Ch.9-Building the Regression Model I: Model Selection & Validation: 4, 5, 9, 16, 17.

Problem 9.4

In forward stepwise regression, what advantage is there in using a relatively small α-to-enter value for adding variables? What advantage is there in using a larger α-to-enter value?

Using relatively small α-to-enter value to add variables helps reduce model complexity by decreasing flexibility to add too many predictor variables into the model. Unnecessary variables won't enter the model, which could also increase the noise. On the other hand, using a larger α-to-enter value increases flexibility and allows for more complex models. This could lead to unnecessary variables being included in the model. Depending on the size of the data, n, tuning this α-limit can either make the model more conservative for small n or liberal for larger n sized model, which would affect the estimate of σ^2.

Problem 9.5

In forward stepwise regression, why should the α-to-enter value for adding variables never exceed the α-to-remove value for deleting variables?

Allowing α-to-enter exceed the α-to-remove value can cause cyclic entry and removal of the same variables. Then, by bounding the α-to-enter value by the α-to-remove value we prevent this potential issue.

Problem 9.9

Refer to **Patient satisfaction** Problem 6.15. The hospital administrator wishes to determine the best subset or predictor variables for predicting patient satisfaction.

a. Indicate which subset of predictor variables you would recommend as best for predicting patient satisfaction according to each of the following criteria (Support your recommendations with appropriate graphs):

   (1) $R^2_{a,p} = 66.1\%$: X_1 (patient age) and X_3 (anxiety level)
   (2) $AIC_p = 215.06$: X_1 (patient age) and X_3 (anxiety level)
   (3) $C_p = 2.8$: X_1 (patient age) and X_3 (anxiety level)
   (4) $PRESS_p = 4902.8$: X_1 (patient age) and X_3 (anxiety level).

   See graphs and outputs in Figures 1a through 2b

Best Subsets Regression: Y.PatientSatisfaction versus ... 3.AnxietyLevel

Response is Y.PatientSatisfaction

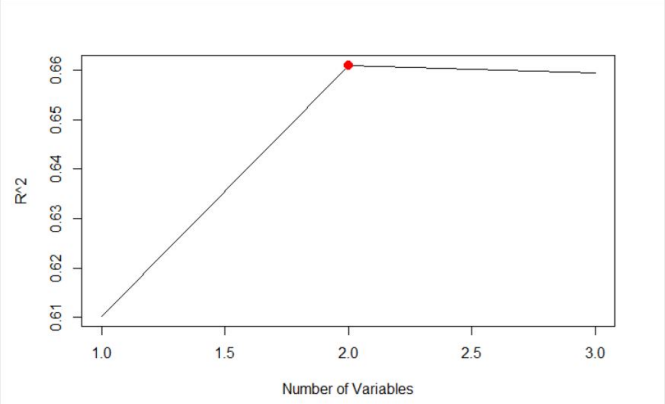
Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	X1. Patient Age	X2. Severity	X3. Anxiety Level
1	61.9	61.0	5569.6	58.3	8.4	10.760	X		
1	41.5	40.2	8451.4	36.8	35.2	13.327			X
2	67.6	66.1	4902.8	63.3	2.8	10.035	X		X
2	65.5	63.9	5235.2	60.8	5.6	10.358	X	X	
3	68.2	65.9	5057.9	62.2	4.0	10.058	X	X	X

(a) "Best" subset output

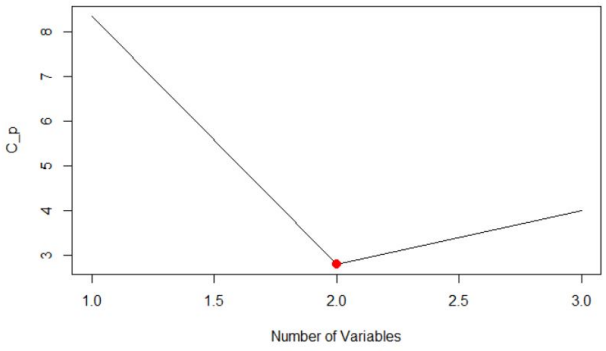
```
  (Intercept)  Age Severity Anxiety
1        TRUE TRUE    FALSE   FALSE
2        TRUE TRUE    FALSE    TRUE
3        TRUE TRUE     TRUE    TRUE
[1] -38.56 -42.20 -39.24
```

(b) AIC output

Figure 1: Graphs for selecting the best subset



(a) $R^2_{a,p}$ output graph



(b) C_p output graph

Figure 2: Graphs for selecting the best subset

b. Do the four criteria in part (a) identify the same best subset? Does this always happen?

Yes, all four criteria in part (a) identify the same best subset. We have seen examples in the book where this is not always the case

c. Would forward stepwise regression have any advantages here as a screening procedure over the all-possible-regressions procedure?

Forward stepwise regression would have only two steps as opposed to five iterations that the all-possible-regressions procedure had evaluated.

Problem 9.16

Refer to **Kidney function** Problem 9.15.

a. Using first-order and second-order terms for each of the 3 predictor variables (centered around the mean) in the pool of potential X variables (including cross products of the first-order terms), find the 3 best hierarchical subset regression models according to the C_p criterion.

The three best hierarchical subset regression models based on the C_p criterion are as follows:
1. X_1, X_2, X_3, X_1X_2;
2. $X_1, X_2, X_3, X_1X_2, X_3^2$;
3. $X_1, X_2, X_3, X_1X_2, X_3^2, X_2^2$.

Response is Y.CreatinineClearance

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	X1. Means	X2. Means	X3. Means	X1. SQRD	X2. SQRD	X3. SQRD	X1X2	X1X3	X2X3
1	64.3	63.1	60.0	48.5	18.896	X								
1	44.6	42.8	36.5	91.2	23.534		X							
2	75.3	73.6	69.8	26.7	15.986	X	X							
2	71.9	70.0	66.7	34.0	17.042	X		X						
3	85.5	84.0	80.8	6.5	12.457	X	X	X						
3	80.9	79.0	74.4	16.4	14.274	X		X		X				
4	87.9	86.2	83.9	3.3	11.582	X	X	X				X		
4	87.3	85.4	83.0	4.6	11.873	X	X	X	X					
5	88.8	86.7	84.0	3.4	11.357	X	X	X			X	X		
5	88.3	86.1	81.7	4.4	11.601	X	X	X		X		X		
6	89.1	86.5	81.7	4.8	11.425	X	X	X		X	X	X		
6	88.9	86.4	83.8	5.0	11.485	X	X	X			X	X	X	
7	89.2	86.2	79.9	6.3	11.547	X	X	X	X	X	X			X
7	89.2	86.2	79.9	6.4	11.551	X	X	X		X	X	X		X
8	89.4	85.9	79.6	8.0	11.699	X	X	X	X	X	X	X		X
8	89.3	85.7	78.6	8.3	11.783	X	X	X	X	X	X		X	X
9	89.4	85.3	78.1	10.0	11.950	X	X	X	X	X	X	X	X	X

Figure 3: C_p output

b. Is there much difference in C_p for the three best subset models?

There is not that much difference between the C_p's, as they are all less than p number of predictor variables (including $X_0 = 1$). Then we need to check, which C_p would give the smallest difference with p.
$p - C_p$:
$5 - 3.3 = 1.7$
$6 - 3.4 = 2.6$
$7 - 4.8 = 2.2$

Problem 9.17

Refer to **Patient satisfaction** Problems 6.15 and 9.9. The hospital administrator was interested to learn how the forward stepwise selection procedure and some of its variations would perform here.

a. Determine the subset of variables that is selected as best by the forward stepwise regression procedure, using F limits of 3.0 and 2.9 to add or delete a variable, respectively. Show your steps.

Using Minitab and setting the α-to-enter equal to 0.04 and α-to-leave to 0.05 (based on the F limits given) we would select `Patient Age` and `Anxiety Level` as the subset of variables selected as best by the forward stepwise regression procedure. (no variables dropped since `Severity` is never added). See Figure 4.

Stepwise Selection of Terms

Candidate terms: X1.PatientAge, X2.Severity, X3.AnxietyLevel

	-----Step 1----		-----Step 2----	
	Coef	P	Coef	P
Constant	119.94		145.9	
X1.PatientAge	-1.521	0.000	-1.200	0.000
X3.AnxietyLevel			-16.74	0.009
S		10.7597		10.0354
R-sq		61.90%		67.61%
R-sq(adj)		61.03%		66.10%
R-sq(pred)		58.34%		63.33%
Mallows' Cp		8.35		2.81

α to enter = 0.04, α to remove = 0.05

Figure 4: Stepwise output in Minitab

b. To what level of significance in any individual test is the F limit of 3.0 for adding a variable approximately equivalent here?

In any individual test the F limit of 3.0 for adding a variable is approximately equivalent to the significance level of 0.1 ($F_{1,44;0.10} = 3$)

c. Determine the subset of variables that is selected as best by the forward selection procedure, using an F limit of 3.0 to add a variable. Show your steps.

Using Minitab and setting the α-to-enter equal to 0.04 (based on the F limits given) we would select `Patient Age` and `Anxiety Level` as the subset of variables selected as best by the forward stepwise regression procedure.

Similar results are achieved when performing similar analysis in R with the following steps:

1. Fit a model without predictors and pick the predictor to add into the model based on the highest F statistic (> 3).

2. After the first run, we would add `Age` as a predictor variable, which would also decrease AIC from 263 to 220.
3. Next we add `Anxiety level` to the model which reduces the AIC to 215 an has higher F statistic.
4. Figure 8 shows that `Severity` should not be added to the model since F-value is not within our F limits, has high P-value, and would increase the AIC.

Forward Selection of Terms

Candidate terms: X1.PatientAge, X2.Severity, X3.AnxietyLevel

	-----Step 1----		-----Step 2----	
	Coef	P	Coef	P
Constant	119.94		145.9	
X1.PatientAge	-1.521	0.000	-1.200	0.000
X3.AnxietyLevel			-16.74	0.009
S		10.7597		10.0354
R-sq		61.90%		67.61%
R-sq(adj)		61.03%		66.10%
R-sq(pred)		58.34%		63.33%
Mallows' Cp		8.35		2.81

α to enter = 0.01

Figure 5: Forward stepwise output in Minitab

```
Single term additions

Model:
Satisfaction ~ 1
         Df Sum of Sq   RSS AIC F Value   Pr(F)
<none>                13369 263
Age       1      8275  5094 220    71.5 9.1e-11 ***
Severity  1      4860  8509 244    25.1 9.2e-06 ***
Anxiety   1      5555  7814 240    31.3 1.3e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Step 1 Forward Stepwise output in R

```
Single term additions

Model:
Satisfaction ~ Age
         Df Sum of Sq  RSS AIC F Value  Pr(F)
<none>                5094 220
Severity  1       481 4613 218    4.48 0.0401 *
Anxiety   1       763 4330 215    7.58 0.0086 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: Step 2 Forward Stepwise output in R

```
Single term additions

Model:
Satisfaction ~ Age + Anxiety
         Df Sum of Sq  RSS AIC F Value Pr(F)
<none>                4330 215
Severity  1      81.7 4249 216   0.807  0.37
```

Figure 8: Step 3 Forward Stepwise output in R

d. Determine the subset of variables that is selected as best by the backward elimination procedure, using an F limit of 2.9 to delete a variable. Show your steps.

Using Minitab and setting the α-to-leave equal to 0.05 (based on the F limits given) we would select `Patient Age` and `Anxiety Level` as the subset of variables selected as best by the backward stepwise regression procedure. `Severity` is dropped on the Step 2.

Similar results are achieved when performing similar analysis in R with the following steps:

1. Fit a model with all predictor variables included and pick the predictor to drop based on the lowest F statistic (< 3).
2. After the first run, we would drop `Severity` because the F-value is 0.81.
3. Figure 11 shows that both `Age` and `Anxiety` remain in the model based on the F-limit.

Backward Elimination of Terms

Candidate terms: X1.PatientAge, X2.Severity, X3.AnxietyLevel

	-----Step 1----		-----Step 2----	
	Coef	P	Coef	P
Constant	158.5		145.9	
X1.PatientAge	-1.142	0.000	-1.200	0.000
X2.Severity	-0.442	0.374		
X3.AnxietyLevel	-13.47	0.065	-16.74	0.009
S		10.0580		10.0354
R-sq		68.22%		67.61%
R-sq(adj)		65.95%		66.10%
R-sq(pred)		62.17%		63.33%
Mallows' Cp		4.00		2.81

α to remove = 0.05

Figure 9: Backward stepwise output in Minitab

e. Compare the results of the three selection procedures. How consistent are these results? How do the results compare with those for all possible regressions in Problem 9.9?

All three procedures give the same result of only including ***`Age`*** *and* ***`Anxiety`*** *in our model with the same regression function:* $Y = 145.9 - 1.200X1 - 16.74X3$. These results are consistent with the possible regressions in Problem 9.9, where based on different criteria, same predictor variables were selected.

```
Single term deletions

Model:
Satisfaction ~ Age + Severity + Anxiety
         Df Sum of Sq  RSS AIC F Value   Pr(F)
<none>                4249 216
Age       1      2858 7106 238   28.25 3.8e-06 ***
Severity  1        82 4330 215    0.81   0.374
Anxiety   1       364 4613 218    3.60   0.065 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: Step 1 Backward Stepwise output in R

```
Single term deletions

Model:
Satisfaction ~ Age + Anxiety
         Df Sum of Sq  RSS AIC F Value   Pr(F)
<none>                4330 215
Age       1      3484 7814 240    34.6 5.4e-07 ***
Anxiety   1       763 5094 220     7.6  0.0086 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Step 2 Backward Stepwise output in R

Additional Problem

Use the **Patient Satisfaction** data (Prob 6.15) to show the equivalence of the t and F tests when testing $H_0 : \beta_2 = \beta_3$.

T-test

$$H_0 : \beta_2 = \beta_3$$
$$H_a : \beta_2 \neq \beta_3$$
$$\beta_2 = -0.442;\ \beta_3 = -13.47$$
$$s_{\beta_2-\beta_3} = \sqrt{s^2_{\beta_2} + s^2_{\beta_3} - 2\hat{\sigma}^2_{\beta_2\beta_3}} =$$
$$= \sqrt{0.24203 + 50.4052 - 2 \cdot (-1.7916)} =$$
$$= 7.36413$$
$$t^* = \frac{-13.028}{7.36413} = -1.769$$
$$P = 2 \cdot P\{t_{42} > 1.769\} = 2 \cdot tcdf(1.769, \infty, 42) =$$
$$= 2 \cdot 0.042068 = 0.08414$$
$$(t^*)^2 = 3.1294 = F^*$$

F-test

Restricted model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 + X_3)$$
$$H_0 : \beta_2 = \beta_3$$
$$H_a : \beta_2 \neq \beta_3$$
$$MSE_F = 101.2;\ SSE_F = 4248.8;$$
$$SSE_R = 4565.5$$
$$F^* = \frac{4565.5 - 4248.8}{101.2} =$$
$$= \frac{316.7}{101.2} = 3.1294$$
$$P = P\{F_{1,42} > 3.1294\} =$$
$$= Fcdf(3.1294, \infty, 1, 42) = 0.08415$$