STA 674

Regression Analysis And Design Of Experiments

Comparing and Selecting Models – Lecture 2

STA 674, RADOE:
Comparing and Selecting Models

- Last time, introduced the topic and discussed the "big dog": model/variable selection using a particular criterion and looking at all possible combinations of predictors.
- This time, we talk about the stepwise methods.

STA 674, RADOE:
Comparing and Selecting Models

Variable Selection

2. Stepwise Selection

- Select subset of predictors by sequentially adding variables to or removing from the model.
- Possible Implementations
    1. Backward elimination
    2. Forward selection
    3. Stepwise regression

STA 674, RADOE:
Comparing and Selecting Models

Variable Selection

2. a) Backward Elimination
    1. Fit model with all predictors.
    2. Remove predictor with t-value closest to 0/largest p-value.
    3. Refit model.
    4. Repeat 2 and 3 until all remaining predictors are statistically significant.

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

- Response

  y = log(Full Expiratory Volume)

- Predictor Variables

  x_1 = height

  x_2 =smoking (0=no,1=yes)

  x_3 =gender (0=female,1=male)

  all pairwise interactions:

  - height and smoking,
  - height and gender,
  - smoking and gender

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 1

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.28744	0.11325	-20.20	<.0001
Ht	1	0.05225	0.00189	27.59	<.0001
Smoke	1	0.38380	0.45971	0.83	0.4041
Gender	1	0.03685	0.14035	0.26	0.7930
ht_x_gender	1	-0.00031873	0.00232	-0.14	0.8908
smoke_x_gender	1	0.03216	0.04910	0.65	0.5128
smoke_x_ht	1	-0.00607	0.00713	-0.85	0.3943

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 2

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.27490	0.06693	-33.99	<.0001
Ht	1	0.05204	0.00111	46.77	<.0001
Smoke	1	0.38292	0.45931	0.83	0.4048
Gender	1	0.01765	0.01263	1.40	0.1628
smoke_x_gender	1	0.03030	0.04718	0.64	0.5209
smoke_x_ht	1	-0.00604	0.00712	-0.85	0.3961

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 3

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.27409	0.06689	-34.00	<.0001
Ht	1	0.05200	0.00111	46.80	<.0001
Smoke	1	0.23778	0.39971	0.59	0.5521
Gender	1	0.01982	0.01216	1.63	0.1037
smoke_x_ht	1	-0.00365	0.00606	-0.60	0.5471

Delete this predictor because we keep the main effect (SMOKE)...and delete the interaction (smoke_x_height) even though it has a slightly higher P value

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 4

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.26723	0.06588	-34.41	<.0001
Ht	1	0.05190	0.00110	47.31	<.0001
Smoke	1	-0.00272	0.02069	-0.13	0.8956
Gender	1	0.01881	0.01204	1.56	0.1188

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 5

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.26498	0.06358	-35.62	<.0001
Ht	1	0.05186	0.00105	49.55	<.0001
Gender	1	0.01901	0.01193	1.59	0.1117

STA 674, RADOE:
Comparing and Selecting Models

Example – Effect of Smoking on Lung Capacity

Backward Elimination – Step 6

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-2.27143	0.06353	-35.75	<.0001
Ht	1	0.05212	0.00103	50.38	<.0001

STA 674, RADOE:
Comparing and Selecting Models

Variable Selection

2. b) Forward Selection
    1. Fit all models with one predictor.
    2. Choose predictor that is most significant and add it to the model.
    3. Examine all remaining predictors; find their partial F statistics when added to the variables currently in the model.
    4. Repeat 2 and 3 until none of the remaining predictors are statistically significant.

Partial F...found in ANOVA...does decrease in in variability merit increase in model complexity

STA 674, RADOE:
Comparing and Selecting Models

Variable Selection

2. c) Stepwise Regression: combine backward elimination and forward selection.
    1. Start by fitting all models with one predictor.
    2. i. Add most significant predictor (if one exists).
    ii. Remove least significant predictor (if any). Remove INsignificant predictor (if any) by using t or P value
    3. Repeat 3.
    4. Stop when:
        - no remaining predictors can be added and
        - no existing predictors can be removed.

STA 674, RADOE:
Comparing and Selecting Models

Stepwise Selection Methods

Advantages

- 1. Don't need to fit all models.
- 2. Objective (once you have selected your criterion).

Disadvantages

- 1. Definition of best model depends on selection strategy.
- 2. Different methods may select different models.
- 3. Selected model is not necessarily optimal for any fit criterion.