HW4 Solutions

For the summation questions, write solutions by hand or typed. For the data analysis exercise, write a brief report addressing the questions. For Question 4, writing down a model mathematically means writing a relevant equation, and defining the quantities in the equation — it is probably best not to use matrix notation for this. Include, as an appendix, the R code you used to generate your analysis. Recall that you are permitted to collaborate, or to use any internet resources, but you must list all sources that make a substantial contribution to your report. As usual, please include *Sources* and *Please explain* statements.

Practicing $\sum$ notation for sums

Solve the following problems, giving explanation for your answer. The explanation should involve expanding a $\sum_{i=1}^{n}$ expression into all n terms of a sum, or contracting such a sum into a $\sum_{i=1}^{n}$ expression.

1. Evaluate $\sum_{a=b}^{c} d$, where b and c are whole numbers with $c \geq b$.

Solution: $b + c$ since a starts at b and ends at c.

2) Evaluate $\sum_{i=1}^{n}(x_i - \bar{x})$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Solution:

$$\sum_{i=1}^{n}(x_i - \bar{x})$$

We can distribute the summation sign to get

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

Since $\bar{x}$ is a constant in the sum we are really summing $\bar{x}$ n times

$$\sum_{i=1}^{n} x_i - n\bar{x}$$

We replace $\bar{x}$ with $\frac{1}{n}\sum_{i=1}^{n} x_i$. Thus,

$$\sum_{i=1}^{n} x_i - n(\frac{1}{n})\sum_{i=1}^{n} x_i$$

The n's cancel and we're left with

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

3. Show that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(\sum_{i=1}^{n} x_i y_i) - \bar{x}\bar{y}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Solution:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

First, we foil the equations

$$\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y})$$

We then distribute the summation and $\frac{1}{n}$

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\bar{y} - \frac{1}{n}\sum_{i=1}^{n} y_i\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y}$$

Now we can take the constants out of the summations

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\bar{y}\sum_{i=1}^{n} x_i - \frac{1}{n}\bar{x}\sum_{i=1}^{n} y_i + \frac{1}{n}(n)\bar{x}\bar{y})$$

We now note that $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y})$$

Finally, canceling like terms gives the equality.

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

4. Let $\mathbf{1} = (1, 1, \ldots, 1)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be two vectors treated as $n \times 1$ matrices. Use $\sum$ notation to evaluate the matrix product $\mathbf{1}^T\ \mathbf{x}$.

Solutions:

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\sum_{i=1}^{n} x_i$$

5. Let $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$ be two vectors, and let $\mathbb{X} = [\mathbf{u}\ \mathbf{v}]$ be an $n \times 2$ matrix binding together $\mathbf{u}$ and $\mathbf{v}$. Use $\sum$ notation to evaluate the matrix $\mathbb{X}^T\mathbb{X}$.

$$\mathbb{X} = \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \\ \vdots & \vdots \\ u_n & v_n \end{bmatrix}$$

Therefore $\mathbb{X}^T\mathbb{X}$ is equal to

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \\ v_1 & v_2 & \dots & v_n \end{bmatrix} \begin{bmatrix} u_1 & v_1 \\ u_2 & v_2 \\ \vdots & \vdots \\ u_n & v_n \end{bmatrix}$$

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} u_1^2 + u_2^2 + \dots + u_n^2 & u_1v_1 + u_2v_2 + \dots + u_nv_n \\ u_1v_1 + u_2v_2 + \dots + u_nv_n & v_1^2 + v_2^2 + \dots + v_n^2 \end{bmatrix}$$

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} \sum_{i=1}^n u_i^2 & \sum_{i=1}^n u_iv_i \\ \sum_{i=1}^n u_iv_i & \sum_{i=1}^n v_i^2 \end{bmatrix}$$

Investigating the regression effect

```
#install.packages("HistData")
library("HistData")
data("Galton")
head(Galton)
```

```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

1. What is the average height of the children to three significant figures?

```
round(mean(Galton$child), 3)
```

```
## [1] 68.088
```

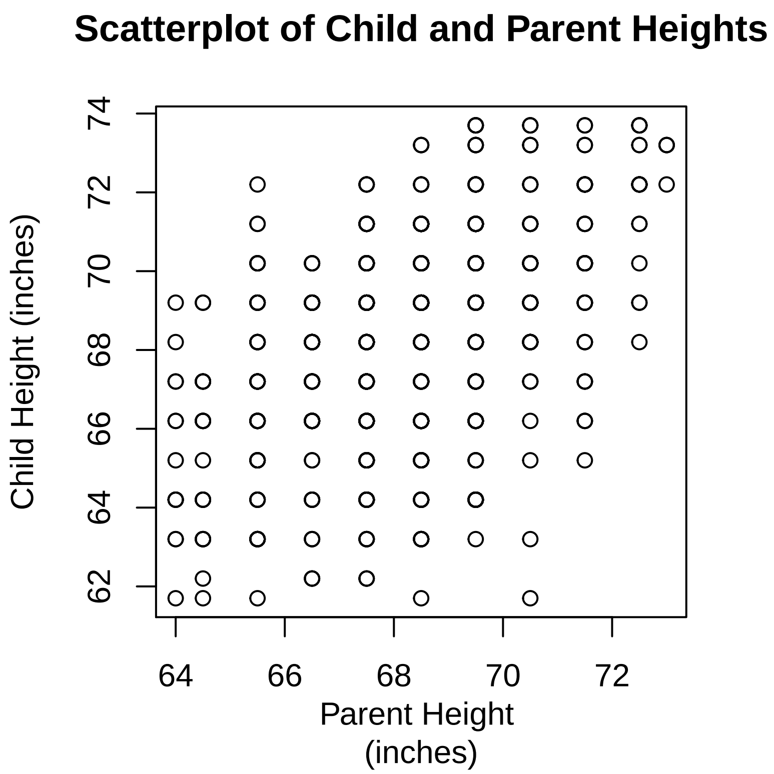
2. What is the average height of children with parent height between 69.0 and 71.0 inches?

```
round(mean(Galton$child[which(Galton$parent >= 69.0 & Galton$parent <= 71.0)]),3)
```

```
## [1] 68.947
```

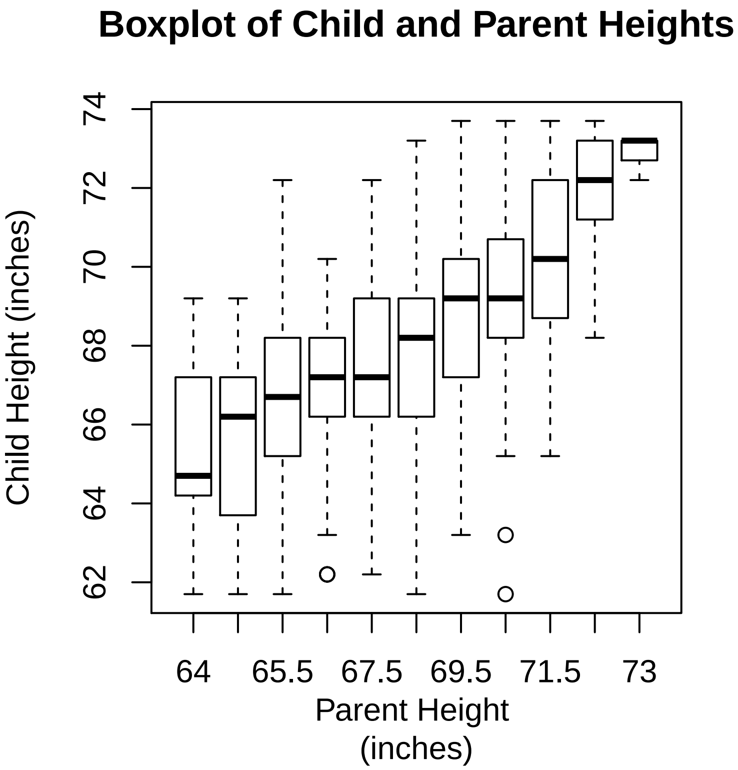
3. Plot the data appropriately. You will have to decide what is “appropriate.”

```
# Scatterplot
plot(Galton$parent, Galton$child, xlab = "Parent Height \n(inches)",
     ylab = "Child Height (inches)",
     main = "Scatterplot of Child and Parent Heights")
```



The scatterplot above shows a general upward trend but it is difficult to see clear patterns. This suggests a different plot would be better.

```
# boxplot
boxplot(Galton$child ~ Galton$parent, xlab = "Parent Height \n(inches)",
        ylab = "Child Height (inches)",
        main = "Boxplot of Child and Parent Heights")
```



The boxplots above confirms the general upward trend we saw from the scatterplot but it is much easier to see the pattern. We can also see the wide range of child heights from parents of the same height across many of the parent heights. In fact, there are a couple of outliers where the parents are tall but the children are short.

4. Write down mathematically a linear model to quantify Galton's observation that the children of tall parents tend to be taller than average yet less tall than their parents and, conversely, the children of short parents tend to be shorter than average yet taller than their parents. This is called the *regression effect.* Find the least squares coefficients of the linear model using R. You can use `{r, eval = FALSE}lm()` rather than writing out the model using matrices. Interpret the estimated coefficients in terms of the regression effect.

We want three different models. One for children of short parents, one for children of average height parents, and another for children of tall parents.

The model for children of short parents is

$$cheight_{i,short} = \beta_{0short} + \beta_{1short} * pheight_i + e_i$$

The model for children of average height parents is

$$cheight_{i,av} = \beta_{0av} + \beta_{1av} * pheight_i + e_i$$

The model for children of tall parents is

$$cheight_{i,tall} = \beta_{0tall} + \beta_{1tall} * pheight_i + e_i$$

We can write these together as

$$cheight_i = \beta_0 + \beta_1 * pheight_i + \beta_3 * z_{i,av} + \beta_4 * pheight_i * z_{i,av} + \beta_5 * z_{i,tall} + \beta_6 * pheight_i * z_{i,tall} + e_i$$

where $z_{i,av}$ is an indicator/dummy variable that takes the value 1 when the parent is considered to be of average height and 0 else. Similarly, $z_{i,tall}$ is an indicator/dummy variable that takes the value 1 when the parent is considered to be tall and 0 else. The $\beta_4 * pheight_i * z_{i,av}$ term is called an *interactive term* which says that I believe the slope corresponding to the parent height is different between short and average height parents. Finally, the last term, $\beta_4 * pheight_i * z_{i,tall}$ states that I believe the slope corresponding to the parent height is different between short and tall parents.

Before creating my linear model, I need to determine which parents are considered to be short, average, and tall.

```
summary(Galton)
```

```
##      parent          child      
##  Min.   :64.00   Min.   :61.70  
##  1st Qu.:67.50   1st Qu.:66.20  
##  Median :68.50   Median :68.20  
##  Mean   :68.31   Mean   :68.09  
##  3rd Qu.:69.50   3rd Qu.:70.20  
##  Max.   :73.00   Max.   :73.70
```

The summary above indicates that the short (bottom 25%) of parents are less than or equal to 67.5 inches, the average height parents are between 67.5 and 69.5 inches, and the tall parents (top 25%) of parents are taller than 69.5 inches.

Next, I create my dummy/indicator variables for average and tall parents.

```
Galton$av_ind <- as.numeric(67.5 <= Galton$parent & Galton$parent < 69.50)
Galton$tall_ind <- as.numeric(Galton$parent >= 69.50)
```

Finally, I can create my linear model for all three parent types.

```
attach(Galton)
child_height <- lm(child ~ parent + av_ind + av_ind*parent + tall_ind + tall_ind*parent, data = Galton)
summary(child_height)
```

```
## 
## Call:
## lm(formula = child ~ parent + av_ind + av_ind * parent + tall_ind + 
##     tall_ind * parent, data = Galton)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.8981 -1.3981  0.1804  1.6042  5.7519 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      19.0839    13.0668   1.460 0.144496    
## parent            0.7231     0.1989   3.635 0.000293 ***
## av_ind           19.4116    19.6343   0.989 0.323090    
## tall_ind        -15.1624    15.8953  -0.954 0.340389    
## parent:av_ind    -0.2921     0.2932  -0.996 0.319455    
## parent:tall_ind   0.2085     0.2370   0.880 0.379349    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.234 on 922 degrees of freedom
## Multiple R-squared:  0.2173, Adjusted R-squared:  0.213
## F-statistic: 51.18 on 5 and 922 DF,  p-value: < 2.2e-16
```

My linear model equation for all three parent types is the following:

$$cheight_i = 19.0839 + 0.7231 * pheight_i + 19.4116 * z_{i,av} + -0.2921 * pheight_i * z_{i,av} + -15.1624 * z_{i,tall} + 0.2085 * pheight_i * z_{i,tall}$$

Therefore, my linear models for the short parents, average parents, and tall parents separately are

(1) $cheight_{ishort} = 19.0839 + 0.7231 * pheight_{ishort}$
(2) $cheight_{iav} = 19.0839 - 0.2921 + 0.7231 * pheight_{iav} - 0.2921 * pheight_{iav} = 18.7918 + .431 * pheight_{iav}$
(3) $cheight_{itall} = 19.0839 - 15.1624 + 0.7231 * pheight_{itall} + 0.2085 * pheight_{itall} = 3.9215 + .9316 * pheight_{itall}$

We can see from these equations that children of short parents are shorter than average but taller than their parents, because they are on average, 72% of their parents height plus 19 inches. Additionally, children of tall parents are taller than average but shorter than their parents as indicated by the coefficient of parent height being less than 1 and the small intercept. In other words, children of short and tall parents "regress towards the mean" or the average height of adults.

Note: We can also construct these equations by subsetting the data into short, average, and tall parents and create separate linear models for each of these datasets. It will produce the same answer.

5. A regression effect for midterm and final scores would be as follows: students who do well in the midterm tend to do above average on the final, yet less well than in the midterm; students who do badly in the midterm tend to do below average in the final, yet better than they did in the midterm. Do you expect to see this regression effect in exam scores? Explain.

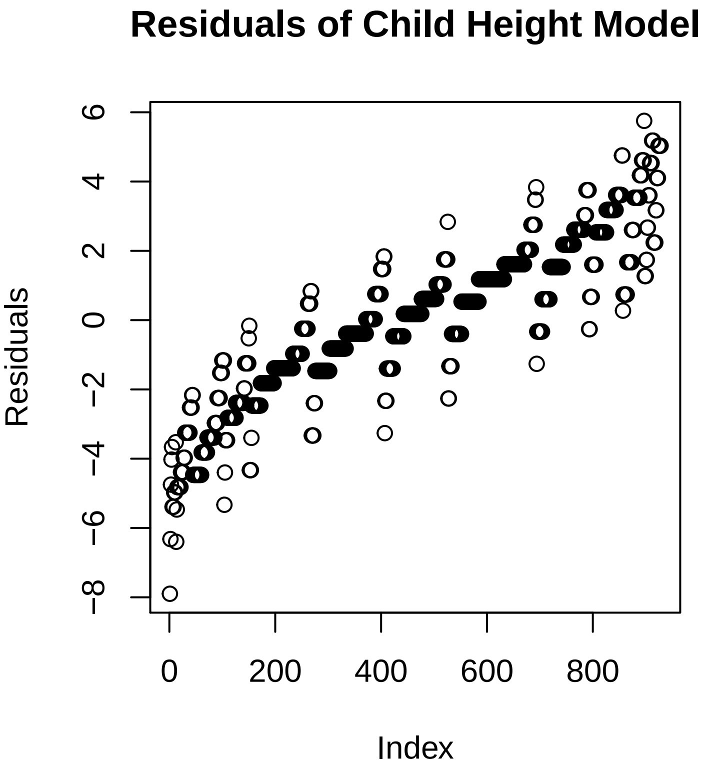
Yes, we would expect to see this regression effect in exam scores because of a similar logic to the height of children of short and tall parents in number 4. Students "regress" towards their mean understanding level so their will be fluctuations around that individual mean (e.g. doing really well on the midterm but comparatively less well on the final). Additionally, students who do really well on the midterm may have a higher individual mean than students who do poorly on the midterm so they will still perform better than average on the final.

6. Which of the following do you feel best describes the residual error in explaining the child's height by the parent's height? Select one of the following choices and explain briefly.

E. Between individual variability: other factors beyond genetics play a role in determining height.

I plot the residuals of the linear model. **Note: You did not have to do this. It is for illustration purposes only.**

```
plot(child_height$residuals, ylab = "Residuals",
     main = "Residuals of Child Height Model")
```



It is unlikely that small fluctuations from measurement error from two different scientists or how tired individuals are would lead to the relatively large residuals that we see (ignoring the trend). Additionally, if participants were not asked to remove their shoes, it would only add some small amount of height to every participant so it evens out. Finally, the round off error is relatively minimal so it also evens out across the board.

More generally, we know that things like environment, diet, and other factors affect childhood development so it would make sense that these other factors also contribute to height of a child.