not prefer to just use the OLS estimates

1. Prediction accuracy:
   To control the high variance due to overfitting
2. Model interpretability:
   To obtain an easy to interpret model by removing irrelevant variables

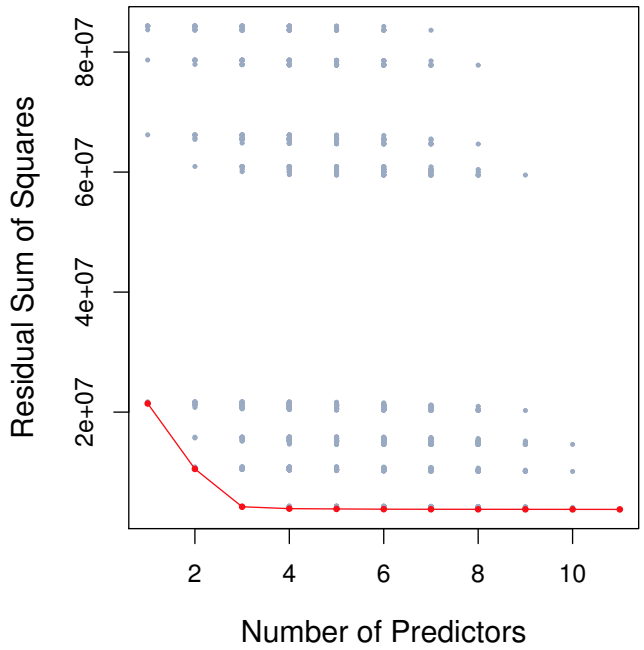
In this chapter, we will discuss *variable selection* methods

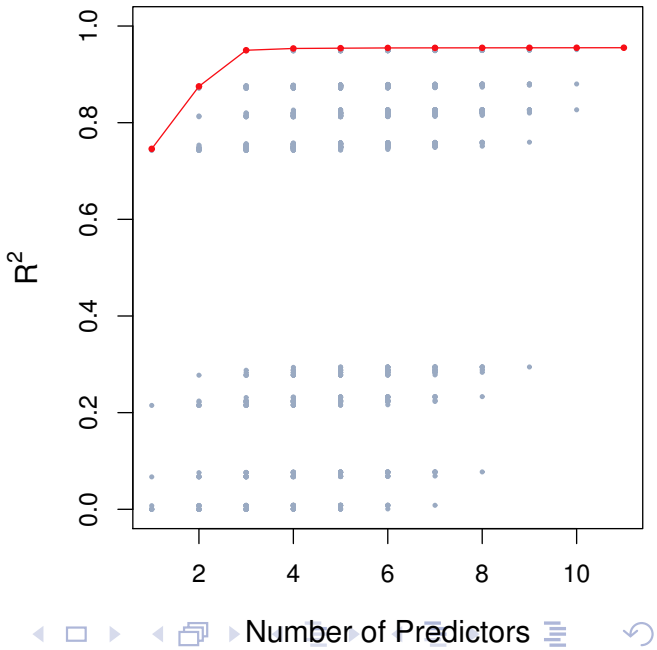
- Also called feature selection, model selection
- We will introduce two types of methods:
  - subset selection
  - shrinkage

Best subset selection

- In this approach, we run a linear regression for each possible combination of the p predictors.
  How many different subset models do we need to consider?
- How do we judge which subset is the "best"?
- The model that includes all the variables (the *full model*) always has the largest R^2 and smallest RSS.

The `Credit` data
$n = 400, p = 11$





Best subset selection: details

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 $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using *cross-validated prediction error, Cp (AIC), BIC, or adjusted* R^2.

Extensions to other models

- Although we have presented best subset selection here for least squares regression, the same ideas apply to other types of models, such as logistic regression.
- The *deviance*

$$D = -2 \log L$$

plays the role of RSS for a broader class of models. Here L is the maximized likelihood function.

$$L = \max_{\theta} f(Y \mid X, \theta) = \max_{\theta} \prod_{i=1}^{n} f(Y_i \mid X_i, \theta)$$

- What is the deviance for the linear regression?
  - Here $\theta = (\beta, \sigma^2)$. What is the MLE of σ^2?

Stepwise selection

- Best subset selection is computationally intensive
- *Stepwise selection*, which explores a far more restricted set of models, is an attractive alternative to best subset selection.

- Forward stepwise selection:
  - Begins with the null model,
  - then adds one predictor at a time that improves the model the most
  - until no further improvement is possible
- Backward stepwise selection:
  - Begins with the full model,
  - then deletes one predictor at a time that improves the model the most
  - until no further improvement is possible

Forward stepwise selection: details

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.

2. For $k = 0, 1, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having *the smallest RSS, or equivalently largest* R^2.

3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using *cross-validated prediction error, Cp (AIC), BIC, or adjusted* R^2.

Forward stepwise selection

- Computational advantage over best subset selection is clear. How many models does it fit?
- It is not guaranteed to find the best possible model out of all 2^p subset models.
- Forward stepwise selection can be used in the $p > n$ case: construct $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}$ only.

The Credit data example

To predict credit card balance:

Number of variables	Best subset	Forward stepwise
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	cards, income, student, limit	rating, income, student, limit

- The first three models are identical,
- but the fourth models differ.

Backward stepwise selection: details

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.

2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having *the smallest RSS, or equivalently largest* R^2.

3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$ using *cross-validated prediction error, Cp (AIC), BIC, or adjusted* R^2.

Backward stepwise selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models.
- Like forward stepwise selection, the backward selection is not guaranteed to find the best possible model out of all 2^p subset models.
- Can we use backward stepwise selection when $p > n$?

Estimating test error: two approaches

Each of the procedures (best subset, forward stepwise, backward stepwise) returns a sequence of models $\mathcal{M}_k$ indexed by model size $k = 0, 1, 2, \ldots, p$.

Among these $p + 1$ models, we should to choose the one with the lowest test error.

- We can indirectly estimate test error by making an *adjustment* to the training error to account for the bias due to overfitting.
  - C_p statistic
  - AIC
  - BIC
  - Adjusted R^2
- We can directly estimate the test error, using cross-validation.

Mallow's C_p statistic

For a subset model containing k predictors, *Mallow's C_p* is defined as

$$C_p = \frac{\text{RSS}}{\hat{\sigma}^2(\mathcal{M}_p)} - n + 2k.$$

- The C_p statistic adds a penalty of $2k$ to the training RSS.
- The model with the smallest C_p is preferred.
- If two models of the same size, them comparing C_p is equivalent to comparing their RSS.
- For two models with the same RSS, C_p prefers the smaller model.

Akaike information criterion (AIC)

For a subset model containing k predictors, *AIC* is defined as

$$\text{AIC} = -2\log L + 2k.$$

- AIC adds a penalty of $2k$ to the goodness-of-fit.
- The model with the smallest AIC is preferred.
- When σ^2 is known, C_p and AIC are the same. Why?

Bayesian information criterion (BIC)

For a subset model containing k predictors, *BIC* is defined as

$$\text{BIC} = -2\log L + k\log(n).$$

- BIC adds a penalty of $k\log(n)$ to the goodness-of-fit.
- The model with the smallest BIC is preferred.
- BIC tends to prefer smaller models than AIC.

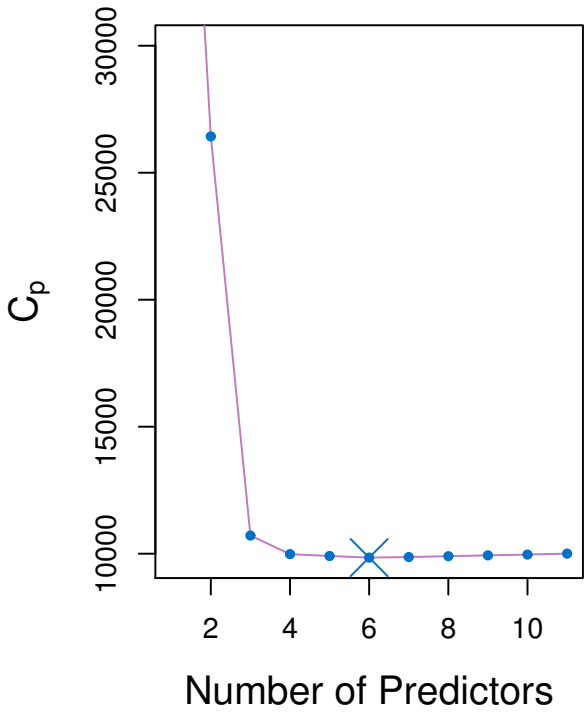
Adjusted R^2

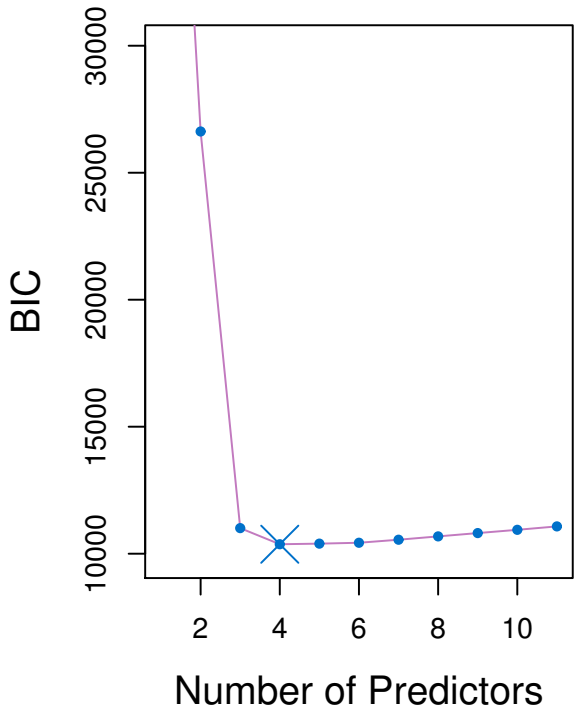
For a subset model containing k predictors, *adjusted* R^2 is defined as

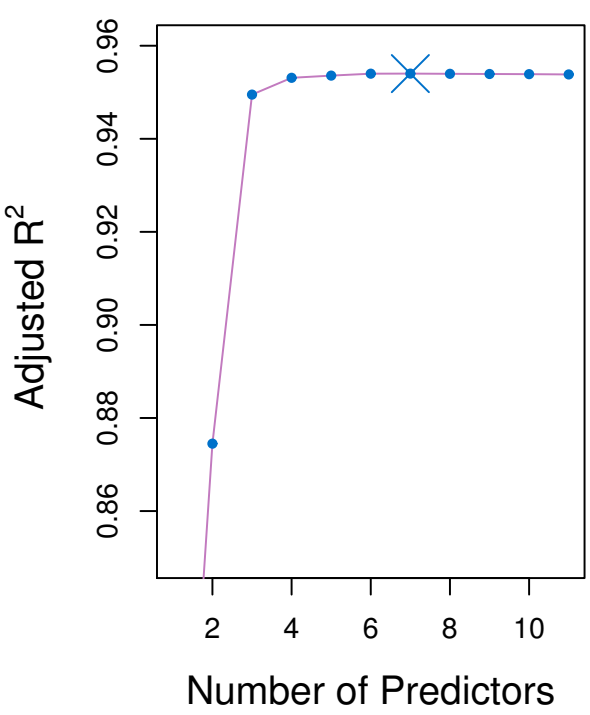
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}.$$

- The model with the largest adjusted R^2 is preferred.
- Maximizing the adjusted R^2 is equivalent to minimizing $\text{RSS}/(n-k-1)$
- Despite its popularity, the adjusted R^2 is not as well motivated in statistical theory as AIC, BIC, and C_p.

The Credit data







Cross validation

- In addition to training error adjustments, we can also use CV to select the optimal model size $\hat{k}$.
- Once selected, we will refit model $\mathcal{M}_{\hat{k}}$ using all training data.