HW4 - Akshaj Kammari

Due: 03/04/2024

```
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(moderndive)
library(rockchalk)
```

```
##
## Attaching package: 'rockchalk'
```

```
## The following object is masked from 'package:dplyr':
##
##     summarize
```

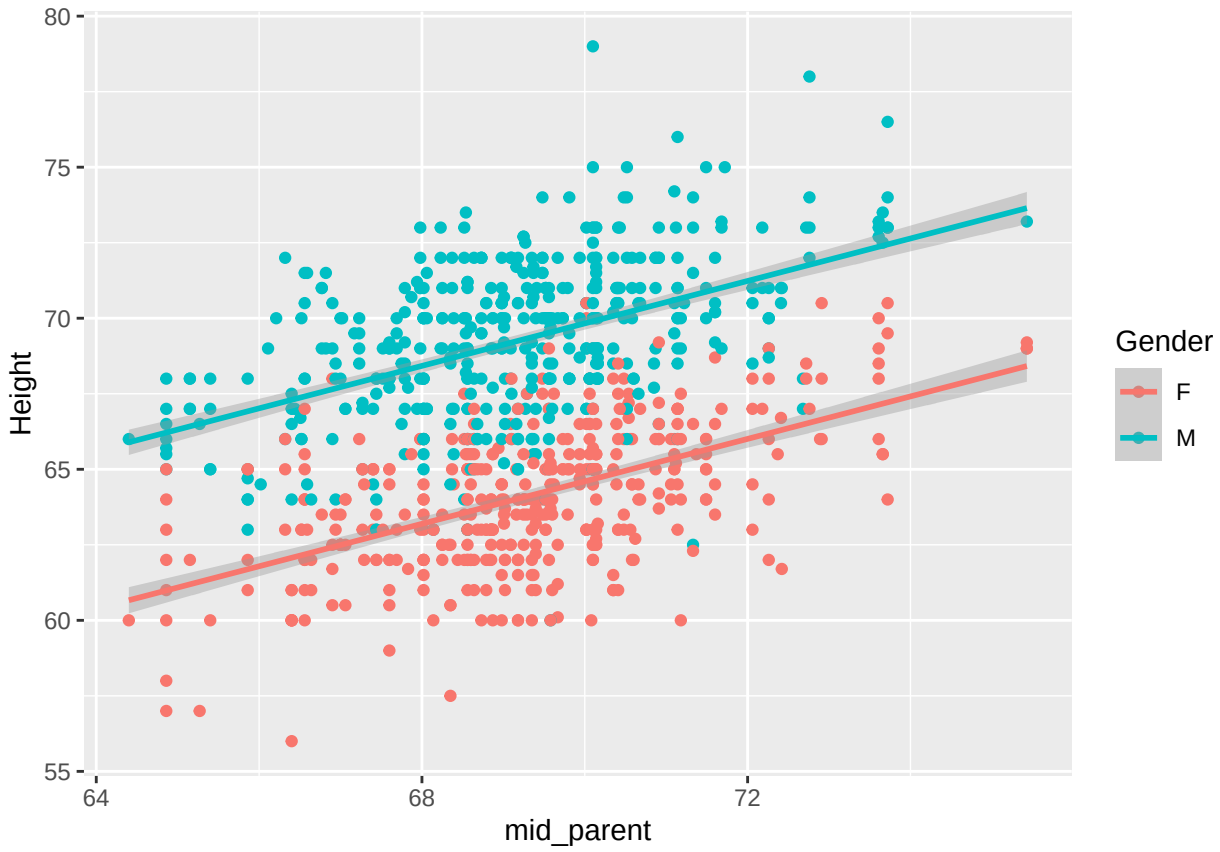
##Part I: Regression of child's height on mid-height of parent and gender. In this part, the dependent variable is the raw height of the child, and the child's gender is one of the independent variables, the other independent one is the mid-height of parent (Father height + 1.08 * Mother Height)/2

```
galton <- read.csv("galton_height.csv") %>% mutate(mid_parent = (Father + Mother*1.08)/2)
```

#Q1

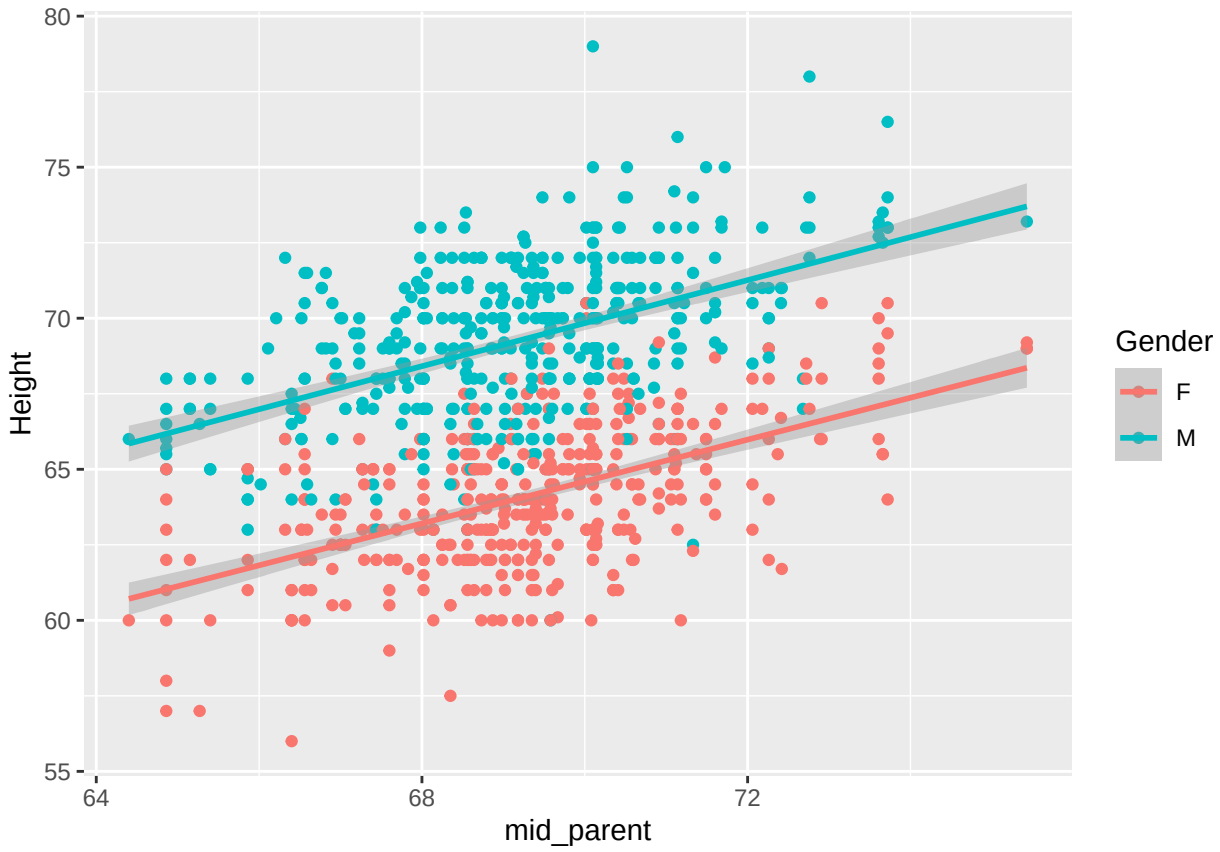
```
#same slope
ggplot(data = galton, aes(x = mid_parent, y = Height, color = Gender)) + geom_point() + geom_parallel_sl
```

```
## Warning: `geom_parallel_slopes()` doesn't need a `method` argument ("lm" is
## used).
```




```
#different slope
ggplot(data = galton, aes(x = mid_parent, y = Height, color = Gender)) + geom_point() + geom_smooth(metl
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



#Q2

```
model1 <- lm(data = galton, Height ~ mid_parent + Gender)
get_regression_table(model1)
```

```
## # A tibble: 3 x 7
##   term        estimate std_error statistic p_value lower_ci upper_ci
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    15.4       2.76       5.59       0   10.0     20.8
## 2 mid_parent    0.703     0.04      17.7        0    0.625    0.781
## 3 Gender: M     5.23      0.144     36.2        0    4.94     5.51
```

###Both independant variables are significant because the p-values are below 0.05.

#Q3 ###The expected increase in the child's height if the mid-parents height increases is 0.703. Gender does not matter because the slopes are parallel.

#Q4

```
model2 <- lm(data = galton, Height ~ mid_parent * Gender)
get_regression_table(model2)
```

```
## # A tibble: 4 x 7
##   term                estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                  <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept             16.1       3.92      4.1     0        8.37    23.7
## 2 mid_parent             0.693     0.056    12.3     0        0.582    0.804
## 3 Gender: M              3.93      5.50      0.714   0.475   -6.87    14.7
## 4 mid_parent:GenderM     0.019     0.08      0.235   0.814   -0.137    0.175
```

#Q5 ###The increase is 0.693 for the daughter and 0.712 for the son. The slopes are similar, but not the same due to the fact that the regressions are done separately, unlike the parallel slope.

#Q6

```
#son
male = galton %>%filter(Gender == "M")
male_model = lm(data = male, Height ~ mid_parent)
get_regression_table(male_model)
```

```
## # A tibble: 2 x 7
##   term       estimate std_error statistic p_value lower_ci upper_ci
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    20.0        4.12      4.85       0   11.9     28.1
## 2 mid_parent    0.712      0.06     12.0        0    0.595    0.829
```

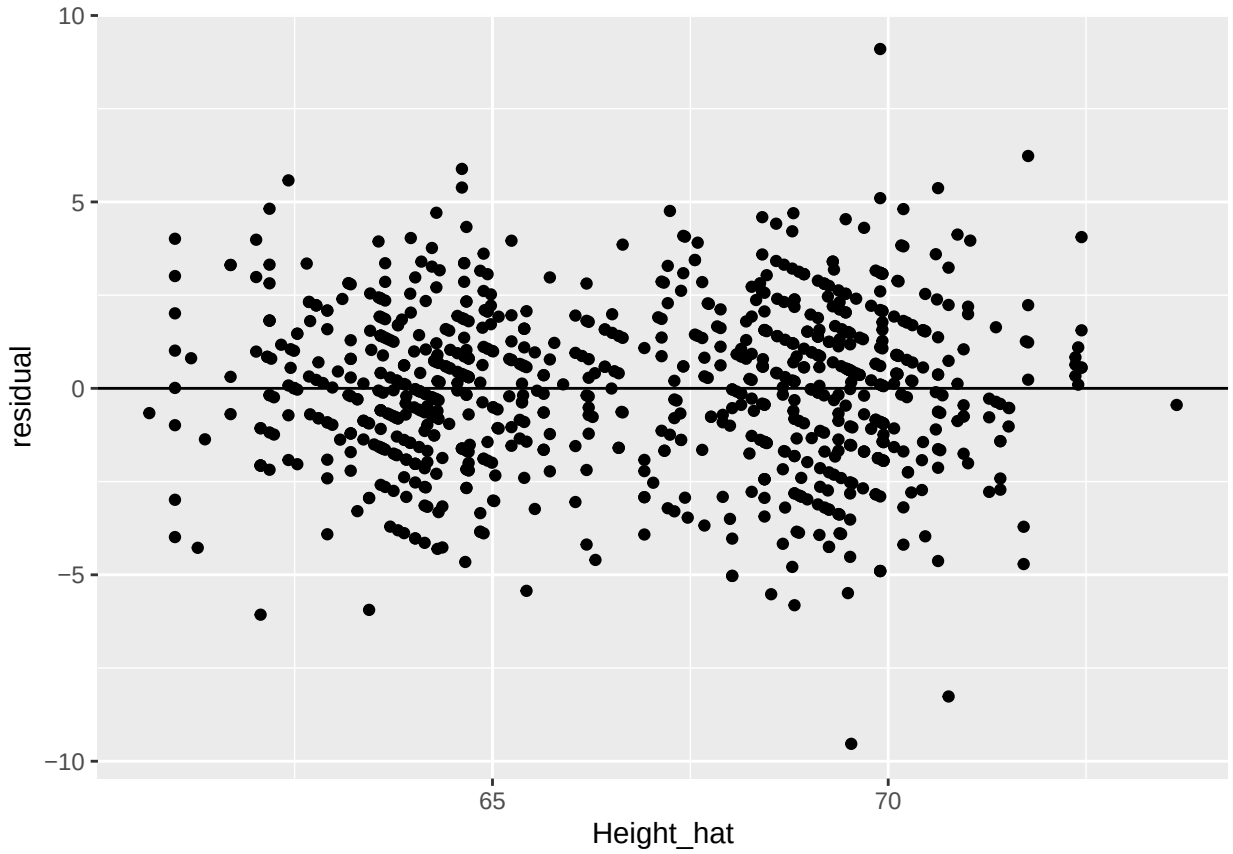
```
#daughter
female = galton%>%filter(Gender == "F")
female_model = lm(data = female, Height ~ mid_parent)
get_regression_table(female_model)
```

```
## # A tibble: 2 x 7
##   term       estimate std_error statistic p_value lower_ci upper_ci
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    16.1        3.62      4.43       0    8.94    23.2
## 2 mid_parent    0.693      0.052    13.3        0    0.591    0.796
```

#Q7 ###I would choose the Q2 model because it is easier to get the result, as I only have to run the regression once versus twice with the Q4 model.

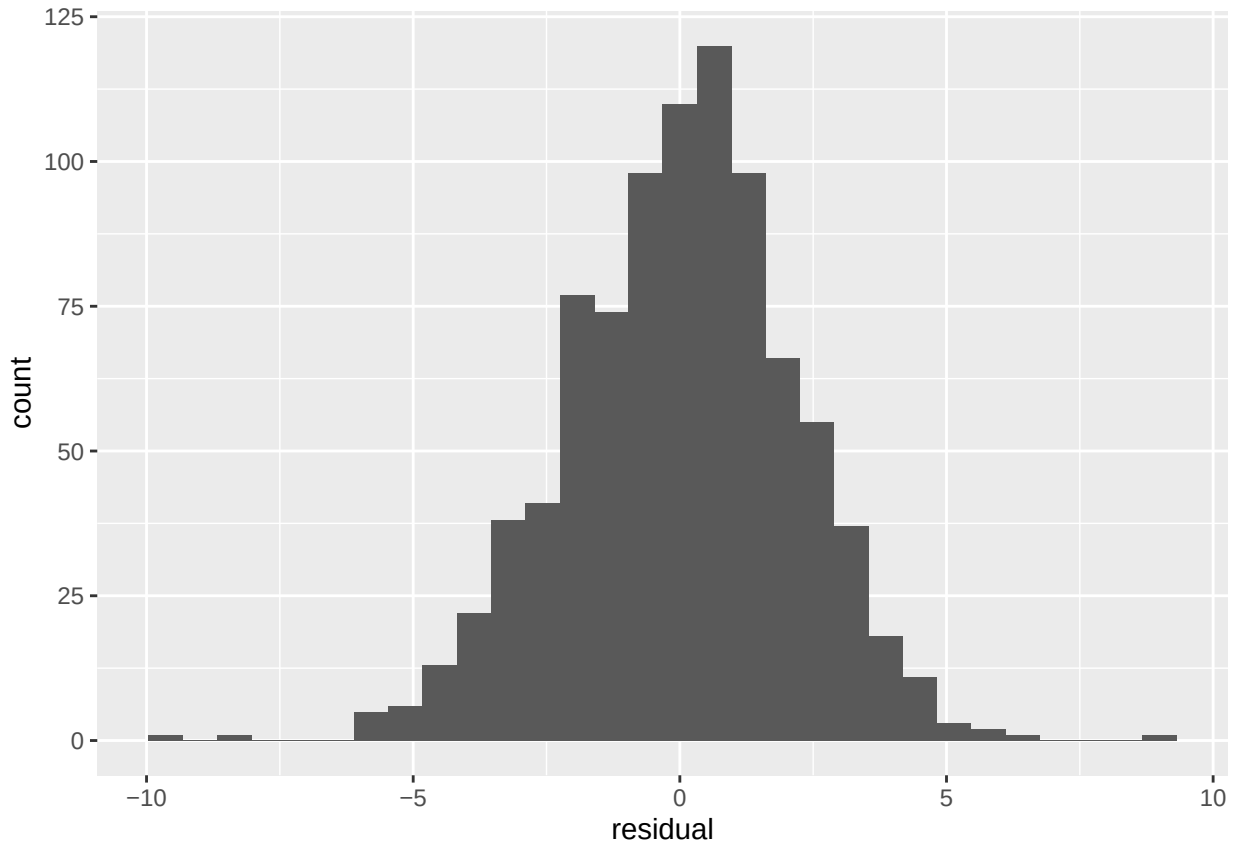
#Q8

```
#scatterplot
points = get_regression_points(model1)
ggplot(data = points, mapping = aes(x = Height_hat, y = residual)) + geom_point() + geom_hline(yintercep
```

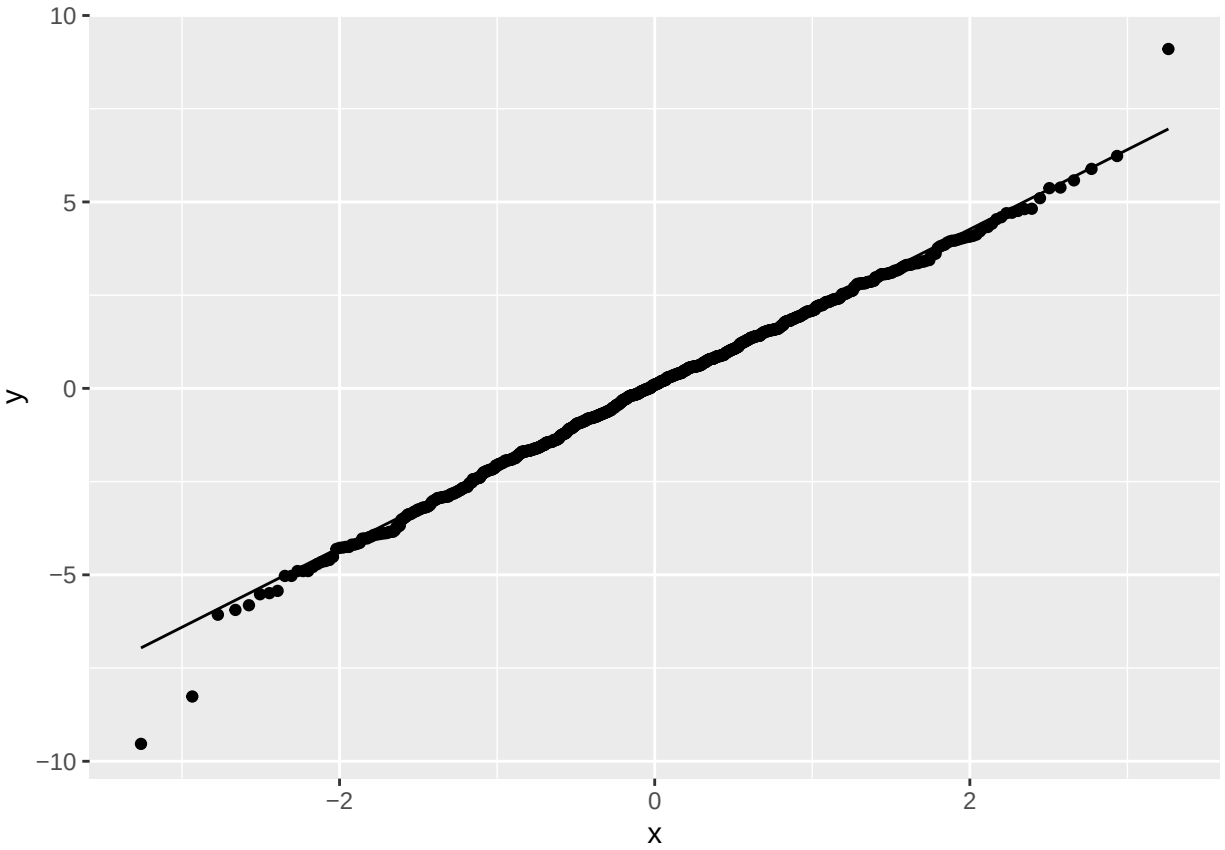



```
#histogram
ggplot(data = points, mapping = aes(x = residual)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#qqplot
ggplot(data = points, aes(sample = residual)) + stat_qq()+ stat_qq_line()
```



###R^2 is 0.637 and follows the LINE property.

##Part II: Regression of child's height on father and mother's height. In this part, the dependent variable is the height of the child (adjusted for gender), and the independent variables are mother's and father's individual raw height.

#Q9

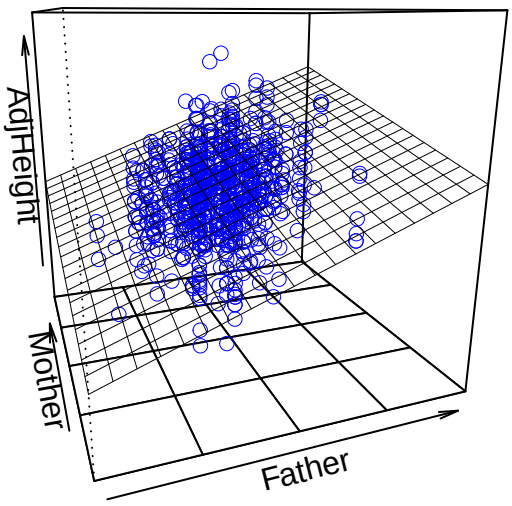
```
galton <- read.csv("galton_height.csv") %>% mutate (AdjHeight = ifelse(Gender == "M",Height, Height*1.08
model3 = lm(data = galton, AdjHeight ~ Father + Mother)
get_regression_table(model3)
```

```
## # A tibble: 3 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   18.7       2.83       6.60       0   13.1     24.3
## 2 Father       0.423     0.03      14.0        0    0.364    0.482
## 3 Mother       0.332     0.032     10.3        0    0.268    0.395
```

#Q10 ###The father's height has no effect on the increment, as the increase is always 0.378.

#Q11

```
plotPlane(model3, plotx1 = "Father", plotx2 = "Mother")
```



#Q12

```
model4 = lm(data = galton, AdjHeight ~ Father * Mother)
get_regression_table(model4)
```

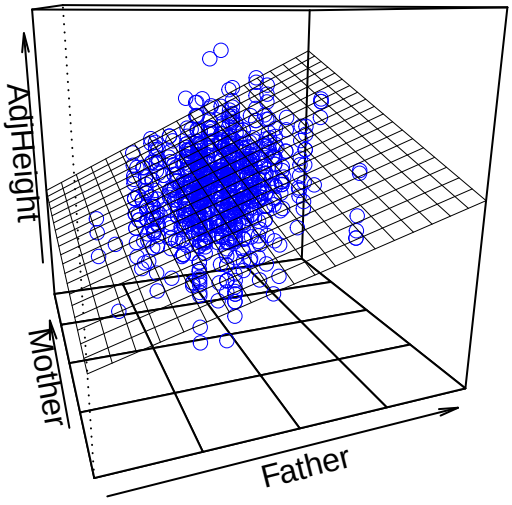
```
## # A tibble: 4 x 7
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       76.4      54.7       1.40    0.163  -31.0    184.
## 2 Father          -0.409     0.788    -0.519   0.604   -1.96     1.14
## 3 Mother          -0.567     0.851    -0.666   0.506   -2.24     1.10
## 4 Father:Mother    0.013     0.012     1.06    0.291   -0.011    0.037
```

###The model is not significant because the p-values are above 0.05.

#Q13 ###68 inches ((-0.567)+(0.013)(68)) = 0.317 ###70 inches ((-0.567)+(0.013)(70)) = 0.343 ###72 inches ((-0.567)+(0.013)(72))= 0.369 ###The difference is 0.026

#Q14

```
plotPlane(model4, plotx1 = "Father", plotx2 = "Mother")
```



#Q15 ###The model from Q12 is insignificant, hence I would choose the model from Q9.

##Part III: Regression of child's height on father and mother's height and gender. In this part, the dependent

variable is the raw height of the child, and the independent. variables are mother's and father's individual raw height, and the child's gender

#Q16

```
model5 <- lm(data = galton, Height ~ Father + Mother + Gender)
get_regression_table(model5)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   15.3       2.75       5.59       0    9.95    20.7
## 2 Father       0.406     0.029     13.9        0    0.349    0.463
## 3 Mother       0.321     0.031     10.3        0    0.26     0.383
## 4 Gender: M    5.23      0.144     36.3        0    4.94     5.51
```

#Q17

```
galton <- galton %>% mutate(mid_parent=(Father+Mother*1.08)/2)
model6 <- lm(data = galton, Height ~ mid_parent + Gender)
get_regression_table(model6)
```

```
## # A tibble: 3 x 7
##   term       estimate std_error statistic p_value lower_ci upper_ci
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    15.4       2.76       5.59       0   10.0     20.8
## 2 mid_parent    0.703     0.04      17.7        0    0.625    0.781
## 3 Gender: M     5.23      0.144     36.2        0    4.94     5.51
```

```
#Comparing coefficients
coefficients_Q16 <- model5$coefficients
coefficients_Q17 <- model6$coefficients

coefficients_Q16
```

```
## (Intercept)      Father      Mother     GenderM
##  15.3447600   0.4059780   0.3214951   5.2259513
```

```
coefficients_Q17
```

```
## (Intercept)  mid_parent     GenderM
##  15.4030108   0.7028176   5.2281684
```

###The coefficients for all 3 are similar. This is because both models are estimating the same linear relationship between the child's height and the independent variables, just using different representations (individual heights in Q16 and mid-parent height in Q17).