HW4

2024-03-02

galton_height <- read.csv(“galton_height.csv”, header = TRUE)

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

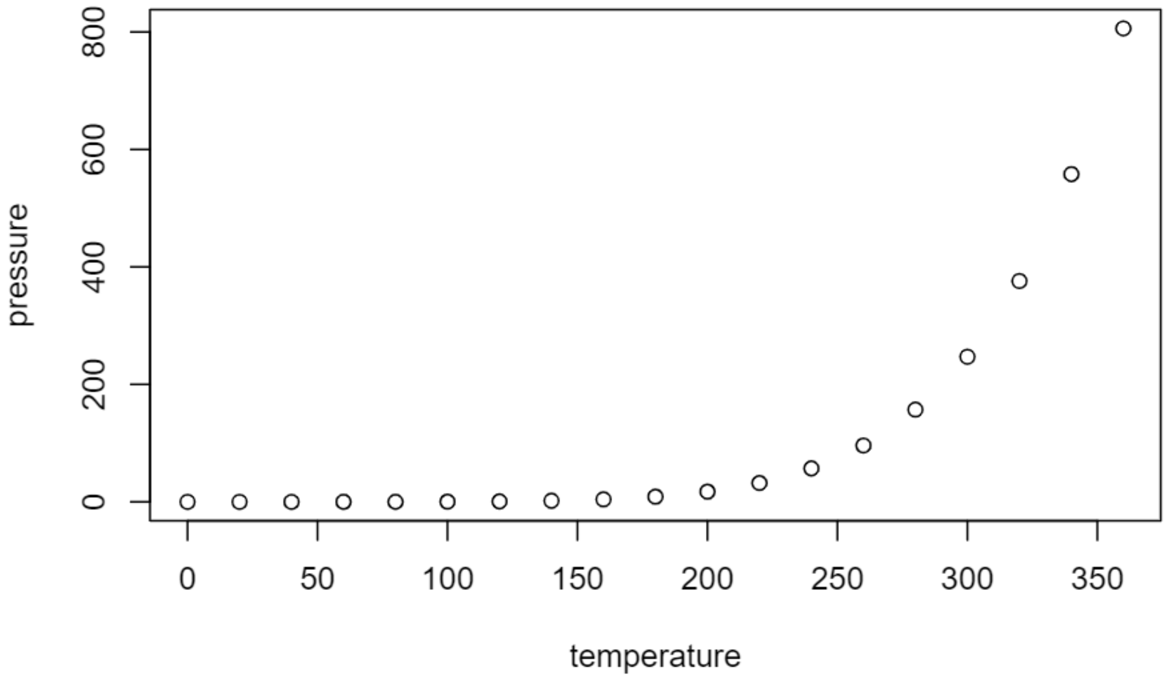
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

##Revisit the regression Galton's data with parents and kids' height

##Part I: Regression of child's height on mid-height of parent and gender

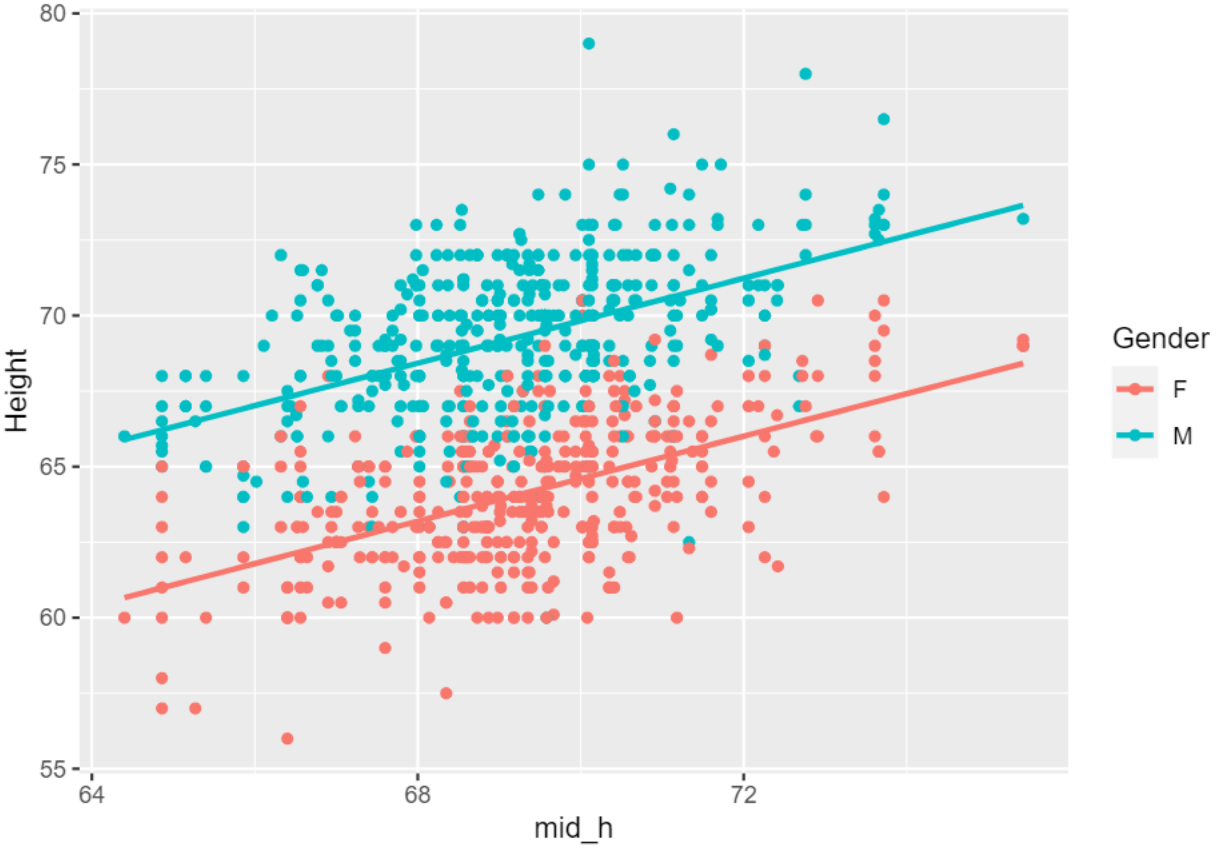
```
galton_height <- read.csv("galton_height.csv", header = TRUE)

mother_h2 <- galton_height$Mother * 1.08
galton_height$mid_h <-(galton_height$Father + mother_h2 ) / 2
```

###Q1:Have a scatterplot of child's height vs. mid height of parents, with different colors for male and female of the child. Add smooth lines for male and female separately in the scatterplot. First, plot the smooth lines with same slope. Secondly with different slopes. By looking at the plots, does it look like that the slopes are different even if you didn't force them to be parallel? ###Q1 Answer: When observing the two plots, it can be seen that the slopes are same when they are not forced to be parallel as well.

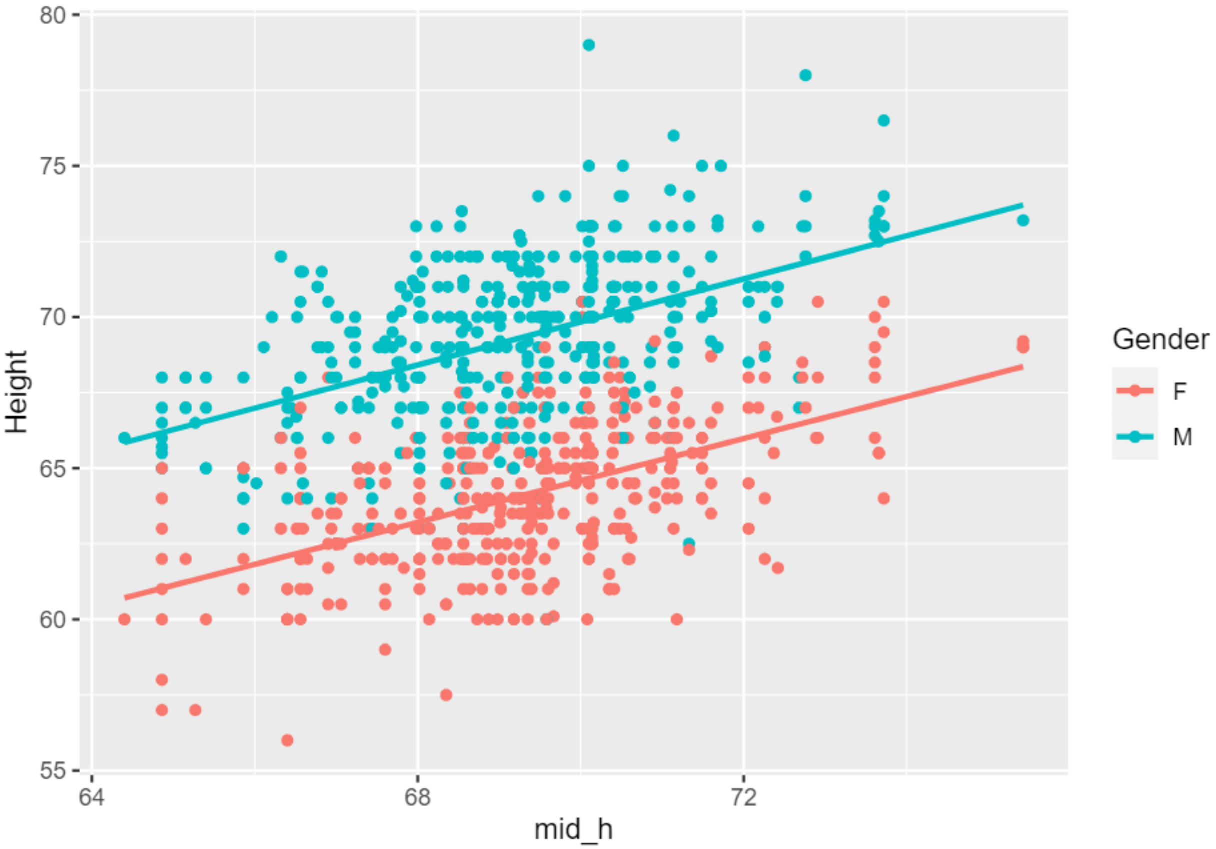
```
ggplot(data = galton_height,aes(x=mid_h, y=Height, color = Gender)) + geom_point()+ geom_parallel_slope
```

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



```
ggplot(data = galton_height,aes(x=mid_h, y=Height, color = Gender)) + geom_point()+  geom_smooth(method
```

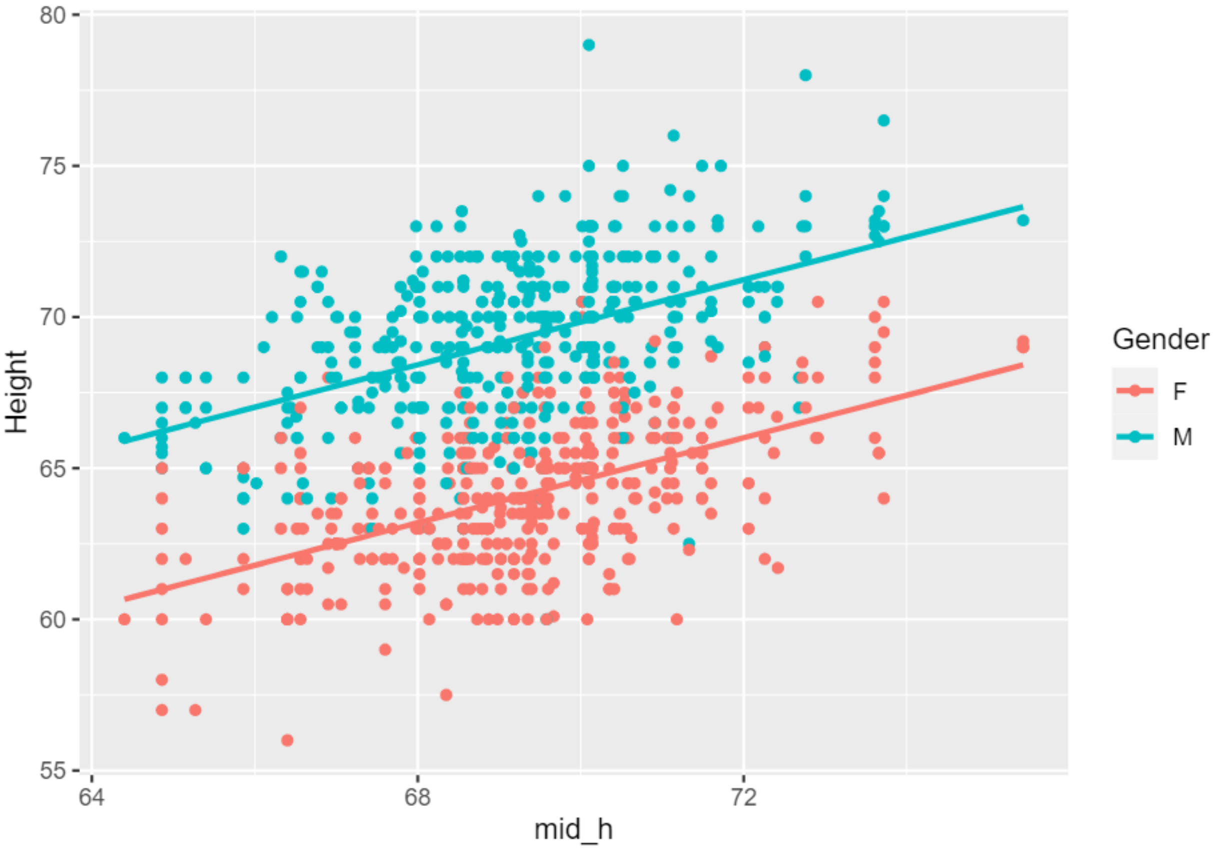
```
## `geom_smooth()` using formula = 'y ~ x'
```



###Q2:Run a regression with parallel slopes. Are both independent variables significant? ###Q2 Answer: Yes, both the independent variables are significant as they consist of a p-value of 0.

```
ggplot(data = galton_height,aes(x=mid_h, y=Height, color = Gender)) + geom_point()+ geom_parallel_slope
```

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



```
model1 <- lm(data=galton_height, Height ~ mid_h + 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_h        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
```

```
model1 <- lm(data=galton_height, Height ~ mid_h)
get_regression_table(model1)
```

```
## # 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   22.4       4.31       5.19       0   13.9     30.8
## 2 mid_h        0.641     0.062     10.3        0    0.519    0.763
```

```
model1 <- lm(data=galton_height, Height ~ Gender)
get_regression_table(model1)
```

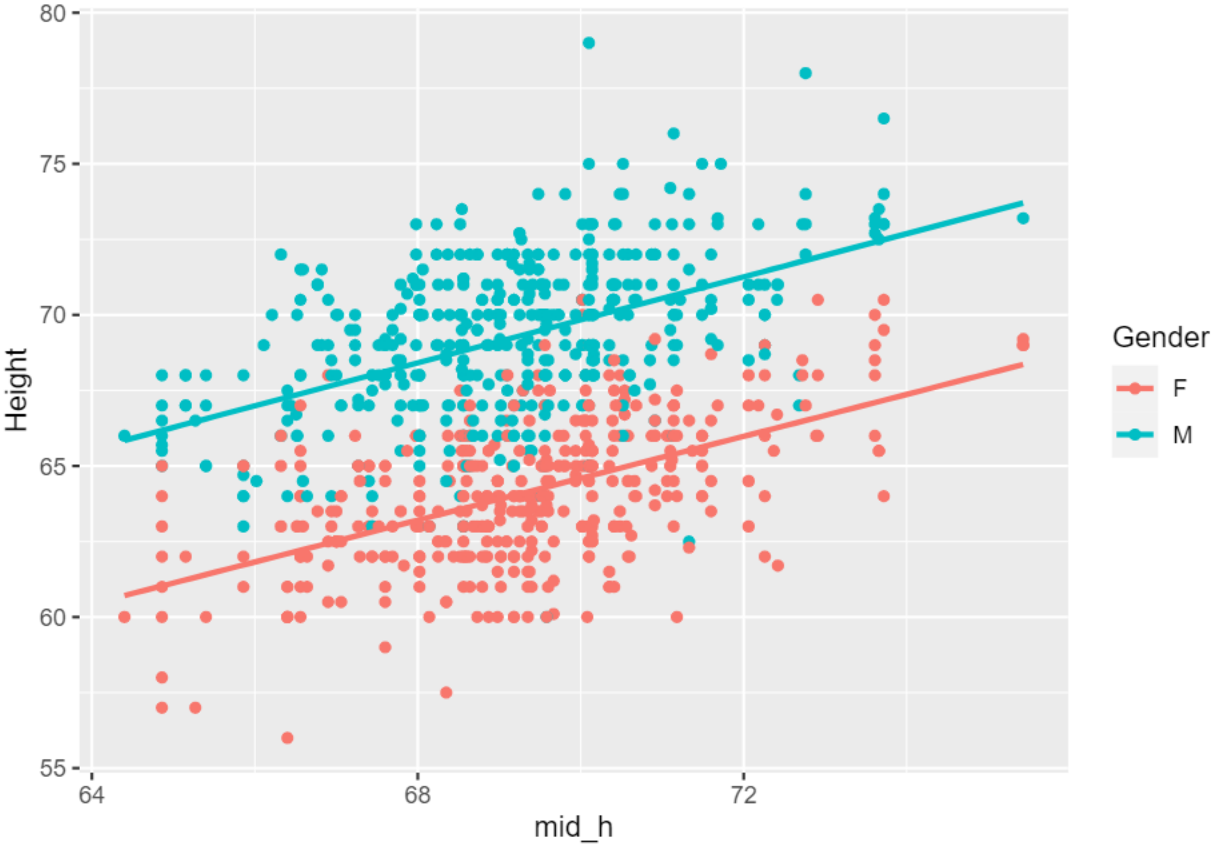
```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   64.1       0.121    532.         0   63.9     64.3
## 2 Gender: M    5.12      0.168     30.5        0    4.79     5.45
```

###Q3:In the linear model in Q2, what is the expected increase in child's height if the mid-height of parents increases by 1 inch? Does this depend on the gender of the child? ###Q3 Answer: The expected increase in child's height if the mid-height of the parents increases by 1 inch is 5.228 inches. No, this does not depend on the child's gender.

###Q4: Run a regression with interaction term. Is the interaction term significant? ###Q4 Answer: The interaction term is not significant as it has a p-value that is greater that 0.05. Furthermore, the slopes of the two genders do not have a significant difference (difference of 0.019).

```
ggplot(data = galton_height,aes(x=mid_h, y=Height, color = Gender)) + geom_point()+ geom_smooth(method
```

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



```
male <- galton_height %>% filter(Gender=="M")
model2 <- lm(data=male, Height ~ mid_h)
get_regression_table(model2)
```

```
## # 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_h        0.712      0.06     12.0        0    0.595    0.829
```

```
female <- galton_height %>% filter(Gender=="F")
model2 <- lm(data=female, Height ~ mid_h)
get_regression_table(model2)
```

```
## # 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_h        0.693     0.052     13.3        0    0.591    0.796
```

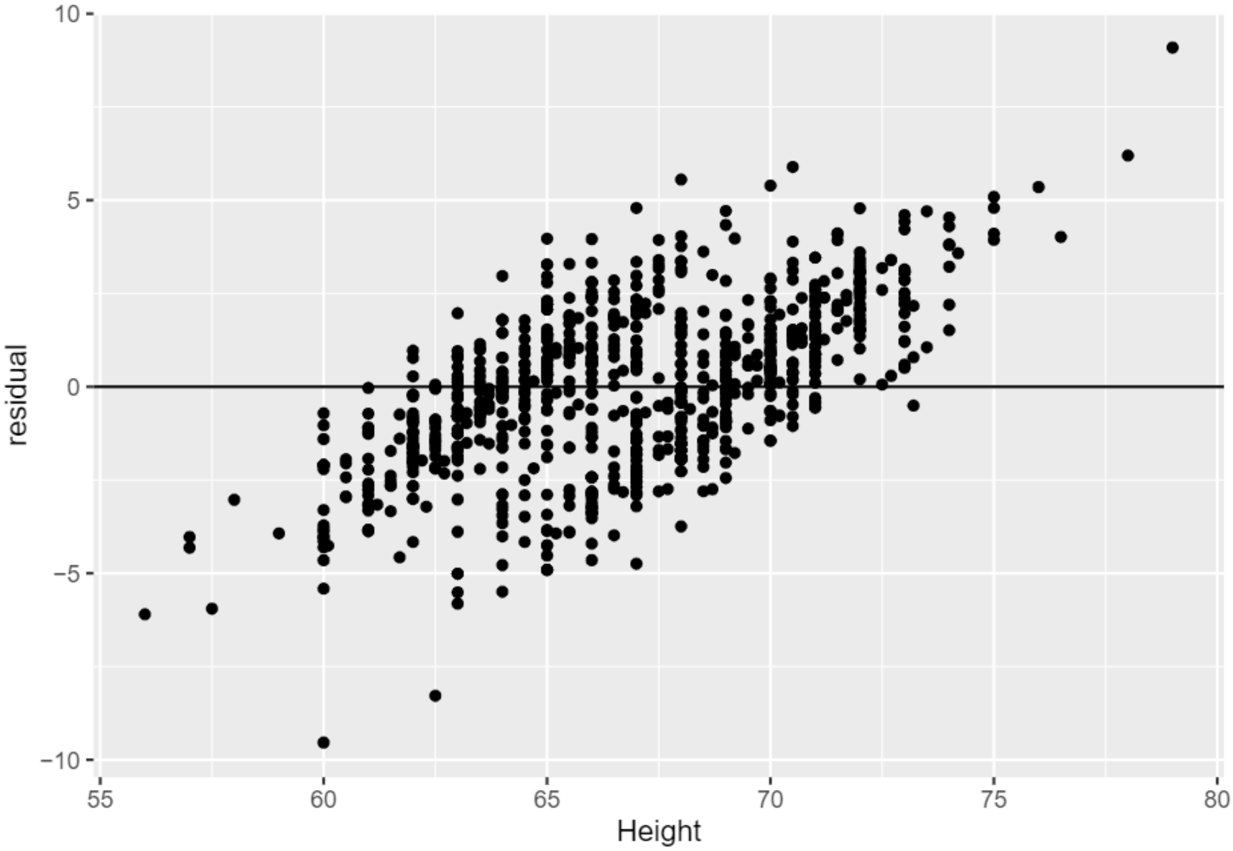
```r
model2 <- lm(data=galton_height, Height ~ mid_h * Gender)
get_regression_table(model2)
```

```
## # 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_h            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_h:GenderM    0.019     0.08      0.235   0.814   -0.137    0.175
```

```r
points <- get_regression_points(model2)
points
```

```
## # A tibble: 898 x 6
##        ID Height mid_h Gender Height_hat residual
##     <int>  <dbl> <dbl> <chr>       <dbl>    <dbl>
##  1     1   73.2  75.4 M             73.7   -0.503
##  2     2   69.2  75.4 F             68.4    0.841
##  3     3   69    75.4 F             68.4    0.641
##  4     4   69    75.4 F             68.4    0.641
##  5     5   73.5  73.7 M             72.4    1.06
##  6     6   72.5  73.7 M             72.4    0.057
##  7     7   65.5  73.7 F             67.1   -1.63
##  8     8   65.5  73.7 F             67.1   -1.63
##  9     9   71    72.1 M             71.3   -0.303
## 10    10   68    72.1 F             66.0    1.98
## # i 888 more rows
```

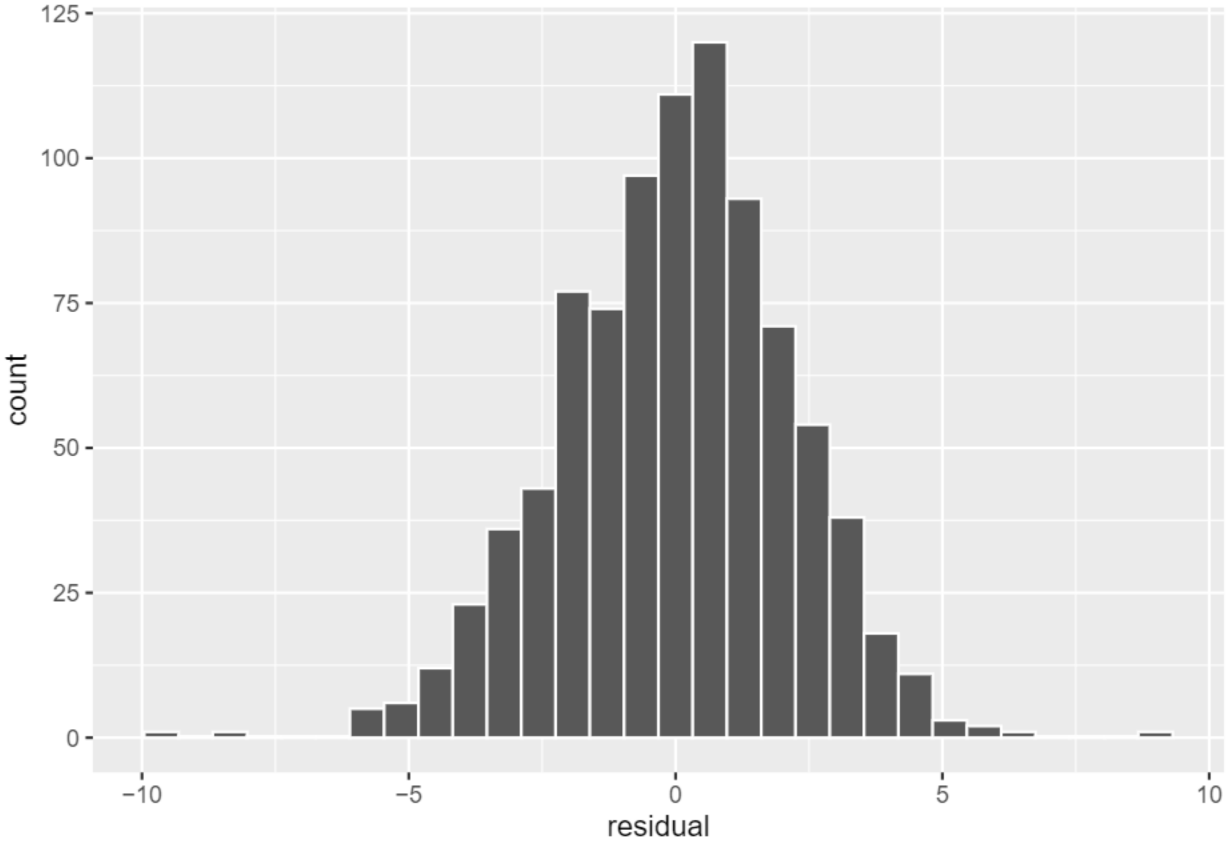
```r
ggplot(data=points, aes(x=Height, y=residual)) + geom_point() + geom_hline(yintercept=0)
```



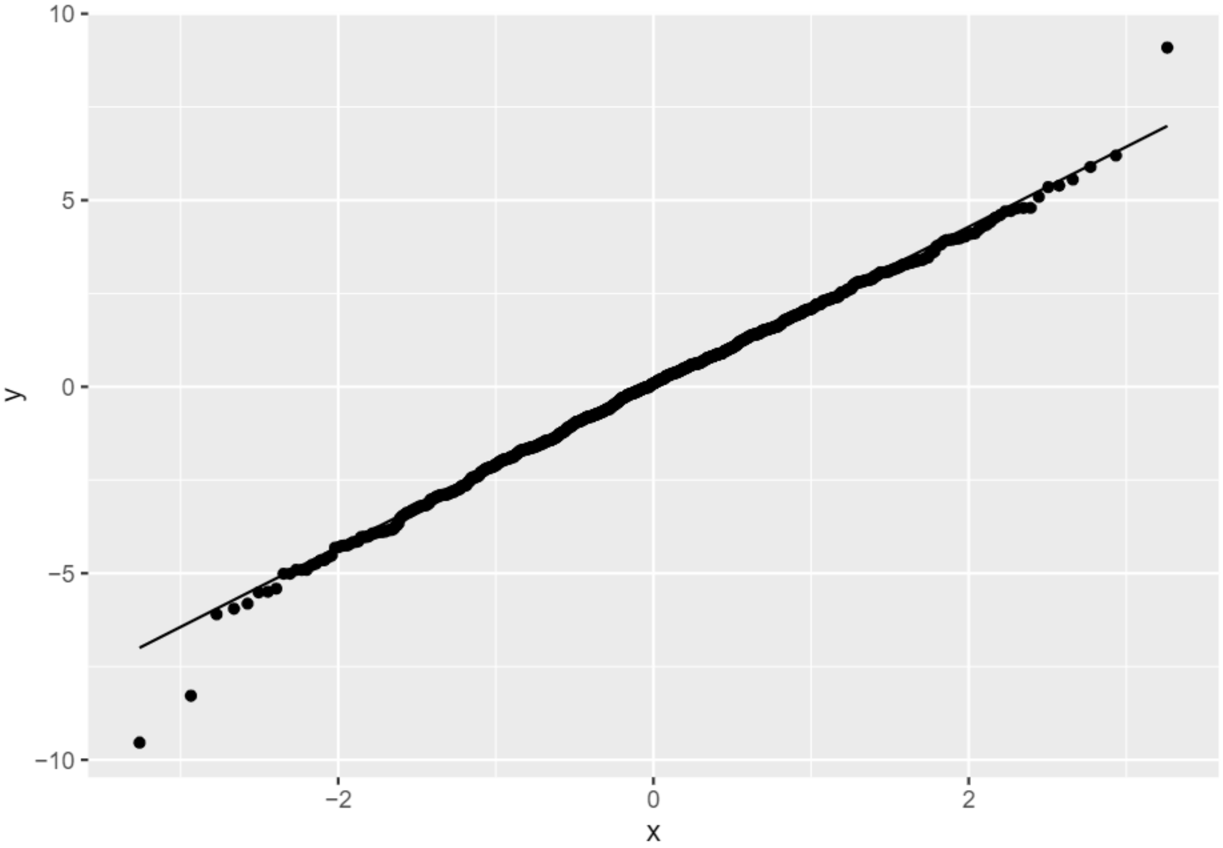
```
ggplot(data=points, aes(x=residual)) + geom_histogram(bindwidth=1, color="white")
```

```
## Warning in geom_histogram(bindwidth = 1, color = "white"): Ignoring unknown
## parameters: `bindwidth`

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



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



```
get_regression_summaries(model2)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.637         0.636  4.65  2.16  2.16      523.       0     3   898
```

###Q5: In the linear model in Q4, what is the expected increase in adult daughter's height if the mid-height of parent increases by 1 inch? What is the expected increase in adult son's height if the mid-height of parent increases by 1 inch? Is there significant difference between these two values? Why? ###Q5 Answer: The expected increase in adult daughter's height it the mid-height of parent increases by 1 inch is 0.693. The expected increase in adult son's height if the mid-height of parent increases by 1 inch is 0.712. These is not significant difference between these two values because they have the very similar slopes which can be observed in the scattersplot.

###Q6: Run two regressions for child's height on mid-height of parents, respectively for adult daughters and sons. Compare the two slopes from the two regressions to the slopes from the interaction model. ###Q6 Answer: The slopes are similar to the two slopes from the two regressions from the interaction model.

```
male <- galton_height %>% filter(Gender=="M")
model2 <- lm(data=male, Height ~ mid_h)
get_regression_table(model2)
```

```
## # 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_h        0.712      0.06     12.0        0    0.595    0.829
```

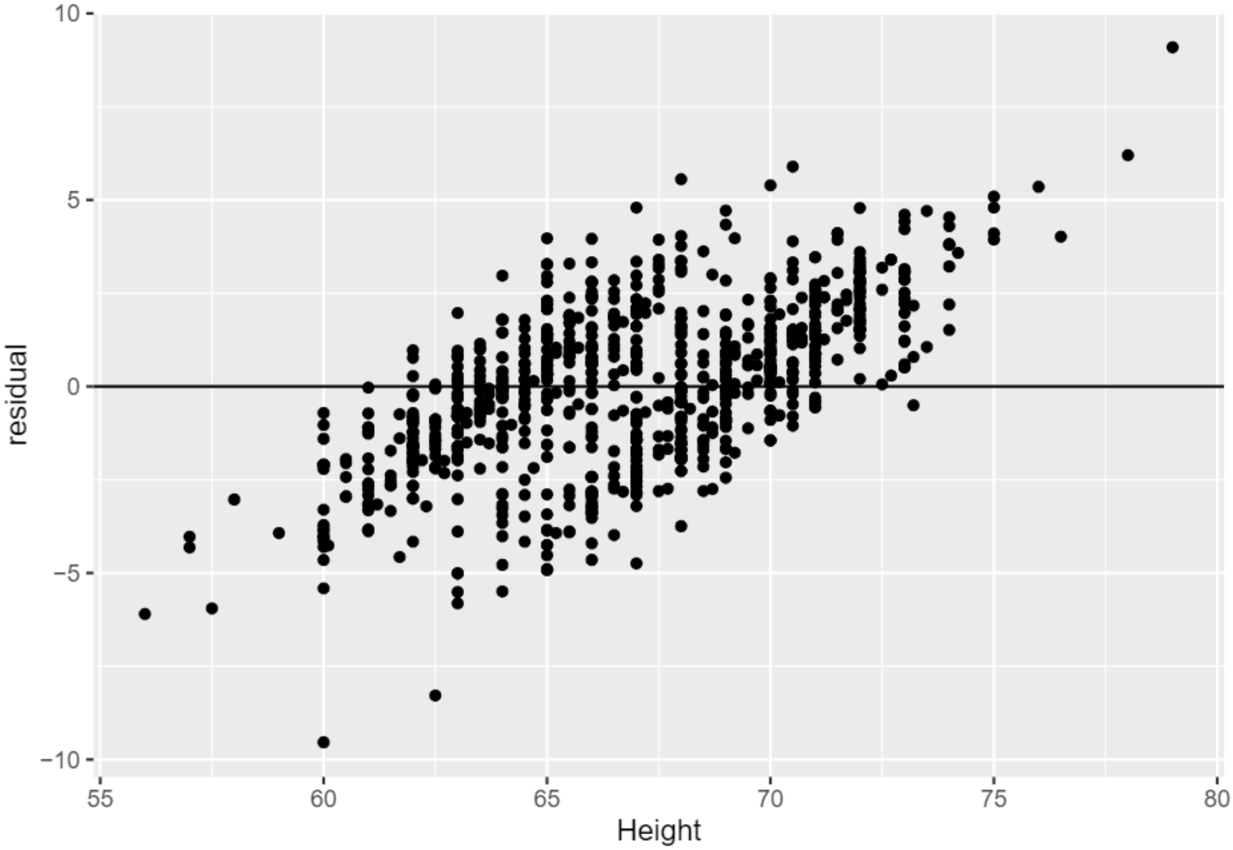
```
female <- galton_height %>% filter(Gender=="F")
model2 <- lm(data=female, Height ~ mid_h)
get_regression_table(model2)
```

```
## # 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_h        0.693     0.052     13.3        0    0.591    0.796
```

###Q7: If you are going to choose the final model between Q2 and Q4, which one will you choose? Why? ###Q7 Answer: I will choose the model from Q2 as it is a significant model with a p-value of 0. The interaction model in Q4 is not significant due to its large p-values and the small difference between the two regression slopes.

###Q8: Have the scatterplot, histogram and QQ plot of the residual from the model you choose in Q7, and comment on the assumptions of the multiple regression. What is the R^2 from the model? ###Q8 Answer: * Mean 0/linearity: the assumption is satisfied. This is because there is linear relationship between y and x, there are both positive and negative residuals. Moreover, the sum of the residiials will be 0. * Constant variance: This is not satisfied becuase there is a cone shape in the scatter plot. Moreover, the variance is larger when the child's height is between 63 and 70 than others, the variances at smaller and larger heights are much smaller. * Normal distribution: This is statisfied as the plot has a bell-shape. However, there seems to be three outliers. * Independency: This satisfied, there is not serious violation of this assumption. The R^2 value is 0.637.

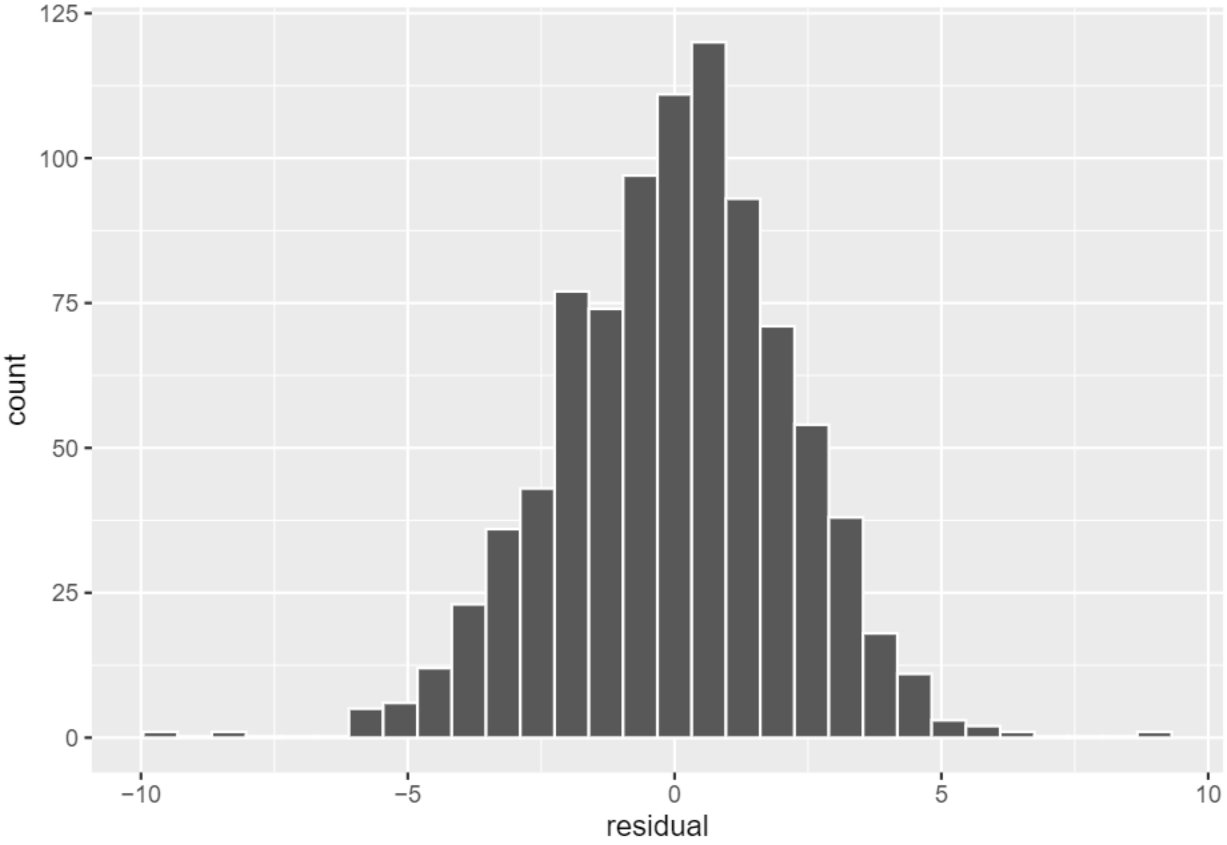
```
ggplot(data=points, aes(x=Height, y=residual)) + geom_point() + geom_hline(yintercept=0)
```



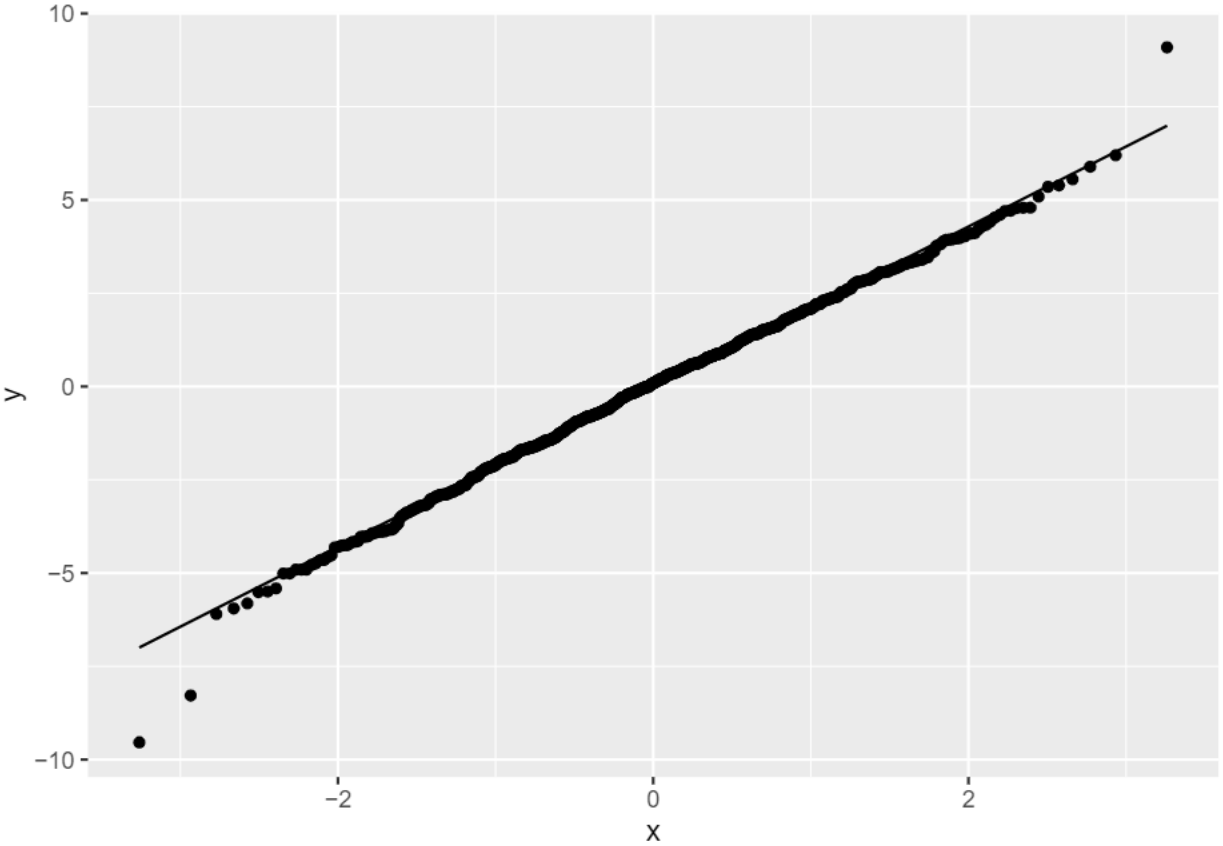
```r
ggplot(data=points, aes(x=residual)) + geom_histogram(bindwidth=1, color="white")
```

```
## Warning in geom_histogram(bindwidth = 1, color = "white"): Ignoring unknown
## parameters: `bindwidth`
```

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



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



##Part II: Regression of child's height on father and mother's height (Height~Father and Mother)

###Q9: Run a regression without interaction term. Are both independent variables significant? ###Q9 Answer: Yes, both indepedent variables are significant.

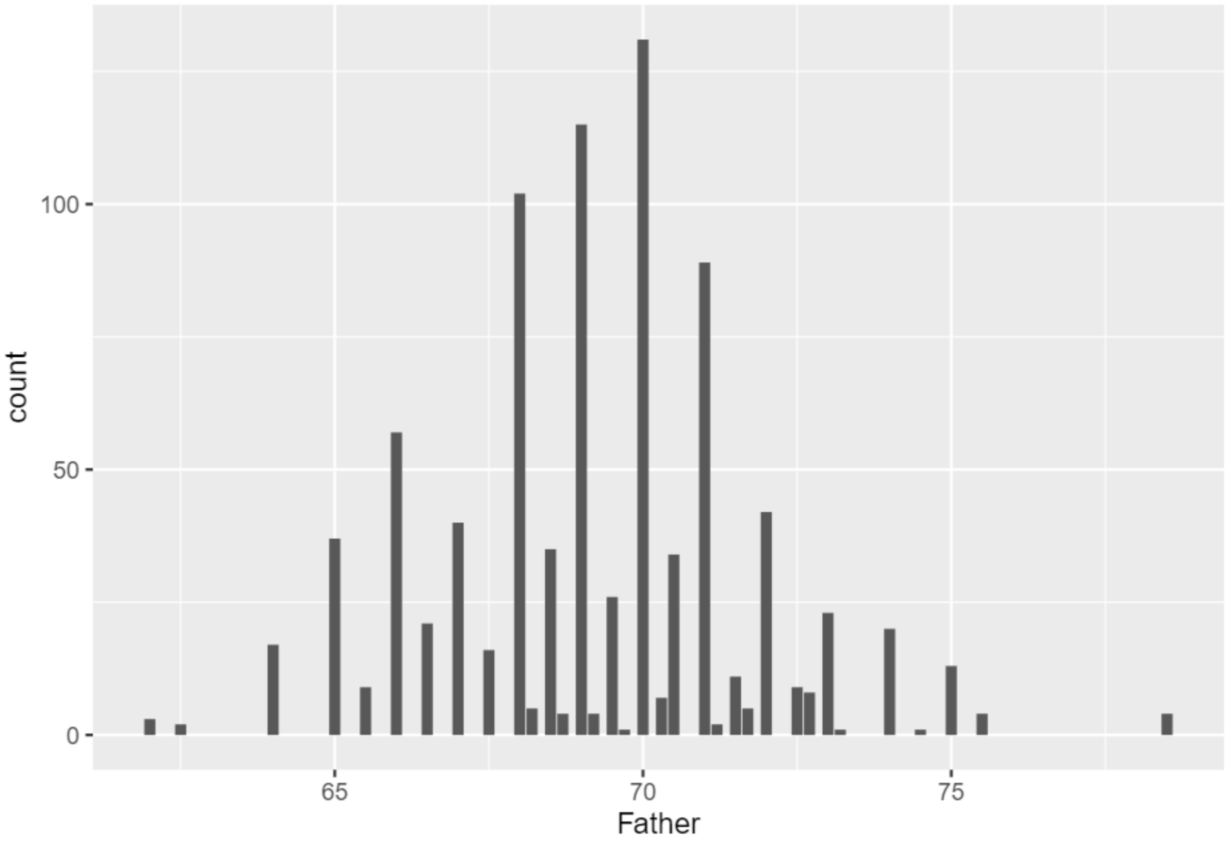
```
galton_height$Height[galton_height$Gender == "F"] <- galton_height$Height[galton_height$Gender == "F"]
cor(galton_height[,c("Height","Father", "Mother")])
```

```
##           Height     Father     Mother
## Height 1.0000000 0.42413624 0.32440954
## Father 0.4241362 1.00000000 0.07366461
## Mother 0.3244095 0.07366461 1.00000000
```

```
table(galton_height$Father)
```

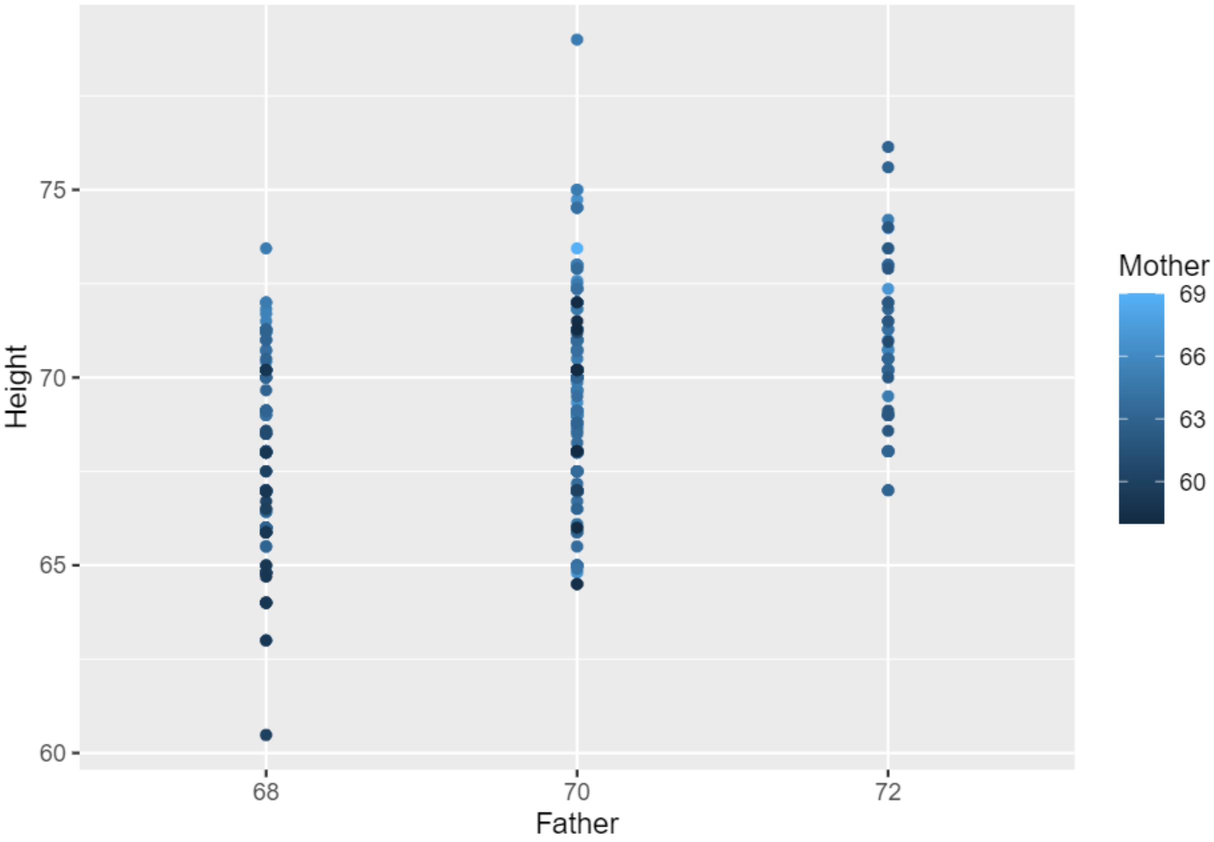
```
##
##   62 62.5   64   65 65.5   66 66.5   67 67.5   68 68.2 68.5 68.7   69 69.2 69.5
##    3    2   17   37    9   57   21   40   16  102    5   35    4  115    4   26
## 69.7   70 70.3 70.5   71 71.2 71.5 71.7   72 72.5 72.7   73 73.2   74 74.5   75
##    1  131    7   34   89    2   11    5   42    9    8   23    1   20    1   13
## 75.5 78.5
##    4    4
```

```
ggplot(data=galton_height, aes(x=Father)) + geom_bar()
```



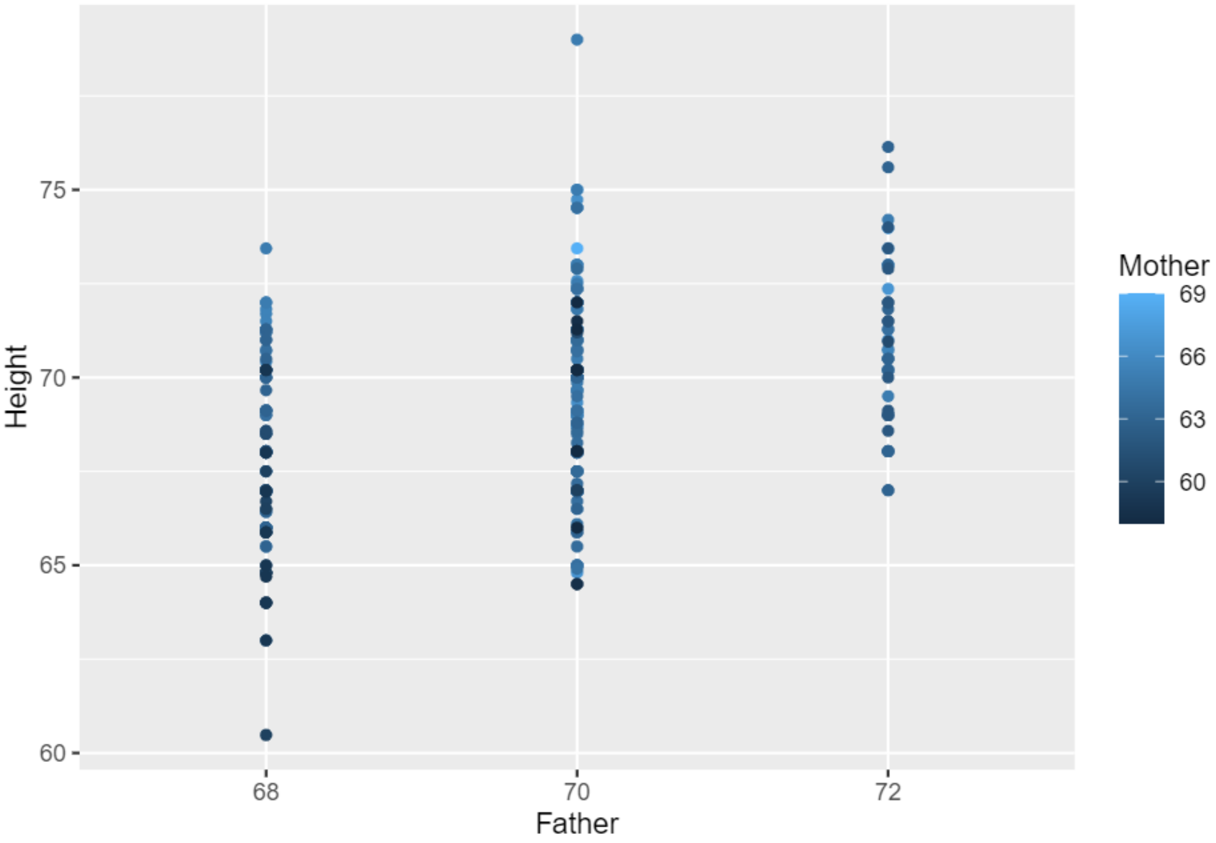
```
sub <- galton_height %>% filter(Father==68 | Father==70 | Father==72) %>% mutate(Father=as.factor(Fathe
ggplot(data=sub, aes(y= Height, color=Mother, x=Father)) + geom_point() + geom_parallel_slopes(method="
```

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



```
ggplot(data=sub, aes(y=Height, color=Mother, x=Father)) + geom_point() + geom_smooth(method="lm", se=FA
```

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



```r
model_21 <- lm(data=galton_height, Height ~ Father)
get_regression_table(model_21)
```

```
## # 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   38.4       2.20       17.4       0   34.1     42.7
## 2 Father       0.446     0.032      14.0       0    0.383    0.508
```

```r
get_regression_summaries(model_21)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1      0.18         0.179  5.52  2.35  2.35      197.       0     1   898
```

```r
model_21 <- lm(data=galton_height, Height ~ Mother)
get_regression_table(model_21)
```

```
## # 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   45.8       2.28       20.1       0   41.4     50.3
## 2 Mother       0.365     0.036      10.3       0    0.295    0.435
```

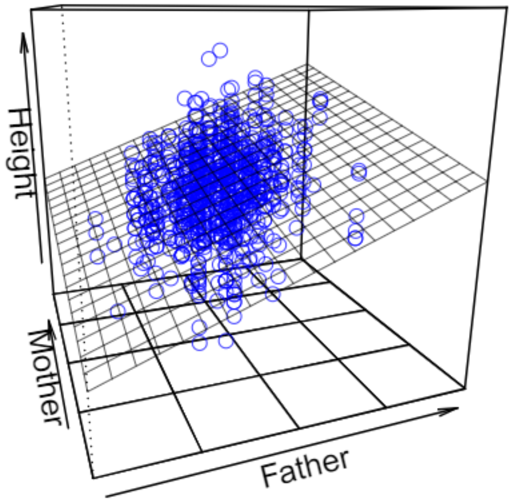
```r
get_regression_summaries(model_21)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.105         0.104  6.02  2.45  2.46      105.       0     1   898
```

```
model_21 <- lm(data=galton_height, Height ~ Father + Mother)
get_regression_table(model_21)
```

```
## # 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
```

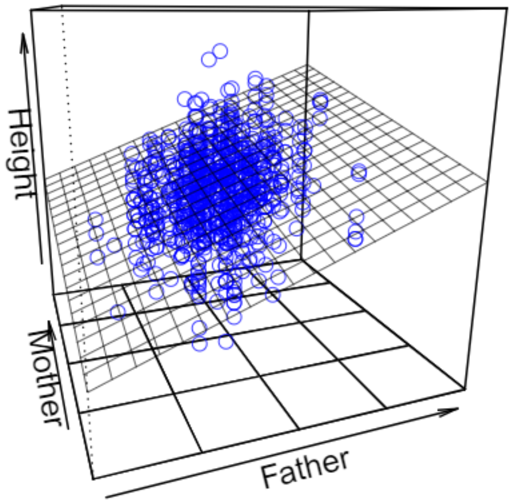
```
plotPlane(model_21, plotx1="Father", plotx2="Mother" )
```



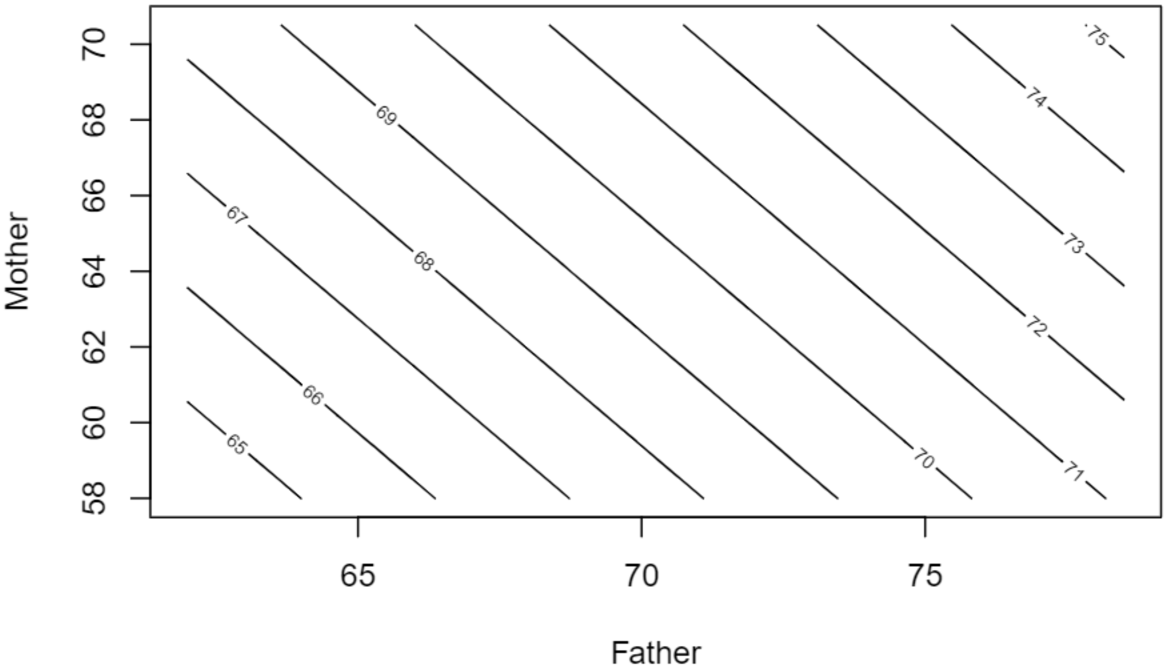
###Q10: In the linear model in Q9, what is the expected increase in child's height if the mother's height increases by 1 inch, while father's height is 68 inches? While father's height is 70 inches? While the father's height is 72 inches? Does the increment depend on the height of the father? ###Q10 Answer: The expected increase in child's height if the mother's height increases by 1 inch, while father's height is 68, 70, and 72 inches is 0.332. The increment does not depend on the height of the father.

###Q11: Get the regression plane and contour plot from the model in Q9 ###Q11 Answer:

```
plotPlane(model_21, plotx1="Father", plotx2="Mother" )
```



```
contour(model_21, Mother ~ Father)
```



###Q12: Run a regression with interaction term. Is the interaction term significant? ###Q12 Answer: The interaction term is not significant due to the p-value being greater that 0.05.

```
cor(galton_height[,c("Height","Father", "Mother")])
```

```
##           Height     Father     Mother
## Height 1.0000000 0.42413624 0.32440954
## Father 0.4241362 1.00000000 0.07366461
## Mother 0.3244095 0.07366461 1.00000000
```

```
model_22 <- lm(data=galton_height, Height ~ Father * Mother)
get_regression_table(model_22)
```

```
## # 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
```

```
sub1<- galton_height %>% filter(Father==68)
model_sub1 <- lm(data=sub1, Height ~ Mother)
get_regression_table(model_sub1)
```

```
## # 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    28.9      6.65       4.34       0    15.7     42.1
## 2 Mother        0.62     0.106      5.86       0     0.41     0.83
```

```
sub1<- galton_height %>% filter(Father==70)
model_sub1 <- lm(data=sub1, Height ~ Mother)
```

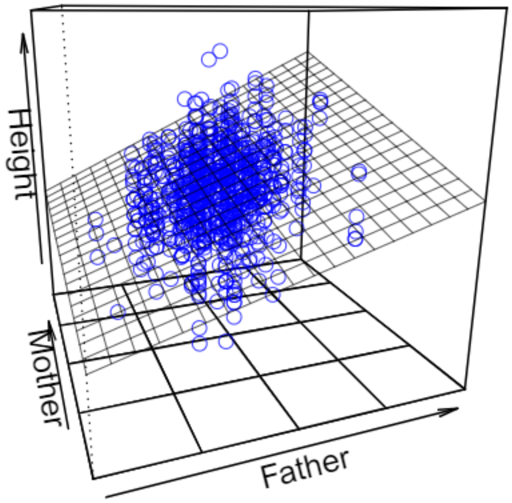
```
get_regression_table(model_sub1)
```

```
## # 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   52.8       5.74       9.20   0        41.5     64.2
## 2 Mother       0.259     0.089      2.91   0.004     0.083    0.436
```

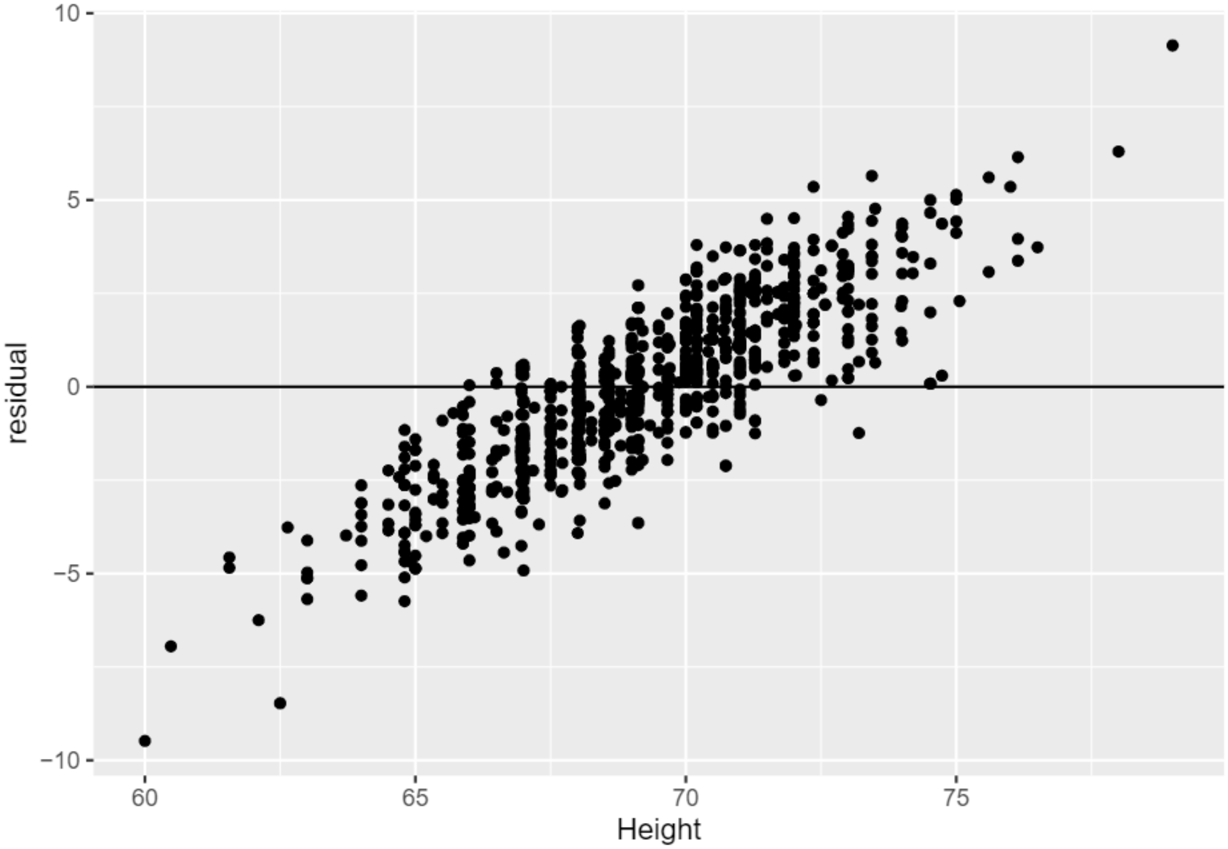
```
sub1<- galton_height %>% filter(Father==72)
model_sub1 <- lm(data=sub1, Height ~ Mother)
get_regression_table(model_sub1)
```

```
## # 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    52.5     11.4        4.61   0        29.5     75.6
## 2 Mother        0.29     0.179      1.62   0.112    -0.071    0.651
```

```
plotPlane(model_22, plotx1="Father", plotx2="Mother")
```



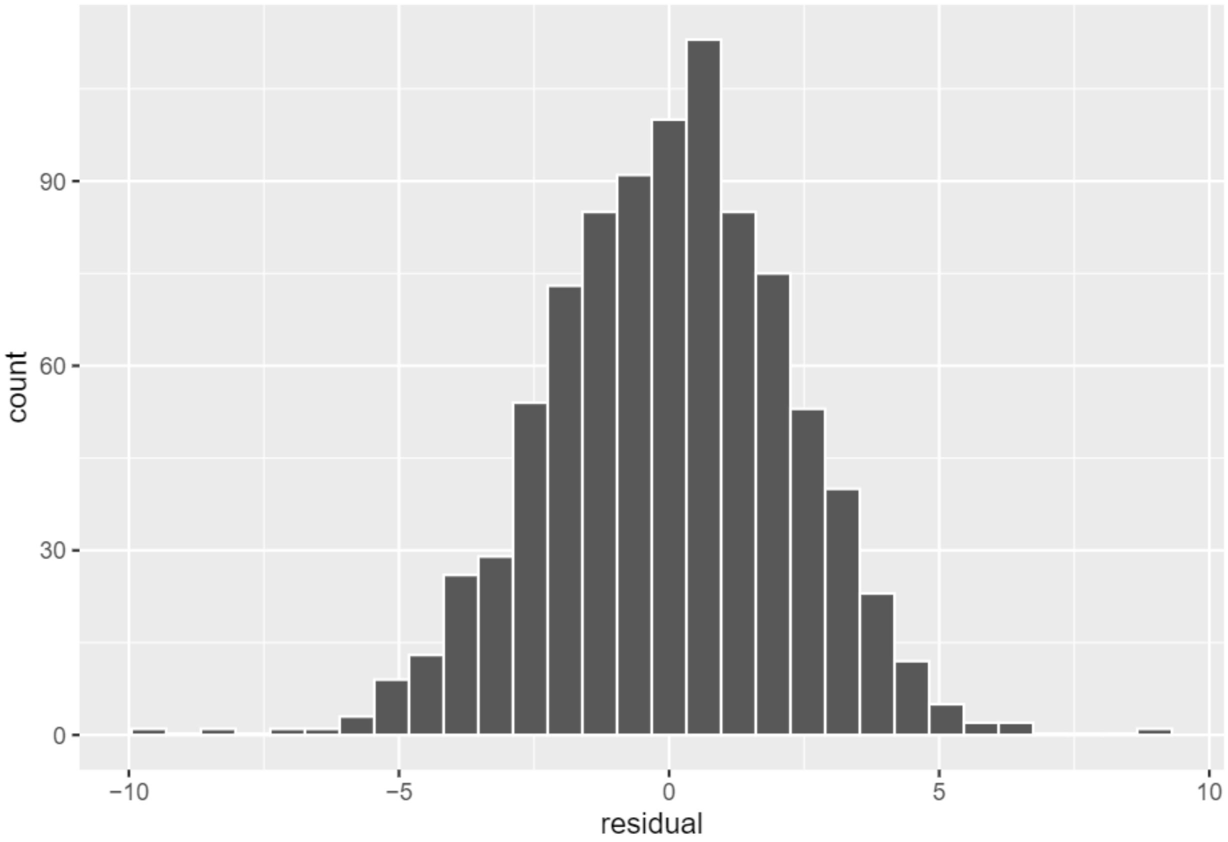
```
points <- get_regression_points(model_22)
ggplot(data=points, aes(x=Height, y=residual)) + geom_point() + geom_hline(yintercept=0)
```



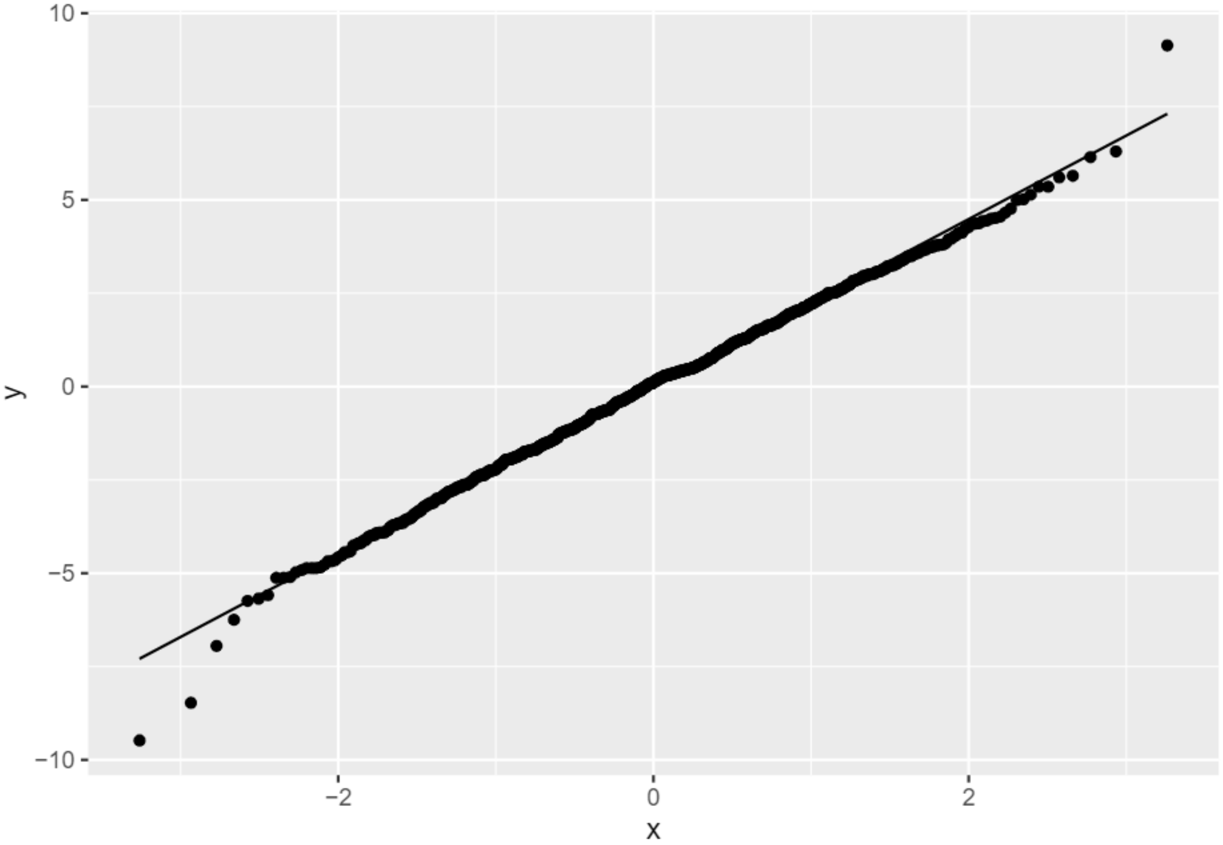
```
ggplot(data=points, aes(x=residual)) + geom_histogram(bindwidth=1, color="white")
```

```
## Warning in geom_histogram(bindwidth = 1, color = "white"): Ignoring unknown
## parameters: `bindwidth`
```

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



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



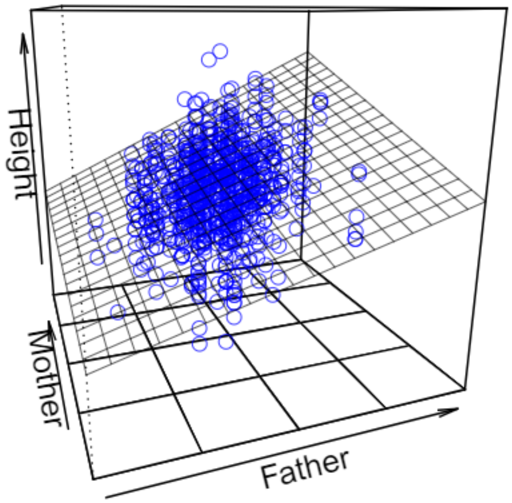
```
get_regression_summaries(model_22)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.267         0.265  4.93  2.22  2.23      109.       0     3   898
```

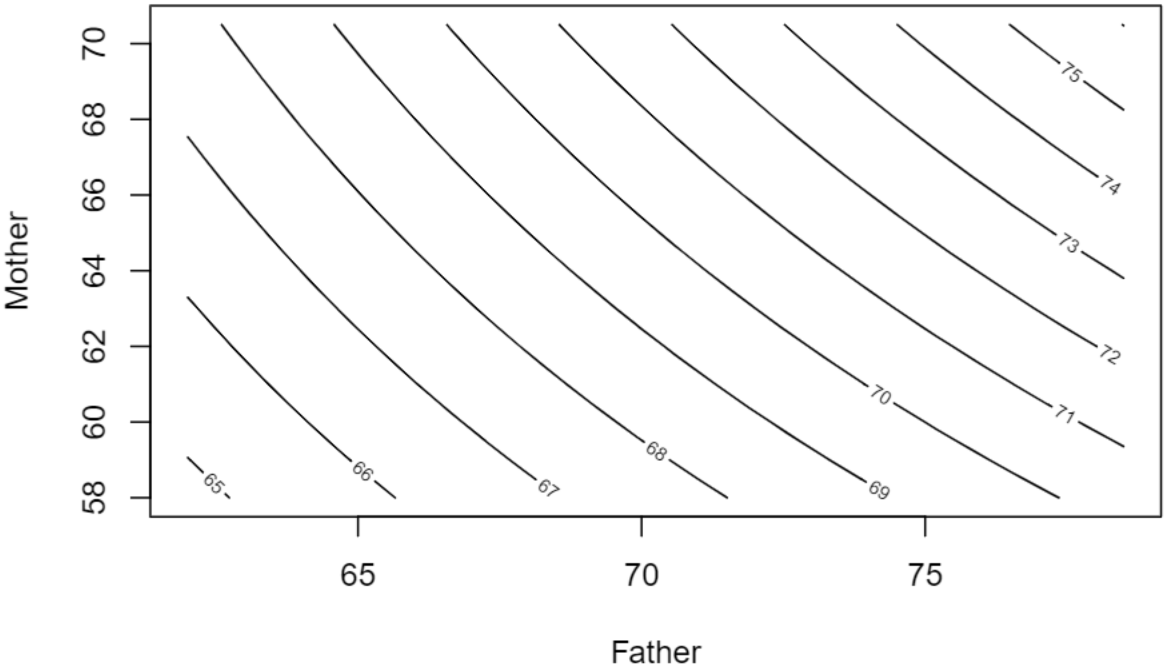
###Q13: In the linear model in Q12, what is the expected increase in child's height if the mother's height increases by 1 inch, while father's height is 68 inches? While father's height is 70 inches? While the father's height is 72 inches? Are there any differences between these increments? Are the differences significant? ###Q13 Answer: The expected increase in child's height if the mother's height increases by 1 inch, while father's height is 68 inches is 0.317, when the father's height is 70 inches it is 0.343, and when the father's height is 72 it is 0.369 increase. The difference between these increments is 0.026. The difference is insignificant.

###Q14: Get the regression plane and contour plot from the model in Q12. ###Q14 Answer:

```
plotPlane(model_22, plotx1="Father", plotx2="Mother")
```



```
contour(model_22, Mother ~ Father)
```



###Q15: If you are going to choose a model between Q9 and Q12, which one will you choose? Why? ###Q15 Answer: I will choose the model from Q9, model with out interaction term, this is because the p-value is 0, therefore the model is significant. Moreover, the slopes are same, therefore, the difference is significant. In Q12 model, model with interaction term, has a p-value that is greater that 0.05, therefore making it insignificant. Moreover, the difference between the slopes is insignificant.

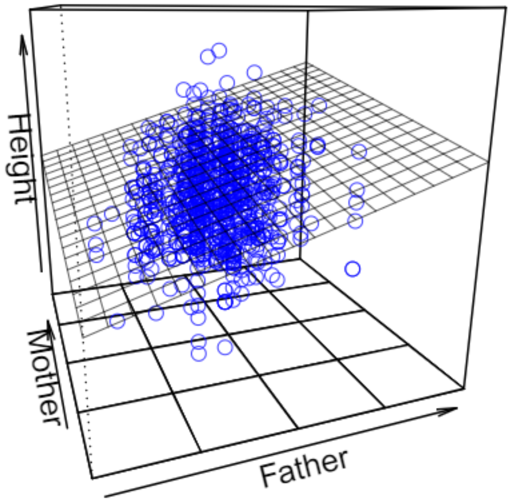
##Part III: Regression of child's height on father and mother's height and gender

###Q16: Run the model without any interaction. Are all independent variables significant? What are the coefficients for each variable? ###Q16 Answer: All the independent variables are significant as they have a p-value of 0. The coefficients for mother's height is 0.3214951, father's height is 0.4059780, child's gender is 5.226, and child's height is 15.345.

```
galton_height <- read.csv("galton_height.csv", header = TRUE)
model_31 <- lm(data=galton_height, Height ~ Father + Mother + Gender)
get_regression_table(model_31)
```

```
## # 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
```

```
plotPlane(model_31, plotx1="Father", plotx2="Mother")
```



```
model_31$coefficients
```

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

###Q17: From the final model you got from Q7, replacing mid parent height by the formula (Father + Mother * 1.08)/2, you can also get the coefficients for Father's height, Mother's height, and Gender. Compare these 3 coefficients to those from Q16. Are they the same? Why? ###Q17 Answer: The coefficient is the final model from Q7 is larger than the coefficient in the model from Q16. They are not the same becuase the lm function in the model from Q7 only consist of the mid-height which is calculated by the formula, therefore, the mutate function must be employed in order to define the mid-height of the parents and then include the new mid-height variable in the lm function. The model from Q16 involves an lm function that has variable that are already defined in the dataset.

```
model1$coefficients
```

```
## (Intercept)     GenderM
##   64.110162    5.118656
```