- Module 7 -
Model Assessment and Selection III

Outline

- Subset Selection
  - Best-Subset Selection
  - Forward-Stepwise Selection
  - Backward-Stepwise Selection
- Shrinkage Methods
  - Ridge Regression
  - The Lasso

Recap

Linear Regression Models: Multiple Features (1)

- Recall, linear regression with multiple inputs (or features)

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}\,, \qquad \forall i = 1, 2, \dots N$$

where

N	number of samples in the data
p	number of features being modeled
$x_{i,j}$	j-th feature of the input variable of the i-th sample
$\hat{\beta}_j$	weight (or parameter or coefficient) that determines how the j-th feature affects the prediction
$\hat{y}_i$	predicted output of the i-th sample

- Least squares estimates: obtained by minimizing the residual sum of squares (RSS)

$$\begin{aligned} RSS(\hat{\beta}_0, \dots, \hat{\beta}_p) &= \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{N} \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}) \right)^2 \end{aligned}$$

Linear Regression Models: Multiple Features (2)

- Consider linear regression models with potentially very large number of features (e.g., DNA microarray)

- Least squares estimates may not be satisfactory:
  - **Prediction accuracy**: tends to overfit the training data, i.e., low bias but large variance

    ➔ may perform poorly on unseen data

  - **Interpretation**: often desired to have a smaller subset of features that exhibit the strongest effects (most informative) on the outcome

    ➔ desire easier interpretation

- Solutions:
  - **Subset Selection**
  - **Shrinkage Methods**

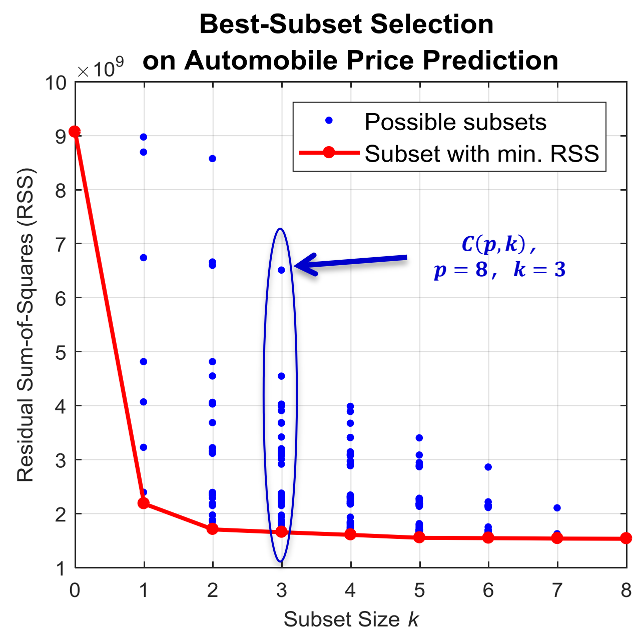
Subset Selection: Best-Subset Selection (1)

- In **best-subset selection,** retain only a subset of the features and eliminate the rest from the model
  - ➔ the goal is to select a subset of k features of $\boldsymbol{X}$, where $k \in \{0, 1, 2, \ldots, p\}$, that best predicts $\boldsymbol{y}$ (smallest RSS)
  - ➔ Steps:

    For each $k \in \{0, 1, 2, \ldots, p\}$,

    1. For every possible subset of size k of p features in $C(p, k)$, perform linear regression on the selected features and evaluate the RSS (**blue dots**)
    2. Find the subset that gives the smallest RSS (**red dot**)

    Determine k using cross-validation



- **Issue:** exhaustive search – only feasible for $p < 40$

Subset Selection: Forward-Stepwise Selection (2)

- Rather than search through all possible subsets (becomes infeasible for $p \gg 40$) ➔ seek a path through them
- **Forward-stepwise selection** starts with the intercept $\hat{\beta}_0$ (an empty model) and then sequentially adds into the model one feature at a time that most improves the fit
- Compared to best-subset selection, it
  - ➔ is sub-optimal but computationally feasible for large p
  - ➔ is a more constrained search (will have lower variance but perhaps more bias)

  - ➔ Steps:
    1. Start with an empty model (no features included)
    2. Choose among the p features to find the best single-feature model that gives the minimum RSS
    3. Choose among the remaining $p-1$ features to find another feature, when included with the previously chosen feature, that gives the minimum RSS
    4. Choose among the remaining $p-2$ features …
    5. …

    Determine k using cross-validation

Subset Selection: Backward-Stepwise Selection (3)

- **Backward-stepwise selection** starts with the full model (all features included), and then sequentially deletes one feature at a time that has the least impact on the fit
  - Note that backward-stepwise selection can only be used when $N > p$, i.e., more observations than features

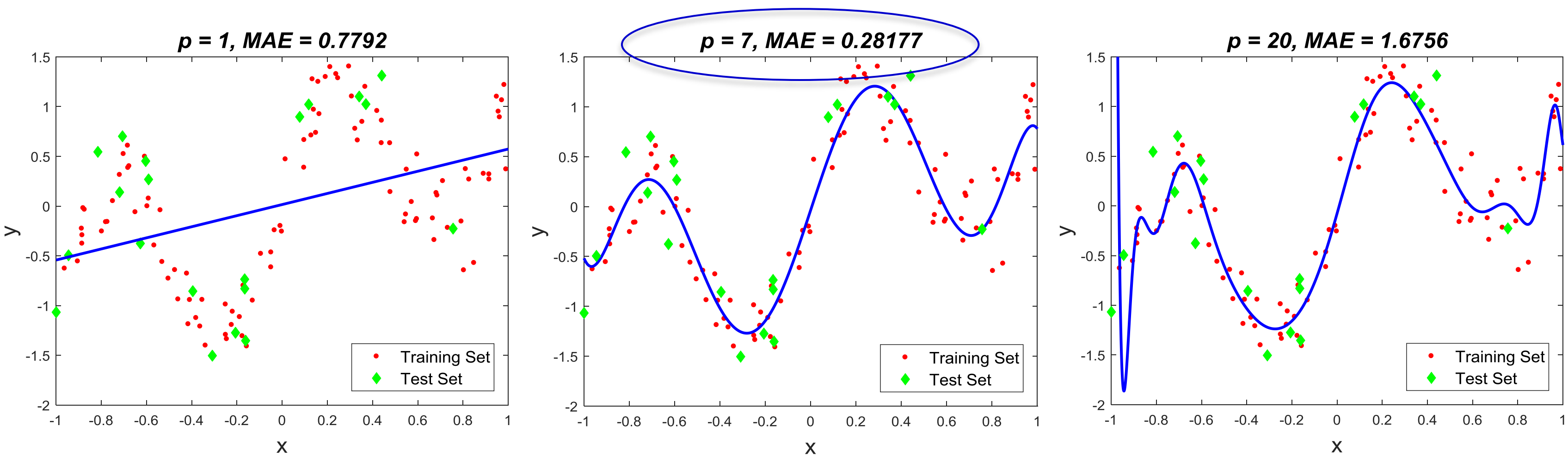
Shrinkage Methods (1)

- By retaining a subset of features and discarding the rest, subset selection produces a model that
  - is more interpretable
  - has possibly lower prediction error than the full model
- However, because subset selection is a discrete process (where features are either retained or discarded), it often exhibits high variance and so does not necessarily reduce the prediction error of the full model

➔ **Shrinkage methods** are more continuous and do not suffer as much high variability

Shrinkage Methods (2)

- Recall in the polynomial regression model, if p is large, least squares estimate results in a model that
  - ➔ tends to overfit the training data
  - ➔ has high variance and may perform poorly on the test data



Shrinkage Methods (3)

- In addition, $\hat{\beta}_j$ can take on large values

Table of coefficients $\hat{\beta}_j$ for polynomial regression of various degree p

p	1	7	20
$\hat{\beta}_0$	0.02	-0.03	-0.07
$\hat{\beta}_1$	0.56	6.91	8.05
$\hat{\beta}_2$		0.05	5.78
$\hat{\beta}_3$		-35.87	-63.94
$\hat{\beta}_4$		-0.12	-119.59
$\hat{\beta}_5$		55.47	271.15
$\hat{\beta}_6$		0.23	1004.98
$\hat{\beta}_7$		-25.85	-886.94
…			…
$\hat{\beta}_{20}$			10444.27

- **Shrinkage methods** (or **regularization**) fits a full model containing all p features, but the estimated coefficients are **constrained** (or **regularized**) such that they are shrunk towards zero in a continuous fashion

  ➔ i.e., shrinkage methods discourage $\hat{\beta}_j$ from reaching large values by imposing a penalty on the magnitude of $\hat{\beta}_j$

Shrinkage Methods (4)

- This is done by adding a **penalty term** to the residual sum of squares (RSS) of the least squares estimate
- This gives the **penalized** (or **regularized**) RSS:

Regularization parameter

$$RSS^{regularized}(\hat{\beta}_0, \dots, \hat{\beta}_p, \lambda) \quad = \underbrace{\textstyle\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}_{\text{Residual sum of squares}} + \lambda \underbrace{R(\hat{\beta}_1, \dots, \hat{\beta}_p)}_{\text{Penalty function on } \hat{\beta}_j}$$

where λ is the **regularization parameter** (or tuning parameter) that controls the amount of shrinkage

- the larger the value of λ, the smaller the magnitude of $\hat{\beta}_j$

- Note (by convention) the intercept, $\hat{\beta}_0$, is omitted from the penalty term

Shrinkage Methods: Ridge Regression (5)

- One common choice of the penalty function is the L^2 norm of $\hat{\beta}_j$

$$R\left(\hat{\beta}_1, \dots, \hat{\beta}_p\right) = \sum_{j=1}^{p} \left|\hat{\beta}_j\right|^2$$

- This leads to the **Ridge Regression**

$$RSS^{ridge}\left(\hat{\beta}_0, \dots, \hat{\beta}_p, \lambda\right) = \sum_{i=1}^{N} \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})\right)^2 + \lambda \sum_{j=1}^{p} \left|\hat{\beta}_j\right|^2$$

- In matrix notation

$$RSS^{ridge}\left(\hat{\beta}, \lambda\right) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}\right) + \boldsymbol{\lambda}\hat{\beta}^{\boldsymbol{T}}\hat{\beta}$$

Shrinkage Methods: Ridge Regression (6)

- The **ridge estimate** $\hat{\beta}^{ridge}$, is obtained by minimizing $RSS^{ridge}(\hat{\beta}, \lambda)$ with respect to $\hat{\beta}$:

$$\hat{\beta}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Ridge Estimate

where

$\boldsymbol{X}$ is **standardized** (or normalized) (i.e., each $x_{i,j}$ is replaced by $(x_{i,j} - \mu_j)/\sigma_j$, where μ_j and σ_j are the mean and standard deviation of the j-th column of $\boldsymbol{X}$, respectively)

I is a $p \times p$ identity matrix

No need to add a column of 1's

$\hat{\beta}_0$ is left out of the regression

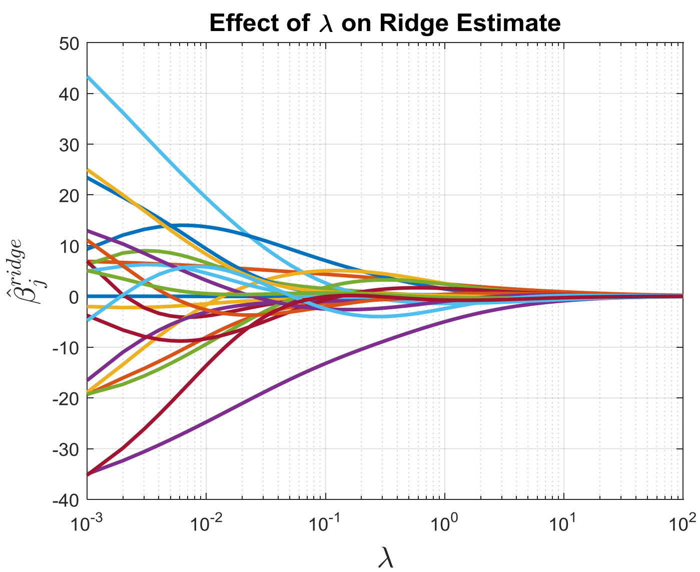
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix} \text{ and } \hat{\beta}^{ridge} = \begin{bmatrix} \hat{\beta}_1^{ridge} \\ \hat{\beta}_2^{ridge} \\ \vdots \\ \hat{\beta}_p^{ridge} \end{bmatrix}$$

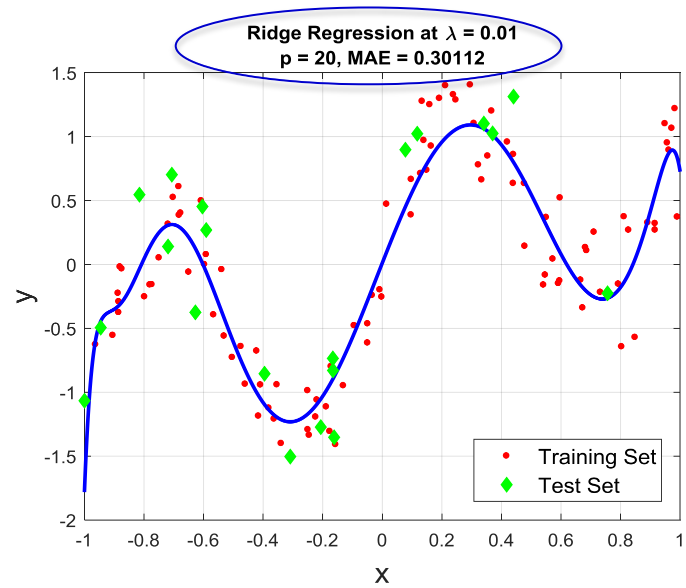
$\boldsymbol{N \times 1}$ $\quad$ $\boldsymbol{N \times p}$ $\quad$ $\boldsymbol{p \times 1}$

- The intercept coefficient $\hat{\beta}_0$, is estimated by $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$

Shrinkage Methods: Ridge Regression (7)

- When $\lambda = 0$, the penalty term has no effect, i.e., the ridge estimate $\hat{\beta}^{ridge}$ is the same as the least squares estimate $\hat{\beta}^{LS}$ (previously denoted $\hat{\beta}$ in Module 4)
- As $\lambda \to \infty$, the impact of the penalty term increases, and the ridge estimate $\hat{\beta}^{ridge}$ approaches zero





- The best value of λ can be determined using cross-validation

Shrinkage Methods: The Lasso (8)

- Another common choice of the penalty function is the L^1 norm of $\hat{\beta}_j$

$$R(\hat{\beta}_1, \dots, \hat{\beta}_p) = \sum_{j=1}^{p} |\hat{\beta}_j|$$

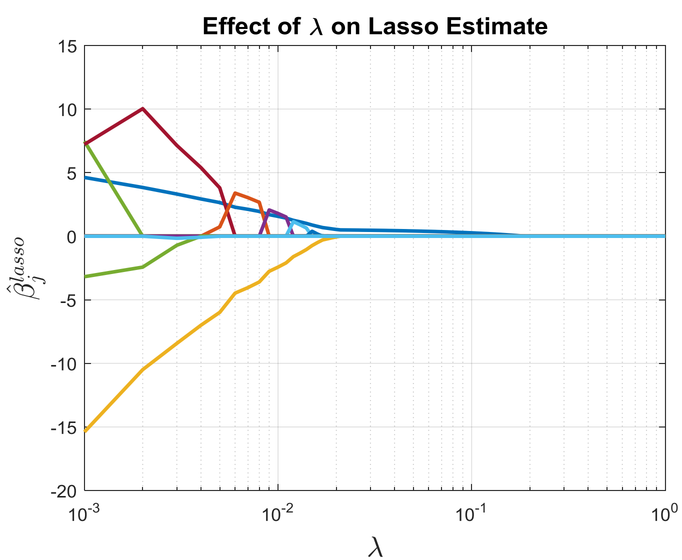
- This leads to **The Lasso** (least absolute shrinkage and selection operator)

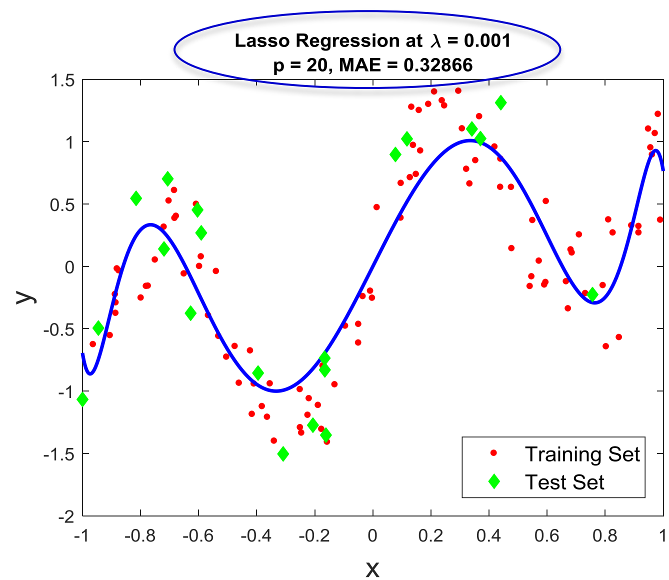
$$RSS^{lasso}(\hat{\beta}_0, \dots, \hat{\beta}_p, \lambda) = \sum_{i=1}^{N} \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})\right)^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j|$$

- Note, however, that the **lasso estimate** $\hat{\beta}^{lasso}$, does not have a closed form expression as in ridge regression

Shrinkage Methods: The Lasso (9)

- As with ridge regression, the lasso shrinks the estimated coefficients $\hat{\beta}^{lasso}$ towards zero
- In addition, when the tuning parameter λ in the lasso is sufficiently large, the L^1 penalty function has the effect of forcing some of the coefficients to be exactly zero, i.e., subset selection





$\hat{\beta}_1$	4.62
$\hat{\beta}_2$	0.03
$\hat{\beta}_3$	-15.39
$\hat{\beta}_5$	7.46
$\hat{\beta}_7$	7.22
$\hat{\beta}_{19}$	-3.18
other $\hat{\beta}_j$	0

- The best value of λ can be determined using cross-validation