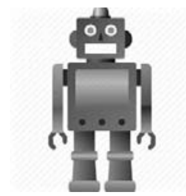
Automated model selection

HL Chapter 4 – part 5



Some options

- Stepwise selection

- Best subsets selection

Before conducting any automated model selection …
… at a minimum …

…perform the following univariate analyses

- Categorical variables
  - Cross-tabulation
  - Collapsing or removal of categories if necessary/possible
- Continuous variables
  - Descriptive statistics and extreme observations
  - Removal/correction of outliers if necessary
  - Checking scale

Stepwise selection

(forwards and backwards)

Main effects model

- List ALL study variables in the model statement

Entering variables

- Select p_{Entry}
  - At step 0, the intercept is added
  - At each subsequent step, all study variables that are not yet part of the model are added (one at a time) to the model from the previous step
  - The new variable with the lowest p-value is selected into the model as long as the p-value is less than p_{Entry}

Removing variables

- Select p_{Exit}
  - At each step, after the new variable has entered the model, the p-values of all model covariates are compared to p_{Exit}
  - Variables with a p-value$\geq$ p_{Exit} are removed from the model
  - p_{Exit} must be greater than p_{Entry}; otherwise the same variable may be repeatedly entered and removed

Categorical variables with > 2 categories

- SAS Class statement
  - Stepwise procedure tests overall statistical significance of the set of design variables
  - If only some of the design variables in a set are significant, the overall test may be non-significant
  - Important design variables may be missed

Categorical variables with > 2 categories

- Design variables created in data step
  - Variables are treated as separate variables
  - It is possible that the final model only includes a subset of a set of design variables
  - This may not be biologically meaningful

Variables with 0 cells

- Interaction terms may have 0 cells
- In the stepwise procedure SAS output, find any table called

  Analysis of Effects Eligible for Entry
- Determine which interaction terms could not be tested due to 0 cells (missing values for test statistic and p-value)

Scale assessment

- If you already assessed scale in a purposeful selection, you are golden
- If you have not assessed scale yet, you must do it now
- In this example, based on the results from purposeful selection, we keep all continuous variables linear

Pros and cons of stepwise selection

Pros

- Quick and easy (in theory)

Cons

- Confounders may be missed
- Biological/clinical importance is ignored
- Model stability is ignored
- Categorical variables with >2 categories may be treated incorrectly

Conclusion

- Avoid automated stepwise model selection like the plague

Example: GLOW500 data set

```
proc logistic descending data=glow500;
  class site_id raterisk/param=ref ref=first;
  model fracture= site_id priorfrac age weight height
        bmi premeno momfrac armassist
        smoke raterisk fracscore
        /stepwise sle=0.15 sls=0.20 details;
run;
```

All study variables

p_{Entry} p_{Exit}

Results from stepwise selection

Step	Variable entered	Model	p-values	Variable removed
1	FRACSCORE	FRACSCORE	<0.0001	None

Results from stepwise selection

Step	Variable entered	Model	p-values	Variable removed
2	RATERISK	RATERISK	0.0131	None
		RATERISK 2 vs. 1	0.0584	
		RATERISK 3 vs. 1	0.0033	
		FRACSCORE	<0.0001	

Results from stepwise selection

Step	Variable entered	Model	p-values	Variable removed
3	HEIGHT	HEIGHT	0.0332	None
		RATERISK	0.0140	
		RATERISK 2 vs. 1	0.0817	
		RATERISK 3 vs. 1	0.0035	
		FRACSCORE	<0.0001	

Results from stepwise selection

Step	Variable entered	Model	p-values	Variable removed
4	PRIORFRAC	PRIORFRAC	0.0990	None
		HEIGHT	0.0355	
		RATERISK	0.0312	
		RATERISK 2 vs. 1	0.0946	
		RATERISK 3 vs. 1	0.0086	
		FRACSCORE	0.0002	

Note that PRIORFRAC is not significant at the 0.05 level; may want to remove this variable from the model

Results from stepwise selection

Step	Variable entered	Model	p-values	Variable removed
5	None			

Recall main effects model from purposeful selection

- Priorfrac
- Age
- Height
- Momfrac
- Armassist
- Raterisk (3 vs. 1,2)

Stepwise selection model
- Height
- Raterisk
- Fracscore

is entirely different

Stepwise selection of interactions

- List the variables in the main effects model (here, height, raterisk and fracscore)
- Use transformed variables if indicated
- Also list all interactions of interest between model covariates
- Tell SAS to include the main effects in the model and to then select interactions

Stepwise selection of interactions

```
proc logistic descending data=glow500;
 model fracture = height raterisk2 fracscore
                  height*raterisk2 height*fracscore
                  raterisk2*fracscore
         /stepwise sle=0.15 sls=0.20 include=3 details;
run;
```

Automatically include the first 3 variables listed; the first 3 variables are the main effects

Results

- In this example, no interactions are significant at the 0.15 level (results not shown)

Recall final model from purposeful selection

- Priorfrac
- Age
- Height
- Momfrac
- Armassist
- Raterisk (3 vs. 1,2)
- Age × Priorfrac
- Momfrac × Armassist

Stepwise selection model
- Height
- Raterisk
- Fracscore

is entirely different

Conclusion repeated

- Avoid automated stepwise model selection like the plague

Best subsets selection

Main effects model

- List ALL study variables in the model statement

Minimum number of covariates

- Select *start*
  - *Start* is the minimum number of model covariates
  - Example: start=3

    With start=3, best subsets selection will not suggest models with only 1 or 2 covariates. There will be at least 3 covariates in each suggested model.

Maximum number of covariates

- Select *stop*
  - *Stop* is the maximum number of model covariates
  - Example: stop=8

    With stop=8, best subsets selection will not suggest models with more than 8 covariates

How many "best" models?

- Select *best*
  - *Best* is the number of models you would like to see for each number of model covariates
  - Example: start=3 stop=8 best=4

    With best=4, best subsets selection will show the best 4 models with 3 covariates, the best 4 models with 4 covariates, …, the best 4 models with 8 covariates

What is "best"?

- How <u>should</u> best subset selection decide which models are best?
  - The models are not nested

    ➔ Likelihood ratio test not appropriate
  - Mallow's C_q should be used
  - Mallow's C_q allows for comparisons between non-nested models with different numbers of covariates

What is "best"?

- How does best subset selection in SAS decide which models are best?
  - SAS calculates Score Chi-square values
  - Score Chi-square values allow for comparisons between non-nested models
  - BUT: Score Chi-square values depend on the number of variables in the model
  - So, Score Chi-square values do not allow for comparisons between models with different numbers of covariates

Categorical variables with > 2 categories

- The SAS class statement does not work with best subsets selection
- Must create your own design variables

- That's a lot of work
- Variables are treated as separate variables
- It is possible that the final model only includes a subset of a set of design variables
- This may not be biologically meaningful NONSENSE

Variables with 0 cells

- Interaction terms may have 0 cells
- Best subsets selection lets you know which interaction terms were not tested

Scale assessment

- If you already assessed scale in a purposeful selection, you are golden
- If you have not assessed scale yet, you must do it now
- In this example, based on the results from purposeful selection, we keep all continuous variables linear

Pros and Cons

Pros

- Quick and easy (in theory)
- Variables missed in purposeful selection may be found

Pros and cons

Cons

- Biological/clinical importance is ignored
- Model stability is ignored
- Categorical variables with >2 categories may be treated incorrectly
- Cannot use class statement in SAS
- Should not be used in the presence of variables with 0 cells

Conclusion

- We will not use best subsets selection to find the best model
- We will use best subsets selection to make sure we didn't miss any important variables during purposeful model selection
- We will only run best subsets selection after completing purposeful selection to double check our variable selection

Example: GLOW500 data set

In the data step
- Create design variables s2-s6 for variable site_id
- Create design variables r2, r3 for variable raterisk

```
proc logistic descending data=glow500;
  model fracture= s2 s3 s4 s5 s6 priorfrac age weight
        height bmi premeno momfrac
        armassist smoke r2 r3 fracscore
        / selection=score start=3 stop=8 best=4;
run;
```

All study variables

Recall main effects model from purposeful selection

- Priorfrac
- Age
- Height
- Momfrac
- Armassist
- Raterisk (3 vs. 1,2)

Results from best subsets selection

Regression Models Selected by Score Criterion		
# of Vars	Score Chi-Square	Variables Included in Model
3	45.2031	HEIGHT r3 FRACSCORE
3	44.4595	PRIORFRAC HEIGHT FRACSCORE
3	44.2335	PRIORFRAC r3 FRACSCORE
3	43.5541	r2 r3 FRACSCORE

FRACSCORE not included in purposeful selection model

Results from best subsets selection

Regression Models Selected by Score Criterion		
# of Vars	Score Chi-Square	Variables Included in Model
4	48.7581	PRIORFRAC HEIGHT r3 FRACSCORE
4	48.3267	WEIGHT BMI r3 FRACSCORE
4	47.9567	HEIGHT r2 r3 FRACSCORE
4	47.5145	PRIORFRAC HEIGHT MOMFRAC FRACSCORE

FRACSCORE, WEIGHT, BMI not included in purposeful selection model

Results from best subsets selection

Regression Models Selected by Score Criterion		
# of Vars	Score Chi-Square	Variables Included in Model
5	51.5791	PRIORFRAC WEIGHT BMI r3 FRACSCORE
5	51.2432	PRIORFRAC HEIGHT r2 r3 FRACSCORE
5	51.1475	PRIORFRAC HEIGHT MOMFRAC r3 FRACSCORE
5	50.9131	WEIGHT BMI r2 r3 FRACSCORE

FRACSCORE, WEIGHT, BMI not included in purposeful selection model

Results from best subsets selection

Regression Models Selected by Score Criterion		
# of Vars	Score Chi-Square	Variables Included in Model
6	54.1579	PRIORFRAC WEIGHT BMI MOMFRAC r3 FRACSCORE
6	53.9346	PRIORFRAC WEIGHT BMI r2 r3 FRACSCORE
6	53.2872	PRIORFRAC HEIGHT MOMFRAC r2 r3 FRACSCORE
6	53.2592	s2 PRIORFRAC WEIGHT BMI r3 FRACSCORE

FRACSCORE, WEIGHT, BMI not included in purposeful selection model

Results from best subsets selection

Regression Models Selected by Score Criterion		
# of Vars	Score Chi-Square	Variables Included in Model
7	56.1607	PRIORFRAC WEIGHT BMI MOMFRAC r2 r3 FRACSCORE
7	55.7467	s2 PRIORFRAC WEIGHT BMI MOMFRAC r3 FRACSCORE
7	55.7156	PRIORFRAC WEIGHT HEIGHT BMI r2 r3 FRACSCORE
7	55.6436	PRIORFRAC WEIGHT BMI MOMFRAC SMOKE r3 FRACSCORE

FRACSCORE, WEIGHT, BMI not included in purposeful selection model

Results from best subsets selection

Regression Models Selected by Score Criterion

# of Vars	Score Chi-Square	Variables Included in Model
8	57.7211	PRIORFRAC WEIGHT HEIGHT BMI MOMFRAC r2 r3 FRACSCORE
8	57.5959	PRIORFRAC WEIGHT BMI MOMFRAC SMOKE r2 r3 FRACSCORE
8	57.5018	s2 PRIORFRAC WEIGHT BMI MOMFRAC r2 r3 FRACSCORE
8	57.4602	PRIORFRAC AGE WEIGHT BMI MOMFRAC ARMASSIST r2 r3

FRACSCORE, WEIGHT, BMI not included in purposeful selection model

Should we change the main effects model from purposeful selection?

- May want to consider FRACSCORE
- However, FRACSCORE is a summary variable that may not be very meaningful
- May want to consider WEIGHT and/or BMI instead of HEIGHT
- May want to consider removing ARMASSIST

Best subsets selection of interactions

- List the variables in the main effects model (let's use the main effects model from purposeful selection)
- Use transformed variables if indicated

- Also list all interactions of interest between model covariates
- Tell SAS to include the main effects in the model and to then select interactions

Best subsets selection of interactions

```
proc logistic descending data=glow500;
  model fracture=priorfrac momfrac armassist raterisk2 height age

priorfrac*momfrac priorfrac*armassist priorfrac*raterisk2
priorfrac*height priorfrac*age momfrac*armassist  momfrac*raterisk2
momfrac*height  momfrac*age armassist*raterisk2 armassist*height
armassist*age raterisk2*height raterisk2*age height*age
          /selection=score start=6 stop=10 best=4 include=6;
run;
```

Automatically include the first 6 variables listed; the first 6 variables are the main effects

Results

- In this example, the interactions most commonly selected are
  - Momfrac*armassist
  - Priorfrac*age
  - Armassist*height

(Results not shown)

Recall final model from purposeful selection

- Priorfrac
- Age
- Height
- Momfrac
- Armassist
- Raterisk (3 vs. 1,2)
- Age × Priorfrac
- Momfrac × Armassist

> - Armassist*height is not statistically significant in this model ($p=0.2533$)
> - Keep model from purposeful selection

Conclusion repeated

- We will not use best subsets selection to find the best model
- We will use best subsets selection to make sure we didn't miss any important variables during purposeful model selection
- We will only run best subsets selection after completing purposeful selection to double check our variable selection