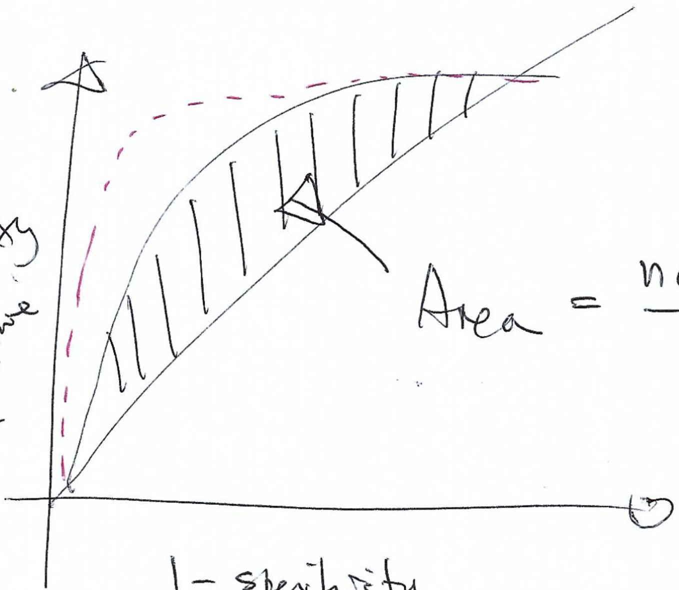
20Nov

ROC

(Receiver operating characteristic curve)



sensitivity = ~~true positive~~ rate

= ~~# of~~

$$= \frac{\text{\# of true positive}}{\text{total \# of positive}}$$

1 − specificity

= false positive rate

$$= \frac{\text{\# of false positive}}{\text{total \# of negative}}$$

$$\text{Area} = \frac{nc + 0.5(t - nc - nd)}{t}$$

nc = # of concordant pairs

nd = # of discordant pairs

t = total # of pairs

0.9	<	area	<	1	excellent
0.8	<	"	<	0.9	good
0.7	<	"	<	0.8	fair
0.6	<	"	<	0.7	Pass
0.5	<	"	<	0.6	Fail

Example 1

$$\log\left(\frac{\hat{P}}{1-\hat{P}}\right) = -5.5784 + 1.14 * \cancel{\text{logload}}\ \ln(\text{load})$$

$$\Rightarrow \hat{P} = \frac{\exp(-5.5784 + 1.14 * \ln(\text{load}))}{1 + \exp(-5.5784 + 1.14 * \ln(\text{load}))}$$

load	90	80	70	35	5	
predicted prob. →	0.38966	0.35824	0.32404	0.17867	0.02312	
1	130	95	189	95	13	522
0	170	205	411	405	587	1778

↑ obs. positive

obs negative

Rule
- 1 true positive
- 0 false negative
- 1 false positive
- 0 true negative

Rule positive if est. prob. ≥ 0.38966

↑ $y=1$

negative if est. prob < 0.38966

\# of true positive $= 130$

\# of false negative $= 95 + 189 + 95 + 13 = 522 - 130 = 392$

\# of false ~~negative~~ positive $= 170$

\# of true negative $= 205 + 411 + 405 + 587 = 1778 - 170 = 1608$

Sensitivity = # of true positive / # of observed positive ~~value~~

$= 130 / 522 = 0.24904$

1 − Specificity = # of false positive / # of observed negative

$= 170 / 1778 = 0.09561$

$nc = \sum_{\text{all rules}}$ # of true positive * # of true negative

$nd = \sum_{\text{all rules}}$ # of false negative * # of false positive

Tied pairs $= 130 * 170 + 95 * 205 + 189 * 411 + 95 * 405 + 13 * 587 =$

$t = nc + nd +$ tied pairs

e.g. area $= 0.72$

Chapter 4 Best model (Back to linear regression)

1. Sequential variable selection procedures

(a). Forward selection

Step 0 β_0

Step 1 Find the most significant variables

$H_0: \beta_j = 0$ ← Model under H_0: $y_i = \beta_0 + e_i$

vs $H_1: \beta_j \neq 0$ " " H_1: $y_i = \beta_0 + \beta_j x_{ij} + e_i$

$j = 1, 2, 3, 4, 5$

$$F = \frac{\underbrace{\text{Reg S.S.}}_{R(\beta_j|\beta_0)}/1}{\text{Res S.S.}|_{H_1}/(n-p')}$$

$$= \frac{\cancel{\text{Res S.S.}}\,(\text{Total S.S.} - \text{Res S.S.}|_{H_1})/1}{\text{Res S.S.}|_{H_1}/(n-p')}$$

Res S.S. | when the model with β_0 only

↑

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

Model with β_0 only $\Rightarrow \hat{\beta}_0 = \bar{y}$

$$= \frac{(\text{Res S.S.}|_{H_0} - \text{Res S.S.}|_{H_1})/1}{\text{Res S.S.}|_{H_1}/(n-\boxed{p'})}$$

← same for $\forall j$ " 2 in Step 1

$$= \left(\frac{\boxed{\text{Res S.S.}|_{H_0}}}{\text{Res S.S.}|_{H_1}} - 1\right) * (n-p')$$

Res S.S. | when the model with β_0 only $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, $\hat{y}_i = \hat{\beta}_0 = \bar{y}$

most significant variable $\Rightarrow$ the largest F

$\Updownarrow$

~~$\Rightarrow$~~ smallest Res S.S. $|_{H_1}$

Forward selection

01/02 Final Exam

3. (15 marks) An experiment was conducted to model Y with five explanatory variables X_1, X_2, X_3, X_4 and X_5. We desire an equation of relating Y to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation. The following table gives RSS for all possible regressions. Total sum of squares is equal to 5.0634 and the number of observations is equal to 20.

$\sum_{i=1}^{n}(y_i-\bar{y})^2$ = Res.S.S. | model with β_0 only

No. of parameters in the model	*RSS*	Model
2	2.0338	X_1
2	5.0219	X_2
2	$\boxed{1.5370}$	X_3
2	2.5044	X_4
2	1.5563	X_5
3	1.5921	X_1, X_2
3	1.4397	X_1, X_3
3	1.7462	X_1, X_4
3	1.4963	X_1, X_5
3	1.4707	X_2, X_3
3	2.4381	X_2, X_4
3	1.4388	X_2, X_5
3	1.4590	X_3, X_4
3	$\boxed{1.0850}$	X_3, X_5
3	1.3287	X_4, X_5
4	1.2582	X_1, X_2, X_3
4	1.4257	X_1, X_2, X_4
4	1.2764	X_1, X_2, X_5
4	1.3894	X_1, X_3, X_4
4	1.0644	X_1, X_3, X_5
4	1.3204	X_1, X_4, X_5
4	1.3900	X_2, X_3, X_4
4	$\boxed{0.9871}$	X_2, X_3, X_5
4	1.2178	X_2, X_4, X_5
4	1.0634	X_3, X_4, X_5
5	1.2199	X_1, X_2, X_3, X_4
5	0.9871	X_1, X_2, X_3, X_5
5	1.1565	X_1, X_2, X_4, X_5
5	1.0388	X_1, X_3, X_4, X_5
5	0.9653	X_2, X_3, X_4, X_5
6	0.9652	X_1, X_2, X_3, X_4, X_5

$$F = \frac{\text{Reg S.S.}/1}{\text{Res.S.S}|H_1/(20-3)} = \frac{R(\beta_5|\beta_3,\beta_0)}{\text{Res.S.S.}|H_1/(20-3)}$$

increase in Reg S.S after adding β_5 into the model when the model has X_3 & intercept.

$$F = \frac{\text{Reg S.S.}/1}{\text{Res.S.S}|H_1/(20-4)} = \frac{R(\beta_2|\beta_0,\beta_3,\beta_5)}{\text{Res.S.S.}|H_1/(20-4)}$$

Step 0 β_0

Step 1

$$F = \left(\frac{5.0634}{1.5370} - 1\right) \times (20-2) = 41.298 > F_{\alpha,1,18}$$

$\Rightarrow$ Add X_3 into model

Step 2

$H_0: \beta_5 = 0$ vs $H_1: \beta_5 \neq 0$

model with X_3 / model with X_3 & X_5

$$F = \left(\frac{1.5370}{1.0852} - 1\right) \times (20-3) = 7.082 > F_{\alpha,1,17}$$

$\Rightarrow$ Add X_5 into model

Step 3

$H_0: \beta_2 = 0$ vs $H_1 = \beta_2 \neq 0$

model with X_2, X_3, X_5

$$F = \left(\frac{1.085}{0.9871} - 1\right) \times (20-4) = 1.587 < F_{\alpha,1,16}$$

$\Rightarrow$ X_2 is not significant after the model has X_3 & X_5

$\boxed{\text{STOP}}$

Find the best model by C_p, forward selection, backward selection and stepwise selection. Write down how to get the best model on details. Choose critical values for both ENTRY and STAY to be 2. Comment the results.

Best model : X_3, X_5

(b) Backward elimination

Step 0 Full model : $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_5 x_{i5} + e_i$

Step 1 ~~B~~ Delete most insignificant variable

$H_0 : \beta_j = 0$ $j = 1, 2, 3, 4, 5$

H_1 : full model ~~with~~

$R(\beta_j \mid \beta_0, \beta_1, \cdots \beta_5 \text{ without } \beta_j)$

$$F = \frac{\text{Reg S.S.} / 1}{\text{Res S.S.}|_{H_1} / (n - \underbrace{p'}_{=6})}$$

$$= \cancel{\frac{(\text{Total S.S.} - \text{Reg S.S.}|_{H_0}}{\quad}}$$

$$= \frac{\text{Reg S.S.}|_{H_1} - \text{Reg S.S.}|_{H_0}}{\text{Res S.S.}|_{H_1} / (n - p')}$$

$$= \frac{(\text{Res S.S.}|_{H_0} - \text{Res S.S.}|_{H_1})}{\text{Res S.S.}|_{H_1} \; (n - p')}$$

$$= \left(\frac{\text{Res S.S.}|_{H_0}}{\boxed{\text{Res S.S.}|_{H_1}}} - 1 \right) \times (n - p')$$

~~Stua~~ same $\forall i$

most insignificant variable $\Rightarrow$ smallest F

$\Uparrow$

$\boxed{\text{smallest}}$ Res S.S $|H_0$

Backward elimination

01/02 Final Exam

3. (15 marks) An experiment was conducted to model Y with five explanatory variables X_1, X_2, X_3, X_4 and X_5. We desire an equation of relating Y to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation. The following table gives RSS for all possible regressions. Total sum of squares is equal to 5.0634 and the number of observations is equal to 20.

Best model = X_3, X_5

No. of parameters in the model	*RSS*	Model
2	2.0338	X_1
2	5.0219	X_2
2	$\boxed{1.5370}$	X_3 ← $H_0: \beta_5 = 0$
2	2.5044	X_4
2	1.5563	X_5 ← $H_0: \beta_3 = 0$
3	1.5921	X_1, X_2
3	1.4397	X_1, X_3
3	1.7462	X_1, X_4
3	1.4963	X_1, X_5
3	1.4707	X_2, X_3 ← $H_0: \beta_5 = 0$
3	2.4381	X_2, X_4
3	1.4388	X_2, X_5 ← $H_0: \beta_3 = 0$
3	1.4590	X_3, X_4
3	$\boxed{1.0850}$	X_3, X_5 ← $H_0: \beta_2 = 0$
3	1.3287	X_4, X_5
4	1.2582	X_1, X_2, X_3
4	1.4257	X_1, X_2, X_4
4	1.2764	X_1, X_2, X_5
4	1.3894	X_1, X_3, X_4
4	1.0644	X_1, X_3, X_5
4	1.3204	X_1, X_4, X_5
4	1.3900	X_2, X_3, X_4 $H_0: \beta_5 = 0$
4	$\boxed{0.9871}$	X_2, X_3, X_5 $H_0: \beta_4 = 0$
4	1.2178	X_2, X_4, X_5 $H_0: \beta_3 = 0$
4	1.0634	X_3, X_4, X_5 $H_0: \beta_2 = 0$
5	1.2199	X_1, X_2, X_3, X_4 ← $H_0: \beta_5 = 0$
5	0.9871	X_1, X_2, X_3, X_5 $H_0: \beta_4 = 0$
5	1.1565	X_1, X_2, X_4, X_5 $H_0: \beta_3 = 0$
5	1.0388	X_1, X_3, X_4, X_5 $H_0: \beta_2 = 0$
5	$\boxed{0.9653}$	X_2, X_3, X_4, X_5 $H_1 = \beta_1 = 0$
6	0.9652	X_1, X_2, X_3, X_4, X_5

Step 4 Model under H_1 has X_3, X_5

$H_0: \beta_5 = 0$

$F = \left(\frac{1.5370}{1.085} - 1\right) * (20-3) = 7.082 > F_{\alpha,1,17}$ STOP

Step 3 model under H_1 has X_2, X_3, X_5

$H_0: \beta_2 = 0$

$F = \left(\frac{1.085}{0.9871} - 1\right) * (20-4) = 1.587 < F_{\alpha,1,16}$

Insignificant ⇒ Drop X_2

Step 2 model under H_1 with X_2, X_3, X_4, X_5 $H_0: \beta_4 = 0$

$F = \left(\frac{0.9871}{0.9653} - 1\right) * (20-5) = 0.339 < F_{\alpha,1,15}$

Insignificant ⇒ Drop X_4

Step 1

$F = \left(\frac{0.9653}{0.9652} - 1\right) * (20-6) = 0.00145 < F_{\alpha,1,14}$

Insignificant ⇒ drop X_1

Step 0

Find the best model by C_p, forward selection, backward selection and stepwise selection. Write down how to get the best model on details. Choose critical values for both ENTRY and STAY to be 2. Comment the results.

(c) Stepwise regression $\Leftarrow$ Forward + backward for each step

Step 0 $\quad \beta_0$

Step 1 Forward selection $\Rightarrow$ choose the most significant variable

X_3 is added $\Leftarrow$ $F = 41.298$

Backward elimination $\Rightarrow$ check ~~with~~ whether the variable(s) in the model is insignificant

Can't drop X_3

Step 2 Model has X_3

Forward

X_5 is added $\Leftarrow$ $F = 7.082$ $\Rightarrow$ model has X_3, X_5

Backward

① $H_0 = \beta_3 = 0$ vs H_1 = model with X_3, X_5

1.5563 Res S.S. for the model with ~~X_3~~, X_5

$$F = \left(\frac{\cancel{1.537}}{\cancel{1.5563}} - 1 \right) \times (20-3)$$

1.085 $\Leftarrow$ Res S.S. for the model with X_3, X_5

$= 7.3844 > F_{\alpha, 1, 17}$

X_3 is significant

$\Rightarrow$ Can't drop X_3

② $H_0 = \beta_5 = 0$ vs H_1 = model with X_3, X_5

model under H_0 (has ~~X_5~~ X_3)

$$F = \left(\frac{1.537}{1.0852} - 1 \right) \times (20-3)$$

model ~~with~~ under H_0 (has X_3, X_5)

$= 7.082$

$\Rightarrow$ Can't drop X_5

Step 3 model has X_3, & X_5.

Forward Can't find any significant variable

$\Rightarrow$ STOP

<u>stepwise</u> ① Can't find any significant variable in forward selection

② Enter a varable by Forward

Drop the same varable by backward in later step

⇒ Add the same varable again

STOP!