Assignment for Monday, November 18:

Use the **Patient Satisfaction** data from Problem 6.15 for this homework.

1. There are three predictor variables available for this data set: Age, Severity and Anxiety.
    a. Examine all subsets of predictor variables, and determine the best model and the worst model based on adjusted R^2. State what variables are in the best model and the worst model, and give the adjusted R^2 values for these two models.

Solution: The best model includes Age and Anxiety, and adjusted R^2 is 0.661. The worst model includes Severity only, and adjusted R^2 is 0.349.
The R command and output are as follows:

```
> leaps(x=Ch6Pr15[,2:4], y=Ch6Pr15[,1], names=names(Ch6Pr15)[2:4],
method="adjr2")
$which
    Age Severity Anxiety
1  TRUE    FALSE   FALSE
1 FALSE    FALSE    TRUE
1 FALSE     TRUE   FALSE
2  TRUE    FALSE    TRUE
2  TRUE     TRUE   FALSE
2 FALSE     TRUE    TRUE
3  TRUE     TRUE    TRUE

$label
[1] "(Intercept)" "Age"         "Severity"    "Anxiety"

$size
[1] 2 2 2 3 3 3 4

$adjr2
[1] 0.6103248 0.4022134 0.3490737 0.6610206 0.6389073 0.4437314 0.6594939
```

    b. Repeat Part (a) but choose the best and worst model based on Cp, and show the Cp values for the best and worst models. Is the value of Cp reasonable for the final model you choose? Explain.

Solution: The R command is the same as above except replace "adjr2" with "Cp". The best model includes Age and Anxiety, and Cp = 2.807. The worst model contains Severity only, and Cp = 42.112. The value of Cp for the best model is reasonable because for that model p = 3 and Cp = 2.807, which is close to 3.

    c. Did you choose the same final model in Parts (a) and (b)? If so, use that final model for Part (d). If not, choose one of them to use in Part (d).

Solution: Yes, the same model was chosen – it includes Age and Anxiety, but not Severity.

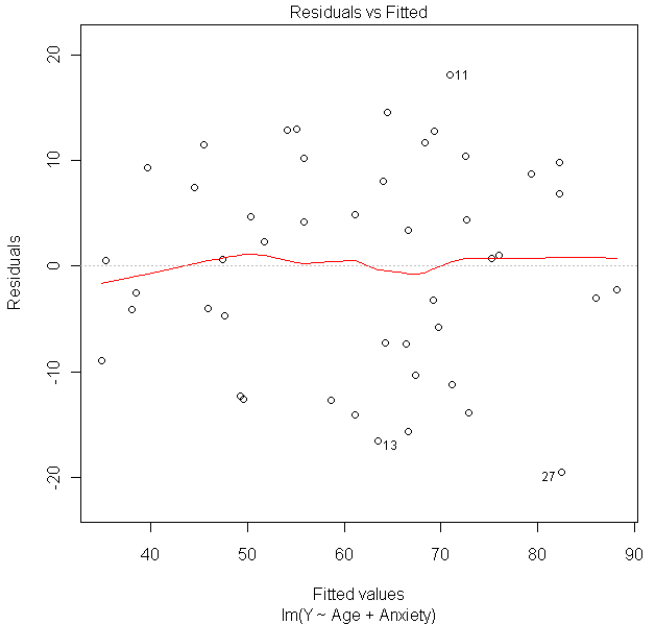
    d. For the final model you choose, give the regression equation, and an appropriate residual plot. Comment on whether you think the model is a good fit.

Solution: The model includes Age and Anxiety, but not Severity. R output is as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.9412    11.5251  12.663 4.21e-16 ***
Age          -1.2005     0.2041  -5.882 5.43e-07 ***
Anxiety     -16.7421     6.0808  -2.753  0.00861 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.04 on 43 degrees of freedom
Multiple R-squared: 0.6761,   Adjusted R-squared: 0.661
F-statistic: 44.88 on 2 and 43 DF,  p-value: 2.98e-11
```

The regression equation is $\hat{Y} = 145.94 - 1.2005(Age) - 16.7421(Anxiety)$. Here is a residual plot, from the command `plot (Model, 1);` the model looks like a good fit.



2. Using the same data, do a stepwise procedures using AIC as the criterion for the best model. Start with the full model and work backwards, but remember that "stepwise" allows variables to re-enter once they have been removed. You can read the instruction for how to do this in R in the R Word document labeled "Model Selection in R," or for specific directions, see below. If you are using R Commander, you can do it this way:
   - Fit the full model with all 3 variables; call the model Full
   - Go to the menu Models – Stepwise model selection. Click on AIC, then choose your method.

   To use stepwise in R, once you fit the full model (call it Full), use the command:

   ```
   > stepwise(Full, direction='backward/forward', criterion='AIC')
   ```

   (If you wanted to use only forward selection or backward elimination, you would change the "direction" to include only one of them. You can also use `'forward/backward'` to start with no variables and work forward, instead of working backwards first).
   a. Determine the best model using stepwise regression, starting with all variables in the model. Show your steps and report what model was ultimately selected.
   b. Explain what has been done by R after each of the first two steps.

Solution in bold, with the explanations for Part b interspersed with the output:

```
> stepwise(Full, direction='backward/forward', criterion='AIC')
Direction:  backward/forward
Criterion:  AIC

Start:  AIC=216.18
Y ~ Age + Anxiety + Severity

           Df Sum of Sq    RSS    AIC
- Severity  1     81.66 4330.5 215.06
<none>                  4248.8 216.19
- Anxiety   1    364.16 4613.0 217.97
- Age       1   2857.55 7106.4 237.84
```

(For part b) The above notation means that if you remove Severity, AIC becomes 215.06, if you remove "none" AIC is 216.19, if you additionally remove Anxiety AIC becomes 217.97, etc. Note that removing Severity (indicated by "– Severity") reduces AIC to 215.06 and nothing else gives a better AIC.

```
Step:  AIC=215.06
Y ~ Age + Anxiety

           Df Sum of Sq    RSS    AIC
<none>                  4330.5 215.06
+ Severity  1      81.7 4248.8 216.19
- Anxiety   1     763.4 5093.9 220.53
- Age       1    3483.9 7814.4 240.21
```

Now notice that adding Severity back in would increase AIC to 216.19, and removing either of the two remaining variables would cause AIC to go up. Therefore, R has quit and shows the final model:

```
Call:
lm(formula = Y ~ Age + Anxiety, data = Ch6Pr15)

Coefficients:
(Intercept)          Age      Anxiety
    145.941       -1.200      -16.742
```

3. Did you choose the same final model in Problem 1 and Problem 2? If not, explain which method you think gave you the best model. If you did choose the same model, explain whether that would always happen, and if not, which method would give you the best model in general.

Solution: For this example, both methods resulted in the same model, which includes Age and Anxiety. In general that will not be the case. The best method is "best subsets" because it considers all models considered by the stepwise methods, plus additional models as well. Stepwise may never find the best subset of variables, because of the complexity of multicollinearity.

Assigned Wed, November 20:

The file Nov21Hmwk.txt contains a subset of the data for the student and parents height data set that we have used in numerous examples. The subset consists of the male students only, with the case removed that had a clearly erroneous mother's height of 80 inches. There are 75 cases in the data set. The variables in the data file are:

ID = the original ID from the full dataset, which you should use to identify cases
Sex = Male for everyone in this data set (and thus you won't need to use it)
momheight, dadheight and Height = heights in inches for mother, father, and student

The variable names are listed at the top of the data file, so you should use the R option that specifies that they are included.

1. Use the variables momheight and dadheight to predict Height.
   Solution:
   The R commands and some output are shown below. (*You don't need to show the full output, just the model is enough.*)

```
> Model <- lm(Height~momheight+dadheight, data=HW201)
> summary(Model)
Call:
lm(formula = Height ~ momheight + dadheight, data = HW201)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.13189    7.11050   2.269  0.02628 *
momheight    0.29072    0.10679   2.722  0.00813 **
dadheight    0.50946    0.07609   6.696    4e-09 ***
Residual standard error: 2.565 on 72 degrees of freedom
Multiple R-squared: 0.4972,      Adjusted R-squared: 0.4832
F-statistic:  35.6 on 2 and 72 DF,  p-value: 1.781e-11
```

2. Find case diagnostic values for the four diagnostic measures discussed in class. (These include t_i, h_{ii}, $(DFFIT)_i$, and Cook's distance.)
   Solution: *You just need to show something verifying that you did this, such as the R commands or the start of the output.*
   The R commands are as follows; output is not shown because it would be too long!
   The following are used to compute t_i, h_{ii}, $(DFFIT)_i$, and Cook's distance, respectively:

```
HW201$rstudent <- rstudent(Model)
HW201$Hii <- hatvalues(Model)
HW201$dffits <- dffits(Model)
HW201$cooks <- cooks.distance(Model)
```

3. For each of the diagnostic measures, identify cases that need to be investigated (if any). Use the variable "ID" to identify them so we know which cases you have identified.
   Solution:
   Here are the values used to identify ("flag") problem cases:
   Studentized deleted residuals t_i: Flag if the absolute value is greater than **3** (or moderate if > 2).
   Leverage h_{ii}: Flag if it's greater than 2p/n, which in this case is 2(3)/75 = **0.08,** extreme is greater than 3p/n = 0.12.

(DFFITS)i: Flag if its absolute value is greater than 1 for small n, otherwise $2\sqrt{\frac{p}{n}}$ = **0.4**

Cook's distance: Flag if it's greater than F(.2, p, n – p), which is F(.2, 3, 72) = **.3358.** You could also use the R defaults of 0.5 (moderate) or 1.0 (extreme).

Here are the cases flagged with each measure, showing values for all of the diagnostic measures:

Studentized deleted residuals with absolute value > 2, extreme > 3 in bold:

ID	momht	dadheight	Height	**rstudent**	Hii	dffits	cooks
36	64	67	75	2.496775	0.018027	0.3382929	0.035562
38	60	67	73	2.146535	0.030468	0.3805236	0.045963
131	61	66	57	**-4.70985**	0.024398	-0.74482	0.142883

Leverage greater than 0.08, extreme greater than 0.12 in bold:

ID	momheight	dadheight	Height	rstudent	**Hii**	dffits	cooks
21	54	68	68	0.645669	**0.158157**	0.279859	0.02632
25	59	60	64	0.059863	0.083805	0.018105	0.000111
86	66	55	65	0.73186	**0.222999**	0.392074	0.051573
122	60	78	70	-1.39415	**0.13065**	-0.54046	0.096108
138	71	76	77	0.623132	0.117608	0.227492	0.017399

DFFITS with absolute value > 0.4 (If you used the greater than 1 rule, nothing is flagged):

ID	momheight	dadheight	Height	rstudent	Hii	**dffits**	cooks
16	62	62	61	-1.9327	0.050057	-0.44366	0.063209
104	65	64	63	-1.88088	0.045058	-0.40856	0.053747
122	60	78	70	-1.39415	0.13065	-0.54046	0.096108
131	61	66	57	-4.70985	0.024398	-0.74482	0.142883

No cases were flagged using Cook's distance. The largest one was .142, which is less than .3358.

4. For each case identified in #3, provide the data values and the diagnostic measure(s) that caused the case to be flagged.
Solution:
Done in #3 above.

5. Choose 4 of the cases flagged and provide an explanation for why that case was flagged as unusual.

Solution: *You can choose any 4 of the flagged cases for your answer.* Below are explanations for why various cases are flagged.

Note: It is useful to plot the two X variables in a scatter plot, to see why certain points have high leverage. There is a plot of momheight vs dadheight on the last page with some of the ID numbers filled in. To create this plot I used:

```
plot(HW201$momheight,HW201$dadheight)
identify(HW201$momheight,HW201$dadheight,labels=HW201$ID)
```

After the plot is created (with no ID labels showing) you can click on points that look like they have high leverage, and R adds the ID number. For the plot shown below, I added some of the ID numbers.

Explanations of flagged cases:

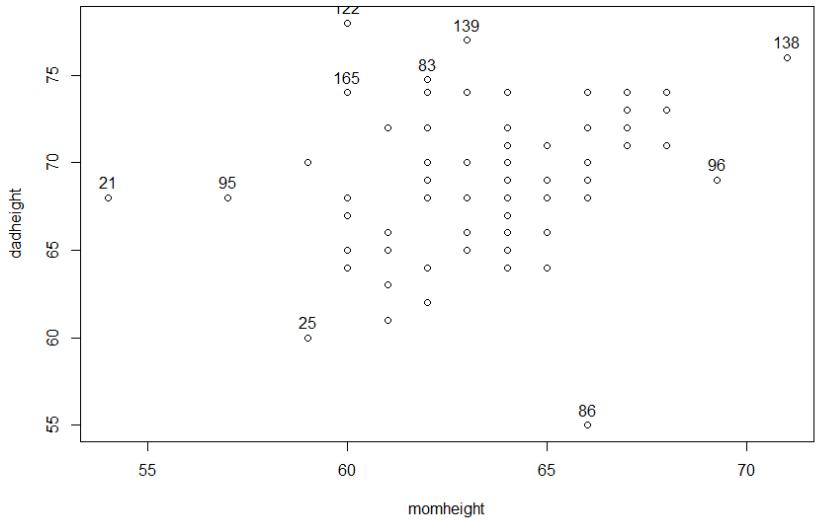
FLAGGED BY t_i:
Cases 36 and 38: Both of these people are much taller than predicted given their parents' heights. Case 36 is predicted to be 68.87 inches tall but is 75 inches, and Case 38 is predicted to be 67.71 inches, but is 73 inches. There is no obvious explanation, other than natural variability.
Case 131: This case was flagged because of the studentized deleted residual, and DFFITS. DFFITS is a combination of residual and leverage, and in this case, it's the residual that caused it to be large. The case had an unusually large, negative residual, indicating that the predicted value was much larger than the actual Y value. You can see why - the actual height for this student is only 57 inches, which is less than 5 feet tall. The parents are short, but not that short. The predicted value for this person is 67.49 inches. Therefore, it is possible that the actual Y value was recorded incorrectly, and should have been 67 inches.

FLAGGED BY LEVERAGE h_{ii}:
Cases 21, 25, 86, 122 and 138 were flagged because of the leverage, h_{ii}, indicating that the X values are an unusual combination. (Case 122 was flagged by DFFITS as well.) Examine the momheight and dadheight for each case to see why. These are easy to find on the plot below.

Case 21: The mom's height is unusually low, at 54 inches.
Case 25: Both parents are short at 59 and 60 inches. The dad's height is particularly low for a male.
Case 86: The dad's height is only 55 inches, which is under 5 feet tall. This is quite likely a mistake, but without being able to check the original data, it should not be removed. This is the most influential case, as indicated by the highest value of hii.
Case 122: The combination of the heights is unusual. The mom's height is only 5 feet, while the dad's height is 6 feet 6 inches. This case was flagged by DFFITS as well, because it has both a large leverage and moderately large residual. The predicted Y is 73.13, but the actual height is only 70.
Case 138: Everyone is tall in this family! The case is flagged because the parents are both so tall.

FLAGGED BY DFFITS (and not already discussed): Cases are flagged by this measure because of a combination of large studentized deleted residual and leverage, or by one of them being unusually large.
Case 16: The leverage on this case is not unusual, so it's the studentized deleted residual that causes it to have large DFFITS. The parents are both 62 inches, but the student is predicted to be 65.7 inches tall and is only 61 inches. In general male students are predicted to be taller than female students, so a male with particularly short parents, like this one, will have a predicted height larger than the parents. In this case, that prediction was off.
Case 104: This case is similar to Case 16. The student is predicted to be 67.6 inches tall, but is only 63 inches.

Plot of momheight vs dadheight. (*Not* required as part of your solution, but useful for identifying points.)



6. Discuss whether any of the identified cases should be removed from the analysis.

Solution:
There is no right or wrong answer here, it's your reasoning that counts. The only reasons to remove cases are if they are clearly mistakes, or if the X variable combination is unusual and the predictions are bad, so that it appears the cases belong to a different population. If cases are removed for that reason, then the regression equation should not be used on similar X values in the future.

Cases with possible errors are Cases 21 (mom is only 54 inches), Case 86 (dad is only 55 inches) and Case 131 (student is only 57 inches). Ideally we would have access to the original records to see if they are likely to be mistakes. You should never remove cases just because they are unusual. If you do, you may be removing natural variability, and your inference results will not be correct.

None of the cases can justifiably be removed as belonging to a different population, unless you think that the very short parents are real values and that equation shouldn't be used in the future for short parents. But that isn't reasonable here, because they don't show particularly bad fits, so they don't seem to require a different model.