Group Number F

Project 6

Height of Children (Project 6, Group Project)

Introduction

The height of parents and their children was recorded among 205 families, one child was randomly selected for families with more than one offspring. The dataset contains mid-parent height[1] (in inches), height of child (in inches) and the gender of the child (male or female).

It is of interest to determine if there exists a relationship between the height of parents and their children – can the height of children be predicted from the mid-parent height? If so, does the relationship depend on the gender of the child.

Exploratory analysis:

Summary statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
57.00	64.50	66.70	66.91	69.50	76.50

Table 1.1: Summary statistics from R

```
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  57.00   62.50   64.00   64.19   65.78   70.50

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  64.00   67.00   69.00   69.31   71.20   76.50
```

Table 1.2: Summary statistics by factor levels, *male* and *female*, from R

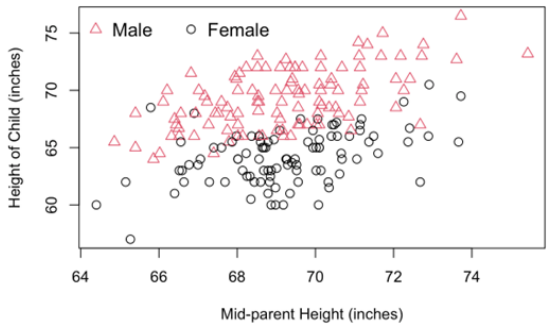

Figure 1.1: Scatterplot of *midparentHeight* versus *childHeight* with *gender* highlighted.

Figure 1.1 suggests a moderately-strong positive linear relationship between mid-parent height and the height of the child, the height of child appears to be systemically higher for male children than female children. There are some potential outliers present in figure 1.1, for example, notice the observation at (roughly) (65, 56).

The correlation coefficients which is for female and for male measure the linear association between variables; hence, these values can give a more precise understanding of the strength of this relationship.

Statistical analysis

The linear model is fitted in R with mid-parent height as the explanatory variable and height of child as the response variable. The following output is obtained in Table 2.1 below.

Table 2.1: Linear Model output of `Model0` from R

[1] Note that mid-parent height is defined as (father's height + 1.08 * mother's height)/2.

```
Call:
lm(formula = childHeight ~ midparentHeight, data = height)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4712 -2.3884 -0.2817  2.5540  6.6611

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      21.8862     8.5110   2.572   0.0108 *
midparentHeight   0.6505     0.1229   5.292 3.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.36 on 203 degrees of freedom
Multiple R-squared:  0.1212,    Adjusted R-squared:  0.1169
F-statistic: 28.01 on 1 and 203 DF,  p-value: 3.122e-07
```

The equation of the linear line from Table 2.1 is:

Child height = 21.8862 + 0.6505 *Mid-parent height* [equation 1]

[please analyse this Xi]

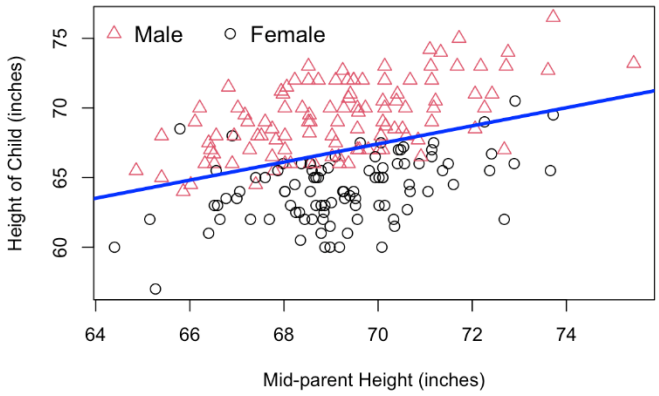

Figure 2.1: Scatterplot of *midparentHeight* versus *childHeight* with fitted line from `Model` superimposed.

Using the fitted model, the point estimate of child height for when the mid-parent height is 68 inches is 66 inches using figure 2.1 and 66.1202 inches using [equation 1]. When mid-parent height is 70 inches, the model estimates child height to be 67 inches using figure 2.1 and 67.4212 inches using [equation 1]. For completeness the prediction interval from `R` is also included below in Table 2.2.

Table 2.2: Prediction Interval output from `R`

	fit	lwr	upr
1	66.11823	59.47149	72.76497
2	67.41917	60.77630	74.06204

Now, it is of interest to determine if there exists a different relationship between the explanatory variable and the response variable for the level of the factor variable *gender*, there are two variables – male and female. Henceforward model selection will be carried out using the hypothesis test at the significance level 0.05.

First consider the fitted linear model `Model1`, which fits different regression lines for each group, that is the intercept and slope value for the linear regression line of the relationship between mid-parent height and child height depends on the child's gender.

Table 2.3: Analysis of Variance Table of `Model1` from `R`

```
Analysis of Variance Table

Response: childHeight
                       Df  Sum Sq Mean Sq  F value    Pr(>F)
midparentHeight         1  316.11  316.11  68.8460 1.529e-14 ***
gender                  1 1363.57 1363.57 296.9769 < 2.2e-16 ***
midparentHeight:gender  1    4.67    4.67   1.0173    0.3144
Residuals             201  922.89    4.59
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From Table 2.3, the interaction term between mid-parent height and gender is `midparentHeight:gender`. Using the hypotheses:

H_0: parallel lines (the parallel regression lines model is to be adopted)

H_1: separate lines (the separate regression lines model is to be adopted)

The interaction term has a p-value of 0.3144 which is greater than the significance level, so the interaction term is not statistically significant. That is, there is a failure to reject the null hypothesis as the sample does not provide sufficient evidence that two different slopes for the regression lines are required to describe the relationship between mid-parent height and child height.

Since the hypothesis test coincides with the confidence interval, the confidence interval for the difference in slopes should also contain zero, for completeness the confidence interval obtained in `R` is included below in Table 2.4.

Table 2.4: Confidence Interval output from `R`

```
midparentHeight midparentHeight
     -0.4718374       0.1519609
```

Now consider the fitted linear model `Model2`, which fits two parallel lines to the data. That is, the two groups will have the same intercept term but a different slope term for the linear regression line of the relationship between mid-parent height and child height.

Table 2.5: Linear Model output of `Model2` from `R`

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    17.51400    5.43465   3.223  0.00148 **
midparentHeight  0.67393    0.07841   8.595 2.28e-15 ***
gendermale      5.16931    0.29998  17.232  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.143 on 202 degrees of freedom
Multiple R-squared:  0.6442,    Adjusted R-squared:  0.6407
F-statistic: 182.9 on 2 and 202 DF,  p-value: < 2.2e-16
```

The equations of the parallel lines from Table 2.5 are:

$$\textit{Child height} = \begin{cases} 22.6833 + 0.6739 * \textit{mid-parent height} & \text{for } \textit{males} \\ 17.5140 + 0.67393 * \textit{mid-parent height} & \text{for } \textit{females}^2 \end{cases} \quad \text{[equation 2]}$$

When the mid-parents' height value is 0, the model expects the child's height to be 22.68331 for males and 17.51400 for females, this has no meaningful purpose and serves to adjust the heights of the regression lines. For each one-inch increase in the mid-parents' height, the child's height increases by 0.67 inches for both males and females.

[2] Note that: For *males*, *child height* = (17.5140 + 5.1693) + 0.6739 * *mid-parent height*.

The coefficient of determination, denoted R^2, is the proportion of the total variation in the response explained by the fitted model[3], which is calculated as 0.6442 with reference to Table 2.5. Hence 64.42% of the variation in child height is explained by the fitted model. Hence, this model provides a fine fit to the data.

Table 2.6: Analysis of Variance Table of `Model2` from `R`

```
Analysis of Variance Table

Response: childHeight
                  Df  Sum Sq Mean Sq F value    Pr(>F)
midparentHeight    1  316.11  316.11   68.84 1.502e-14 ***
gender             1 1363.57 1363.57  296.95 < 2.2e-16 ***
Residuals        202  927.56    4.59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the hypotheses:

H_1: single line (the single regression lines model is to be adopted)

H_0: parallel lines (the parallel regression lines model is to be adopted)

The p-value for the factor `gender` is very small[4] and is less than the significance level, so the factor `gender` is statistically significant. That is, the alternative hypothesis is accepted as the sample does provide sufficient evidence that two difference intercepts are required to describe the relationship between mid-parent height and child height. Again, for completeness the 95% confidence interval for the difference between the lines is obtained in `R` is included below in Table 2.7.

Table 2.7: Confidence Interval output comparing `Model` and `Model2` from `R`

```
as.factor(gender)male as.factor(gender)male
          -5.760712            -4.577910
```

It seems that male children are systematically taller than female children. With 95% confidence, the height of male children is likely more than female children by between 4.5779 inches and 5.7607 inches.
Therefore, the parallel regression lines model is adopted and examined for this point onwards.

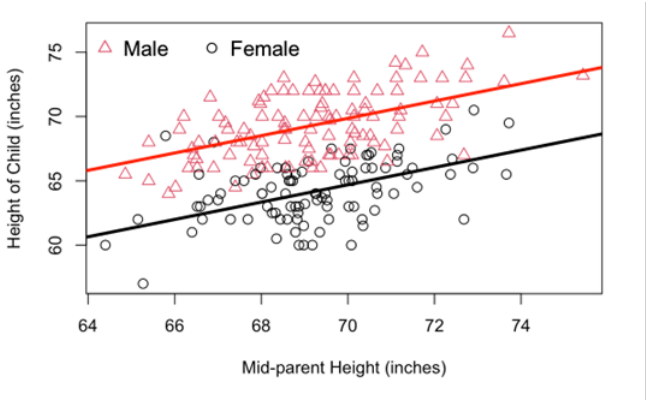

Figure 2.2: Scatterplot of *mid-parentHeight* versus *childHeight* with *gender* highlighted

with fitted lines from `Model2` superimposed.

Using the fitted model, the point estimate for a male and female child's height when the mid-parent height is 66 inches is expected to be 67.16 and 61.99 inches respectively using [equation 2]. The point estimate for when the mid-parent height is 74 inches is expected to be 73 inches for males 68 inches for females using Figure 2.2. For completeness the prediction interval from `R` is also included below in Table 2.8.

Table 2.8: Prediction Interval output from `R` (left-hand side for *male*, right-hand side for *female*)

[3] R^2 can be obtained by subtracting the residual sum of squares from the total sum of squares and dividing it by the total sum of squares

[4] Note: p-value $< 2.2e-16$.

	fit	lwr	upr
1	67.16253	62.88940	71.43566
2	72.55395	68.24464	76.86326

	fit	lwr	upr
1	61.99322	57.71630	66.27013
2	67.38464	63.07451	71.69476

Assumptions

Before using the fitted model to draw conclusions, model adequacy should be examined, this is checked using informal graphical based methods below. The assumptions will be checked for `Model2` only as it is the best fitting model for the data.

Each observations corresponds to a different child from a different family, so it is highly reasonable to assume that the observations are independent, it follows that the errors should also be independent. Judging the key features of the data, it is justifiable to assume that the values of mid-parent heights and gender are recorded without error[5].

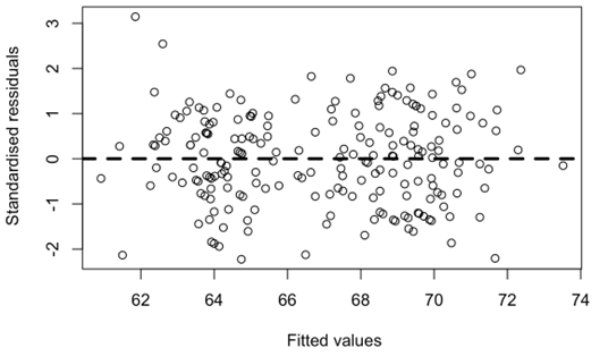

Figure 3.1 checks if the mean of the errors is zero and to detect changes in the error variance. Points are randomly scattered above and below the zero line. Vertical variation of points is constant across the range of fitted values, but at lower fitted values there are some points with a greater vertical variation. However, it is reasonable to assume that errors have a mean zero with constant variance.

Figure 3.1: Standardised residuals versus fitted values from Model2.

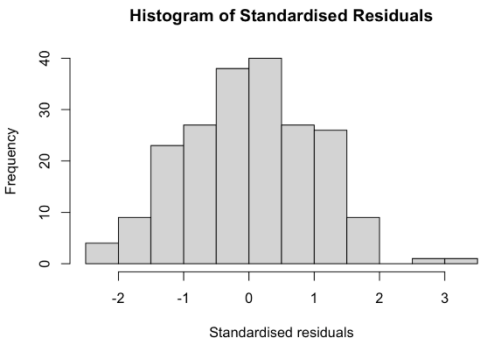

Figure 3.2: Histogram of the standardised residuals obtained from `Model2`.

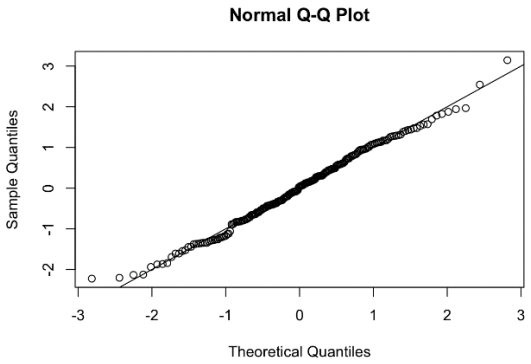

Figure 3.3: Normal probability (Q-Q) plot from `Model2`.

Figures 3.2 and 3.3 checks if the errors are normally distributed. By inspection of figure 3.2, the histogram has a clear bell-shape, moreover in figure 3.3, the points follow the straight diagonal line on the normal (Q-Q) plot with slight deviation from the line, y=x, at the ends, which suggests a thin-tailed (normal) distribution. So it is reasonable to assume that errors are normally distributed.

[5] Along with this assumption, there is an implicit assumption that height measuring methods were consistent and no errors where made when computing the mid-parent height value.

Conclusion

The main result from the formal analysis was that the height of children can be predicted from the mid-parents' height using a simple linear regression model , moreover `model2` highlights that this relationship changes depending on the levels of the factor variable, *gender*. That is, a collection of parallel regression lines provides the best fit to the data. The contrast in R^2 between `model0` and `model2`, highlights the importance of considering the levels of gender when predicting the height of a child. Thus supporting the initial impression of data gathered from figure 1.1. The model seems suitable as all assumptions – mean zero, constant variance, normality of errors, independence of errors, and values of the exploratory variables are recorded without error - appear to satisfactory.

As mentioned previously, there appears to be some potential outliers present in figure 1.1, so further research may investigate these, following this, these outliers may be deleted or corrected legitimately without changing the results. Furthermore, it may be of interest to consider if the socioeconomic background of a child will affect their height. That is, those from more affluent backgrounds may be shorter or taller than their peers from poorer backgrounds for every mid-parent height.

[]