



Statistical Learning

Lecture 2: High-Dimensional Linear Regression

Yi He

January 11, 2023

Plan for Today

1. Introduction
2. Stepwise Regression
3. Model Selection
4. Problems in High Dimensions

Introduction

Linear Regression Model

- Target y_i , features $x_i = (x_{i,1}, \dots, x_{i,p})^T$, $i = 1, \dots, n$.
- Linear regression model

$$y_i = \beta_0 + x_i^T \beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | x_i] = 0.$$

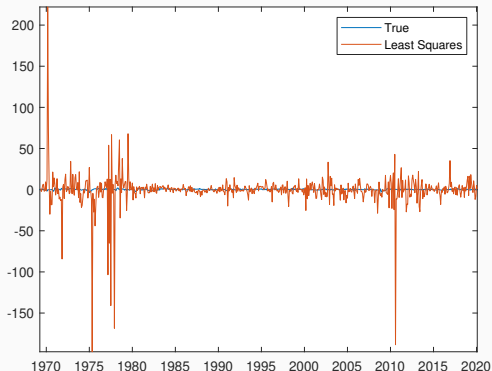
- The least squares estimator $\hat{\beta}_0$ and $\hat{\beta}$ minimizes the empirical squared risk

$$\frac{1}{n} \sum_{i=1}^n \left(y_i - \beta_0 - x_i^T \beta \right)^2.$$

- There is no regularization yet.

Curse of Dimensionality: FRED-MD

- Response = monthly growth rate US industrial production index $\log(IP_t/IP_{t-1})$
- Features = one-month lagged observations of $p \approx 100$ economic variables from FRED-MD database
- Rolling-window forecasts with size $n = 120$



Stepwise Regression

Regularization under Complexity Constraint

- Now minimize the mean squared error (empirical risk) subject to an additional complexity constraint

$$\# \text{ active features} = \sum_{j=1}^p \mathbb{1}(\beta_j \neq 0) = d,$$

where d is some hyperparameter.

- It generates a restricted model $\mathcal{M}_d = \{0\} \cup \left\{ 1 \leq j \leq p : \widehat{\beta}_j^{(d)} \neq 0 \right\}$ of size $1 + d$ (including intercept)
- The estimator is equivalent to the least-squares estimator under the restricted model \mathcal{M}_d .

Best Subset Selection: Comments

- Stepwise: we need to choose among $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_D$, for some $D \leq \min\{p, n - 1\}$.
- Non-nested: $\mathcal{M}_1 \not\subset \mathcal{M}_2 \dots$
- Exhaustive search is expensive for (relatively) large D

Hastie, Tibshirani and Tibshirani (2017) for $n = 500$ and $p = 100$:
At 3 minutes per value of k [= d in our slides], if we wanted to use 10-fold cross-validation to choose between the subset sizes $k = 0, \dots, 50$, then we are already facing 25 hours of computation time.

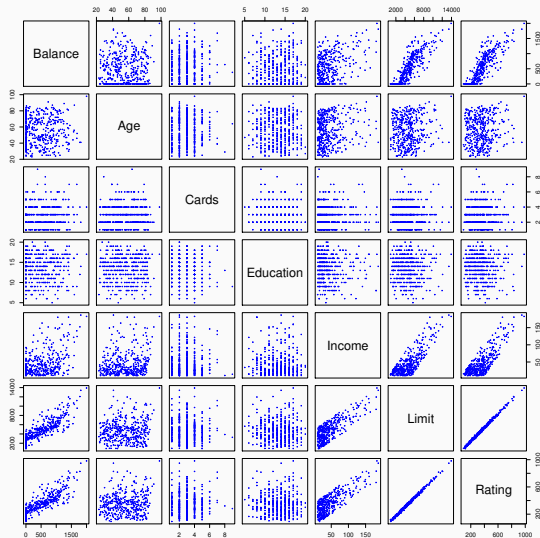
Textbook Example: Credit data set

- balance: average credit card debt for a number of individuals

Features

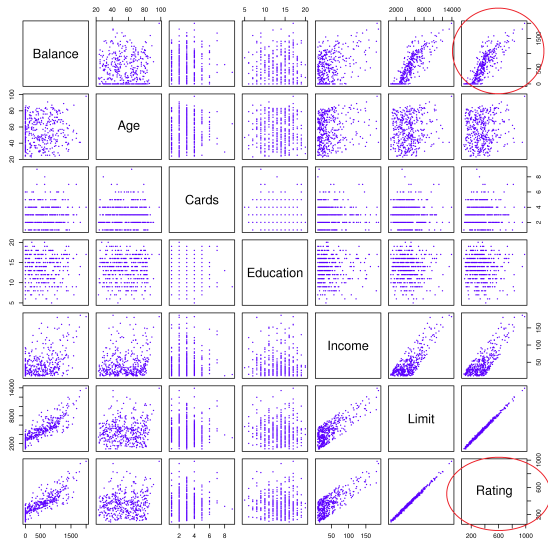
- age
- cards: number of credit cards
- education: years of education
- income in thousands of dollars
- limit: credit limit
- rating: credit rating
- student: dummy variable
- ... see ISLR, page 83

Credit Data Set: Correlation Plot



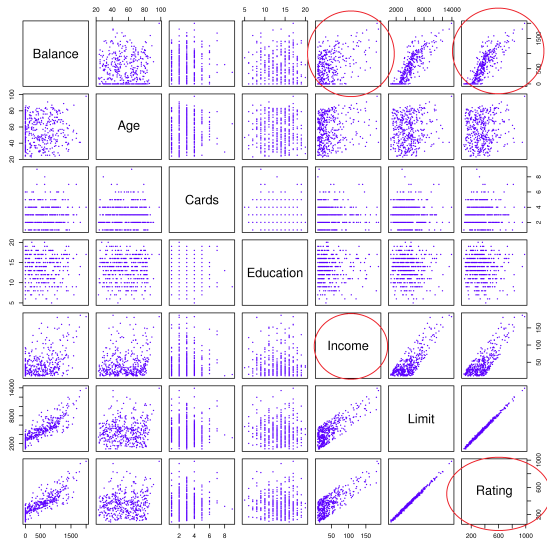
ISLR Figure 3.6

Credit Data Set: ISLR Table 6.1 $d = 1$



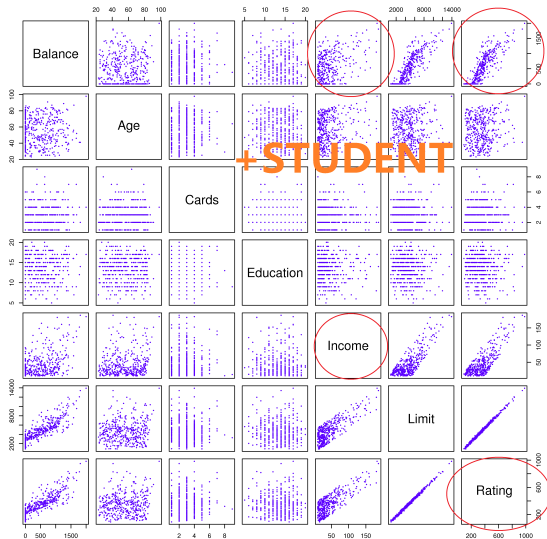
ISLR Figure 3.6

Credit Data Set: ISLR Table 6.1 $d = 2$



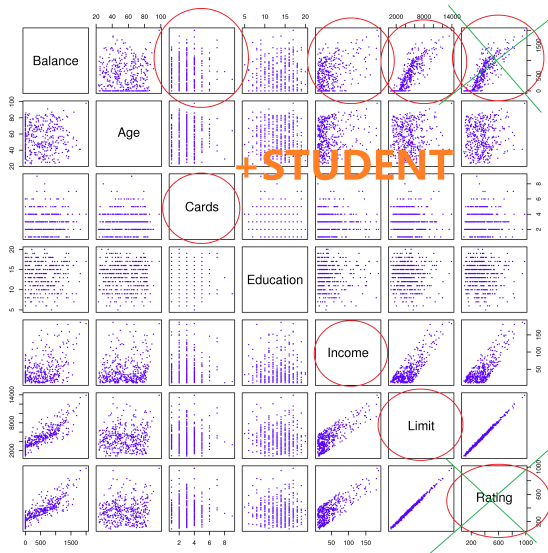
ISLR Figure 3.6

Credit Data Set: ISLR Table 6.1 $d = 3$



ISLR Figure 3.6

Credit Data Set: ISLR Table 6.1 $d = 4$



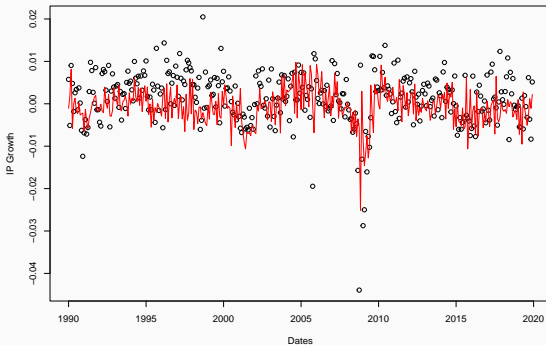
ISLR Figure 3.6

Forward Stepwise Regression

- **Nested:** $\mathcal{M}_0 \subset \mathcal{M}_1 \subset \mathcal{M}_2 \dots$
- \mathcal{M}_0 : only intercept
- $\mathcal{M}_{d+1} = \mathcal{M}_d + \text{one predictor}$ by minimizing the mean squared error (= empirical risk)
- Credit data set: same as best subset for $d = 1, 2, 3$ but different for $d = 4$.
- Computationally more attractive than best subset
- In practice, we also need to choose among $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_D$, $D \leq \min\{p, n - 1\}$.

Forward Stepwise Regression: IP Example

- Prior-1990 sample as training set: choose $d = 9$ variables



Forward Selection Improves Out-of-sample Rolling-window Forecasts

Model Selection

Empirical VS Population Risk

- Consider a sequence of nested models (such as that from forward selection):

$$\mathcal{M}_0 \subset \mathcal{M}_1 \subset \dots \subset \mathcal{M}_D$$

- Choose \mathcal{M}_d with optimal criterion value $C(\mathcal{M}_d)$.
- Can we use the empirical risk as criterion? Consider the training mean squared error given by

$$\frac{1}{n} \text{RSS}(\mathcal{M}_d) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}_d(x_i))^2,$$

where $\hat{f}_d(x_i)$ are the estimated means using model \mathcal{M}_d .

Go to www.menti.com and use the code 15 97 72 6

Can we use RSS as model selection criterion?

 Mentimeter

0

Yes, that is what use in the forward selection, isn't it?

0

No, RSS always prefers the most flexible model

0

I do not know



Mallows' C_p

- True means $\mu(x_i) = \beta_0 + \sum_{j \in \mathcal{M}_d} \beta_j x_{i,j} + \sum_{j \notin \mathcal{M}_d} \beta_j x_{i,j}$
- Mean squared estimation error of the true means on the training set

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \left(\mu(x_i) - \hat{f}_d(x_i) \right)^2 \\ & \approx \frac{1}{n} \left\{ \text{RSS}(\mathcal{M}_d) + \underbrace{2d\sigma^2}_{\text{penalty for complexity}} \right\} \underbrace{- \frac{n-2}{n} \sigma^2}_{\text{constant}} \end{aligned}$$

where $\sigma^2 = \text{var}(\varepsilon_i)$ denotes the true error variance.

- Mallows' $C_p = \frac{1}{n} \{ \text{RSS}(\mathcal{M}_d) + 2d\hat{\sigma}^2 \}$ with some appropriate estimator $\hat{\sigma}^2$

Bayesian Information Criterion

- Suppose that the model is uncertain before observing the data: prior probabilities $\mathbb{P}(\mathcal{M}_d)$, $d = 0, \dots, D$.
- After observing the data we update the distribution:

$$\mathbb{P}(\mathcal{M}_d | \text{data}) \propto \exp\left(-\frac{1}{2} \text{BIC}(\mathcal{M}_d)\right) \mathbb{P}(\mathcal{M}_d) \quad d = 0, \dots, D.$$

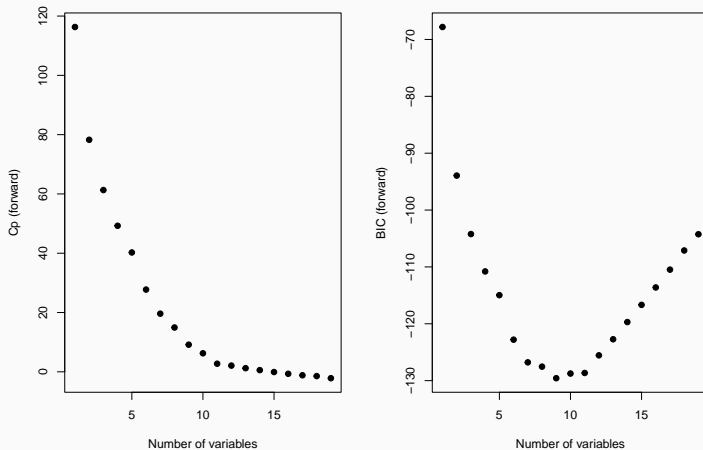
where

$$\text{BIC}(\mathcal{M}_d) \equiv \frac{1}{n} \left(\text{RSS}(\mathcal{M}_d) + \underbrace{\log(n) d \sigma^2}_{\text{penalty for complexity}} \right)$$

for Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent of x_i .

- Replace σ^2 with $\hat{\sigma}^2$ in practice.
- Mallows' C_p uses a smaller factor $2 < \log(n)$ when $n \geq 8$
- BIC favors a less flexible model

FRED-MD Example



BIC chooses $d = 9$ while Mallows' C_p favors a larger model

Validation Approach

- Treat dimension d as a hyperparameter.
- Divide the data into a *training* set and a *validation* set
- Fit the model \mathcal{M}_d using the training set
- Predict using the feature values on validation set but the fitted coefficients $\hat{\beta}_0^{(d)}$ and $\hat{\beta}^{(d)}$ from the training set
- Calculate the test mean squared error on validation set

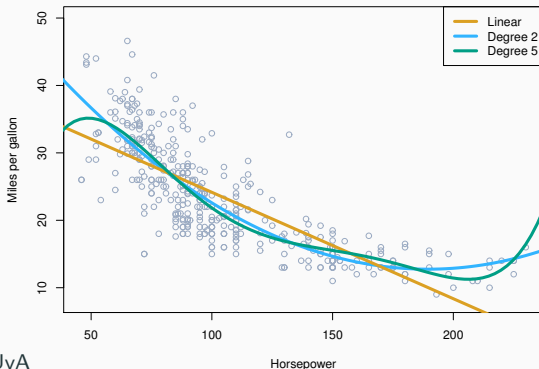
$$\text{MSE}_d = \text{Avg}_{i \in \text{validation set}} (y_i - \hat{\beta}_0^{(d)} - x_i^T \hat{\beta}^{(d)})^2$$

- Compare the test MSE_d between models
- Cross-validation: leave-one-out or k -fold

Textbook Example: Auto Dataset

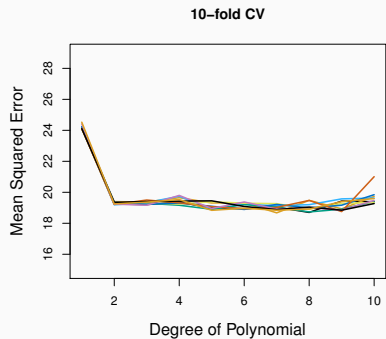
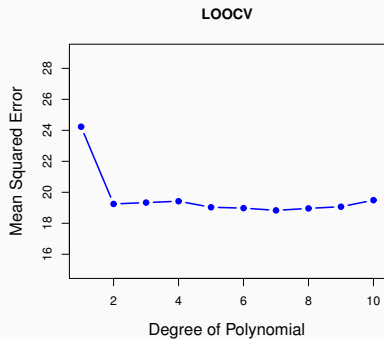
- mpg: gas mileage in miles per gallon
- horsepower
- Polynomial regression

$$\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \dots + \beta_d \text{horsepower}^d + \epsilon$$



ISLR
Figure 3.8

Degree Selection



ISLR Figure 5.4

Problems in High Dimensions

Why Cross Validation

- High dimensional dataset: p/n is large especially when $p > n$
- Perfectly fit all points in training data: useless!
- C_p and BIC are **in**appropriate: sample variance $\hat{\sigma}^2 = 0$
- Better use cross validation: model complexity $\max_d |\mathcal{M}_d|$ cannot be too large for forward selection
- CV becomes expensive quickly as the model complexity grows

Sparse and Dense Models

- Removing noise features may improve forecasting
- If the **true** model is *sparse*:

$$p_0 = \#\{\beta_i : \beta_i \neq 0, i = 1, \dots, p\}$$

is small (relative to n)

- Selecting a useful model may reduce the dimensionality problem
- Economic data tend to be *dense*, however
- To be discussed further

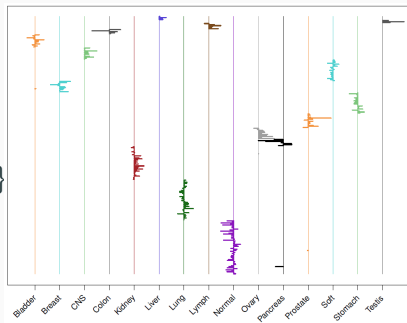


Figure 1.1, *Statistical Learning with Sparsity*: estimated nonzero feature weights 15-class gene expression cancer data