# Unit 13 Pre Class

w203

*Adam Yang*

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.

```
Galton <- GaltonFamilies
Galton["female"] <- as.integer(Galton$gender == "female")
Galton["male"] <- as.integer(Galton$gender == "male")
model1 <- lm(childHeight ~ female, data = Galton)
summary(model1)
```

```
##
## Call:
## lm(formula = childHeight ~ female, data = Galton)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.234 -1.604 -0.104  1.766  9.766
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.2341     0.1139  608.00   <2e-16 ***
## female       -5.1301     0.1635  -31.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 932 degrees of freedom
## Multiple R-squared:  0.5137, Adjusted R-squared:  0.5132
## F-statistic: 984.4 on 1 and 932 DF,  p-value: < 2.2e-16
```

When we use the linear model to predict group means, we see that, not holding other factors constant, female children are around 5 inches shorter than male children. The t-test provided lets us know that this is a highly significant t-statistic and therefore, we can reject the null hypothesis that male and female heights are equal.

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?

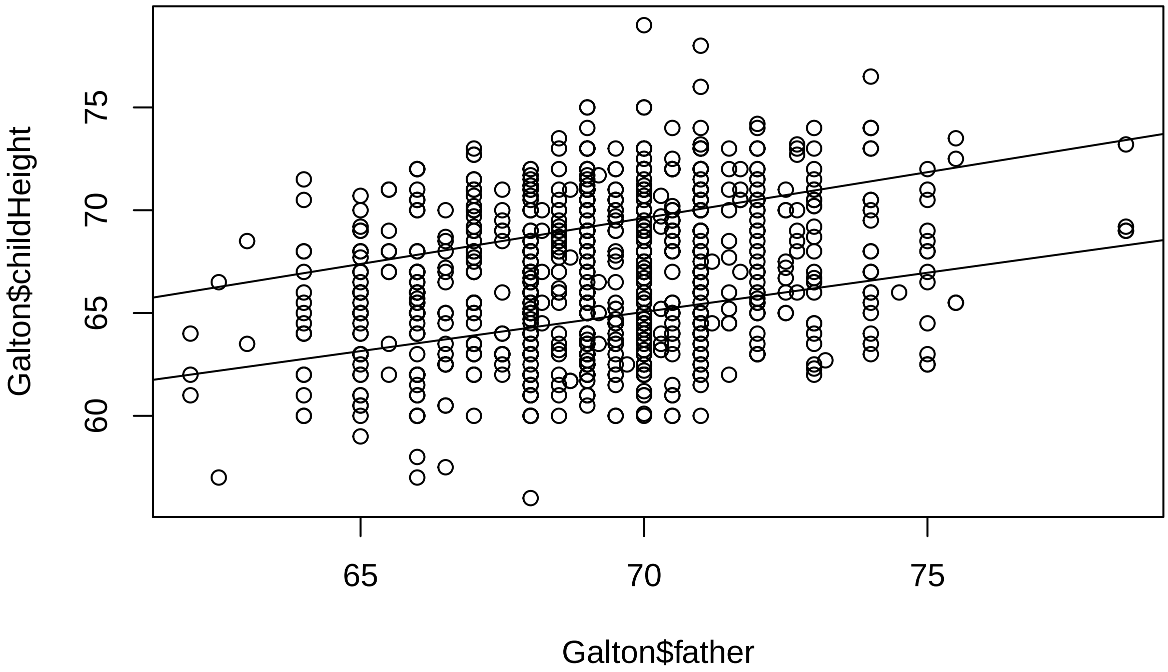
```r
model2 <- lm(childHeight ~ female+father*female, data = Galton)
summary(model2)
```

```
##
## Call:
## lm(formula = childHeight ~ female + father * female, data = Galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3959 -1.5047 -0.0047  1.5913  9.3808
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.36258    3.12513  12.276   <2e-16 ***
## female        -0.66761    4.20332  -0.159    0.874
## father         0.44652    0.04518   9.884   <2e-16 ***
## female:father -0.06522    0.06071  -1.074    0.283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.282 on 930 degrees of freedom
## Multiple R-squared:  0.5948, Adjusted R-squared:  0.5935
## F-statistic:   455 on 3 and 930 DF,  p-value: < 2.2e-16
```

The slope coefficient for father of 0.447 is the effect on child's height for males. The hypothesis that is tested for this coefficient is if there's a difference between the effect of the father's height on a male child vs a female child. The result is highly significant so the answer is we can reject the null hypothesis that the father's height has the same effect on a male child and a female child. The interaction term's coefficient is the difference between the impact on males and on females.

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?

```r
plot(Galton$father, Galton$childHeight)
abline(model2$coefficients[1], model2$coefficients[3])
abline(model2$coefficients[1], model2$coefficients[3]+model2$coefficients[4])
```



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

I violates the assumption of random sampling because there is clustering occuring within the data set. There are multiple children per family that would have the same genetic make up so each children from the same family would be part of the same cluster inherently.