SDS 439 - Homework 01

Joe Guinness

2025-01-22

Homework Exercises

This homework concerns analysis of the Galton families dataset, which we will use several times throughout the course to demonstrate various models and concepts. Each row of the dataset describes data for one child. It has columns identifying which family the child belongs to (`family`), the father's height (`father`), the mother's height (`mother`), the number of children in the family (`children`), the ordering within the family of the present child (`childNum`), the sex of the child (`sex`), and the child's height (`childHeight`). It also has a derived variable called `midparentHeight` that is equal to `(father + 1.08*mother)/2`.

The dataset is located in the course repository in the file `datasets/galton_families.csv`.

1. Read in the Galton Families dataset and print out the first six rows of the dataset

```
galton <- read.csv("../datasets/galton_families.csv")
head(galton)
```

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

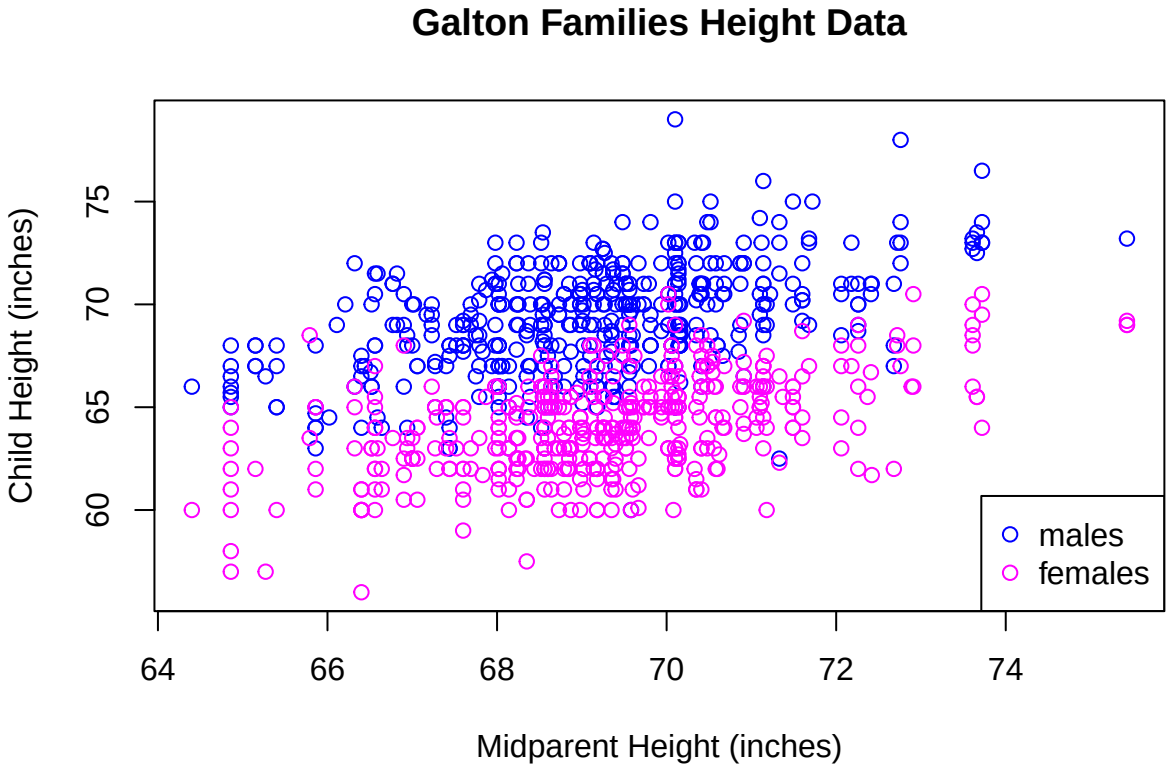
2. Demonstrate that `midparentHeight = (father + 1.08*mother)/2`

```
range( galton$midparentHeight - (galton$father + 1.08*galton$mother)/2 )
```

```
## [1] -1.421085e-14  1.421085e-14
```

3. Make a plot of the data, showing information about the midparent height, the child height, and if you can, information about the sex. Try using a different plotting symbol for the two sexes. Make sure that your plot looks nice in the final pdf. This is what we will grade.

```
plot(
    galton$midparentHeight, galton$childHeight, type = "n",
    xlab = "Midparent Height (inches)",
    ylab = "Child Height (inches)",
    main = "Galton Families Height Data"
)
males <- galton$sex == "male"
points( galton$midparentHeight[males], galton$childHeight[males], col = "blue" )
points( galton$midparentHeight[!males], galton$childHeight[!males], col = "magenta" )
legend("bottomright", c("males","females"), col = c("blue","magenta"), pch = 1 )
```



4. Let's focus on one of the sexes, then we'll come back and analyze both of them. Create a new dataframe that has data for only the male children, and print out the first six rows.

```
galton_males <- galton[ galton$sex == "male", ]
head( galton_males )
```

```
##    family father mother midparentHeight children childNum  sex childHeight
## 1     001   78.5   67.0           75.43        4        1 male        73.2
## 5     002   75.5   66.5           73.66        4        1 male        73.5
## 6     002   75.5   66.5           73.66        4        2 male        72.5
## 9     003   75.0   64.0           72.06        2        1 male        71.0
## 11    004   75.0   64.0           72.06        5        1 male        70.5
## 12    004   75.0   64.0           72.06        5        2 male        68.5
```

5. Use the male dataframe to perform a regression of child height on midparent height "by hand", which means doing all of the calculations without using the `lm` function. Print out your estimates for the regression coefficients, their standard errors, their t-statistics, and their p-values for the two sided test for whether the parameters equal 0. Also print out your estimate of the residual variance.

```
x <- galton_males$midparentHeight
y <- galton_males$childHeight
n <- nrow( galton_males )
xbar <- mean(x)
ybar <- mean(y)
sxx <- sum( (x-xbar)^2 )
sxy <- sum( (x-xbar)*(y-ybar) )
b1hat <- sxy/sxx
b0hat <- ybar - b1hat*xbar
sigmahat <- sqrt( 1/(n-2)*sum( (y - b0hat - b1hat*x )^2 ) )
seb0 <- sigmahat*sqrt(1/n + xbar^2/sxx)
seb1 <- sigmahat/sqrt(sxx)
```

```
t0 <- b0hat/seb0
t1 <- b1hat/seb1
p0 <- 2*pt( -abs(t0), n - 2 )
p1 <- 2*pt( -abs(t1), n - 2 )
print( c( sigmahat = sigmahat ) )
```

```
## sigmahat
##   2.3003
```

```
print( c(b0hat = b0hat, seb0 = seb0, t0 = t0, p0 = p0) )
```

```
##        b0hat         seb0           t0           p0
## 1.991346e+01 4.089427e+00 4.869500e+00 1.522193e-06
```

```
print( c(b1hat = b1hat, seb1 = seb1, t1 = t1, p1 = p1) )
```

```
##        b1hat         seb1           t1           p1
## 7.132745e-01 5.912179e-02 1.206449e+01 1.890477e-29
```

6. Why do you think that the intercept estimate has such a large standard error? Type your answer here

Answer: The midparent heights are never close to zero, and the intercept is essentially a prediction of childHeight when midparent height is zero, so it is an extrapolation and should have a large amount of uncertainty.

7. Use the `lm` function to confirm your "by hand" calculations. Print the summary table from the lm fit. Use the full dataset, but tell lm to analyze only the males by specifying the `subset` argument of `lm`.

```
fit_male <- lm( childHeight ~ midparentHeight, data = galton, subset = sex == "male" )
print( summary( fit_male ) )
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = galton, subset = sex ==
##     "male")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5431 -1.5160  0.1844  1.5082  9.0860
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     19.91346    4.08943   4.869 1.52e-06 ***
## midparentHeight  0.71327    0.05912  12.064  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.3 on 479 degrees of freedom
## Multiple R-squared:  0.2331, Adjusted R-squared:  0.2314
## F-statistic: 145.6 on 1 and 479 DF,  p-value: < 2.2e-16
```

8. Perform two more regressions using `lm` by changing the covaraiate to `mother` and then to `father`. Print out the summary table for these two regressions.

```
summary( lm( childHeight ~ father, data = galton, subset = sex == "male" ) )
```

```
##
## Call:
## lm(formula = childHeight ~ father, data = galton, subset = sex ==
```

```
##     "male")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3959 -1.5122  0.0413  1.6217  9.3808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.36258    3.30837  11.596   <2e-16 ***
## father       0.44652    0.04783   9.337   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.416 on 479 degrees of freedom
## Multiple R-squared:  0.154,  Adjusted R-squared:  0.1522
## F-statistic: 87.17 on 1 and 479 DF,  p-value: < 2.2e-16
```

```
summary( lm( childHeight ~ mother, data = galton, subset = sex == "male" ) )
```

```
##
## Call:
## lm(formula = childHeight ~ mother, data = galton, subset = sex ==
##     "male")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4045 -1.6569  0.2305  1.6829  9.4130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.85804    3.13150   14.64  < 2e-16 ***
## mother       0.36506    0.04887    7.47 3.84e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.486 on 479 degrees of freedom
## Multiple R-squared:  0.1043, Adjusted R-squared:  0.1025
## F-statistic:  55.8 on 1 and 479 DF,  p-value: 3.838e-13
```

9. Use the information in the summary tables to argue which of the three models for `childHeight` is best. Is there a biological reason why this model is best?

Answer: I would argue that the midparent height model is best because it has the lowest residual standard error (estimate of $\widehat{\sigma}$), which measures how far on average the data is from the predicted value. This makes sense biologically because both parents' heights contribute genetically to the children's heights.

10. Perform a regression of `childHeight` on `midparentHeight` for the female children and print the summary matrix.

```
fit_female <- lm( childHeight ~ midparentHeight, data = galton, subset = sex == "female" )
print( summary( fit_female ) )
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = galton, subset = sex ==
##     "female")
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.207 -1.412 -0.045  1.365  6.696
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     18.33348    3.60497   5.086 5.38e-07 ***
## midparentHeight  0.66075    0.05202  12.701  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.024 on 451 degrees of freedom
## Multiple R-squared:  0.2634, Adjusted R-squared:  0.2618
## F-statistic: 161.3 on 1 and 451 DF,  p-value: < 2.2e-16
```

11. What are the major differences between the female model and the male model?

Answer: The female model has a smaller intercept, smaller slope, and a smaller residual standard error. To see the effect of the different intercept and slope, refer to the plot below showing the regression lines. The small residual standard error is likely due to the fact that the female heights are less variable to begin with:

```
sd( galton$childHeight[ galton$sex == "female" ] )
```

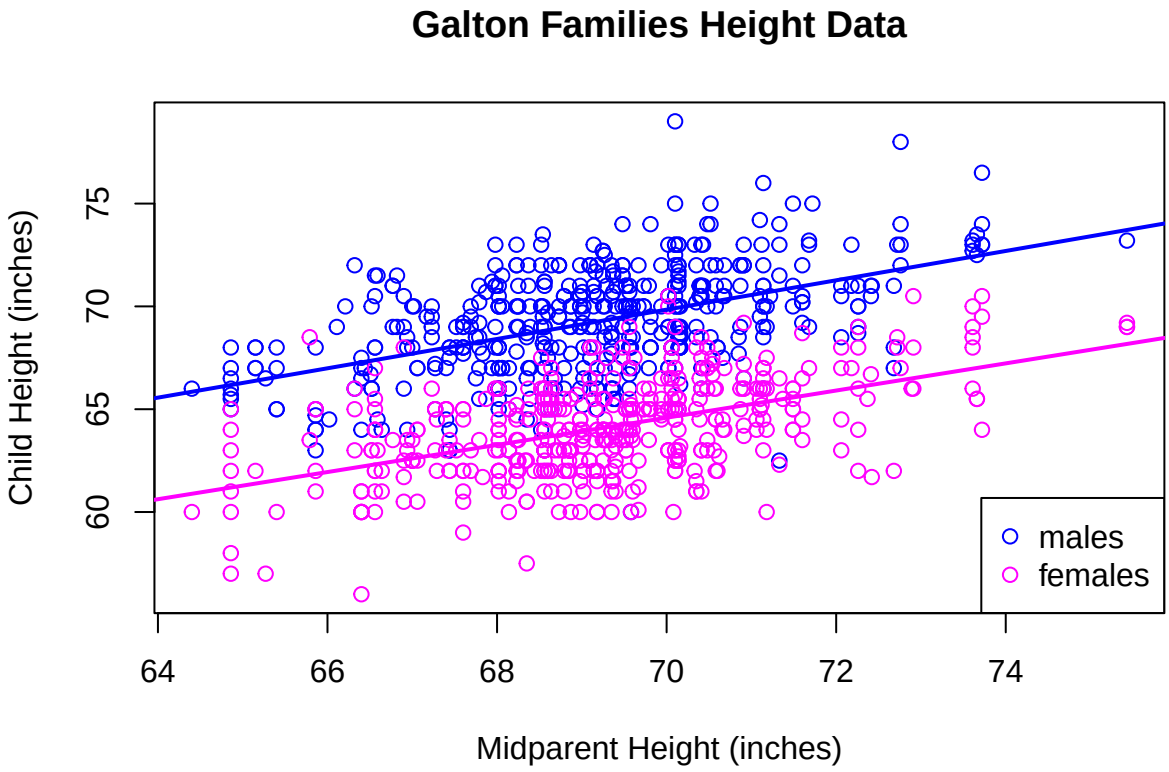
```
## [1] 2.355653
```

```
sd( galton$childHeight[ galton$sex == "male" ] )
```

```
## [1] 2.623905
```

12. Finally, make a plot of childHeight against the midparentHeight, with the two estimated trend lines for males and females added to the plot.

```
plot(
    galton$midparentHeight, galton$childHeight, type = "n",
    xlab = "Midparent Height (inches)",
    ylab = "Child Height (inches)",
    main = "Galton Families Height Data"
)
males <- galton$sex == "male"
points( galton$midparentHeight[males], galton$childHeight[males], col = "blue" )
points( galton$midparentHeight[!males], galton$childHeight[!males], col = "magenta" )
abline( fit_male$coefficients, col = "blue", lwd = 2 )
abline( fit_female$coefficients, col = "magenta", lwd = 2 )
legend("bottomright", c("males","females"), col = c("blue","magenta"), pch = 1 )
```


Galton Families Height Data
Child Height (inches)
60
65
70
75
64
66
68
70
72
74
Midparent Height (inches)
males
females