LAB3

RAJ SHAH

2025-03-01

```r
# Load necessary libraries
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
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.2

##
## Attaching package: 'car'

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

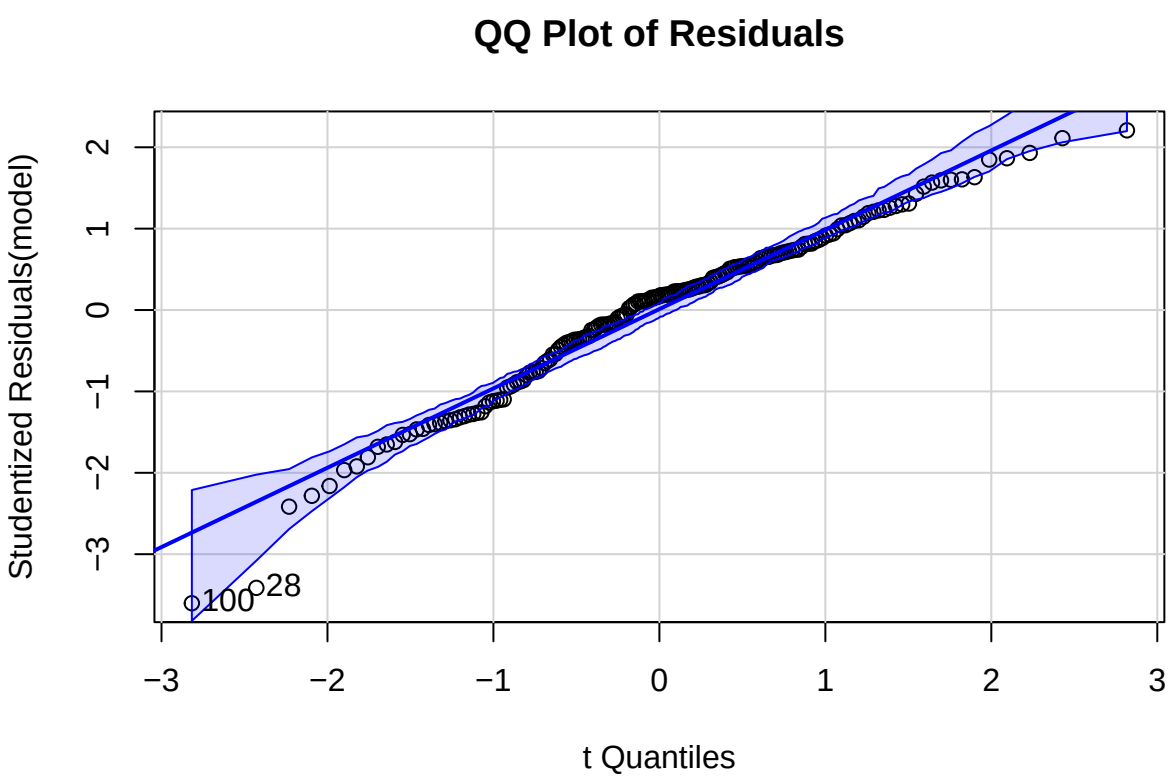
```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

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

# Load the dataset
worldlife <- read.csv("C:\\Users\\rajsh\\OneDrive\\Desktop\\Inference Data Science 291\\LAB3\\Worldlife

# Perform linear regression: Life expectancy in 2023 ~ Life expectancy in 1923
model <- lm(life2023 ~ life1923, data = worldlife)
summary(model)
```

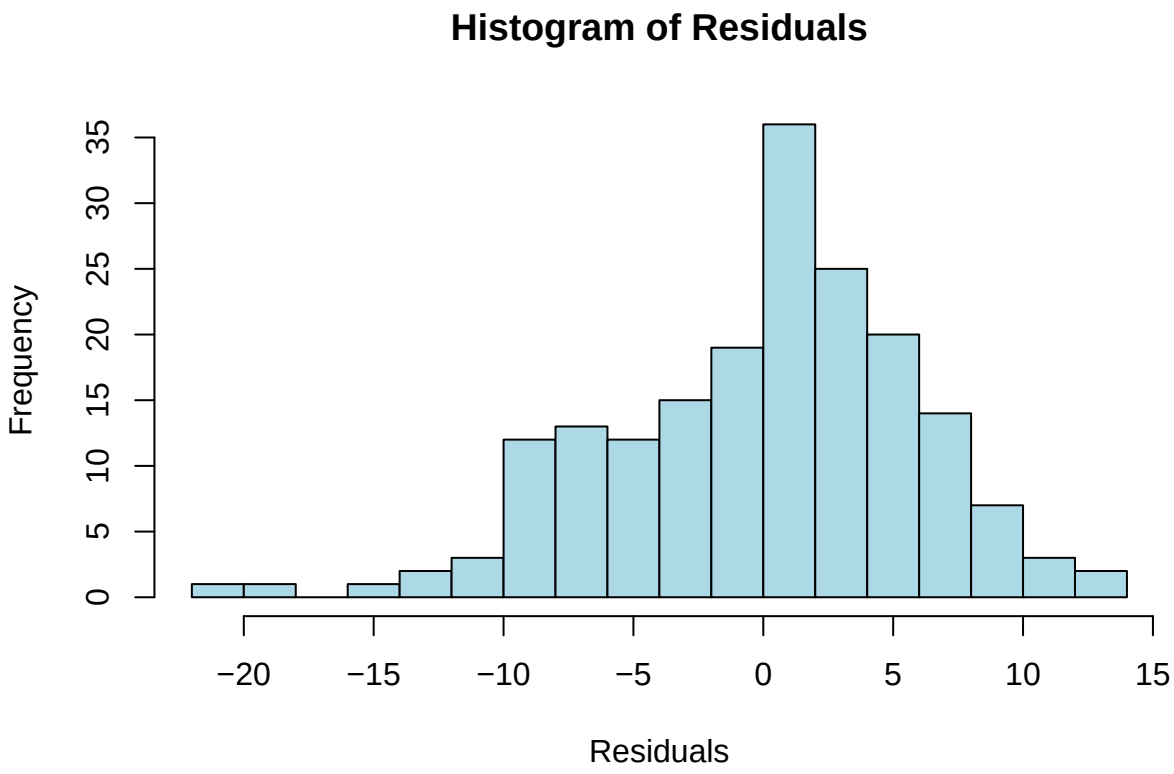
```
##
## Call:
## lm(formula = life2023 ~ life1923, data = worldlife)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.710  -3.687   1.020   3.961  12.933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.40509    1.90355  31.208  < 2e-16 ***
## life1923     0.37837    0.04923   7.685 8.83e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.949 on 184 degrees of freedom
## Multiple R-squared:  0.243,  Adjusted R-squared:  0.2389
## F-statistic: 59.06 on 1 and 184 DF,  p-value: 8.834e-13
```

```r
# 1. Get the QQ plot of the residual
qqPlot(model, main="QQ Plot of Residuals")
```

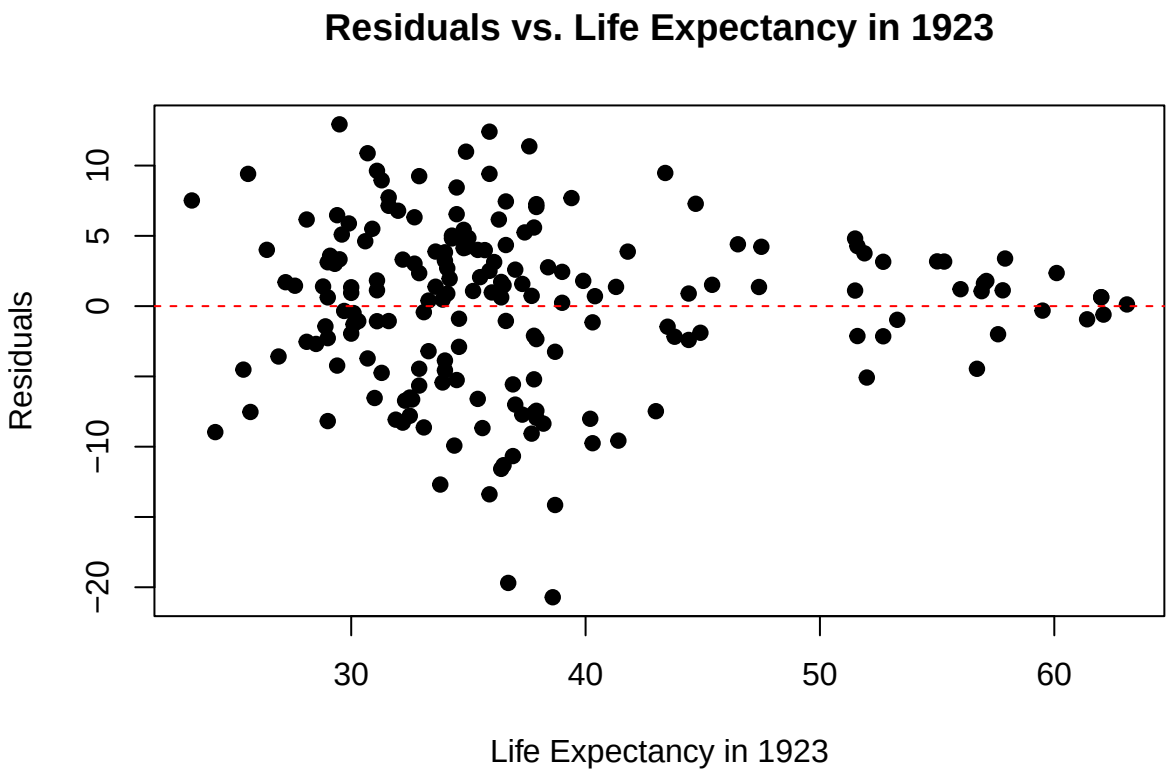



```
## [1]  28 100
```

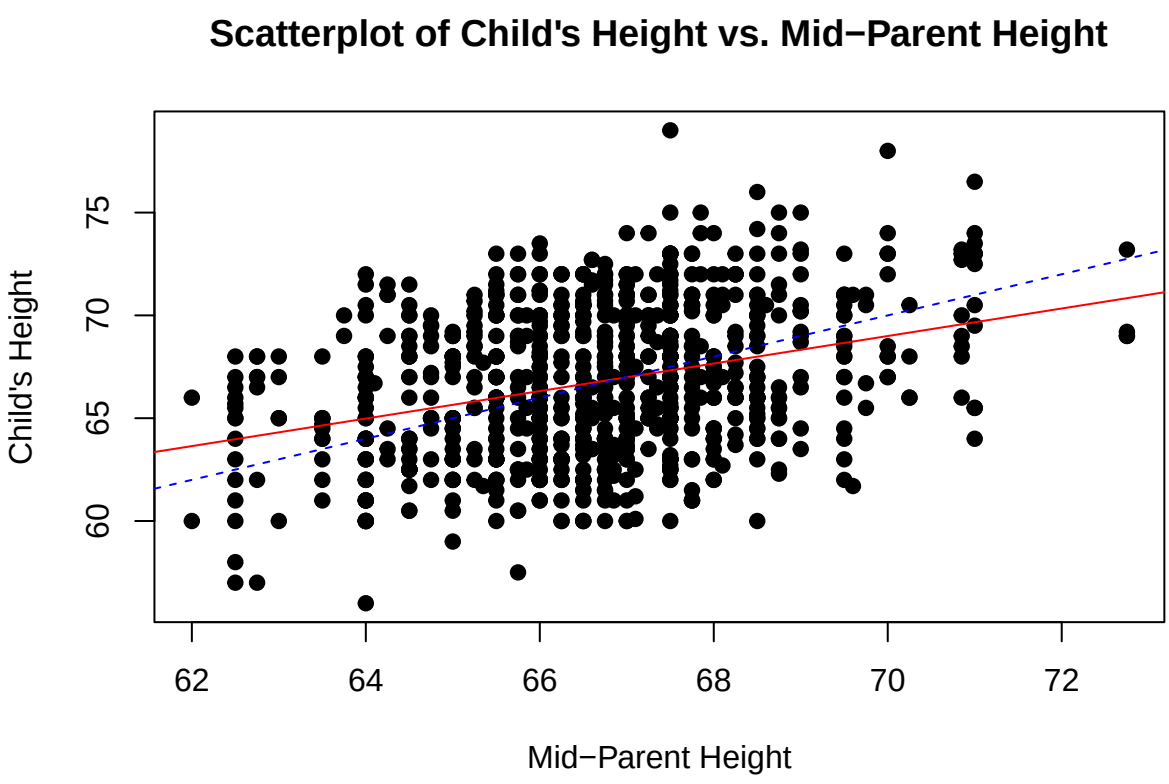
```
# The QQ plot checks if residuals follow a normal distribution.
# If the points align closely with the 45-degree line, the residuals are normally distributed,
# satisfying the assumption of normality. Minor deviations from the line suggest slight departures
# from normality, but there are no severe violations.

# 2. Together with histogram of residual and scatter plot of residual vs. x, check the four assumptions

# Histogram of residuals
hist(residuals(model), breaks=20, col="lightblue", main="Histogram of Residuals", xlab="Residuals")
```




```
# The histogram appears roughly bell-shaped, indicating a nearly normal distribution of residuals.
# Some skewness may be present, but it is not extreme.

# Scatter plot of residuals vs. x
plot(worldlife$life1923, residuals(model), main="Residuals vs. Life Expectancy in 1923", xlab="Life Exp
abline(h=0, col="red", lty=2)
```




```r
# The scatter plot shows that residuals are randomly scattered around zero, indicating homoscedasticity
# If residuals exhibit a funnel shape or pattern, it would indicate heteroscedasticity, which violates
# In this case, no strong pattern is observed, suggesting the model satisfies the assumption of homosce

# Conclusion on Regression Assumptions:
# - Linearity: The relationship between life expectancy in 1923 and 2023 appears linear.
# - Normality of Residuals: Mostly satisfied, with minor deviations.
# - Homoscedasticity: No clear pattern in the residual plot suggests this assumption holds.
# - Independence: Assuming data collection was done independently, this assumption should hold.

# Galton's height data

# Load Galton height dataset
galton <- read.csv("C:\\Users\\rajsh\\OneDrive\\Desktop\\Inference Data Science 291\\LAB3\\Cleaned_Galt

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

# 1. Obtain a scatterplot of y vs. x. Add regression line and y=x diagonal
plot(galton$Mid_Parent_Height, galton$Height, main="Scatterplot of Child's Height vs. Mid-Parent Height
abline(lm(Height ~ Mid_Parent_Height, data=galton), col="red") # Regression line
abline(a=0, b=1, col="blue", lty=2) # y=x diagonal
```




```
# The scatterplot shows a positive correlation between mid-parent height and child's height.
# The regression line (red) represents the best fit, while the diagonal line y = x (blue, dashed) is a
# Since the regression line has a lower slope than y=x, this indicates regression to the mean (extreme

# 2. Compute averages
mean_child_height <- mean(galton$Height)
mean_mid_parent_height <- mean(galton$Mid_Parent_Height)

# The average child's height in the dataset is approximately 66.76 inches.
# The average mid-parent height is approximately 66.66 inches.

# 3. Average height of children for mid-parent height between 72 and 73
subset_galton <- subset(galton, Mid_Parent_Height >= 72 & Mid_Parent_Height <= 73)
mean(subset_galton$Height)
```

```
## [1] 70.1
```

```
# The average child's height for parents with a mid-height between 72 and 73 inches is approximately 70

# 4. Run regression and check significance
model_galton <- lm(Height ~ Mid_Parent_Height, data=galton)
summary(model_galton)
```

```
##
## Call:
```

```
## lm(formula = Height ~ Mid_Parent_Height, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9814 -2.6604 -0.1642  2.7795 11.6762
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        22.1488     4.3076   5.142 3.34e-07 ***
## Mid_Parent_Height   0.6693     0.0646  10.360  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.388 on 896 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 107.3 on 1 and 896 DF,  p-value: < 2.2e-16
```

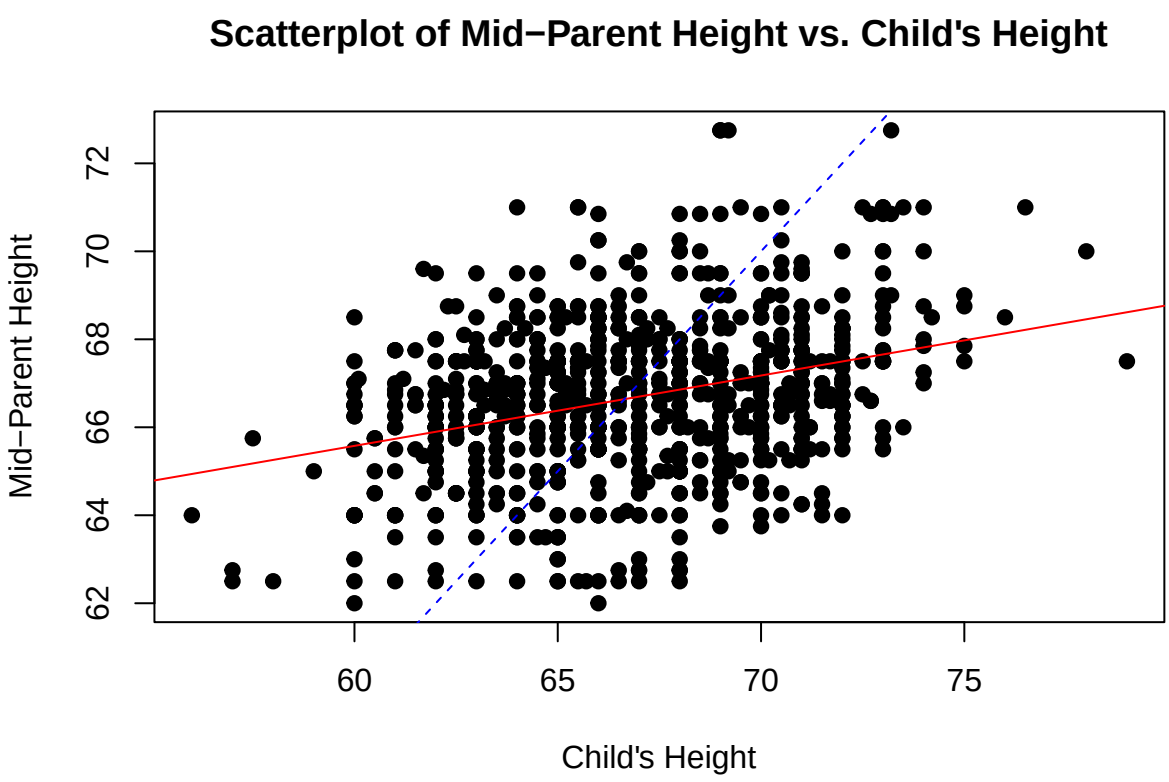
```
# The regression model is significant, as indicated by the very low p-value (< 0.001).
# The R² value is 0.107, meaning about 10.7% of the variability in child height is explained by mid-par

# 5. If the parents' mid-height increases by 1 inch, what is the expected increase in child's height? I
coef(model_galton)[2]
```

```
## Mid_Parent_Height
##         0.6692589
```

```
# The expected increase is approximately 0.67 inches, which is less than 1 inch, demonstrating regressi

# 6. Estimate child's height for specific mid-parent heights
new_data <- data.frame(Mid_Parent_Height = c(64, 68, 70, 72, 76))
predict(model_galton, new_data)
```

```
##        1        2        3        4        5
## 64.98138 67.65841 68.99693 70.33545 73.01249
```

```
# The predicted heights for different mid-parent heights show variation but tend to regress towards the

# Regression of mid-height of parent on child's height

# 1. Scatterplot with regression and y=x diagonal
plot(galton$Height, galton$Mid_Parent_Height, main="Scatterplot of Mid-Parent Height vs. Child's Height
abline(lm(Mid_Parent_Height ~ Height, data=galton), col="red") # Regression line
abline(a=0, b=1, col="blue", lty=2) # y=x diagonal
```




```
# The scatterplot confirms the positive relationship between child's height and mid-parent height.
# The regression line (red) is again flatter than y=x, further indicating regression to the mean.

# 2. Mean mid-parent height for children between 72 and 73 inches
subset_child <- subset(galton, Height >= 72 & Height <= 73)
mean(subset_child$Mid_Parent_Height)
```

```
## [1] 67.64068
```

```
# The mean mid-parent height for children between 72 and 73 inches is approximately 67.64 inches.

# 3. Run regression and check significance
model_rev <- lm(Mid_Parent_Height ~ Height, data=galton)
summary(model_rev)
```

```
##
## Call:
## lm(formula = Mid_Parent_Height ~ Height, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5370 -1.0370  0.0822  0.9706  5.7334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 55.98730    1.03151   54.28   <2e-16 ***
## Height       0.15984    0.01543   10.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.656 on 896 degrees of freedom
## Multiple R-squared:  0.107,  Adjusted R-squared:  0.106
## F-statistic: 107.3 on 1 and 896 DF,  p-value: < 2.2e-16
```

```
# The regression model is significant, with a very low p-value (< 0.001), indicating a meaningful relat

# 4. Expected increase in mid-parent height per 1 inch increase in child's height
coef(model_rev)[2]
```

```
##    Height
## 0.1598445
```

```
# The expected increase in mid-parent height is approximately 0.16 inches, which is much smaller than 1

# 5. Estimate the parent's mid-height if the child's height is 64, 68, 70, 72, 76 respectively, and che
new_child_data <- data.frame(Height = c(64, 68, 70, 72, 76))
predict(model_rev, new_child_data)
```

```
##        1        2        3        4        5
## 66.21735 66.85673 67.17642 67.49611 68.13548
```

```
# The predicted mid-parent heights regress toward the mean, demonstrating regression to the mean in the
```