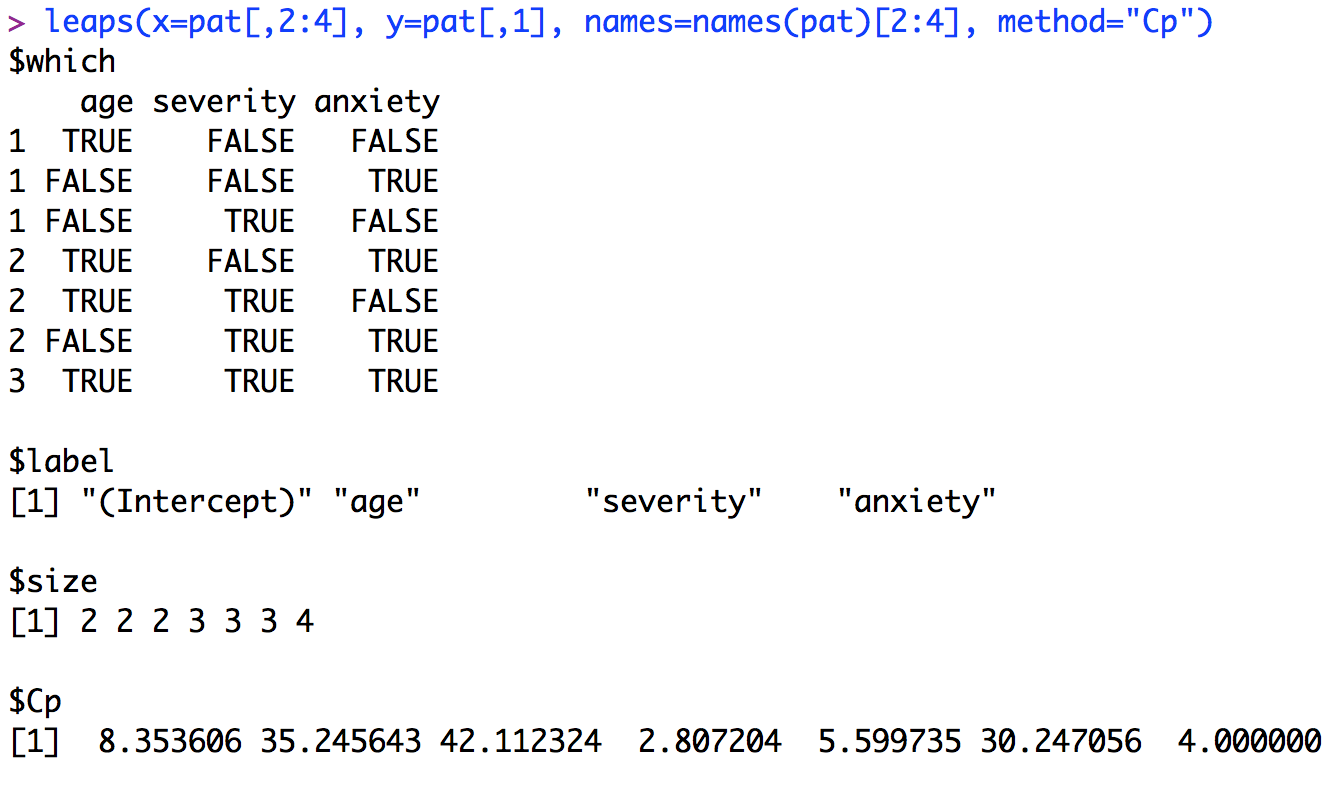
**Stats 201 HW8**

Answers for Nov. 18

1.

a) Best model: Age+Anxiety, adjusted R2 = 0.661; worst model: Severity, adjusted R2 = 0.3491

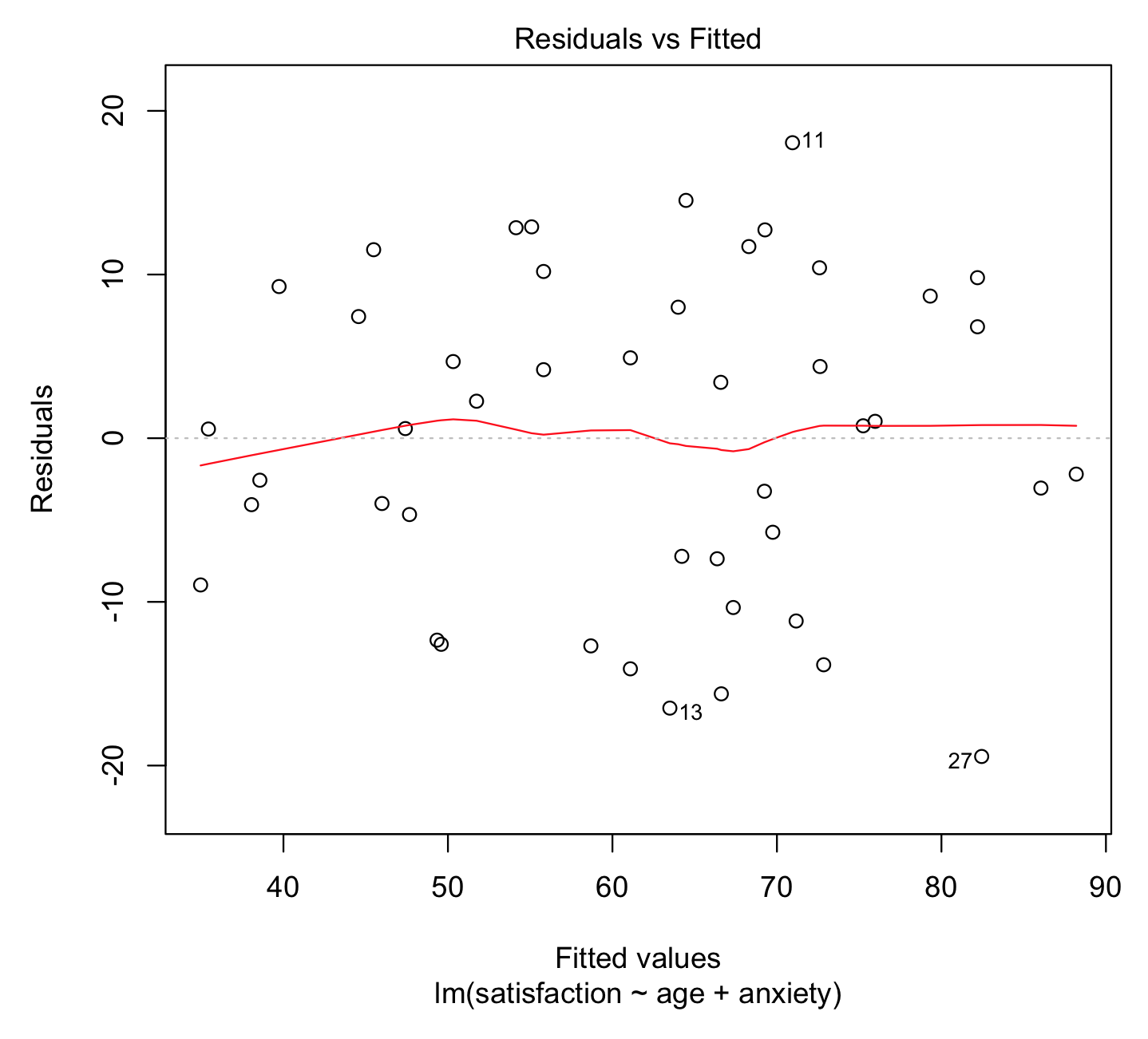
b) Best model: Age+Anxiety, Cp = 2.807; worst model: Severity, Cp = 42.112



It is reasonable to choose the model with predictors Age and Anxiety because the model has a Cp of 2.807, which is less than 3 (the number of predictors in the full model).

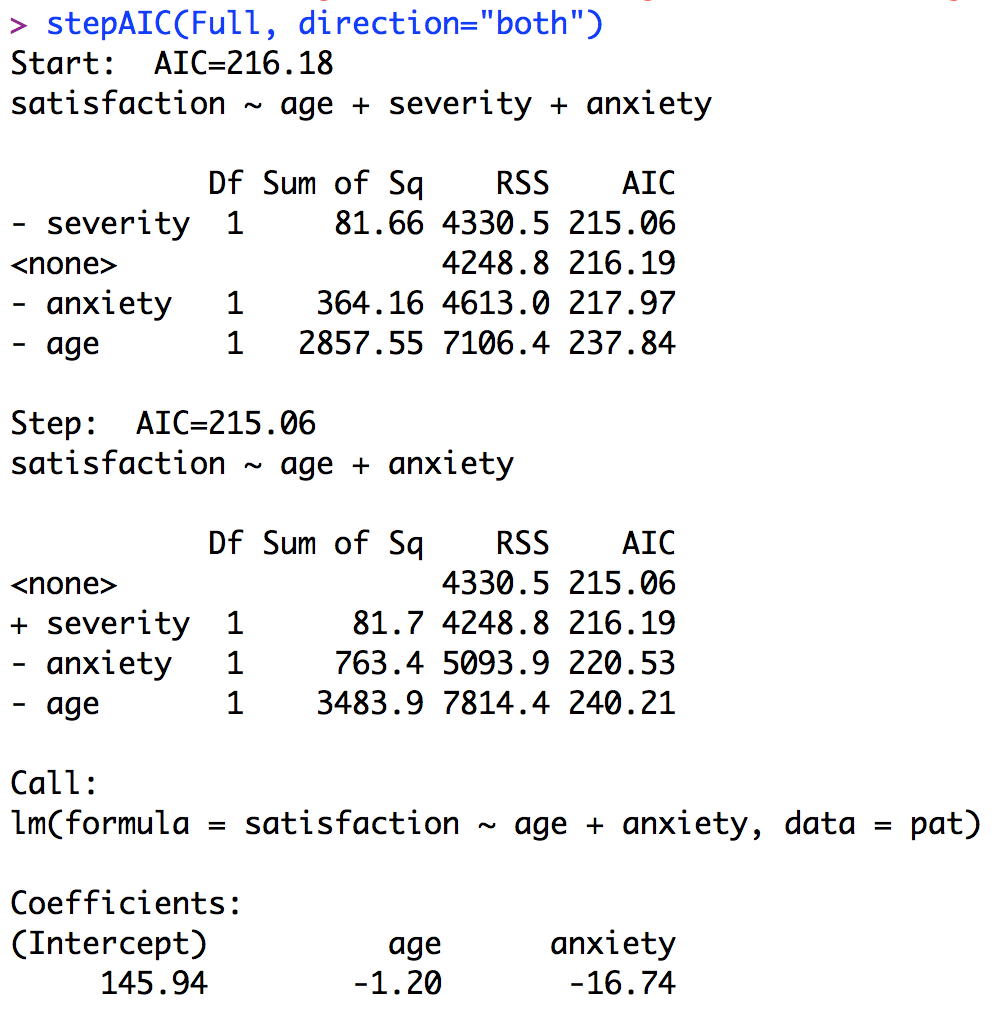
c) Yes, same model: Age + Anxiety.

d) Ŷ = 145.9412 – 1.2005\*X1 – 16.7421\*X2, where Ŷ is the predicted satisfaction, X1 is the age, and X2 is the anxiety. It appears to be a good fit because the line of the averages of the residuals is closed to a horizontal line passing through zero.



2. Best model: Age + Anxiety

a)



b)

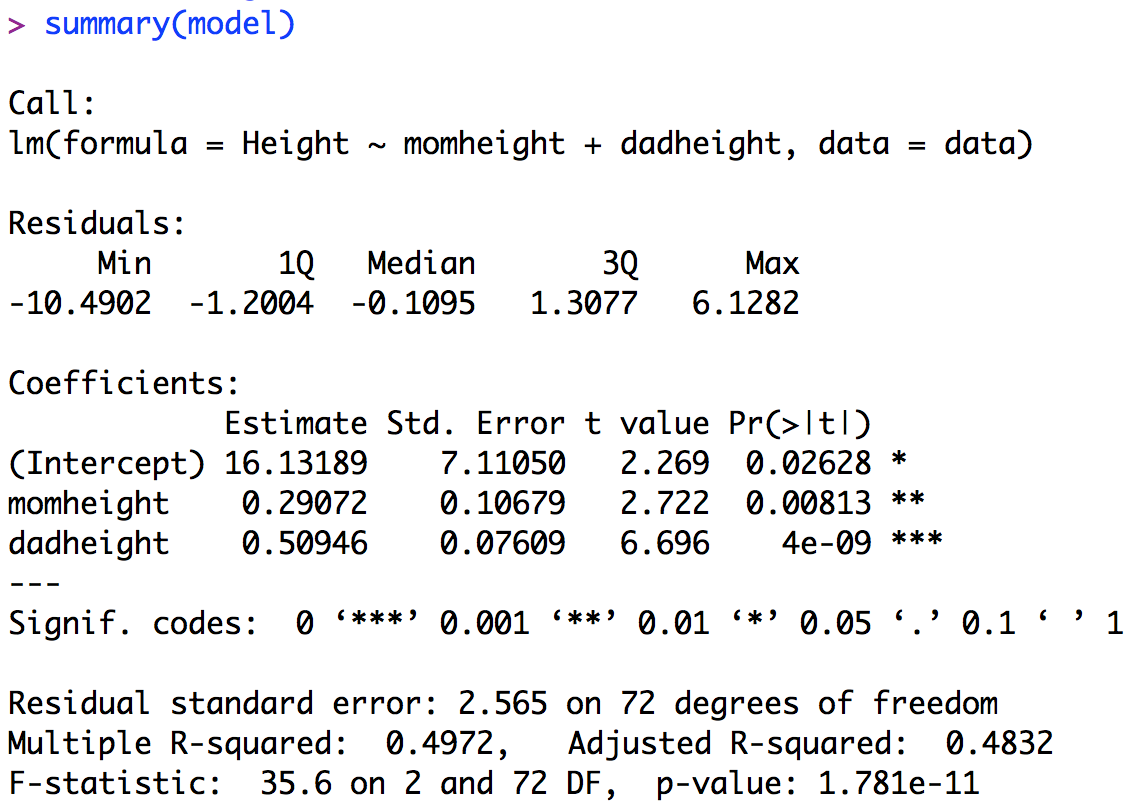
Step 1: Start from the Full model, compute the AIC of the models by removing each of the three predictors, comparing with the full model, and choose the model with the smallest AIC. If after removing we have smaller AIC, we then decide to remove the predictor and use the model with the smallest AIC

Step 2: Start from the model from step1, compute the AIC of the models by removing predictors or add a predictor back, choose the model with the smallest AIC, and decide whether to add or remove a predictor. Here none of the models built by adding or removing a predictor has smaller AIC than the starting model, so we do nothing and decide we have found the best model.

3. Yes, the two methods gave the same model. However, this would not always happen because the best subsets method can always find the best model, but stepwise regression only gets the local minimum.

Answers for Nov. 20

1.



2.

data$rstudent <- rstudent(model)

data$hii <- hatvalues(model)

data$dffitsi <- dffits(model)

data$cooks <- cooks.distance(model)

3. & 4.

a) Flag cases if |ti|> 3

> data[abs(data$rstudent)>3,]$ID

[1] 131

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | momheight | dadheight | Height | rstudent | hii | dffitsi | cooks |
| 131 | 61 | 66 | 57 | -4.7098 | 0.0243 | -0.7448 | 0.1428 |

b) Flag cases if hii > 2\*3/75 = 0.08

> data[data$hii>2\*3/75,]$ID

[1] 21 25 86 122 138

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | momheight | dadheight | Height | rstudent | hii | dffitsi | cooks |
| 21 | 54 | 68 | 68 | 0.6456 | 0.1581 | 0.2798 | 0.0263 |
| 25 | 59 | 60 | 64 | 0.0598 | 0.0838 | 0.0181 | 0.0001 |
| 86 | 66 | 55 | 65 | 0.7318 | 0.2229 | 0.3920 | 0.0515 |
| 122 | 60 | 78 | 70 | -1.3941 | 0.1306 | -0.5404 | 0.0961 |
| 138 | 71 | 76 | 77 | 0.6231 | 0.1176 | 0.2274 | 0.017 |

c) Flag cases if |DFFITS|i > 2\*sqrt(3/75) = 0.4

> data[abs(data$dffitsi)>2\*sqrt(3/75),]$ID

[1] 16 104 122 131

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID | momheight | dadheight | Height | rstudent | hii | dffitsi | cooks |
| 16 | 62 | 62 | 61 | -1.9327 | 0.05 | -0.4436 | 0.0632 |
| 104 | 65 | 64 | 63 | -1.8808 | 0.045 | -0.4085 | 0.0537 |
| 122 | 60 | 78 | 70 | -1.3941 | 0.1306 | -0.5404 | 0.0961 |
| 131 | 61 | 66 | 57 | -4.7098 | 0.0243 | -0.7448 | 0.1428 |

d) Flag cases if Cook’s distance > F(0.2; 2, 73) = 0.223827

> data[data$cooks>qf(0.2,2,73),]$ID

integer(0) (no such cases)

5.

a) ID 131: it is flagged because of low studentized residual and DFFITSi. The predict value of it is 67.49, but its actual height is 57.

b) ID 21: it is flagged because of high leverage (momheight = 54 is too low).

c) ID 25: it is flagged because of high leverage (momheight = 59 and dadheight = 60 are too low).

d) ID 122: it is flagged because of high leverage (momheight = 60 is too low).

6. It depends on what kind of data we conclude as errors / different population. For example, if we identify heights less than 60 inches as errors, we might remove cases with ID 21, 25, 86, and 131. However, we cannot be sure about whether they are errors or true values in the population, so we need a way to justify our assumptions of the errors to decide whether or not to remove the cases.