HW3 - Akshaj Kammari

Due: 02/15/2024

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

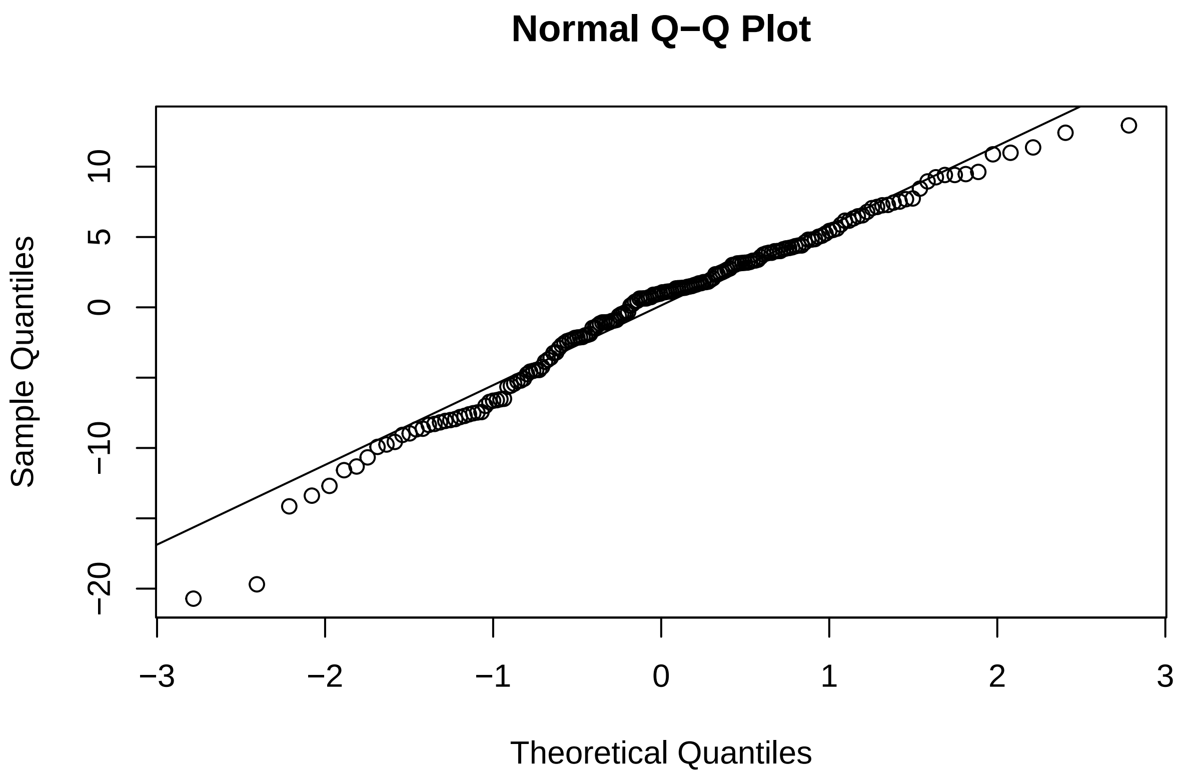
```
##
## Attaching package: 'dplyr'

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

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

```
library(moderndive)
lifedata <- read.csv("Worldlife.csv")
```

##Revisit the regression model of life expectancy in 2023 on life expectancy in 1923 in homework 2

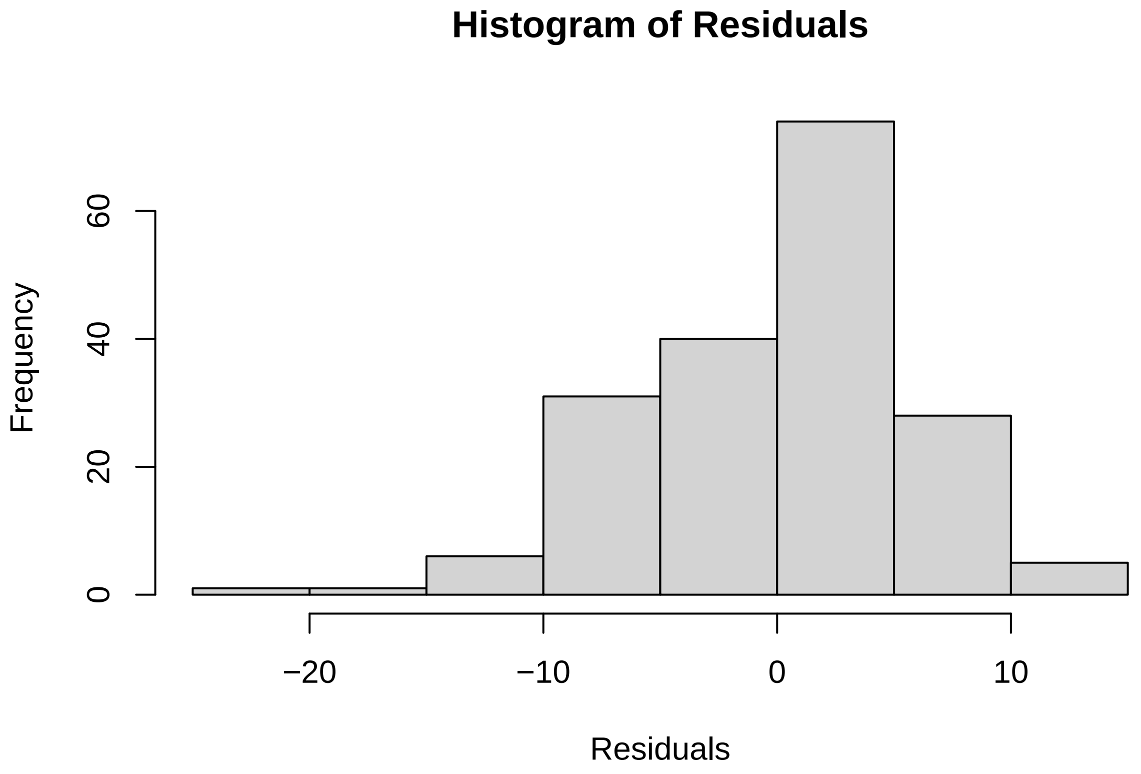
###1. Get the QQ lot of the residual

```
model <- lm(life2023 ~ life1923, data = lifedata)
qqnorm(residuals(model))
qqline(residuals(model))
```

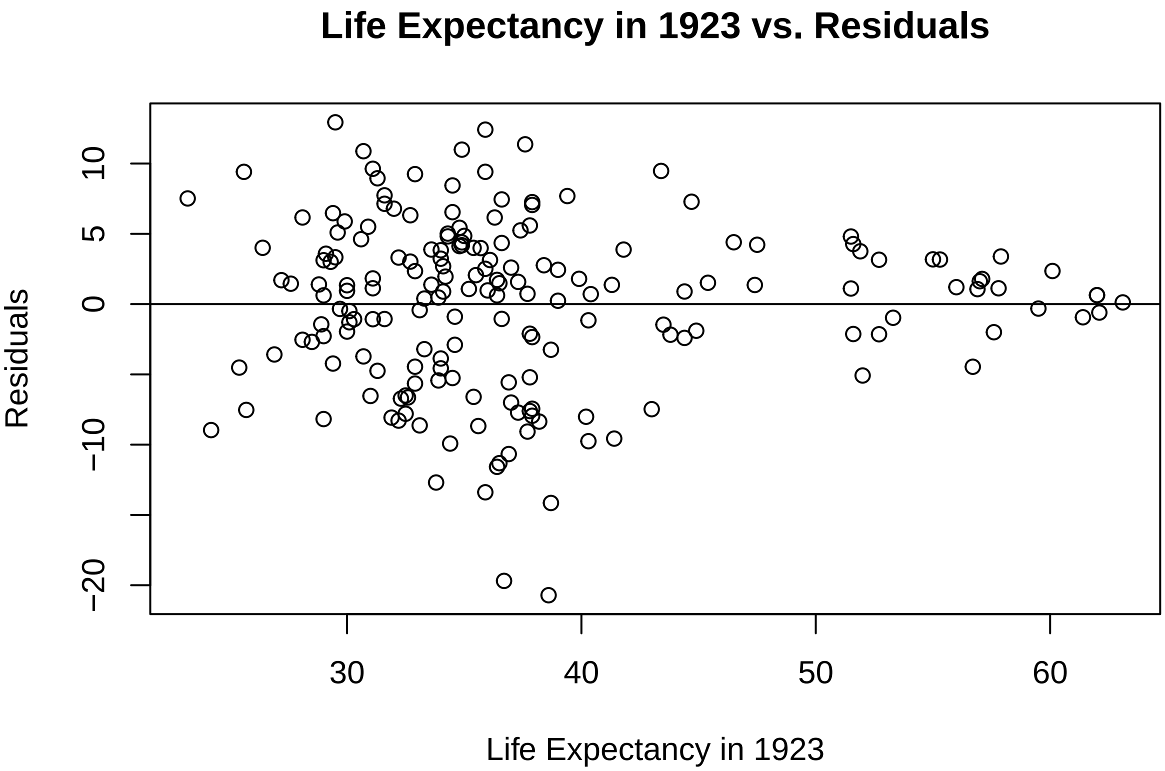


###2. Together with histogram of residual and scatter plot of residual vs. x, check the four assumptions in regression. Explain why or why not for violations in regression

```
hist(residuals(model), main = "Histogram of Residuals", xlab = "Residuals")
```



```
plot(lifedata$life1923, residuals(model),
     main = "Life Expectancy in 1923 vs. Residuals",
     xlab = "Life Expectancy in 1923", ylab = "Residuals")
abline(h = 0)
```



#Linear Pattern between X and Y - The slope suggests a positive linear relationship, supporting the linearity assumption #Independent - The data collected for 1923 and 2023 can be assumed to be independent, as there doesn't seem to be clustering in the data that is being ignored. #Normal Distribution - The plot shows most of the points following the line, and there don't seem to be any extreme deviations, which is a common occurance. #Equal Variance - According to the scatter plot image, it seems that the residuals are roughly evenly distributed, which suggests that equal variance may be satisfied.

##Galton's height data

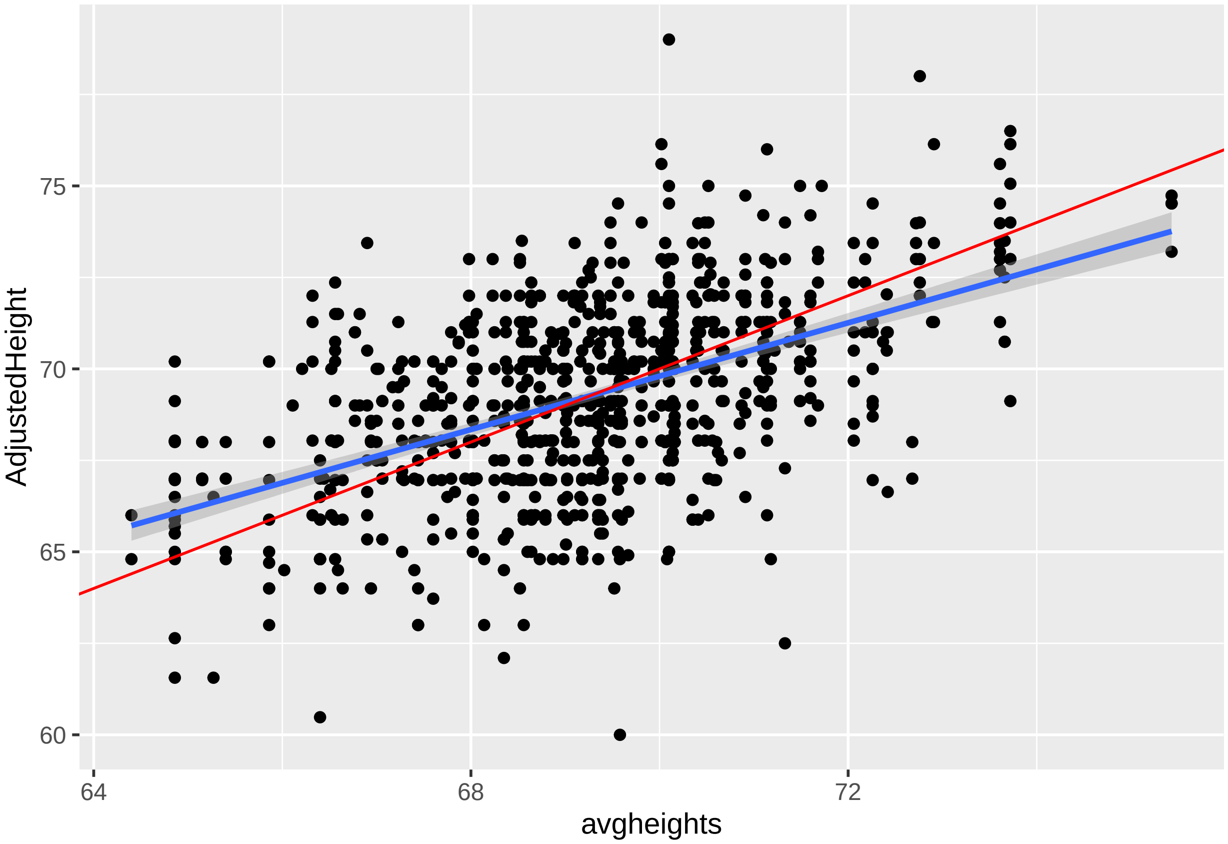
```
galton_data = read.csv("galton_height.csv")
```

##Regression of child's height (gender adjusted) on mid-height of parent

###1. Have a scatterplot of y vs. x. On top of scatterplot, add the regression line and the diagonal line y=x, with different colors

```
galton_data$adjustedMother = galton_data$Mother * 1.08
galton_data$avgheights = (galton_data$Father + galton_data$adjustedMother) / 2
galton_data = galton_data %>% mutate(AdjustedHeight = ifelse(Gender == "F", Height * 1.08, Height))

ggplot(data = galton_data, aes(x = avgheights, y = AdjustedHeight)) +  geom_point() +  geom_smooth(meth
```

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



###2. What is the average children's height in the data? What is the average mid-height of the parent?

```
mean(galton_data$AdjustedHeight)
```

```
## [1] 69.23371
```

```
mean(galton_data$avgheights)
```

```
## [1] 69.22201
```

###3. Among parents whose mid-height between 72 and 73 inches, what is the average height of their children?

```
between7273 = subset(galton_data, galton_data$avgheights < 73 & galton_data$avgheights > 72)
mean(between7273$AdjustedHeight)
```

```
## [1] 71.3178
```

###4. Run regression, is the model significant?

```
galton_height_model = lm(AdjustedHeight ~ avgheights, data = galton_data)
get_regression_table(galton_height_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    18.8       2.84       6.61       0   13.2      24.3
## 2 avgheights    0.729     0.041     17.8        0    0.649     0.81
```

#The model is significant because the p-values are between 0 and 0.05.

###5. If the parents' mid-height increases by 1 inch, what is the expected increase in child's height? Is the expected increase larger or smaller than 1 inch? #The increase in the child's height is 0.729 in, which is less than 1 inch.

###6. Estimate the child's height if the mid-height of parent is 64, 68, 70, 72, 76 respectively, and check their "closeness" to the mean height of all children #Adding the 5 values to a vector and then applying the slope and y-int to calculate the height for those values.

```
vec = c(64,68,70,72,76)
vec = vec * 0.729 + 18.8
vec
```

```
## [1] 65.456 68.372 69.830 71.288 74.204
```

#Finding the mean height and then subtracting it from each of the values to find the "closeness."

```
mn = mean(galton_data$AdjustedHeight)
vec = vec - mn
vec
```

```
## [1] -3.7777149 -0.8617149  0.5962851  2.0542851  4.9702851
```

##Regression of mid-height of parent on child's height (gender adjusted)

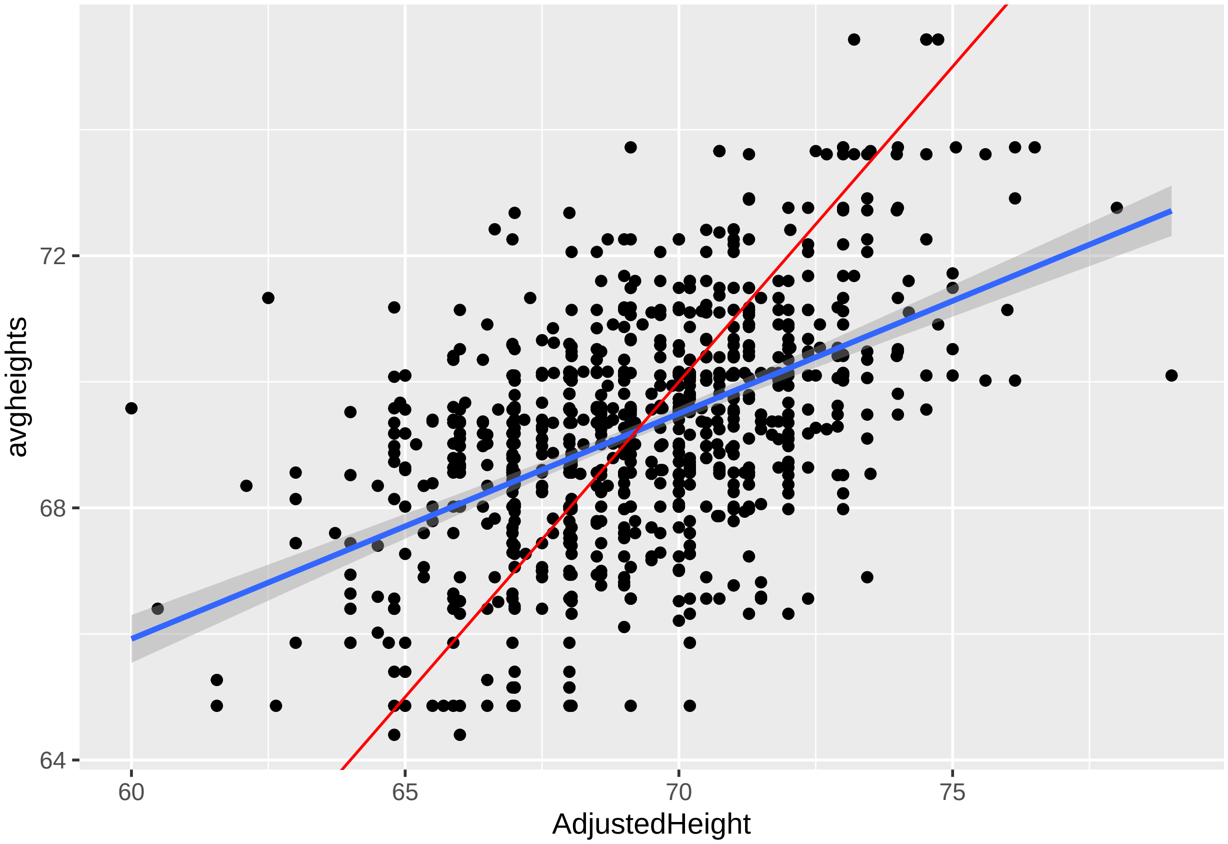
###1. Have a scatterplot of y vs. x. On top of scatterplot, add the regression line and the diagonal line y=x, with different colors

```
model2 = lm(avgheights ~ AdjustedHeight, data = galton_data)
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        44.5        1.39      31.9       0   41.7     47.2
## 2 AdjustedHeight    0.357      0.02      17.8       0    0.318    0.397
```

```
ggplot(data = galton_data, mapping = aes(AdjustedHeight, avgheights)) + geom_point() + geom_smooth(meth
```

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



###2. Among all children with height between 72 and 73 inches, what is the mean mid-height of their parents?

```
between7273 = subset(galton_data, galton_data$AdjustedHeight < 73 & galton_data$AdjustedHeight > 72)
mean(between7273$avgheights)
```

```
## [1] 70.48417
```

###3. Run regression, is the model significant?

```
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        44.5        1.39      31.9       0   41.7     47.2
## 2 AdjustedHeight    0.357      0.02      17.8       0    0.318    0.397
```

#The model is significant because the p-values are between 0 and 0.05.

###4. If the child's height increases by 1 inch, what is the expected increase in parent's mid-height? Is the expected increase larger or smaller than 1 inch? #The increase in the parent's mid-height is 0.357 in, which is less than 1 inch.

###5. Estimate the parent's mid-height if the child's height is 64, 68, 70, 72, 76 respectively, and check their "closeness" to the mean mid-height of all parents #Adding the 5 values to a vector and then applying the slope and y-int to calculate the height for those values.

```
vec = c(64,68,70,72,76)
vec = vec * 0.357 + 44.5
vec
```

```
## [1] 67.348 68.776 69.490 70.204 71.632
```

#Finding the mean height and then subtracting it from each of the values to find the “closeness.”

```
mn = mean(galton_data$avgheights)
vec = vec - mn
vec
```

```
## [1] -1.8740067 -0.4460067  0.2679933  0.9819933  2.4099933
```

##Use the above results to explain regression to the mean #The average mean mid-height of the parents is 69.2. Regression to the mean states that the parent min height decreases if the child’s height is below the average.