



Stat 401 B – Lecture 26



Forward Selection

- The Forward selection procedure looks to add variables to the model.
- Once added, those variables stay in the model even if they become insignificant at a later step.


1



Backward Selection

- The Backward selection procedure looks to remove variables from the model.
- Once removed, those variables cannot reenter the model even if they would add significantly at a later step.

2



Mixed Selection

- A combination of the Forward and Backward selection procedures.
- Starts out like Forward selection but looks to see if an added variable can be removed at a later step.

3

Stat 401 B – Lecture 26

Mixed – Set up

Response: MDBH

Stepwise Regression Control

Prob to Enter 0.250
Prob to Leave 0.250
Direction Mixed

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
10.3855	19	0.5466053	0.0000	0.0000	102.4885	-11.1064

Lock Entered Parameter Estimate nDF SS "F Ratio" "Prob>F"

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	6.265	1	0	0.000	1.0000
<input type="checkbox"/>	<input type="checkbox"/>	X1	0	1	6.207045	26.739	0.0001
<input type="checkbox"/>	<input type="checkbox"/>	X2	0	1	0.616027	1.135	0.3008
<input type="checkbox"/>	<input type="checkbox"/>	X3	0	1	7.335255	43.287	0.0000

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
------	-----------	--------	------------	--------	---------	----	---

4

Stepwise Regression Control

- Direction – Mixed
- Prob to Enter – controls what variables are added.
- Prob to Leave – controls what variables are removed.
- Prob to Enter = Prob to Leave

5

Current Estimates

- The current estimates are exactly the same as with the Forward selection procedure.
- Clicking on Step will initiate the Mixed procedure that starts like the Forward procedure.

6

Stat 401 B – Lecture 26

Stepwise Fit
Response: MDBH

Stepwise Regression Control
Prob to Enter 0.250
Prob to Leave 0.250
Direction **Mixed**

Current Estimates

	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	3.0502449	18	0.1694581	0.7063	0.6900	19.387747	-33.6102

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	3.8956688	1	0	0.000	1.0000
<input type="checkbox"/>	<input type="checkbox"/>	X1	0	1	1.000159	8.294	0.0104
<input type="checkbox"/>	<input type="checkbox"/>	X2	0	1	0.403296	2.590	0.1259
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X3	32.9371533	1	7.335255	43.287	0.0000

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	X3	Entered	0.0000	7.335255	0.7063	19.388	2

7

Current Estimates – Step 1

- X_3 is added to the model
- Predicted MDBH = $3.896 + 32.937 \cdot X_3$
 - $R^2 = 0.7063$
 - $RMSE = \sqrt{MSE} = \sqrt{0.1694581} = 0.4117$

8

Current Estimates – Step 1

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Because X_3 is statistically significant and is the only variable in the model, clicking on Step will not do anything.

9

Stat 401 B – Lecture 26

Current Estimates – Step 1

- Of the remaining variables not in the model X_1 will add the largest sum of squares if added to the model.
 - SS = 1.000
 - "F Ratio" = 8.294
 - "Prob>F" = 0.0104

10

JMP Mixed – Step 2

- Because X_1 will add the largest sum of squares and that addition is statistically significant, by clicking on Step, JMP will add X_1 to the model with X_3 .

11

Stepwise Fit
Response: MDBH

Stepwise Regression Control
 Prob to Enter 0.250
 Prob to Leave 0.250
 Direction Mixed

Current Estimates

	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	2.0500859	17	0.1205933	0.8026	0.7794	9.7842935	-39.557


Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	3.14321066	1	0	0.000	1.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X1	0.03136639	1	1.000159	8.294	0.0104
<input type="checkbox"/>	<input type="checkbox"/>	X2	0	1	0.670967	7.784	0.0131
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X3	22.953752	1	2.128369	17.649	0.0006

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	X3	Entered	0.0000	7.335255	0.7063	19.388	2
2	X1	Entered	0.0104	1.000159	0.8026	9.7843	3


12

Stat 401 B – Lecture 26

 **Current Estimates – Step 2**


- X_1 is added to the model
- Predicted MDBH = $3.143 + 0.0314 * X_1 + 22.954 * X_3$
 - $R^2 = 0.8026$
 - $RMSE = \sqrt{MSE} = \sqrt{0.1205933} = 0.3473$

13

 **Current Estimates – Step 2**

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Because X_3 and X_1 are statistically significant, clicking on Step will not do anything.

14

 **Current Estimates – Step 2**

- Of the remaining variables not in the model X_2 will add the largest sum of squares if added to the model.
 - $SS = 0.671$
 - "F Ratio" = 7.784
 - "Prob>F" = 0.0131

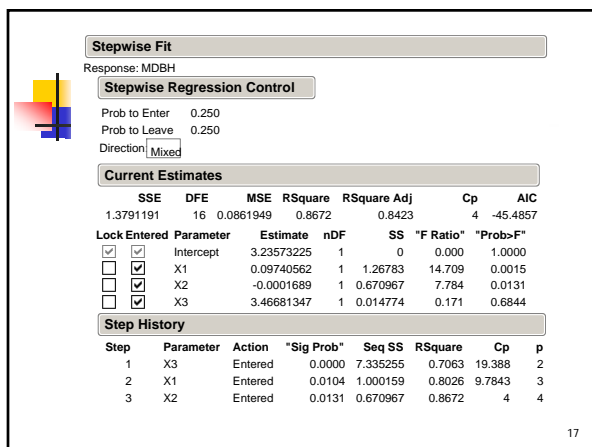
15

Stat 401 B – Lecture 26

JMP Mixed – Step 3

- Because X_2 will add the largest sum of squares and that addition is statistically significant, by clicking on Step, JMP will add X_2 to the model with X_3 and X_1 .

16



The screenshot shows the 'Stepwise Fit' dialog box in JMP. The response variable is 'MDBH'. Under 'Stepwise Regression Control', 'Prob to Enter' and 'Prob to Leave' are both set to 0.250, and 'Direction' is set to 'Mixed'. The 'Current Estimates' table shows the following values:

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
1.3791191	16	0.0861949	0.8672	0.8423	4	-45.4857

The 'Lock Entered' table lists parameters with their estimates, degrees of freedom, sums of squares, F ratios, and p-values:

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	3.23573225	1	0	0.000	1.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X1	0.09740562	1	1.26783	14.709	0.0015
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X2	-0.0001689	1	0.670967	7.784	0.0131
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X3	3.46681347	1	0.014774	0.171	0.6844

The 'Step History' table shows the sequence of parameter additions:

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	X3	Entered	0.0000	7.335255	0.7063	19.388	2
2	X1	Entered	0.0104	1.000159	0.8026	9.7843	3
3	X2	Entered	0.0131	0.670967	0.8672	4	4

17

Current Estimates – Step 3

- X_2 is added to the model
- Predicted $MDBH = 3.236 + 0.0974 \cdot X_1 - 0.000169 \cdot X_2 + 3.467 \cdot X_3$
 - $R^2 = 0.8672$
 - $RMSE = \sqrt{MSE} = \sqrt{0.0861949} = 0.2936$

18

Stat 401 B – Lecture 26

Current Estimates – Step 3

- By clicking on Step you will invoke the Backward part of the Mixed procedure.
- Note that variable X_3 is no longer statistically significant and so it will be removed from the model when you click on Step.

19

Stepwise Fit
Response: MDBH

Stepwise Regression Control
Prob to Enter 0.250
Prob to Leave 0.250
Direction Mixed

Current Estimates

	SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
	1.3938931	17	0.0819937	0.8658	0.8500	2.1714023	-47.2726

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	Intercept	3.26051366	1	0	0.000	1.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X1	0.10691347	1	8.37558	102.149	0.0000
<input type="checkbox"/>	<input checked="" type="checkbox"/>	X2	-0.0001898	1	2.784561	33.961	0.0000
<input type="checkbox"/>	<input type="checkbox"/>	X3	0	1	0.014774	0.171	0.6844

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	X3	Entered	0.0000	7.335255	0.7063	19.388	2
2	X1	Entered	0.0104	1.000159	0.8026	9.7843	3
3	X2	Entered	0.0131	0.670967	0.8672	4	4
4	X3	Removed	0.6844	0.014774	0.8658	2.1714	3

20

Current Estimates – Step 4

- X_3 is removed from the model
- Predicted $MDBH = 3.2605 + 0.1069 \cdot X_1 - 0.0001898 \cdot X_2$
 - $R^2 = 0.8658$
 - $RMSE = \sqrt{MSE} = \sqrt{0.0819937} = 0.2863$

21

Stat 401 B – Lecture 26

Current Estimates – Step 4

- Because X_1 and X_2 add significantly to the model they cannot be removed.
- Because X_3 will not add significantly to the model it cannot be added.
- The Mixed procedure stops.

22

Response MDBH

Summary of Fit

RSquare 0.865785
RSquare Adj 0.849995
Root Mean Square Error 0.286345
Mean of Response 6.265
Observations (or Sum Wgts) 20

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.991607	4.49580	54.8311
Error	17	1.393893	0.08199	Prob > F
C. Total	19	10.385500		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3.2605137	0.333024	9.79	<.0001*
X1	0.1069135	0.010578	10.11	<.0001*
X2	-0.00019	3.256e-5	-5.83	<.0001*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
X1	1	1	8.3755800	102.1490	<.0001*
X2	1	1	2.7845615	33.9607	<.0001*

23

Finding the “best” model

- For this example, the Forward selection procedure did not find the “best” model.
- The Backward and Mixed selection procedures came up with the “best” model.

24

Stat 401 B – Lecture 26

Finding the “best” model

- None of the automatic selection procedures are guaranteed to find the “best” model.
- The only way to be sure, is to look at all possible models.

25

All Possible Models

- For k explanatory variables there are $2^k - 1$ possible models.
- There are k 1-variable models.
- There are $\binom{k}{2}$ 2-variable models.
- There are $\binom{k}{3}$ 3-variable models.


26

All Possible Models

- When confronted with all possible models, we often rely on summary statistics to describe features of the models.
- R^2
- $\text{adj}R^2$
- RMSE

27


Stat 401 B – Lecture 26

 **All Possible Models**

- Another summary statistic used to assess the fit of a model is Mallows C_p .


$$C_p = \left(\frac{SSE_p}{MSE_{Full}} \right) - (n - 2p)$$
$$p = k + 1$$

28

 **All Possible Models**

- The smaller C_p is the “better” the fit of the model.
- The full model will have $C_p = p$.

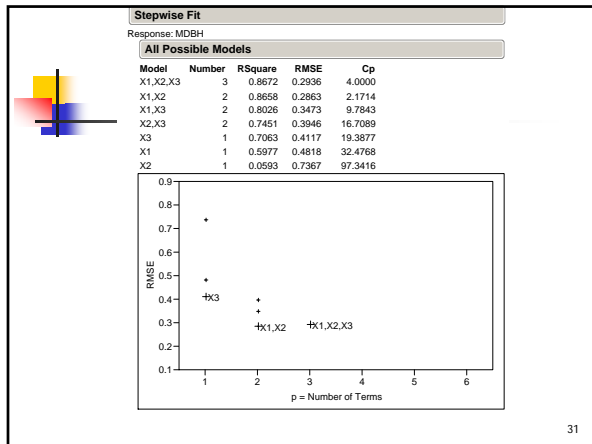
29

 **JMP – Fit Model**

- Personality – Stepwise
- Red triangle pull down – All Possible Models
- Right click on table – Columns
- Check C_p

30

Stat 401 B – Lecture 26



All Possible Models

- Lists all 7 models.
- 3-variable (full) model first.
- 2-variable models – listed in order of the R^2 value.
- 1-variable models – listed in order of the R^2 value.

32

All Possible Models

- Model with X_1, X_2, X_3 – Highest R^2 value.
- Model with X_1, X_2 – Lowest RMSE and lowest C_p .
- Which is “best”? Can’t tell until you look at significance of the variables in the model.

33
