Linear Methods for Regression

Linear Regression model has the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \qquad (3.1)$$

X_j can come from different sources

1. quantitative inputs
2. transformation of quantitative inputs, such as log, square-root, or square
3. numeric or dummy coding of the levels of qualitative inputs. EX, low, medium, high - 1,2,3
4. interaction between variables, X = X1*X2

The most popular estimation method is **least squares**, in which we pick the coefficients β = (β0, β1, . . . , βp)T to **minimize the residual sum of squares**

Simple Linear Regression

RSS, ESS, TSS

Residual sum of squares - RSS:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

$\hat{y}_i$ - predicted y (at each point)

$\bar{y}$ - mean of y (single value, mean of all ys)

TSS: total variance of inherent in the response **before the regression is performed.**

RSS: measure the amount of variability that is left **unexplained** after performing the regression.

ESS: measures the amount of variability that is **explained** after performing the regression.

$TSS = RSS + ESS$

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

R^2 - measures the proportion of the variability in Y that can be explained using X.

R^2 vs Correlation?

Multiple Linear Regression model

linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{3.19}$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response. We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*.

F statistics

- measures the change of residual sum of squares (RSS) per additional parameter in the bigger model, and it is normalized by an estimate of σ^2

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)},$$

- p = number of predictors

on the values of n and p. When n is large, an F-statistic that is just a little larger than 1 might still provide evidence against H_0. In contrast, a larger F-statistic is needed to reject H_0 if n is small. When H_0 is true

When to use F-test.

- both p-value and f-statistics need to be significant

Subset Selection

- Variable subset selection with linear regression

Best-subset Selection

- ISL P205
- very computationally expansive / brute force finding the best combination of features from feature space 2^p (p=total num of features)
- choose k that minimizes the estimate of the expected prediction error

Stepwise Selection

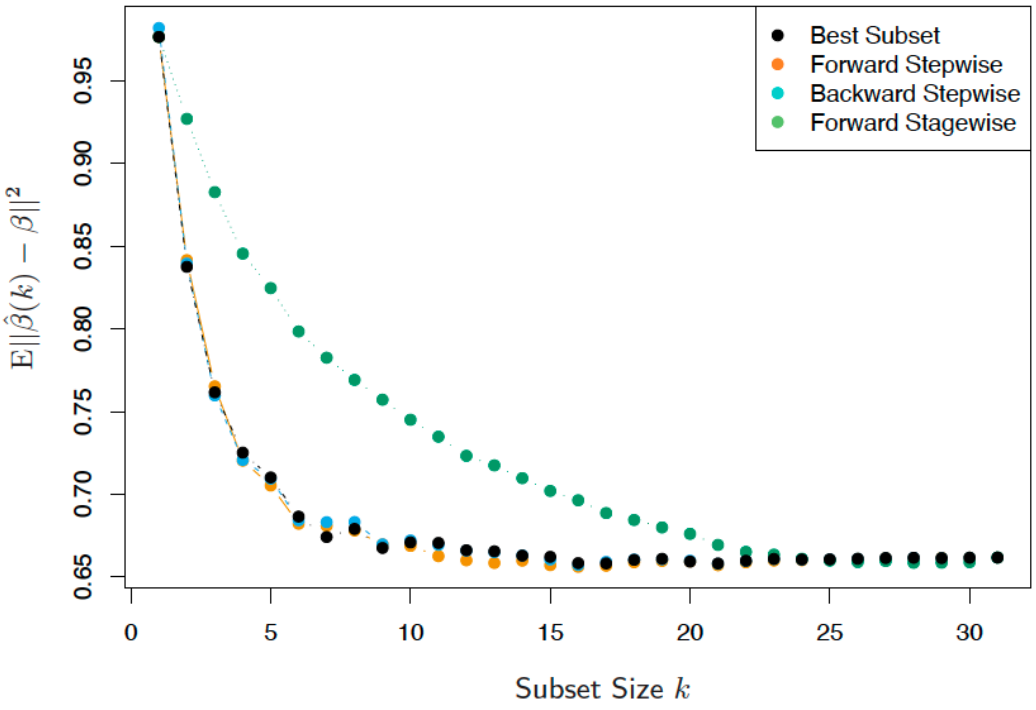
Forward-stepwise selection

- forward stepwise can always be used whether n>>p or n<<p
- starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit.
- a greedy algo, large computation
- pros:
  - less expensive than best-subset approach, even when p>>N
  - will have lower variance, but perhaps more bias

Backward-stepwise selection

- Backward-stepwise selection starts with the full model, and sequentially deletes the predictor that has the least impact on the fit.
- The candidate for dropping is the variable with the smallest Z-score (Exercise 3.10). Backward selection can only be used when N > p, while **forward stepwise can always**
  be used.

Simulation to compare selection methods



- y - MSE
- forward and backward stepwise performs the same
- forward stagewise takes longer to reach the min error.

Others to select:

- AIC criterion: (not always used anymore)
    - AIC criterion for weighing the choices, which takes proper account of the number of parameters fit; at each step an add or drop will be performed that minimizes the AIC score.
- F-Statistics: (not always used anymore)

Stagewise regression

- Steps:

1. it starts like forward-stepwise regression, with an intercept equal to mean y, and centered predictors with coefficients initially all 0.
2. At each step the algo identifies the variable most correlated with the current residual
3. it then computes the simple linear regression coefficient of the residual on this chosen variable, and then adds it to the current coefficient for that variable. This is continued till none of the variables have correlation with the residual.

- works well with high dimensional problems.
-

Lasso and Ridge

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

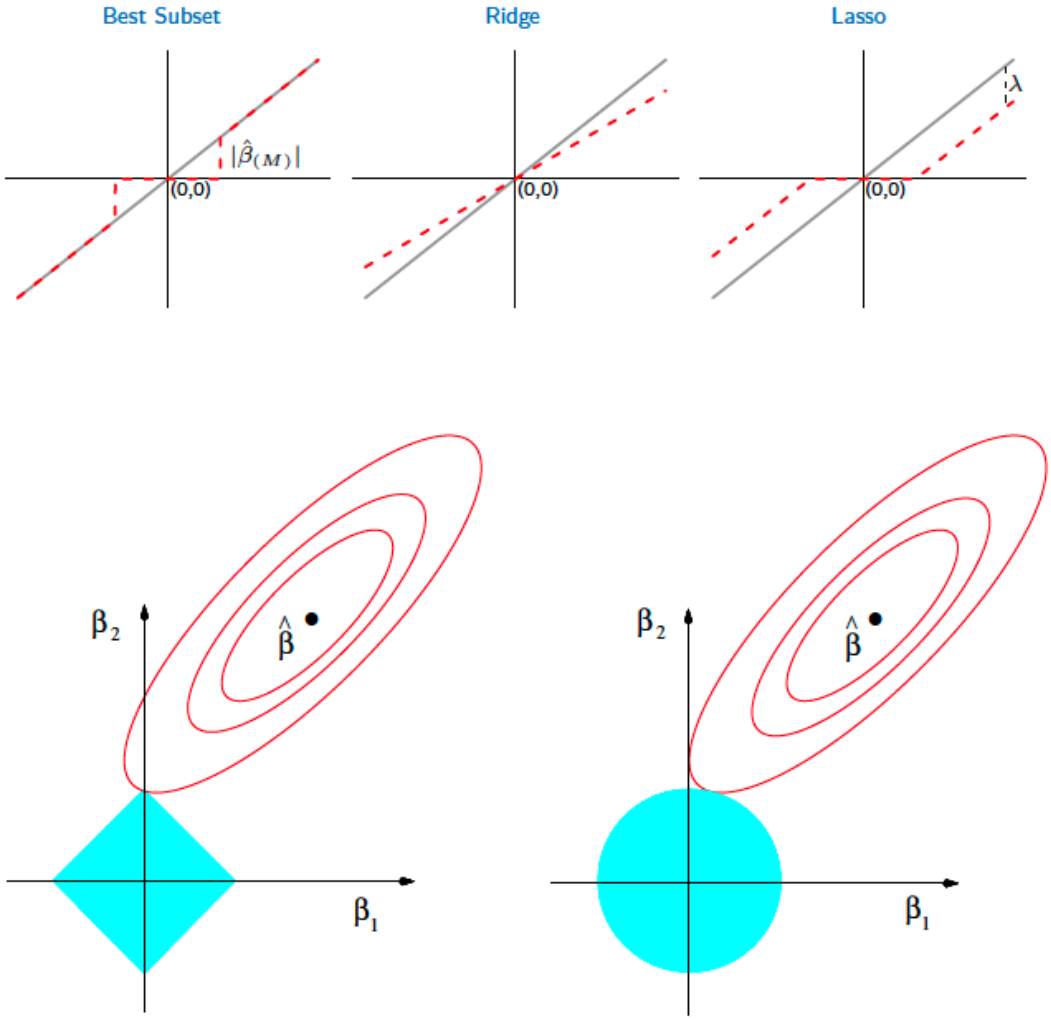


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

region for ridge regression is the disk $\beta_1^2 + \beta_2^2 \leq t$, while that for lasso is the diamond
$|\beta_1| + |\beta_2| \leq t$.