01 Model Selection

Simple Linear Regression

$Y = \beta_0 + \beta_1 X + \epsilon_i$

Assumption: the relationship between Y and $X_1/X_2/X_3$ are linear.

- Fitted model will give a prediction of Y: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual: $e_i = y_i - \hat{y}_i$

Multiple Linear Regressin

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$

- 各predictor互相uncorrelated是理想情况
- 各predictor之间有correlation会导致variance of each coefficient tends to high. Can not easily interpret.

Lest Square Approach

Minimize $RSS = \sum_{i=1}^{n} \left[y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \right]^2$ to find a set of estimates of βs.

Confidence Interval of β Estimate

CI: $\hat{\beta}_j \pm t_{\frac{1-\alpha}{2}} SE(\hat{\beta}_j)$

T-Test

H₀: $\beta_j = 0$ Hₐ: $\beta_j \neq 0$

$$t - stat = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t_{n-2}$$

p-value is the probability that we observe a value that is more extreme than t-stat.

RSS (Residual Sum of Squares)

$$RSS = e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2$$

TSS (Total Sum of Squares)

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R^2 (R-squared)

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

- describe the portion of variance that explained by the model

RSE (Residual Standard Error)

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

Predictor (Variable) Type

- Continuous variable
- Ordinal variable
- Nominal variable

Non-Linear Effect of Predictor

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^2 + \epsilon$
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

Linear Model Selection

Subset Selection

Select a subset of predictors to use.

Forward Stepwise Selection

从null开始, 每次add一个predictor, 选择那个add之后RSS最小的. Until some stopping rule satisfied (eg. all remaining predictors are insignificant)

Backward Stepwise Selection

Contain all predictors. 每次remove the predictor with largest p-value, and fit a new model. Until some stopping rule is reached. (eg. all predictors remain in the model is significant)

- only good when p<n

More Systematic Criteria

AIC, BIC, adjusted R^2

- Used to choose an “optimal” model in the path of forward and backward stepwise selection.

REMEMBER:

- 包括了所有predictors的model永远有最小的RSS (training set error) 和最大的R^2. 所以RSS和R^2不能用来比较predictors数量不同的model们.
- training set error小没用, 我们本质上是需要test set error小.

Shrinkage

Fit a model contain all predictors but has constraint the value of coefficients (shrinkage the coefficients to zero)

- Shrinkage the coefficient can significantly reduce the variance

Lasso Regression

We want to minimize $$\sum_{i=1}^{n}\left[y_i-(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})\right]^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

- shrinkage the estimate value of βs to be small. That is, all predictors will be included even some of them has a minor coefficient.
- λ is the tuning parameter, select a good value. CV is used to evaluate how good a value is. For example, use p-fold-CV we split the dataset to p subset. For each subset, leave is as validation set and use other p-1 subset to train, get a model with a λ value with is best for the p-1 subsets. Totally we will get p λ values.
- Work well is p is just a few thousand and if X-variables are uncorrelated each other.

Ridge Regression

We want to minimize $$\sum_{i=1}^{n}\left[y_i-(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij})\right]^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

- shrinkage the estimate value of βs toward zero. That is, may reduce the number of parameters.
- λ is the tuning parameter, select a good value. CV is used to evaluate how good a value is.

Dimension Reduction

Covert **p** predictors to **m** linear combinations of the **p** predictors where m < p. Then build models (least square/ lasso/ ridge/ …) using the **m** new predictors.

Principle Components Regression

- First do principle component analysis. The 1st PC has largest variance, the second….
- PCs are uncorrelated with each other
- Then fit a regression model using these PCs.

Note:

- PCs are created in an unsupervised way (response Y is not used in PCA).
- There is no guarantee that the PCs best explained original predictors will also explained the response.

Partial Least Squares

- PLS identify new predictors in a supervised way (response Y is used)
- Then fit a regression model using these new predictors.

Note:

PLS approach attempt to find new predictors (directions) that explained both original predictors and response.