ACTL3142

Week 3: Linear Regression 2

Tetian Madfouni

t.madfouni@unsw.edu.au

Announcements

Assignment

- This week covers predictor selection, important for assignment
- Assignment may have some bonus marks allocated through the Kaggle competition format that they’re trying to set up.

Overview: Linear Model Selection

What methods can we use?

- Subset Selection
- Indirect Methods
- Shrinkage (or regularisation)
- Dimension Reduction

For this week we focus on subset selection, what is the general method subset selection methods follow?

- Pick a subset
- Compare multiple other subset(s), based on whether its forward, hybrid, best subset or backward selection
- Compare some selection criteria between the two C_p, BIC, AIC
- Pick the one with the most improvement/best selection criteria

Overview: Selection Criteria/Metrics

R^2 isn't the best solution. Why? And what are the general alternative methods?

- It's effectively a measure on the training set, will always be better with more flexible model and more parameters
- Can either use indirect methods or direct

What are the main indirect methods and why do they work better?

- C_p, AIC, BIC, $\mathrm{Adjusted\ R}^2$ – try to estimate test error and penalise model for using more parameters to varying degrees

Which is best?

- As always, it depends

Conceptual Q1

We perform best subset, forward stepwise and backward stepwise selection on a dataset. For each approach we have $p + 1$ models with $0, 1, 2, \ldots, p$ predictors

a. Which of the three models with k predictors has the smallest *training* RSS?

b. Which of the three models with k predictors has the smallest *test* RSS?

c. True or False

   i. The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.

   ii. The predictors in the *k*-variable model identified by backward stepwise are a subset of the predictors in the (*k*+1)- variable model identified by forward stepwise selection.

   iii. The predictors in the *k*-variable model identified by best subset are a subset of the predictors in the (*k*+1)-variable model identified by best subset selection.