Today

1. Why consider alternatives?
2. ~~Best~~ Subset Selection 3. Choose the optimal model
4. Shrinkage (Regularization)

Announcement : ① Ass 3 is available online
② R codes for today's lecture is available online
③ Test is arranged

1. Why consider the alternatives?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

⇒ improve the model fitting by replacing plain least squares fitting with some alternative fitting procedures.

① better prediction
- $n \gg p$, LSE has low ~~var~~ variance
- $n \approx p$, LSE has a lot of variability (overfitting, poor prediction)
- $n < p$, LSE has no unique solutions (the variance is infinite)

② Model Interpretability
- variable selection (important variables) by removing not important variables

* Alternative Methods

① subset selection { best subset selection / stepwise

② Shrinkage (Regularization) < Ridge Regression / The Lasso

③ Dimension Reduction < Principal Component Regression (PCR) / Partial Least Squares

2. Subset Selection

① Best Subset Selection

Algorithm 6.1

Step 1: Let M_0 denote the null model, which contains no predictors.

Step 2: For $k = 1, 2, \cdots p$

(a) fit $\binom{p}{k}$ models that contain exactly k predictors

(b) Pick the best among $\binom{p}{k}$ models, and call it M_k based on smallest RSS or highest R^2.

Step 3: Select a single best model from $M_0, \cdots M_p$ using cross-validated prediction error, C_p(AIC), BIC or adjusted R^2.

Disadvantages:

① Computational intensive for large p, (2^p)

② Suffer statistical problems when p is large: large the search space, the higher the chance of finding models that look good for the training data, even though they may not have any predictive power for future data.

③ Lead to overfitting and high variance of the coefficient estimates.

② Stepwise selection

△ Forward Stepwise Selection

- begins with a model containing no predictors, and then adds predictors to the model, one at a time, until all the predictors are in the model

⇒ at each step, the variable that gives the greatest improvement to the fit is added to the model

* Algorithm 6.2

Step 1: Let M_0 denote the null model, which contains no predictors

Step 2: For $k=0, 1, \cdots, p-1$;

2.1 consider all $p-k$ models that argument the predictors in M_k with one additional predictor.

2.2 Choose the best among these $p-k$ models, and call it M_{k+1}, (here best is defined as having the smallest RSS or highest R^2)

Step 3: select a single best model from among $M_0 \cdots M_p$ using cross-validated prediction error, C_p (AIC), BIC or adjusted R^2.

Cons & pros:

- computationally feasible
- not guaranteed to find the best possible model out of 2^p models, containing subsets of the p predictors.
- can use for $p > n$.

△ Backward stepwise selection

- begins with the full least squares model containing all p predictors and then iteratively removes the least ~~square~~ useful predictor, one-at-a-time.

Algorithm 6.3

Step 1: Let M_p denote the full model, which contains all p predictors

Step 2: For $k = p, p-1, \cdots, 2, 1$

2.1 Consider all k models that contain all but one of the predictors, in M_k. for a total of $k-1$ predictors.

2.2 Choose the best among these k models, and call it M_{k-1}. (Smallest RSS or highest R^2)

Step 3: Select a single best model from among $M_0, \cdots M_p$ using cross-validated prediction error, C_p (AIC), BIC or adjusted R^2.

Cons & pros:

- search only $1 + \frac{p(p+1)}{2}$ models when p is too large
- cannot guarantee to yield the best model
- require $n > p$

△ Hybrid method

add variable - remove least useful variables at each stage.

3. Choose the optimal model

Goal: low test error, not a model with low training error

⇒ estimating test error

$\begin{cases} \text{Direct: CV validation} \\ \text{Indirect: adjustment to the training error to account for the bias due to overfitting} \end{cases}$

I: Indirect methods: adjust the training error for the model size, and can be used to select among a set of models with different number of variables

Mallow's $C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$

d: # of parameters used

$\hat{\sigma}^2$: the estimate of the variance of ε

$AIC = -2\log L + 2 \cdot d$

$BIC = \frac{1}{n}(RSS + \log(n) \cdot d \cdot \hat{\sigma}^2)$

Adjusted $R^2 = 1 - \frac{RSS/n-d-1}{TSS/n-1}$ d: # of predictors

Notes: ① BIC statistically generally places a heavier penalty on models with many variables, and hence, results in the selection of smaller models than C_p

② In the case of the linear model with Gaussian errors, maximum likelihood and least squares are the same thing, and C_p and AIC are equivalent.

③ C_p, BIC, AIC, a small value indicates a model with a low test error

④ A large value of adjusted R^2 indicates a model with a small test error.

⑤ $\frac{RSS}{n-1-d}$, RSS always decreases, while $\frac{RSS}{n-1-d}$ can be decreasing or increasing.

⑥ Adjusted R^2 pays a price for including unnecessary variables in the model.

⑦ Cp, AIC, BIC all have rigorous theoretical justifications

⑧ Despite its popularity and even though it is quite intuitive, the adjusted R^2 is not as well motivated in statistical theory as AIC BIC and Cp.

II. Direct Methods : Cross-validation or validation set

— provides a direct estimate of the test error, and don't require an estimate of the error variance σ^2

— used in a wider range of model selection tasks, even in cases where it is hard to pinpoint the model degrees of freedom and hard to estimate the error variance σ^2.

— Each of the procedures returns a sequence of models M_k indexed by model size $k=0, 1, 2, \ldots$ choose $\hat{k}$ $M_{\hat{k}}$

— We compute the ~~variance~~ validation error or the cross-validation error for each model M_k under consideration, and then select the k for which the resulting estimated test error is smallest.

* one - standard error rule = select the smallest model for which the estimated test error is within one standard error of the lowest point on the curve.

4. Shrinkage Methods

I. Ridge Regression

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately.

Notes :

① $\hat{\beta}^R$ minimize $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$

- $\lambda \sum_{j=1}^{p} \beta_j^2$ is called shrinkage penalty, is small when $\beta_1 \cdots \beta_p \approx 0$

- λ serves to control the relative impact of these two terms on the regression coefficient estimates.

II. Lasso Regression

$\hat{\beta}^L$ are the values that minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

* From another perspective to understand Ridge Regression & The Lasso

Ridge Regression

minimize $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ subject to $\sum_{j=1}^{p} \beta_j^2 \le s$

Lasso Regression

minimize $\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ subject to $\sum_{j=1}^{p} |\beta_j| \le s$