MULTIPLE REGRESSION EXAMPLE

For a sample of n = 166 UC Davis students, the following variables were measured:

Y = height (inches) (“Height”) X_1 = mother’s height (“momheight”) X_2 = father’s height (“dadheight”) X_3 = 1 if male, 0 if female (“male”)	$\underline{Y} = \begin{array}{c} \text{Height} \\ \begin{bmatrix} 66 \\ \vdots \\ 64 \end{bmatrix} \end{array}, \quad X = \begin{array}{cccc} & \text{Mom} & \text{Dad} & \text{Male} \\ \multicolumn{4}{c}{\begin{bmatrix} 1 & 66 & 71 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 62 & 66 & 1 \end{bmatrix}} \end{array}$ $\underline{Y}$ and ***X*** each have 166 rows

Our goal is to predict student’s height using the student’s mother’s and father’s heights, and sex, where sex is categorized using the variable “male” = 1 if male, 0 if female. The population model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \text{ where } \varepsilon_i \text{ are independent, } N(0, \sigma^2).$$

First let’s look at some plots of the original data to see if there are outliers, and if the patterns look linear. See plots in extended handout on website for the plots and the R commands. The plots are:
Graph 1. Side by side boxplots of female and male students’ heights.
Graph 2. Stem and leaf plot of mothers’ heights.
Graph 3. Stem and leaf plot of fathers’ heights.

Then, after removing one outlier (see discussion of outliers below):
Graph 4. A “scatter plot matrix” (see p. 232 of text), with separate symbols and regression lines shown for males and females. The picture also includes a smoothed histogram (density plot) for each variable, although the one for “Height” doesn’t make much sense as it combines males and females.

Outliers: Examining the plots, a few possible outliers are evident:

- A case with *momheight* = 80 inches. This is almost surely a mistake – it’s a female height of 6 ft, 8 inches. So it is legitimate to remove it, since we can’t recover what it should be. The case is in row 129. We need to change the 80 to the missing value code, which is NA in R. To do this, the R command `>fix(UCDavis1)` opens the data editor. Find the 80 (in row 129) and change it to NA.
- A case with *height* = 57 inches for a male. While unusual (4 ft, 9 inches) it is possible. Do not remove.
- A case with *dadheight* = 55 inches. Again this is very unusual (4 ft, 7 inches) but is possible. Do not remove. One option is to run the analysis with and without it, and see what difference it makes. (Of course you would always report that you had done that, don’t just chose which one you like best!)

From the scatter plot matrix, we see that the relationships between the response variable height and the explanatory variables momheight and dadheight look linear, at least from what we can tell from such tiny pictures.

Now let’s run the linear model. The original data has a text (categorical) variable called “Sex” with two values ‘Male’ and ‘Female’. When you put “Sex” in the model, R knows to assign 0 and 1 to the two categories. (If using R Commander, use *Fit models -> Linear model* instead of *Linear regression*)

```
ParentHt <- lm(Height ~ momheight + dadheight + Sex, data=UCDavis1)
summary(ParentHt)
anova(ParentHt)
```

See the next page for the results.

Follow that up with two plots. R has a simple way of generating these two plots after you have fit a linear model.. `>plot(ParentHt, 1)` plots residuals vs fitted values, and `plot(ParentHt, 2)` gives the normal probability plot. These also show the case numbers for unusual points. So, we have:

Graph 5: Plot of the residuals versus fitted values; **Graph 6:** Normal probability plot of the residuals.

(Partial) regression results (remember that we now have n = 165 cases; we removed one outlier):

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.96746    4.65831   3.642 0.000364 ***
momheight    0.29962    0.06876   4.357 2.34e-05 ***
dadheight    0.41213    0.05107   8.069 1.54e-13 ***
Sex[T.Male]  5.29822    0.36377  14.565  < 2e-16 ***
---

Residual standard error: 2.316 on 161 degrees of freedom
  (8 observations deleted due to missingness) [There already were 7 cases with missing information]
Multiple R-squared: 0.6604,   Adjusted R-squared: 0.6541
F-statistic: 104.4 on 3 and 161 DF,  p-value: < 2.2e-16

Analysis of Variance Table [Note this does not show MSR; we will learn more about this later.]
Response: Height
           Df  Sum Sq Mean Sq F value    Pr(>F)
momheight   1  301.31  301.31  56.192 4.146e-12 ***
dadheight   1  240.38  240.38  44.828 3.405e-10 ***
Sex         1 1137.49 1137.49 212.130 < 2.2e-16 ***
Residuals 161  863.32    5.36
```

- The regression equation (rounding coefficients to 2 decimal places) is:
  Predicted height = 16.97 + 0.30 (momheight) + 0.41 (dadheight) + 5.30 (male)

- The coefficient for the variable “male” tells us that for a *fixed combination* of momheight and dadheight, on average males will be about 5.30 inches taller than females with that same combination
.

- The coefficient of about 0.30 for momheight tells us that for a *given* dadheight and sex, the predicted student’s height increases by about 0.30 inches for every 1.0 inch increase in momheight. For example, for male students whose dads are 70 inches tall, those whose moms are 65 inches tall are on average predicted to be about 0.30 inches taller than those whose moms are 64 inches tall.

- For *each* of the coefficients, a test for H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ has p-value of 0 to several decimal places (column headed `P>|t|`.) These are *conditional* hypotheses. They test whether or not each explanatory variable needs to be in the model, *given* that the others are already there. Therefore, in this example, the tests tell us that all 3 of the explanatory variables are useful in the model, even after the others are already in the model. In other words, even with (for example) mom’s height and student’s sex in the model, dad’s height still adds a substantial contribution to explaining student’s height.

- R^2 = .6604 or 66.04%, which is good. Later we will learn about “Adjusted R^2” which is more useful in multiple regression, especially when comparing models with different numbers of X variables.

- Residual standard error = $\sqrt{MSE} = \sqrt{5.36} = s$ = estimate of σ = 2.316 inches, indicating that *within* every *combination* of momheight, dadheight and sex, the standard deviation of heights is about 2.32 inches.

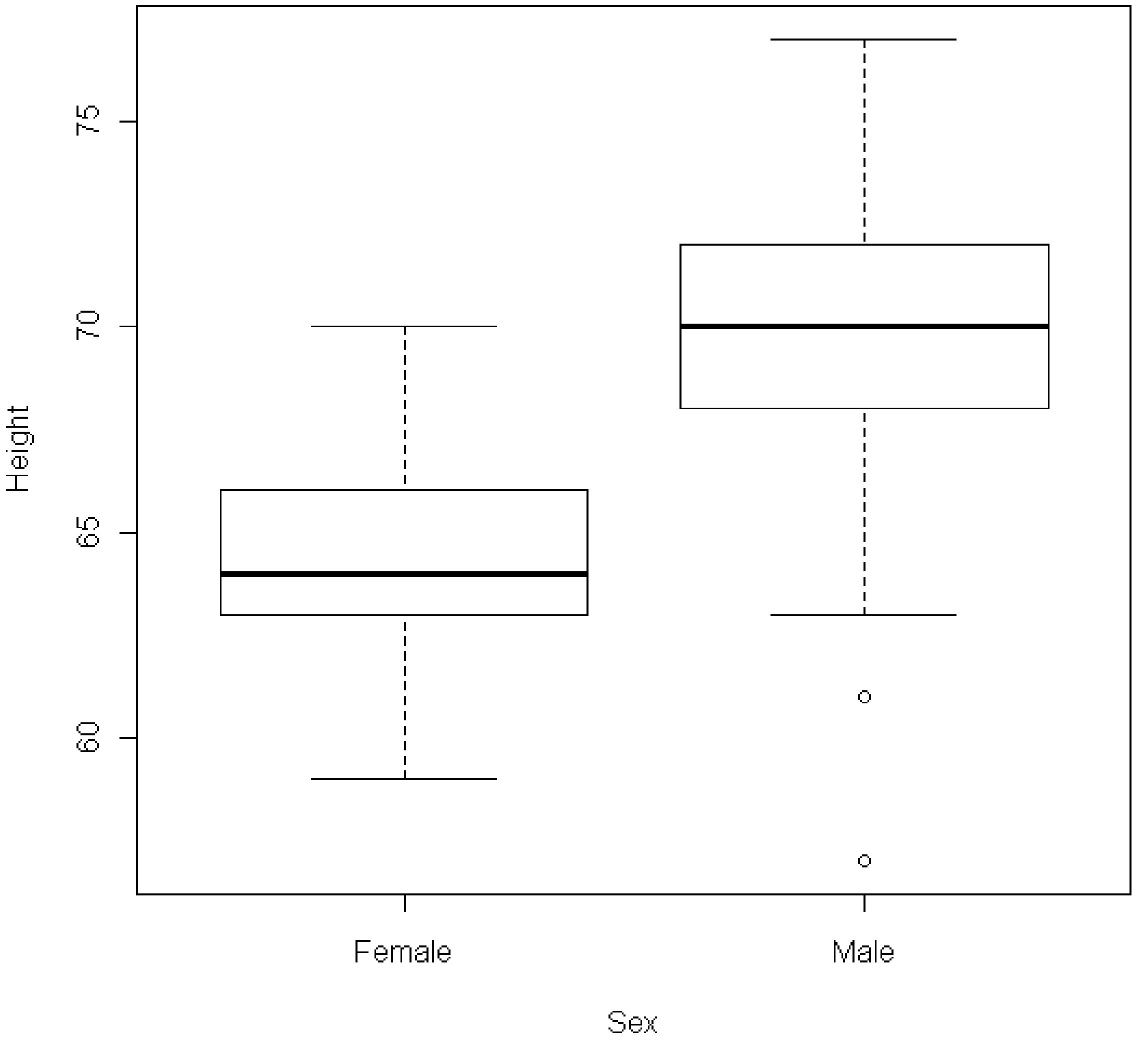
- The F(3, 161) = 104.4 is F* for testing the full model versus the reduced model $E\{Y_i\} = \beta_0$. In other words, it is *simultaneously* testing H_0: β_1, β_2, β_3 all = 0 versus H_a: At least one is not 0. The p-value is given as 2.2 x 10^{-16}, so clearly we can reject the null hypothesis and we can conclude that at least one of the explanatory variables is useful.

- The residual versus fitted values and normal probability plot both show the outlier (height = 57”) but otherwise they look like we hope they would.

Graph 1: Side by side boxplots of female and male students' heights

R Command: `boxplot(Height~Sex, ylab="Height", xlab="Sex", data=UCDavis1)`

Note that there are a few short males, but not short enough to rule out that they are real values.



Graph 2: Stem and leaf plot of mothers' heights. Notice the outlier at 80 inches, an obvious mistake. The shape for the X variables does not matter, so although this looks bell-shaped, it's not really relevant. One more note: There are 3 students with missing values for momheight as indicated by NA's: 3.

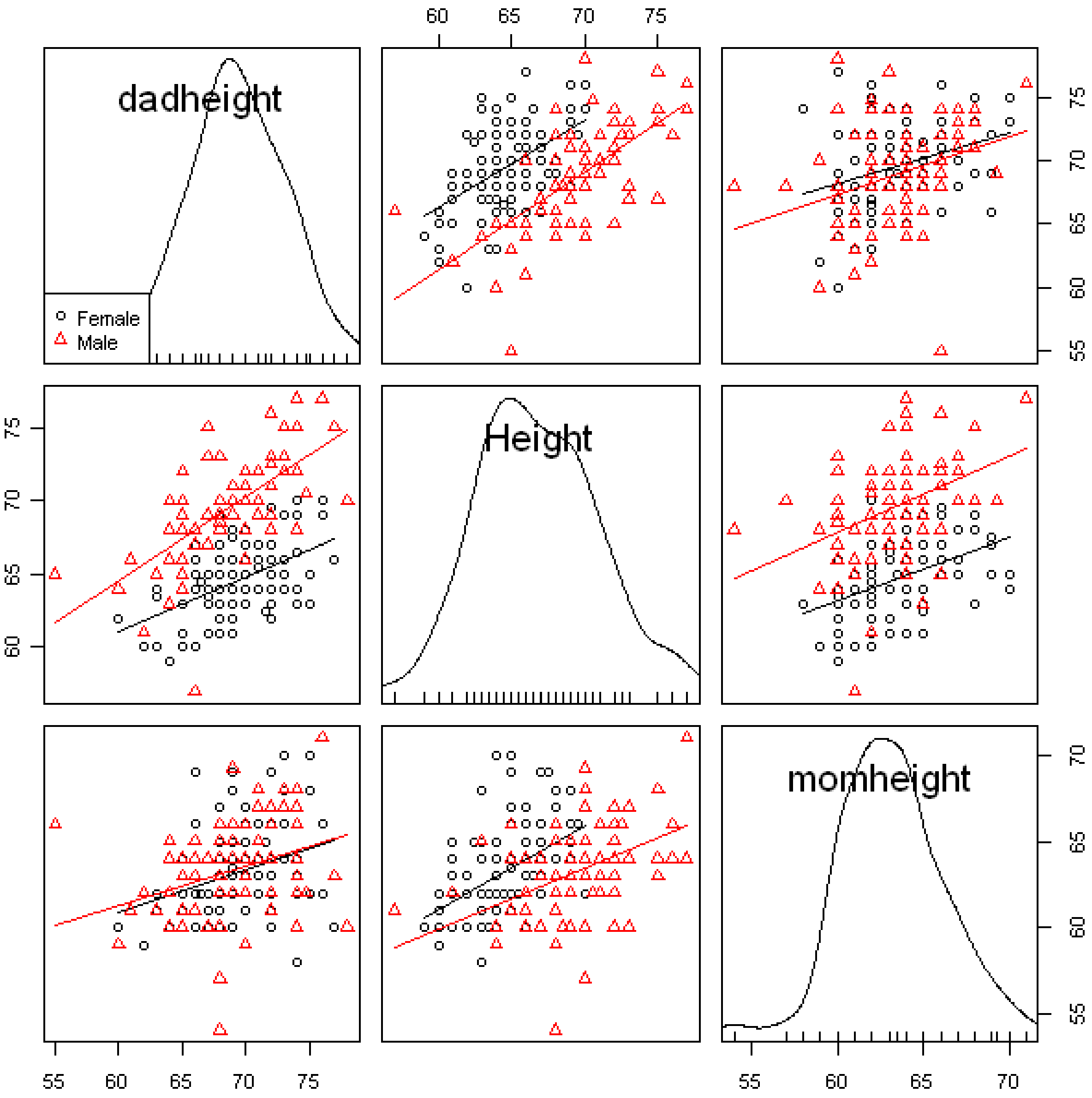
```
> stem.leaf(UCDavis1$momheight, trim.outliers=FALSE, depths=FALSE)
1 | 2: represents 12
 leaf unit: 1
            n: 170
    f | 4
    s | 7
   5. | 8999
   6* | 00000000000000000000000111111111111111111
    t | 22222222222222222222222222222333333333333333333
    f | 444444444444444444444444444555555555555
    s | 66666666666667777777777
   6. | 8888889999
   7* | 001
    t |
    f |
    s |
   7. |
   8* | 0
NA's: 3
```

Graph 3: Stem and leaf plot of fathers' heights. Notice the outlier at 55 inches, not an obvious mistake, so we do not remove it. Also notice that there are 6 missing values..

```
> stem.leaf(UCDavis1$dadheight, trim.outliers=FALSE, depths=FALSE)
1 | 2: represents 12
 leaf unit: 1
            n: 167
    f | 5
    s |
   5. |
   6* | 001
    t | 223333
    f | 44444455555555555
    s | 66666666666777777777777
   6. | 8888888888888888888888899999999999999999
   7* | 0000000000000000000111111111111
    t | 2222222222222223333333
    f | 444444444444444555
    s | 66677
   7. | 8
NA's: 6
```

Graph 4: Scatter plot matrix, separately for Females and Males.
Notice the mother's height of 80 has been removed.
Here is the R command (or you can use R Commander; graph -> Scatterplot matrix, easier in this case!):

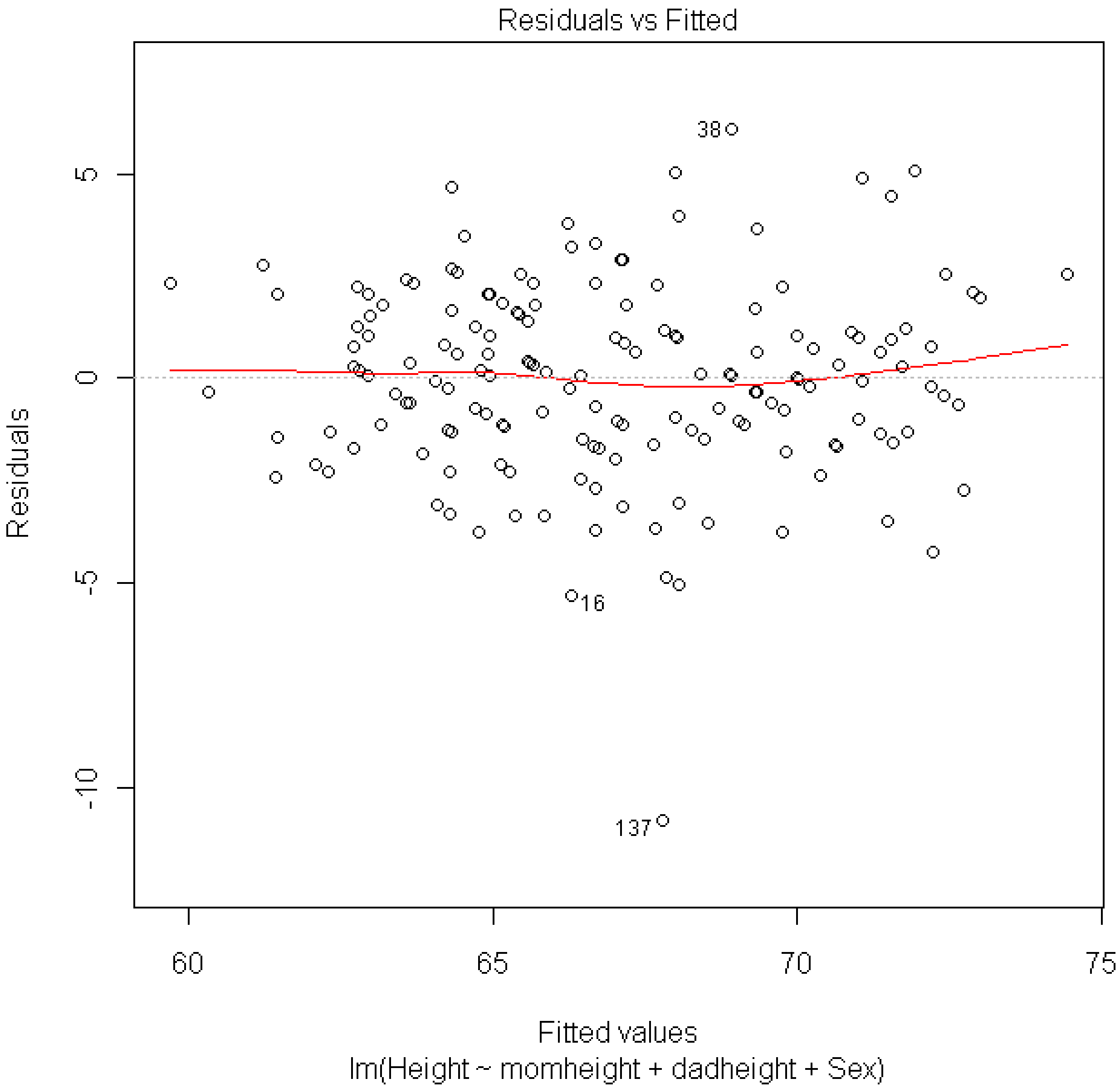
```
scatterplotMatrix(~dadheight+Height+momheight | Sex, reg.line=lm, smooth=FALSE,
spread=FALSE, span=0.5, diagonal= 'density', by.groups=TRUE, data=UCDavis1)
```



Graph 5: Residuals versus fitted values

```
>plot (ParentHt, 1)
```

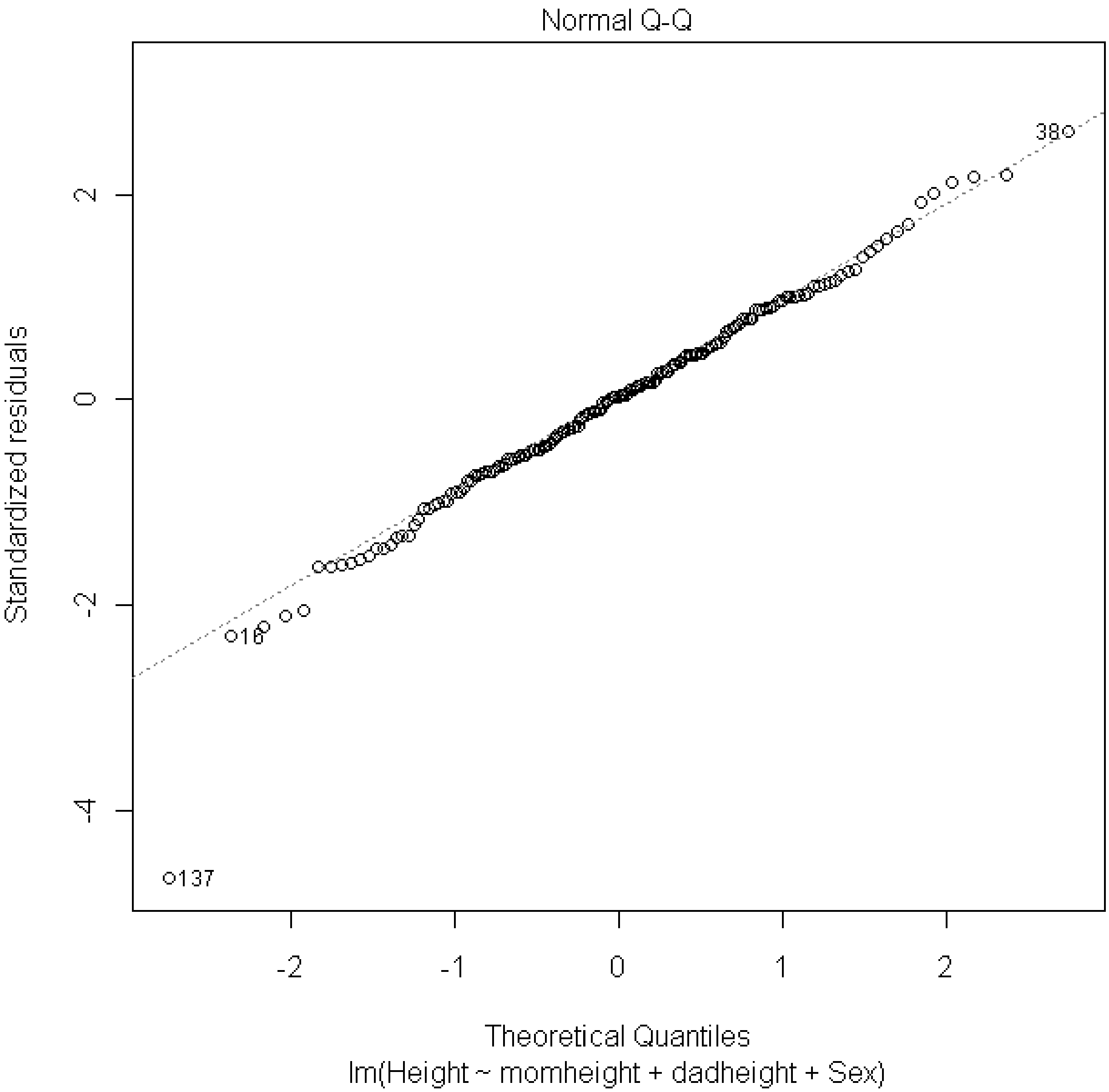
Notice that the very short male (57 inches) is predicted to be much taller, so he has a large negative residual. (His parents were 61 and 66 inches tall, so they were short too.)



Graph 6: Normal probability plot of the residuals.

```
>plot (ParentHt, 2)
```

It shows the outlier too, but otherwise looks good.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Height ~ momheight + dadheight + Sex)