MULTIPLE REGRESSION EXAMPLE

For a sample of n = 166 college students, the following variables were measured:
Y = height
X_1 = mother's height ("momheight")
X_2 = father's height ("dadheight")
X_3 = 1 if male, 0 if female ("male")

Our goal is to predict student's height using the 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 $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$ where ε_i are independent, $N(0, \sigma^2)$.

NOTE: The original data has a text variable called "sex" with two values 'Male' and 'Female'. To create the **indicator variable** "Male" in Stata, the commands are:

```
. generate male = 1
. replace male = 0 if sex=="'Female'"
```

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. The plots are:
G1. Separate histograms of male and female students' heights. Stata: `histogram height, by(sex)`
G2. Histogram of mothers' heights. Stata: `histogram momheight`
G3. Histogram of fathers' heights. Stata: `histogram dadheight`
G4. A "scatter plot matrix" (see p. 232 of text), separately for males and females.
Stata: `graph matrix height momheight dadheight, by(sex)` or Graphics -> Scatterplot matrix

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 value to the missing value code, which is a period in Stata:
  `replace momheight = . in 129`
- 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, not 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 regress command:

```
. regress height momheight dadheight male
```

We will follow that up with some plots:
G5: A plot of the residuals versus fitted values:

```
rvfplot, yline(0)
```

G6: A normal probability plot (we need to create the variable "residuals" first):

```
predict residuals, residuals
qnorm residuals
```

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

```
      Source |       SS       df       MS              Number of obs =     165
-------------+------------------------------           F(  3,   161) =  104.38
       Model |  1679.17741     3  559.725803            Prob > F      =  0.0000
    Residual |   863.31653   161  5.36221447            R-squared     =  0.6604
-------------+------------------------------           Adj R-squared =  0.6541
       Total |  2542.49394   164  15.5030118           Root MSE      =  2.3156

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   momheight |   .2996154   .0687614     4.36   0.000     .1638247     .435406
   dadheight |    .412135   .0510733     8.07   0.000      .311275     .512995
        male |   5.298218   .3637717    14.56   0.000     4.579839    6.016597
       _cons |   16.96746   4.658309     3.64   0.000     7.768196    26.16673
```

- 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" has a specific interpretation. It says 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 of momheight and dadheight.

- 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.

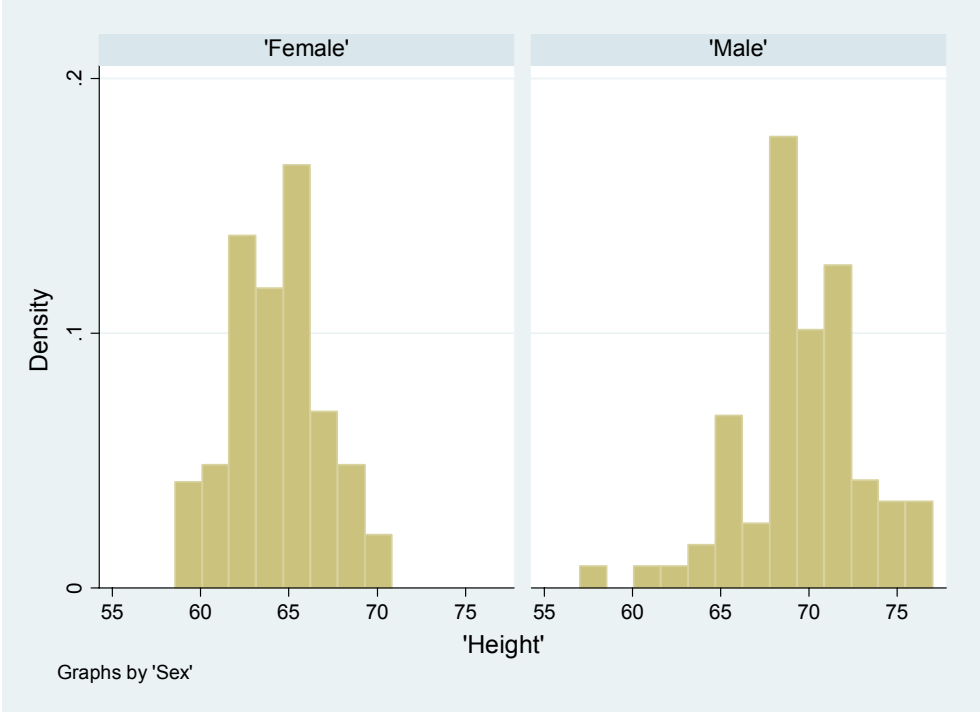
- For *each* of the coefficients, a test for H_0: $\beta = 0$ versus H_a: $\beta \neq 0$ has p-value of 0.000. (See the column headed `P>|t|`.) These are *conditional* hypotheses. They are testing whether or not each explanatory variable needs to be in the model, *given* that the others are already there. 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.

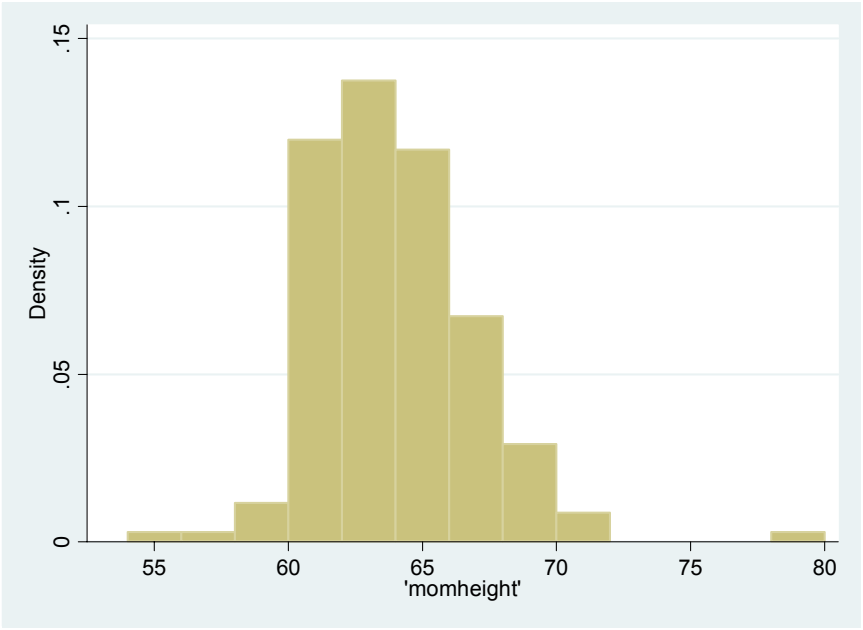
- R^2 = 66.04%, which is pretty good. Later we will learn about "Adjusted R^2" which can be more useful in multiple regression, especially when comparing models with different numbers of X variables.

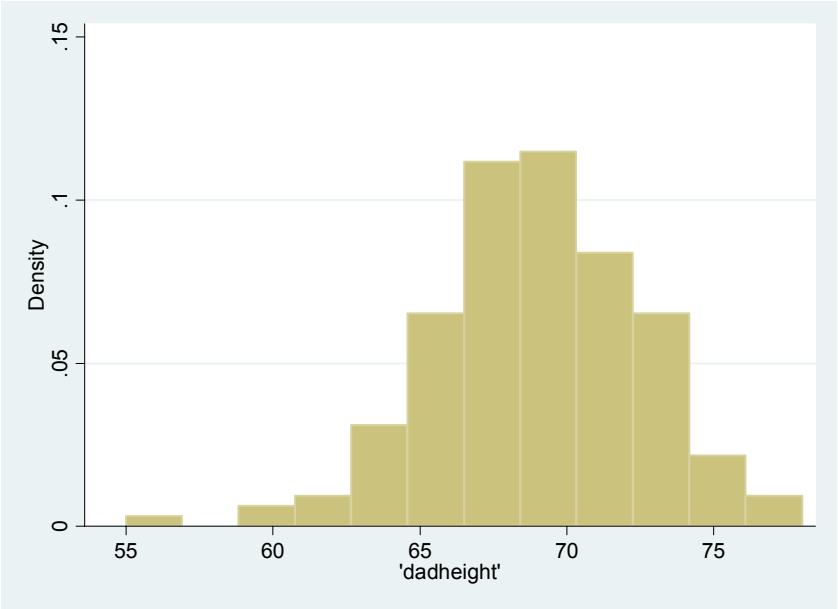
- Root MSE = *s* = our estimate of σ = 2.32 inches, indicating that *within* every *combination* of momheight, dadheight and sex, the standard deviation of heights is about 2.32 inches. In other words, that's the estimate of the standard deviation for the population of all male students with momheight = 64 and dadheight = 70, for the population of all female students with those parents' heights, etc, for any combination of the 3 explanatory variables.

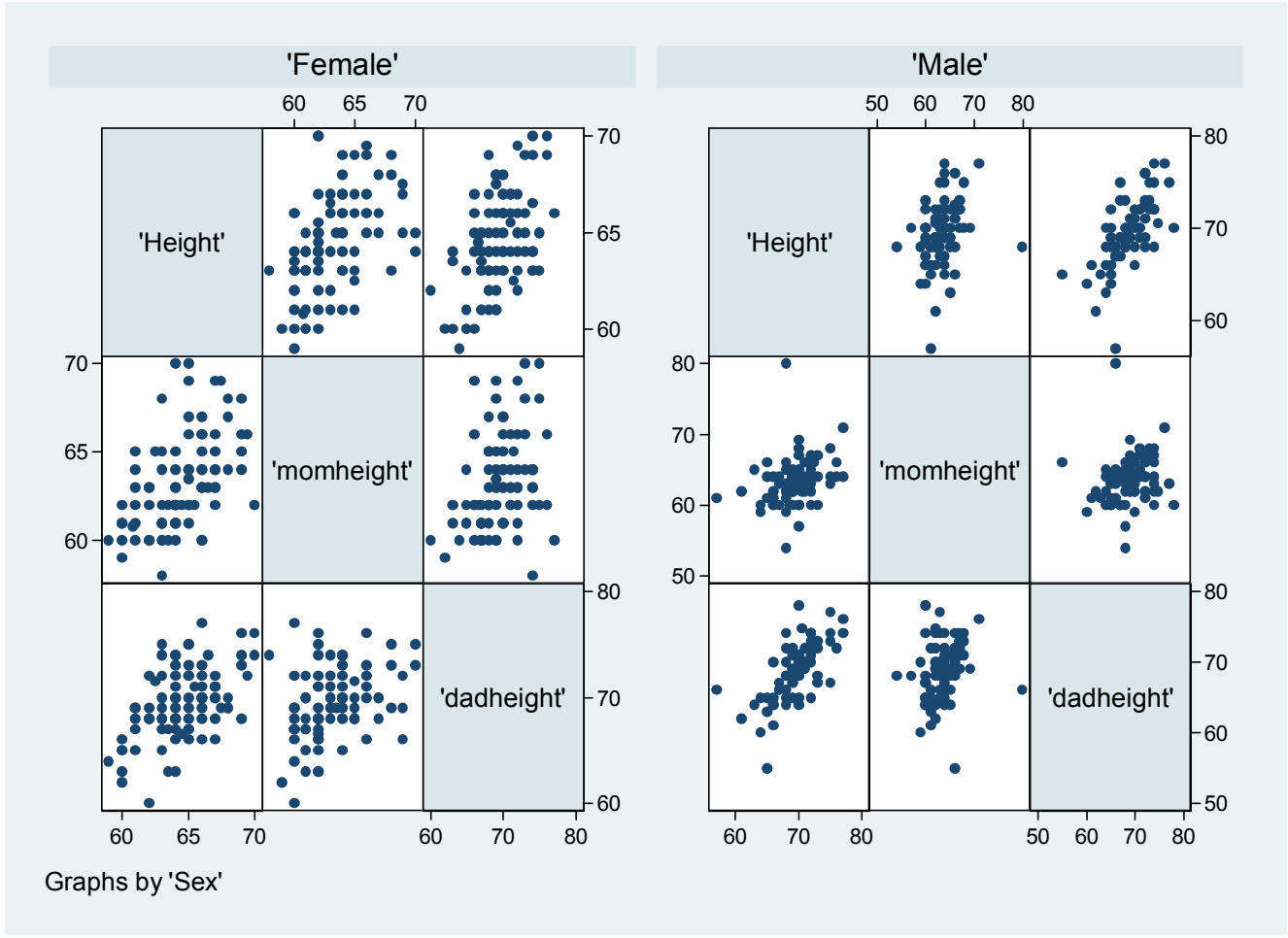
- The F(3, 161) = 104.38 is F* for testing the full model versus the reduced model $E\{Y_i\} = \beta_0$. In other words, it is *simultaneously* testing H_0: β_1, β_2, β_3 all = 0 versus H_a: At least one is not 0. The p-value is given as 0.0000, so clearly we can reject the null hypothesis and we can conclude that at least one of the explanatory variables is useful.

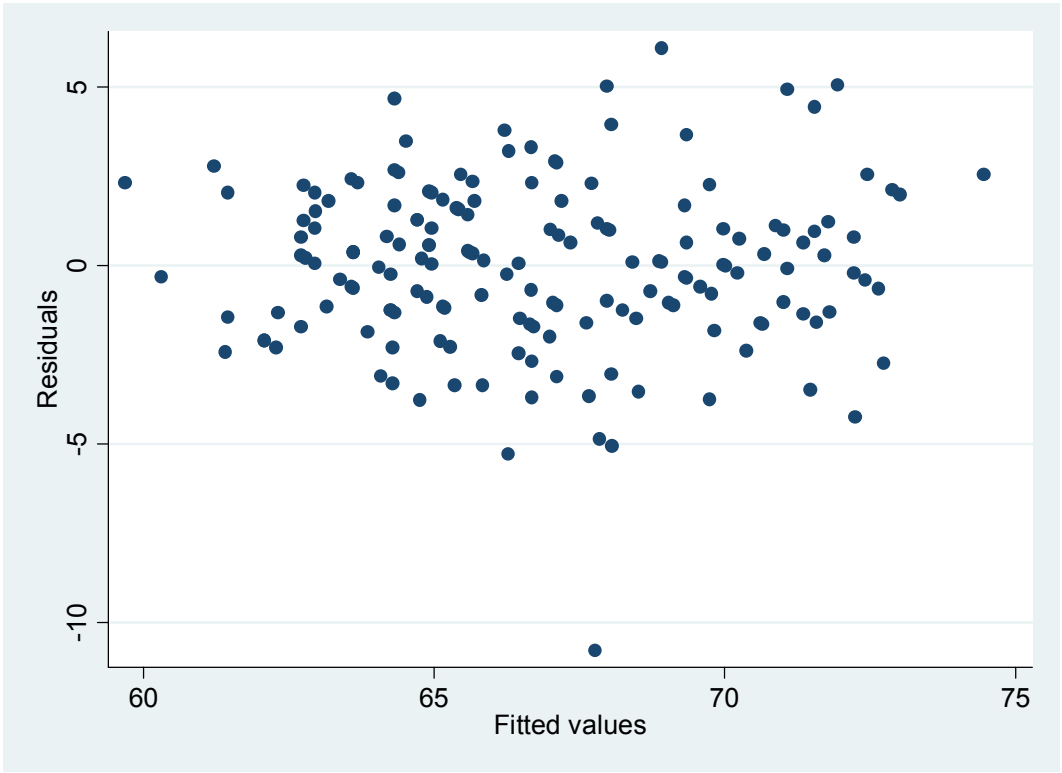
- The residual versus fitted values and normal probability plot both show the outlier (height = 57") but otherwise they look like we hope they would.

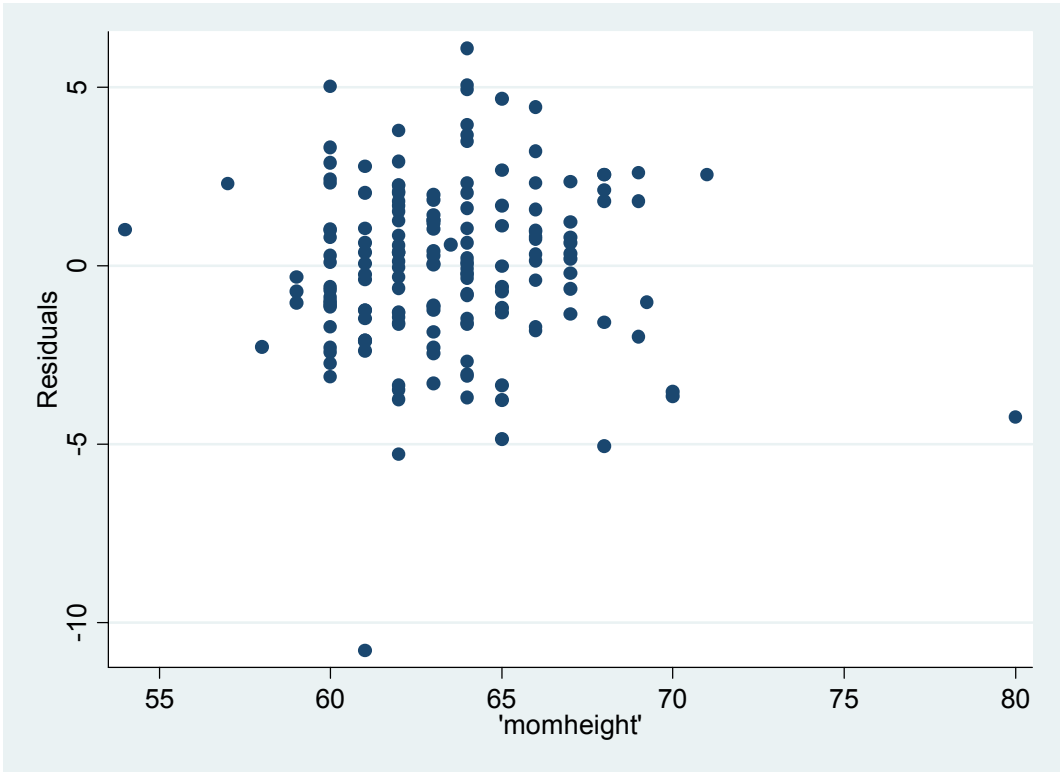

'Female'
'Male'
Density
0
.1
.2
55
60
65
70
75
'Height'
Graphs by 'Sex'


Density
0
.05
.1
.15
55
60
65
70
75
80
'momheight'

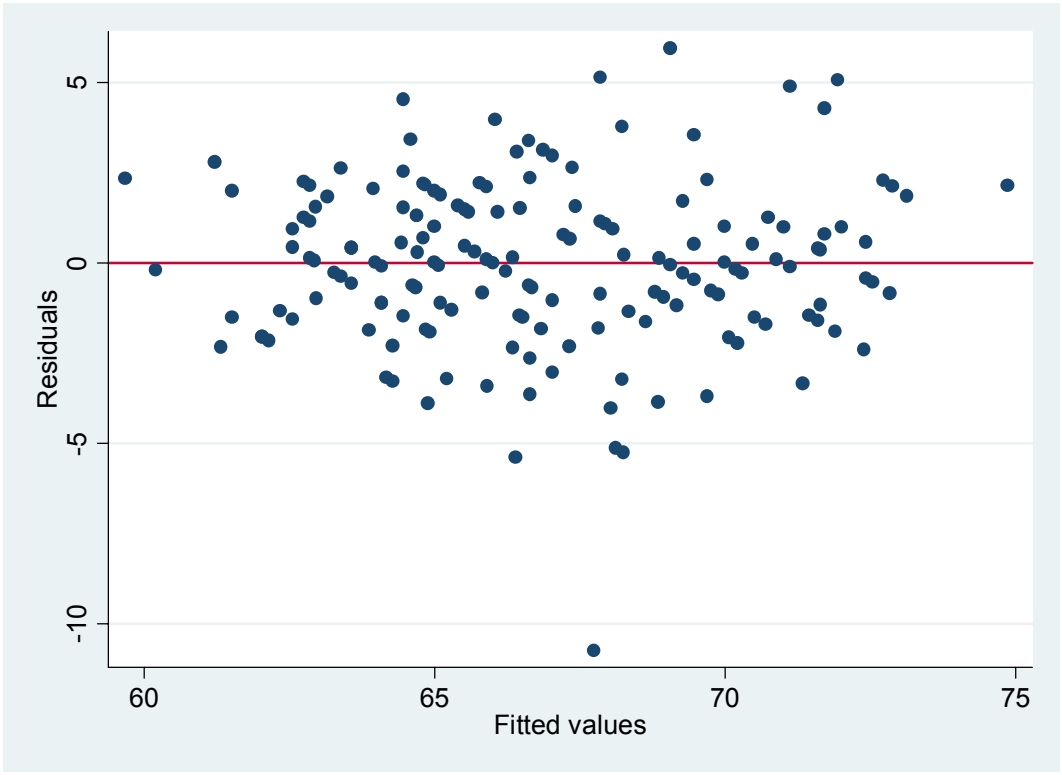

Density
0
.05
.1
.15
55
60
65
70
75
'dadheight'


'Female'
'Male'
'Height'
'momheight'
'dadheight'
Graphs by 'Sex'





Mom height outlier removed:



The remaining outlier is a male 57 inches tall with parents' heights of 61 and 66 inches (mom and dad). This could be a legitimate point so it's not okay to remove it.

Normal probability plot shows the outlier too, but otherwise looks good: