# Unit 13 Pre Class

*w203 Instructional Team*

*3/31/2018*

The `GaltonFamilies` dataframe, which comes with the `HistData` package, reports the height of parents and their children.

```
##install.packages('HistData')
library(HistData)
head(GaltonFamilies)
```

```
##   family father mother midparentHeight children childNum gender
## 1    001   78.5   67.0           75.43        4        1   male
## 2    001   78.5   67.0           75.43        4        2 female
## 3    001   78.5   67.0           75.43        4        3 female
## 4    001   78.5   67.0           75.43        4        4 female
## 5    002   75.5   66.5           73.66        4        1   male
## 6    002   75.5   66.5           73.66        4        2   male
##   childHeight
## 1        73.2
## 2        69.2
## 3        69.0
## 4        69.0
## 5        73.5
## 6        72.5
```

A simple scatter plot shows that father's height (measured by `father`) is a strong predictor of child's height. (See `?Galton` and `?GaltonFamilies` for the original uses of the data.) Let's use this data set to explore indicator variables, interactions, and the classical linear model assumptions.

Q1. The `gender` variable reports the child's gender. The linear model allows us to test the simple hypothesis that female children are taller than males. In the language of regression, `female` would be called the omitted category or excluded category. Define an indicator variable and use it to test the hypothesis described above. [Note: **R** will also accept factor variables as arguments to linear models, and these can be quite usefull.] Describe your results carefully.

Q2. Linear regression also allows us to test a different sort of hypothesis - is the reltaionship between parent's height and child's height different for female than for male children. Specify a model to test this hypothesis. Remember, the model should include not only the interaction, but also both of the constituent terms. Which hypothesis does the coefficient on `father` now test? What about the interaction term? Something strange has happened to the coefficient on `female`. Can you understand why?

Q3. One interpretation of the model you created above is that it estimates two separate regression slopes. Can you superimpose the two corresponding regression lines on the scatterplot?

Q4. Think carefully about this data set. Which one of the classical linear assumptions does it violate?