Unit 13 Pre Class

Joanna Yu (W203 Tuesday 4pm Fall 2018)

12/4/2018

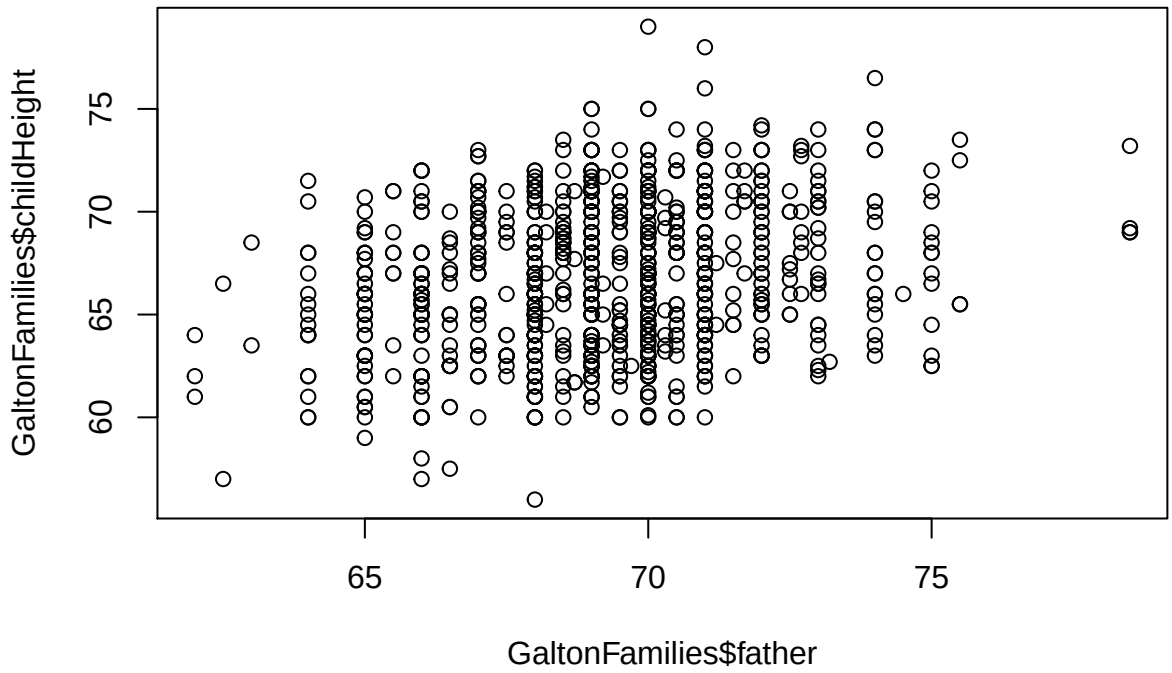
```
library(HistData)
head(GaltonFamilies)
```

```
##   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
```

```
summary(GaltonFamilies)
```

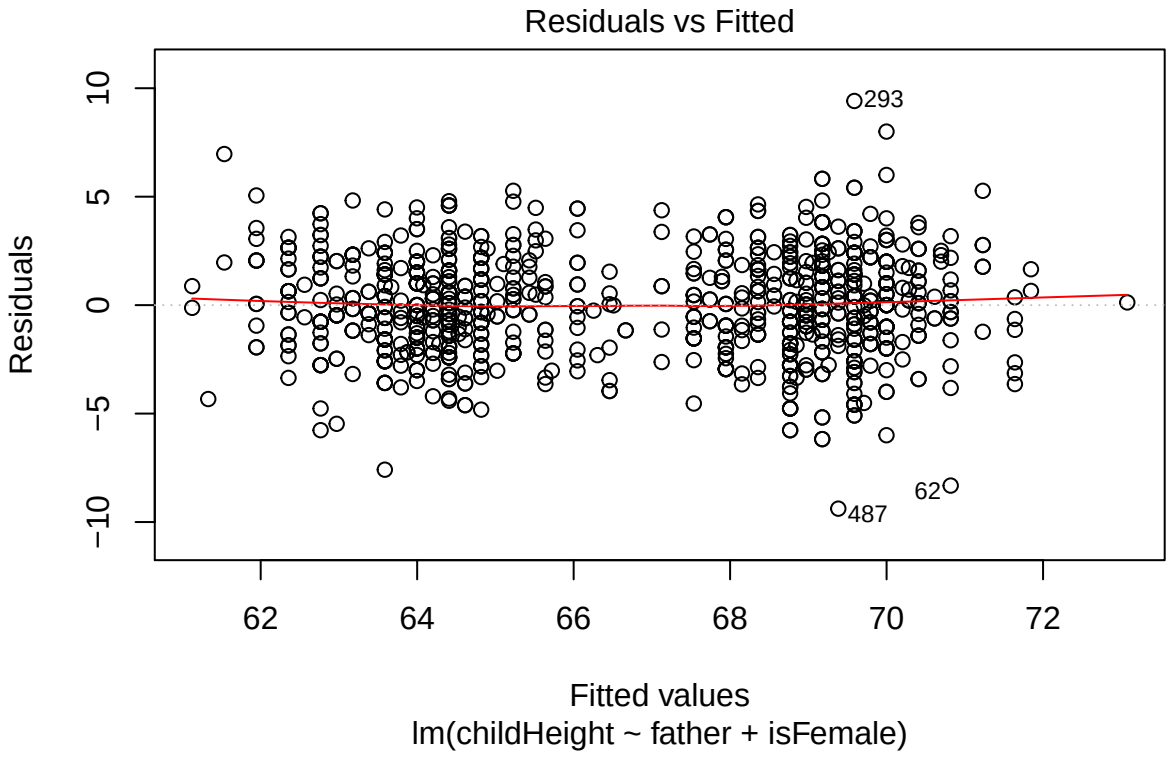
```
##      family        father          mother      midparentHeight
##  185    : 15   Min.   :62.0   Min.   :58.00   Min.   :64.40
##  066    : 11   1st Qu.:68.0   1st Qu.:63.00   1st Qu.:68.14
##  120    : 11   Median :69.0   Median :64.00   Median :69.25
##  130    : 11   Mean   :69.2   Mean   :64.09   Mean   :69.21
##  166    : 11   3rd Qu.:71.0   3rd Qu.:65.88   3rd Qu.:70.14
##  097    : 10   Max.   :78.5   Max.   :70.50   Max.   :75.43
##  (Other):865
##     children         childNum         gender     childHeight
##  Min.   : 1.000   Min.   : 1.000   female:453   Min.   :56.00
##  1st Qu.: 4.000   1st Qu.: 2.000   male  :481   1st Qu.:64.00
##  Median : 6.000   Median : 3.000                Median :66.50
##  Mean   : 6.171   Mean   : 3.586                Mean   :66.75
##  3rd Qu.: 8.000   3rd Qu.: 5.000                3rd Qu.:69.70
##  Max.   :15.000   Max.   :15.000                Max.   :79.00
##
```

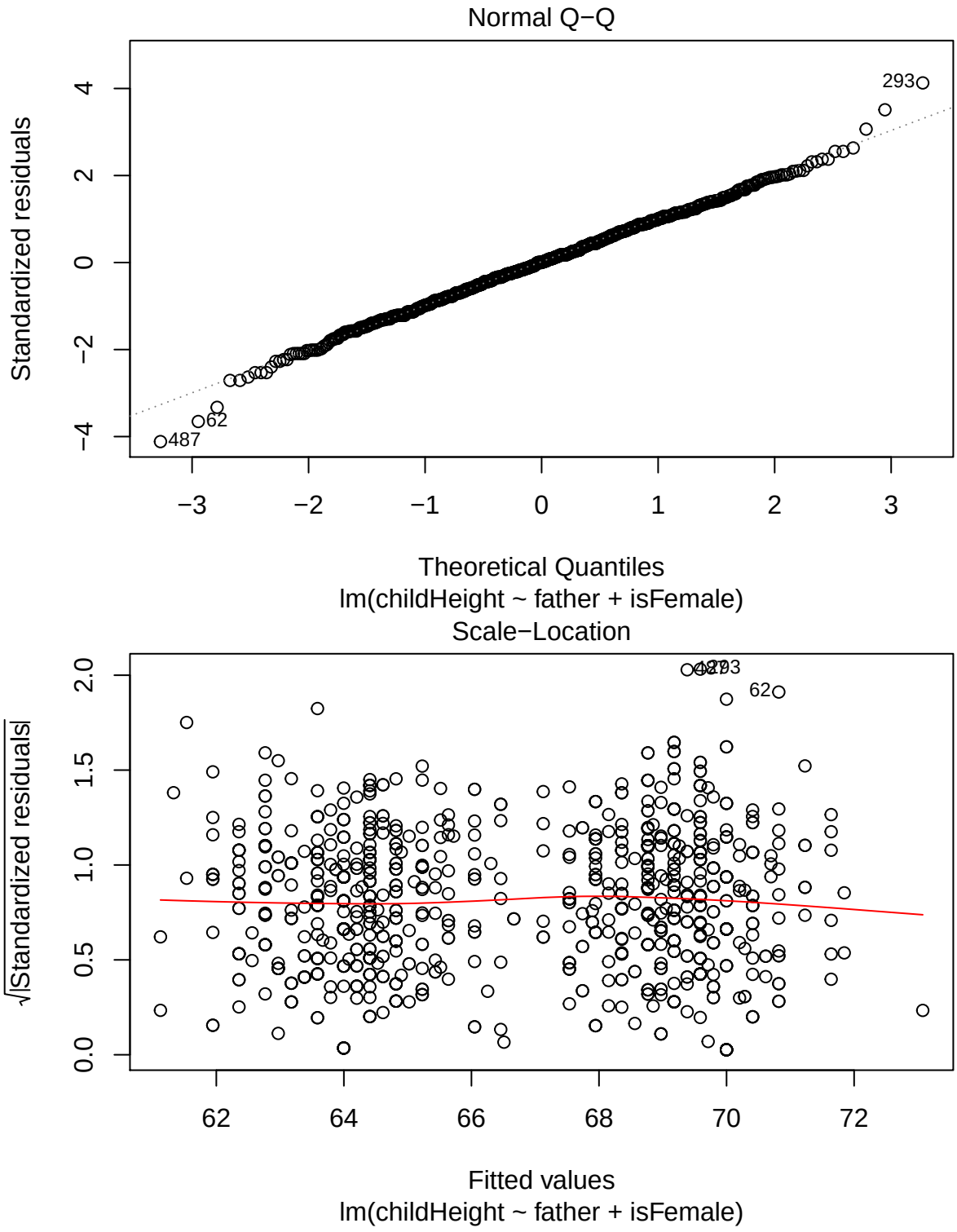
```
plot(GaltonFamilies$father, GaltonFamilies$childHeight)
```

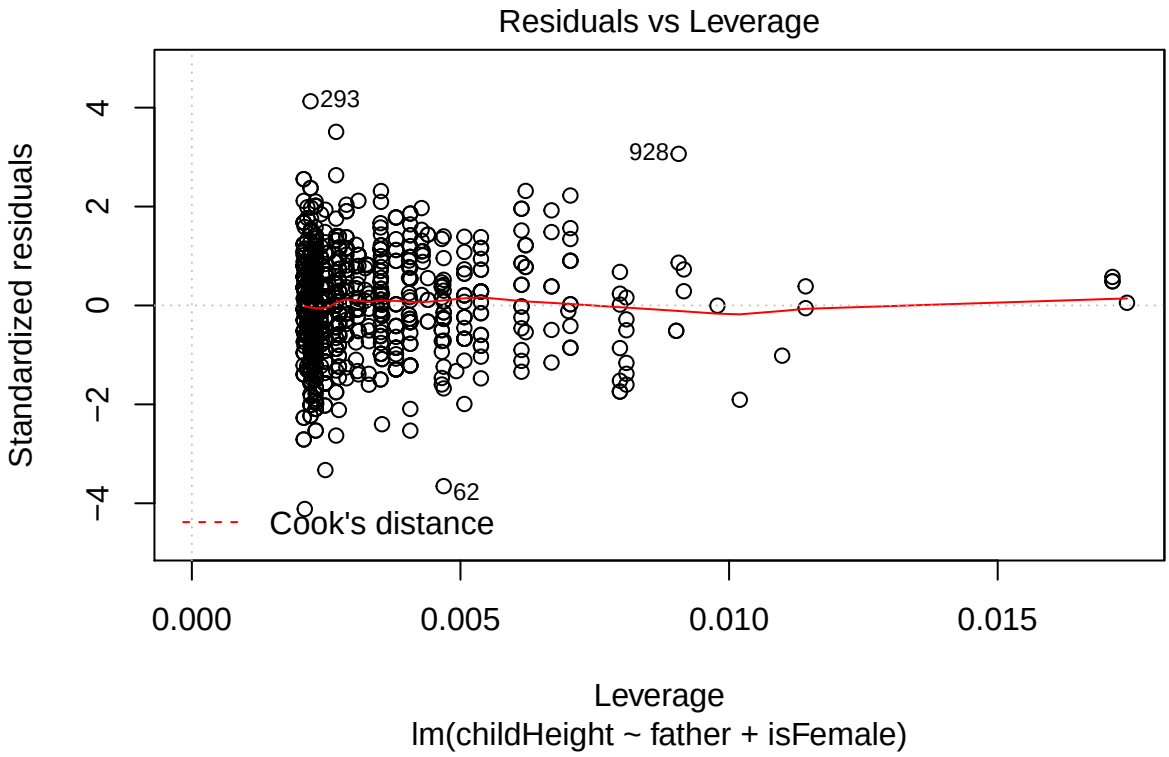


Q1. The gender variable reports the child's gender. The linear model allows us to test the simple hypothesis that female children are taller than males. In the language of regression, female would be called the omitted category or excluded category. Define an indicator variable and use it to test the 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.

```
GaltonFamilies$isFemale = GaltonFamilies$gender == 'female'
model_fheight=lm(childHeight ~ father + isFemale, data = GaltonFamilies)
plot(model_fheight)
```




Normal Q−Q
Standardized residuals
293
62
487
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
lm(childHeight ~ father + isFemale)
Scale−Location
√|Standardized residuals|
293
62
0.0
0.5
1.0
1.5
2.0
62
64
66
68
70
72
Fitted values
lm(childHeight ~ father + isFemale)




```
summary(model_fheight)
```

```
##
## Call:
## lm(formula = childHeight ~ father + isFemale, data = GaltonFamilies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3828 -1.4981  0.0028  1.5924  9.4120
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.85955    2.08928   19.56   <2e-16 ***
## father        0.41041    0.03018   13.60   <2e-16 ***
## isFemaleTRUE -5.18045    0.14948  -34.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.282 on 931 degrees of freedom
## Multiple R-squared:  0.5943, Adjusted R-squared:  0.5934
## F-statistic: 681.8 on 2 and 931 DF,  p-value: < 2.2e-16
```

The result suggests that female children are actually shorter than male children.

Q2. Linear regression also allows us to test a different sort of hypothesis - is the relationship 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?

```
model_male = lm(childHeight ~ father + gender + father*gender, data=GaltonFamilies)
summary(model_male)
```

```
##
## Call:
## lm(formula = childHeight ~ father + gender + father * gender,
##     data = GaltonFamilies)
##
## 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)       37.69497    2.81095  13.410   <2e-16 ***
## father             0.38130    0.04056   9.402   <2e-16 ***
## gendermale         0.66761    4.20332   0.159    0.874
## father:gendermale  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
```

Now the result is no longer statistically significant.

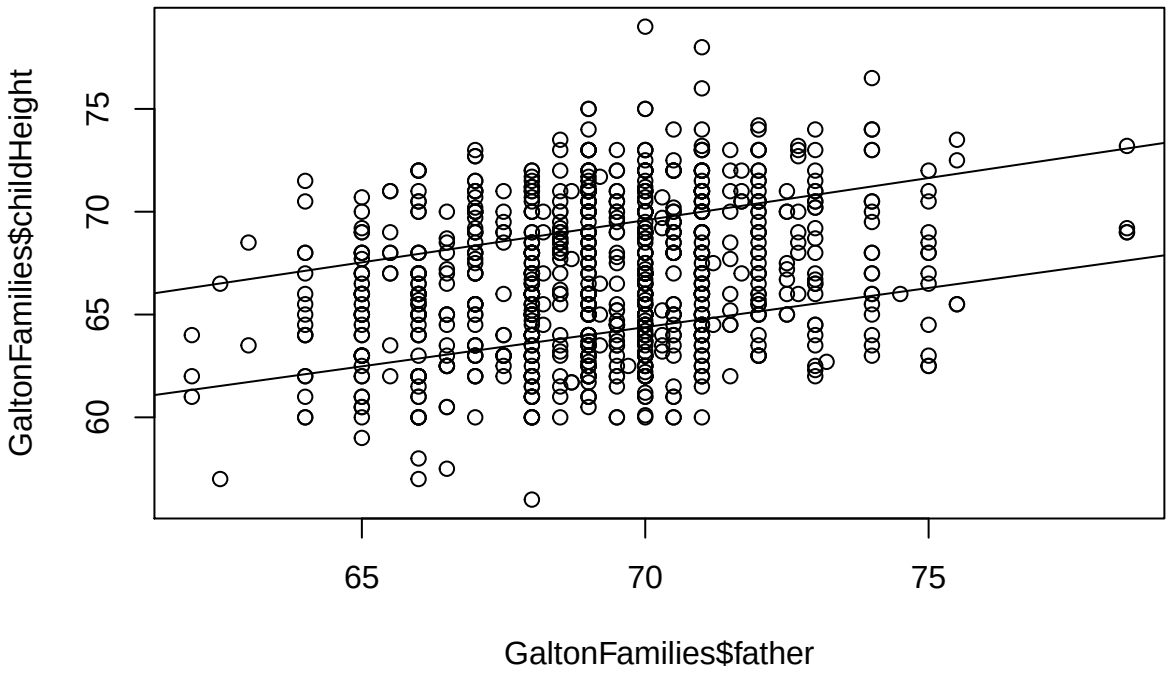
Q3. One interpretation of the model you created above is that it estimates two separate regression slopes. Can you superimpose the two corresponding regression lines on the scatterplot?

```
plot(GaltonFamilies$father, GaltonFamilies$childHeight)
abline(model_fheight)
```

```
## Warning in abline(model_fheight): only using the first two of 3 regression
## coefficients
```

```
abline(model_male)
```

```
## Warning in abline(model_male): only using the first two of 4 regression
## coefficients
```



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

There is high level of colinearity between female and male.