9.9 Refer to patient satisfaction problem 6.15. The hospital administrator wishes to determine the best subset of 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., $R^2_{a,p}$, AIC_p, C_p, and $PRESS_p$. Support your recommendations with appropriate graphs.

Answer: We note that

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/n-1}$$

where SSE_p is the sum of squares when fitting linear model with subset with $p-1$ predictor variables (p is number of parameters which includes the intercept). The Akaike's information criterion is

$$AIC_p = n\ln SSE_p - n\ln n + 2p$$

The prediction sum of squares is

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

where h_{ii} are the diagonal elements of the hat matrix, **H**.

Finally, with P-1 potential X variables, and p parameters (2 if simple linear regress, as β_0 and β_1)

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{p-1})} - (n-2p) = \frac{SSE(X_{p-1})}{SSE(X_1,\ldots X_{P-1})/(n-P))} - (n-2p)$$

We can construct 2^{p-1} alternative models. These are described in table 1

The best subset is X_1, X_3.

Table 1: Results from question 1.

Variables	p	SSE_p	$R^2_{a,p}$	AIC_p	$PRESS_p$	C_p
none	1	13,369.300	0	262.915	13790.098	88.16
X_1	2	5,093.915	0.610	220.529	5569.562	8.353
X_2	2	8,509.044	0.349	244.131	9254.489	42.112
X_3	2	7,814.391	0.402	240.214	8451.432	35.245
X_1, X_2	3	4,613.000	0.639	217.968	5235.192	5.560
✓X_1, X_3	3	4,330.500	0.661	215.061	4902.751	2.807
X_2, X_3	3	7,106.394	0.444	237.84	8115.912	30.247
all	4	4,248.841	0.659	216.185	5057.886	4.00

Answer: Figure (1) graphically shows again the best subset is X_1, X_3.

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

Answer: In this case, they do. This is not always the case, as they describe different aspects of the model.

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

Answer: Here, since there are not very many variables, forward stepwise regression seems less useful. One issue with forward stepwise regression is that MSE will tend to be inflated in initial steps, but since there are not that many steps, this could be an issue.

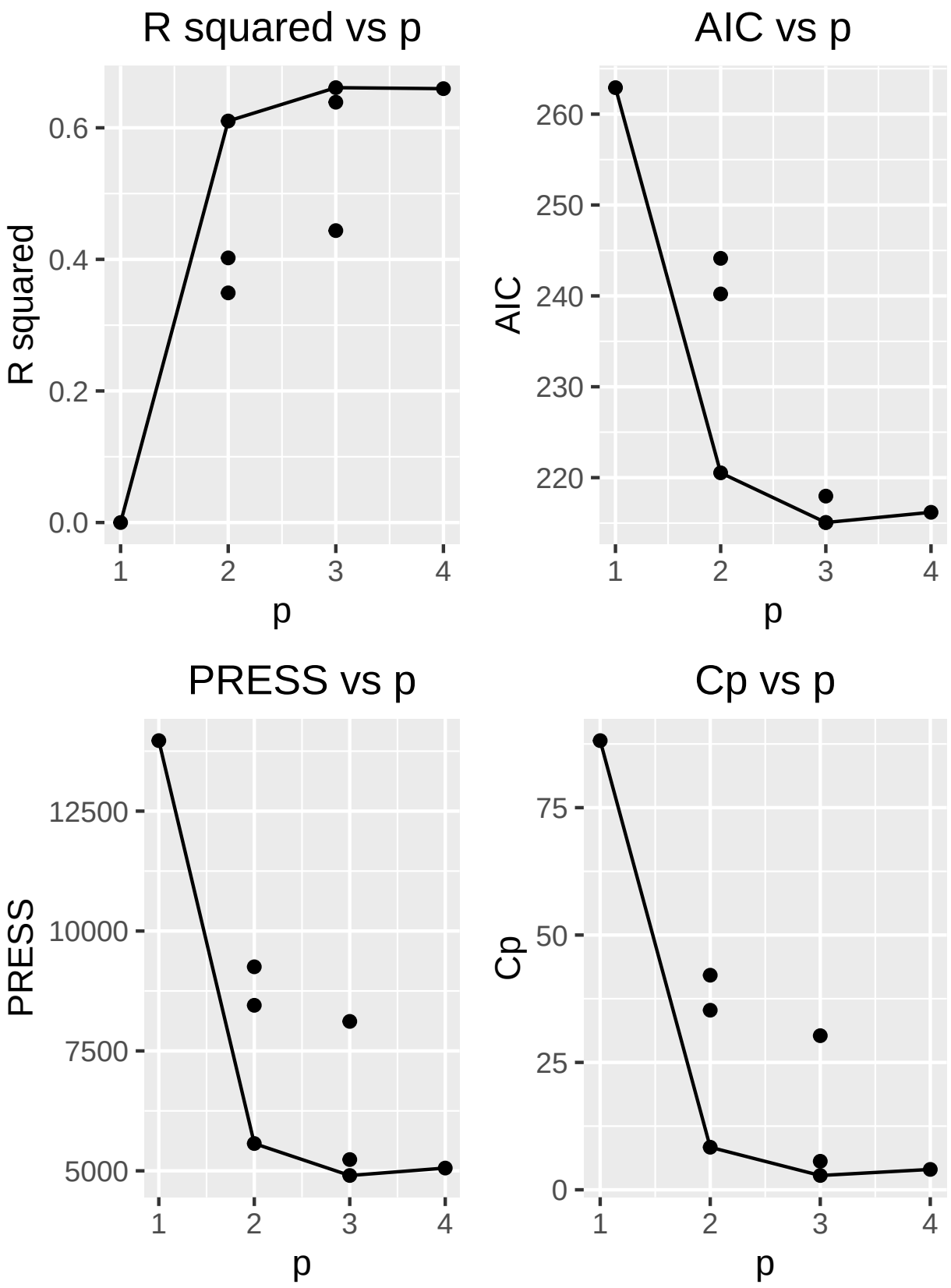

Figure 1: Plots for different criterion in model selection.

9.11 Refer to job proficiency problem 9.10

(a) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to $R^2_{a,p}$ criterion.

Answer: Illustrated in table (2). The four best subsets are, in order from best to worst,

i. X_1, X_3, X_4
ii. X_1, X_2, X_3, X_4
iii. X_1, X_3
iv. X_1, X_2, X_3

Table 2: The best 4 have checkmark.

Variables	p	SSE_p	$R^2_{a,p}$
None	1	9,054	0
X_1	2	6,658.145	0.248
X_2	2	6,817.529	0.230
X_3	2	1,768.023	0.800
X_4	2	2,210.689	0.750
X_1, X_2	3	4,851.180	0.439
✓X_1, X_3	3	606.657	0.930
X_1, X_4	3	1,672.585	0.807
X_2, X_3	3	1,755.813	0.797
X_2, X_4	3	1,962.072	0.773
X_3, X_4	4	1111.31	0.872
✓X_1, X_2, X_3	4	596.7207	0.929
X_1, X_2, X_4	4	1400.128	0.834
✓X_1, X_3, X_4	4	348.197	0.959
X_2, X_3, X_4	4	1095.808	0.870
✓X_1, X_2, X_3, X_4	5	335.98	0.959

(b) Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model?

Answer: We could try to do the Mallow's criterion as well. Additionally, we could also try AIC_p or SBC_p, the Schwarz Bayesian criterion, which penalizes heavier when $n \geq 8$ than the Akaike's information criterion, which is the case.

10.11 Refer to patient satisfaction problem 6.15.

(a) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.10$. State the decision rule and conclusion.

Answer: The null hypothesis is that there are no outliers, and the alternative is there are outliers in our set. To test this, we compare the studentized deleted residuals vs the Bonferroni critical value, noting that we include n tests for all deleted residuals, because if the regression model is appropriate, so that no case is outlying because of a change in the model, then each studentized deleted residual will follow the t distribution with $n-p-1$ degrees of freedom. The studentized deleted residuals are calculated from

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

The Bonferroni simultaneous test procedure with family significance level $\alpha = 0.10$ requires

$$t(1-\alpha/2n; n-p-1) = t(0.9989, 46-4-1) = 3.271$$

Therefore, we compare the studentized residuals vs 3.271, ie

$$|t_i| \leq 3.271$$

we conclude no outlier because none of the t_i are greater than 3.271. (The max is 1.835).

(b) Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

Answer: The diagonal elements are obtained from

$$h_{ii} = X_i^T (X^T X)^{-1} X_i$$

where

$$X_i = \begin{pmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,p-1} \end{pmatrix}$$

In this case, there are 46 diagonal elements, so we just list the first two and last two

$$0.078, 0.0671, \ldots, 0.073, 0.083$$

Leverage values larger than $2p/n$ are considered to indicate outlying cases with regard to their X values. In this case, $p = 4$, since we have 3 predictor variables, and $n = 46$, the number of observations. So our outlier indicator is 8/46=0.173. Index 9=0.184, index 28=0.186, and index 39=0.181 are thus outliers.

(c) Hospital management wishes to estimate mean patient satisfaction for patients who are $X_1 = 30$ years old, whose index of illness severity is $X_2 = 58$, and whose index of anxiety level is $X_3 = 2.0$. Use 10.29 to determine whether this estimate will involve a hidden extrapolation.

Answer: Equation 10.29 is

$$h_{\text{new,new}} = X_{\text{new}}^T (X^T X)^{-1} X_{\text{new}}$$

We note that X_{new} is of the form in 10.18(a).

$$X_{\text{new}} = \begin{pmatrix} 1 \\ 30 \\ 58 \\ 2 \end{pmatrix} \qquad X_{\text{new}}^T = \begin{pmatrix} 1 & 30 & 58 & 2 \end{pmatrix}$$

The computer tells us 10.29 yields

$$h_{\text{new,new}} = 0.3267$$

This seems like an extrapolation, as the old X_1, X_2, and X_3, all had leverages below 0.10, so this seems to be a pretty big jump.

(d) The largest absolute studentized deleted residuals are for cases 11,17, and 27. Obtain the DFFITS, DFBETAS, and Cook's distance values for this case to assess its influence. What do you conclude?

Answer: These have studentized residuals of 1.836, 1.807, and -1.97 respectively. DFFITS gives influence on single fitted value. The $(\text{DFFITS})_i$ value is given by

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_i h_{ii}}}$$

The values are

$$\text{Case } 11 \longrightarrow 0.569$$
$$\text{Case } 17 \longrightarrow 0.666$$
$$\text{Case } 27 \longrightarrow -0.609$$

Cook's distance is the influence on all fitted values, and is found from

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE}$$

The Cook's distances are respectively

$$\text{Case } 11 \longrightarrow 0.077$$
$$\text{Case } 17 \longrightarrow 0.105$$
$$\text{Case } 27 \longrightarrow 0.087$$

Finally, the influence on the regression coefficients, DFBETAS is

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}$$

This is seen in table (3) Using dfbetas function gives a scaled version in table (4)

Table 3: DFBETA

β_0	β_1	β_2	β_3
1.748	-0.0759	-0.091	2.694
-7.930	-0.0096	0.212	0.617
-0.302	0.087	-0.119	1.108

Table 4: DFBETAs

β_0	β_1	β_2	β_3
0.099	-0.363	-0.190	0.390
-0.449	-0.4711	0.443	0.089
-0.0172	0.417	-0.250	0.161

(e) Calculate the average absolute percent difference in the fitted values with and without each of these cases. What does this measure indicate about the influence of each of these cases?

Answer: The average of the absolute percent differences is (for case 11 as an example)

$$\frac{1}{n-1}\sum\left|\frac{\hat{Y}_{i,11} - \hat{Y}_i}{Y_i}\right| \cdot 100\%$$

where $\hat{Y}_{i,11}$ is the fitted values from running a regression without the 11th entry in the data, and $\hat{Y}_i$ is the fitted values from running linear regression with all values, where we ignore the 11th case when comparing the two regressions (as it does not exist in the omitted case). This leads to values of

$$\text{Case } 11 \longrightarrow 1.064\%$$
$$\text{Case } 17 \longrightarrow 1.274\%$$
$$\text{Case } 27 \longrightarrow 1.099\%$$

(f) Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

Answer: The Cook's distances are displayed in the plot in figure(2). There are no obvious outliers.

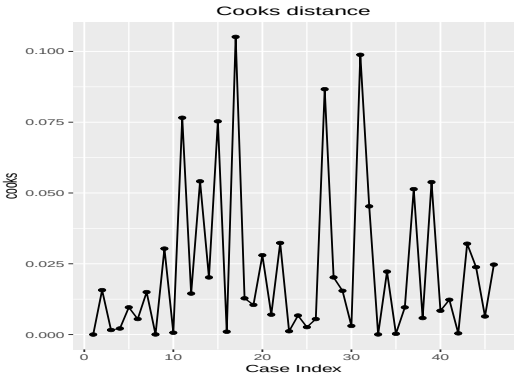

Figure 2: Cook's distances plotted.