Statistics 201 Homework
Assigned Mon, Oct 28 and Due Wed, Nov 6

For this assignment use the data set UCD.txt (linked to the website under data sets and in the list of assignments). The file ***does*** include a row with variable names, and the columns are separated with tabs. The data set, collected on 173 UC Davis students (self-reported), includes the following variables:

ID = numbers from 1 to 173, the ID for that student and the row number with the data
alcohol = average number of alcoholic drinks consumed per week
exercise = average hours per week the student exercises
height = the student's height (in inches)
male = indicator variable, 1 if male and 0 if female
dadht = the student's father's height
momht = the student's mother's height

For this assignment, we will use only the last 4 variables, but retain the file for later use with other variables. Note that there are some missing values, designated NA, but cases with NA for the variables used in this assignment have been removed, so some ID numbers are missing.

1. Fit the model predicting height using male, dadht and momht. (This is the model we went over in discussion.) *Call this model Full.* Show the summary. Write a sentence interpreting the coefficient for "dadht" for this model.

```
> Full <- lm(height~male+dadht+momht, data=UCD)
> summary(Full)

Call:
lm(formula = height ~ male + dadht + momht, data = UCD)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7431  -1.4537   0.0191   1.5299   5.9459

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.96746    4.65831   3.642 0.000364 ***
male         5.29822    0.36377  14.565  < 2e-16 ***
dadht        0.41213    0.05107   8.069 1.54e-13 ***
momht        0.29962    0.06876   4.357 2.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 161 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.6604,     Adjusted R-squared: 0.6541
F-statistic: 104.4 on 3 and 161 DF,  p-value: < 2.2e-16
```

The coefficient for dadht is 0.41 (rounded off). Interpretation: This is an estimate of the average difference in height for two students of the same sex whose mothers are the same height, and whose fathers differ in height by one inch (with the taller father predicted to have the taller child).

2. Fit the model predicting height using male and momht only. *Call this model MomOnly.* Show the summary. Test the null hypothesis that "momht" is useful in this model. Show the hypotheses, the *p*-value, and a conclusion in the context of the situation.

```
> MomOnly <- lm(height~male+momht, data=UCD)
> summary(MomOnly)

Call:
lm(formula = height ~ male + momht, data = UCD)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4918  -1.4455   0.0776   1.6008   7.0776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.38884    4.87674   7.052 4.85e-11 ***
male         5.01498    0.42777  11.724  < 2e-16 ***
momht        0.47685    0.07698   6.194 4.65e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.736 on 162 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.5231,     Adjusted R-squared: 0.5172
F-statistic: 88.85 on 2 and 162 DF,  p-value: < 2.2e-16
```

The requested test: Hypotheses are H_0: $\beta_2 = 0$ vs H_a: $\beta_2 \neq 0$. (Or β_1 if you put momht in first, before male). The p-value (given by R) is 4.65×10^{-9}, so we can clearly reject the null hypothesis and conclude that knowing the mother's height does help predict the student's height.

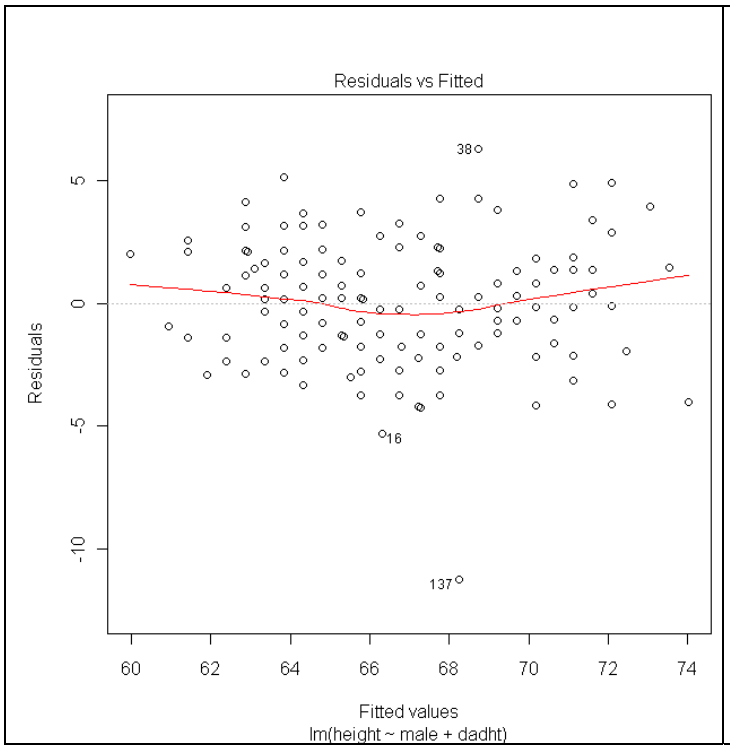
3. Fit the model predicting height using male and dadht only. *Call this model DadOnly.* Show the summary. Write a sentence interpreting the coefficient for "male" for this model.

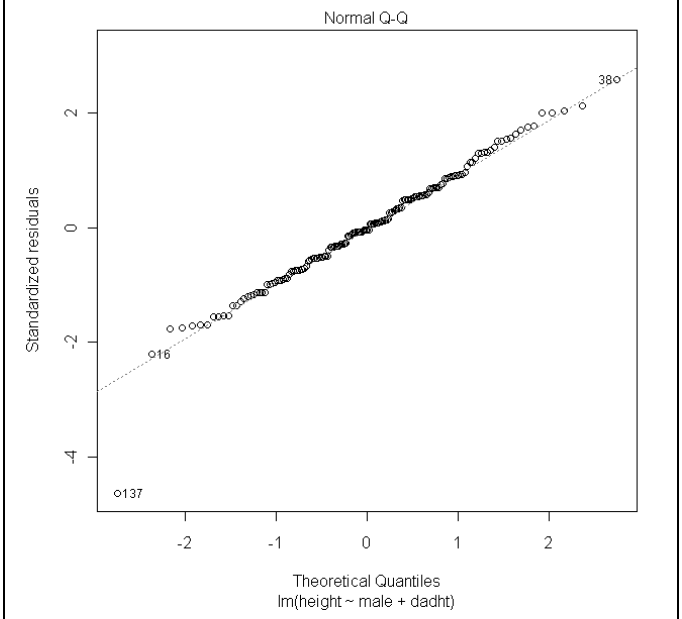
```
> DadOnly <- lm(height~male+dadht, data=UCD)
> summary(DadOnly)

Call:
lm(formula = height ~ male + dadht, data = UCD)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2355  -1.6516  -0.1012   1.4491   6.2813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.97678    3.55308   8.718 3.20e-15 ***
male         5.36620    0.38308  14.008  < 2e-16 ***
dadht        0.48322    0.05101   9.472  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.441 on 162 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.6204,     Adjusted R-squared: 0.6157
F-statistic: 132.4 on 2 and 162 DF,  p-value: < 2.2e-16
```

The coefficient of 5.366 for "male" is an estimate of the difference between the average height for males and the average height for females who have the same Dad's height.

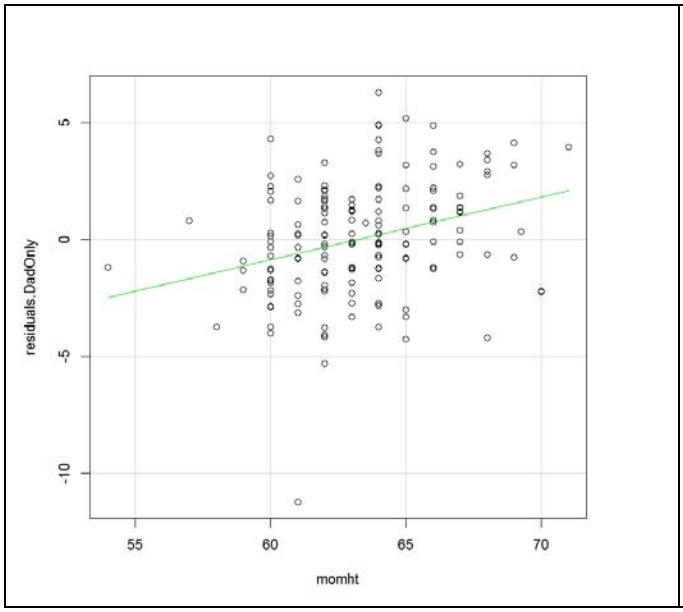
4. Create two diagnostic plots for the *DadOnly* model and comment on them.

Use the commands plot(DadOnly, 1) and plot (DadOnly, 2). Results are shown below. In both plots the residuals appear to fit the assumptions except for the outlier identified as Case 137. That case probably needs further investigation to see if it should be discarded.





5. Plot the residuals from the *DadOnly* model versus "momht." Include the least squares line. What do you learn from this graph?



There is clearly a linear relationship between the residuals from the DadOnly model, and mom's height. In other words, the residuals are linearly related to mom's height, so it makes sense to add mom's height to the model.

6. Compare the three models using two summary measures discussed in class. Which model is best? Which model is second best? Explain.

   The two measures to compare are adjusted R^2 and MSE (or its square root, which is called Residual standard error in R). Here are the values:

Model	*Adjusted R^2*	*Square root of MSE*
Full	*.6541*	*2.316*
MomOnly	*.5172*	*2.736*
DadOnly	*.6157*	*2.441*

   From these values, we can see that the Full model is best, followed by DadOnly, then MomOnly. We want high adjusted R^2 and low MSE.

7. Using the *Full* model, find a 95% confidence interval for the coefficient for "male" and write a sentence interpreting the interval in the context of this situation.

   The confidence interval (from R) is 4.5798389 to 6.016597, or round off, 4.58 to 6.02. Interpretation: We are 95% confident that the difference in average heights for males and females whose parents are the same height is somewhere between 4.58 and 6.02 inches (for the population represented by this sample).