Exploration of Different Coloured Cockatoos

Arian Kalantari

2024-05-24

Random sample of 234 cockatoos have been collected, and the variables recorded for each subject are listed below:

Variable	Description
id	Subject ID
colour	Either "black" or "white"
wingspan	The wingspan of the subject
weight	The weight of the subject in grams
bodylength	The body length of the subject (unknown whether this includes the head)

```
library(knitr)

cockatoo <- read.csv('48549819_data_StatReport.csv', header = TRUE)
kable(head(cockatoo))
```

ID	colour	wingspan	weight	bodylength
subj1	white	99.4	40.85	64.8
subj2	white	101.2	37.38	63.2
subj3	white	92.3	67.45	62.9
subj4	white	91.9	53.93	61.6
subj5	white	96.0	47.67	64.2
subj6	black	68.9	67.54	40.5

With the provided dataset, we are interested in statistical answers to below questions:

(a) Is there any difference in the average weight of black and white cockatoos?
(b) What is the relation between the wingspan of cockatoos and the body length?

Brief Summary:

(a) After applying two-sample t test, we do not have evidence suggesting there is statistically significant difference in average weight between black and white cockatoos.
(b) There are certainly some linear relationship between the wingspan and the body length. However, we found that black and white cockatoos have non-identical linear pattern, so we explored the wingspan vs body_length relationship for 2 x groups.

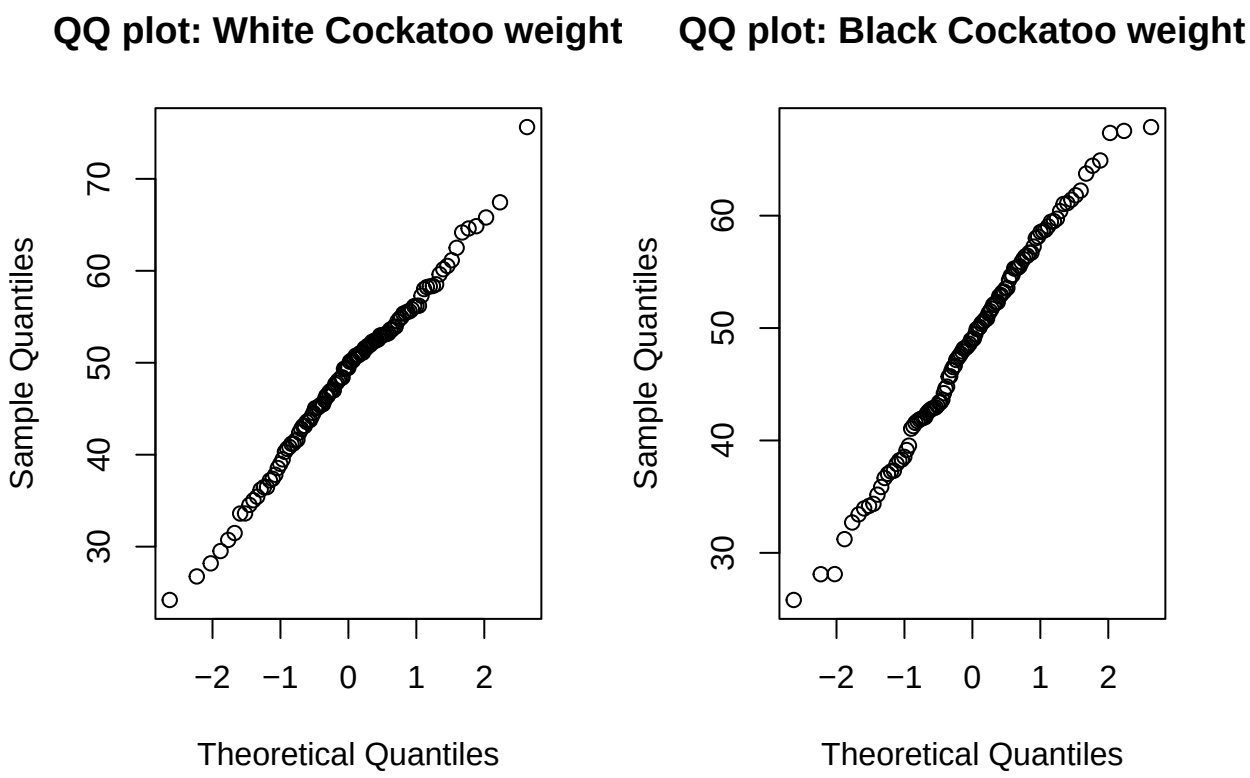
Difference in the Average Weight Between Black and White Cockatoos.

Since we are comparing 2 x groups of cockatoos and our interest is whether these 2 x groups have different mean weight, we can apply the **two sample** t.test

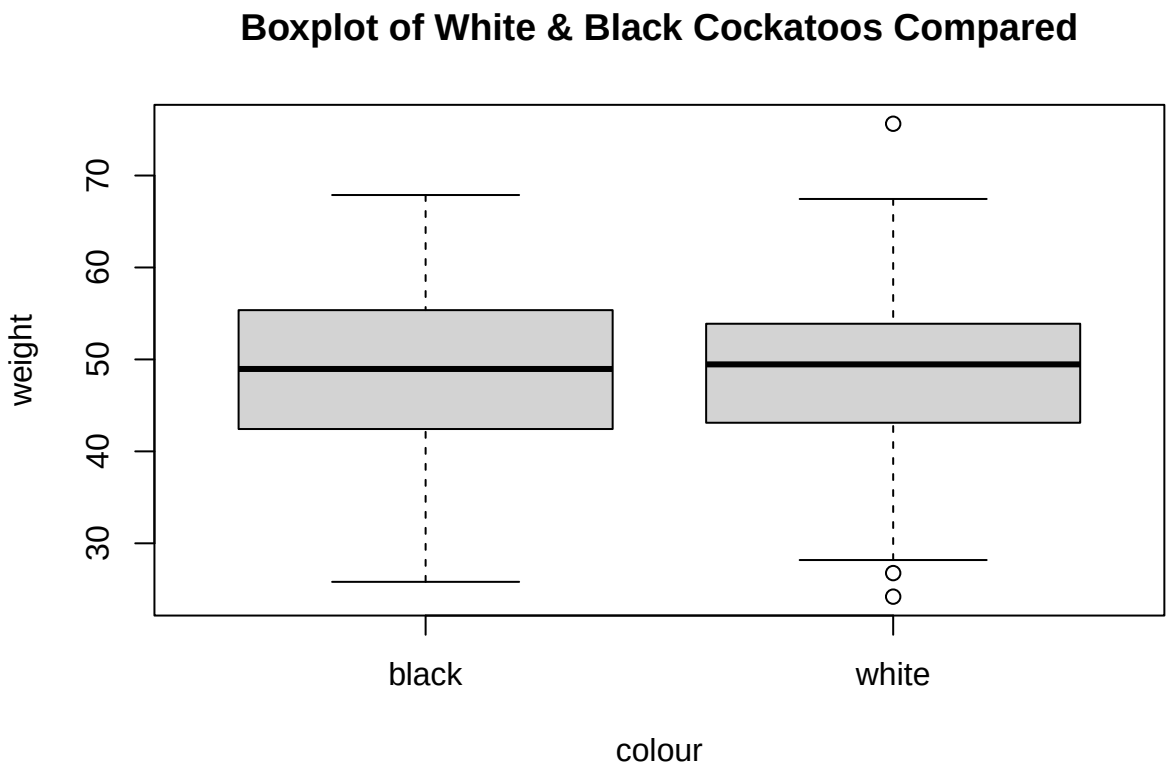
Before conducting the test, we need to check assumptions:

1. **Independence between observations:** This is not able to be verified on face value of the data. The data gathering methodology would need to be checked to ensure there was no bias. Here, we can assume this is met.
2. **Observations close to normally distributed:** QQ-plot for both groups show normality in their distribution. There seems a few outliers in the weight of white cockatoos, however both groups show patterns close to normality.
3. **Equal variance within groups:** The fact that their boxplot lengths appear to be nearly same as well as the calculation of their standard deviations are nearly equal suggest that we can assume have equal variance.

```
white.weight <- cockatoo$weight[cockatoo$colour == 'white']
black.weight <- cockatoo$weight[cockatoo$colour == 'black']

par(mfrow=c(1,2))
qqnorm(white.weight, main="QQ plot: White Cockatoo weights")
qqnorm(black.weight, main="QQ plot: Black Cockatoo weights")
```



```r
boxplot(weight~colour, data=cockatoo, main="Boxplot of White & Black Cockatoos Compared")
```




```r
c(sd(white.weight), sd(black.weight))
```

```
## [1] 9.126537 9.165771
```

Now we checked all assumptions are met, we prepare for the test.

- **Parameter Definition:**
  - μ_1 = mean weight of white cockatoos (grams)
  - μ_2 = mean weight of black cockatoos (grams)
- **Hypothesis:**
  - $H_0 : \mu_1 = \mu_2$ (Both types of cockatoos have the same average weight)
  - $H_1 : \mu_1 \neq \mu_2$ (There is no difference i.e. they are from same distribution.)
- **T-Test:**

```r
t.test(white.weight, black.weight, paired=FALSE, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
```

```
## data:  white.weight and black.weight
## t = -0.18333, df = 232, p-value = 0.8547
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.575262  2.136801
## sample estimates:
## mean of x mean of y
##  48.46966  48.68889
```

- **Test Statistic:**

From the test, we attained the statistic t = -0.18333, which is given by

$$t_{obs} = \frac{\overline{x_1} - \overline{x_2}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- **Null Distribution:**

Under the Null Hypothesis (H_0: $\mu_1 = \mu_2$), test statistic t_{obs} follows $t_{n_1+n_2-2}$.

That is, $t_{obs} \sim t(232)$

- **P-Value:**

We use two-sided P-value = $0.8547 > \alpha = 0.05$

- **Conclusion**

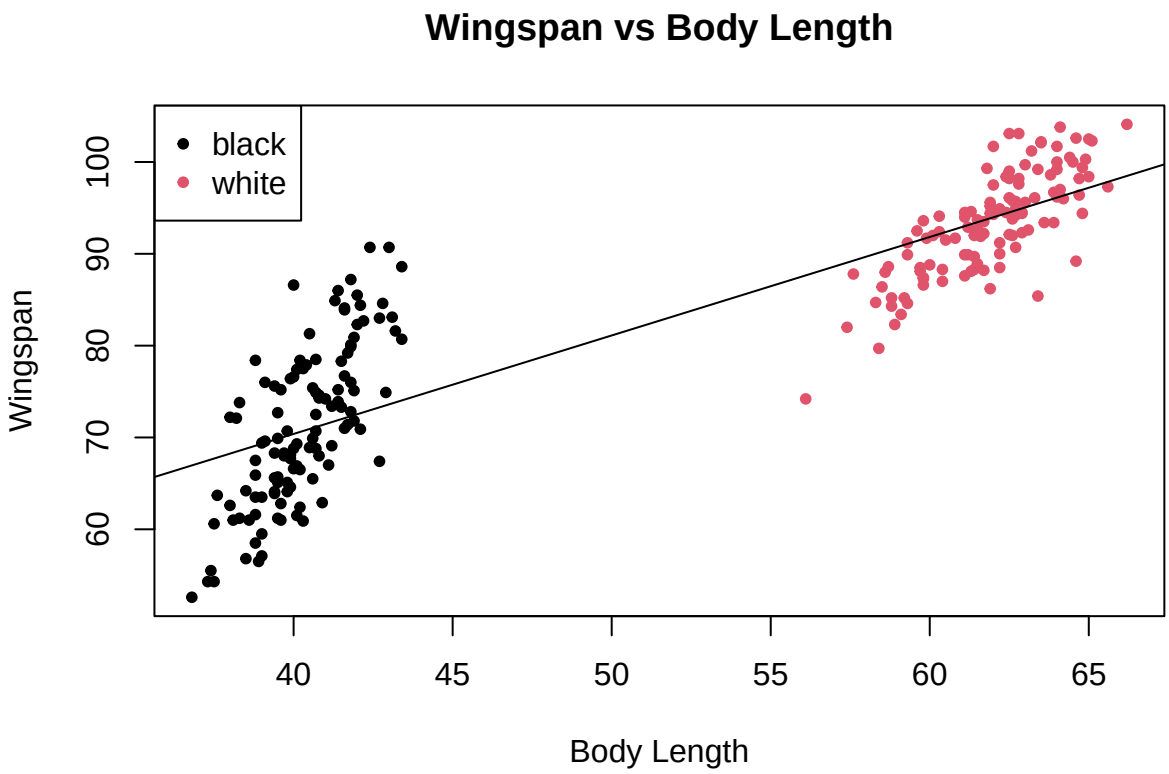
Since our P-Value is bigger than 0.05, we do not reject H_0. Meaning that we do not have evidence against H_0 suggesting that their average weights are not significantly different between cockatoos from 2 x groups (white or black).

Relation Between Wingspan and Body Length.

We first generate a scatterplot to check if there is any linear relation between Wingspan and Body Length.

Data-points seem clearly divided into 2 x groups. Thanks to the colour, we can see there is a difference in distribution depends on the cockatoo's colour.

```
lm.1 = lm(wingspan~bodylength, data = cockatoo)

plot(cockatoo$bodylength, cockatoo$wingspan,
     xlab = "Body Length", ylab = "Wingspan",
     main='Wingspan vs Body Length',
     pch = 20, col=factor(cockatoo$colour))
abline(lm.1)

legend("topleft",
       legend = levels(factor(cockatoo$colour)),
       pch = 20,
       col = factor(levels(factor(cockatoo$colour))))
```



Findings:

- There seems difference between black and white cockatoos. (Test result shows their bodylength is completely different, we should assume they are different.)
- There seems to be 2 x linear relationships by looking at the scatterplot.
- Missing data-points ranging 45 < bodylength < 55 is not ideal for simple linear regression application.

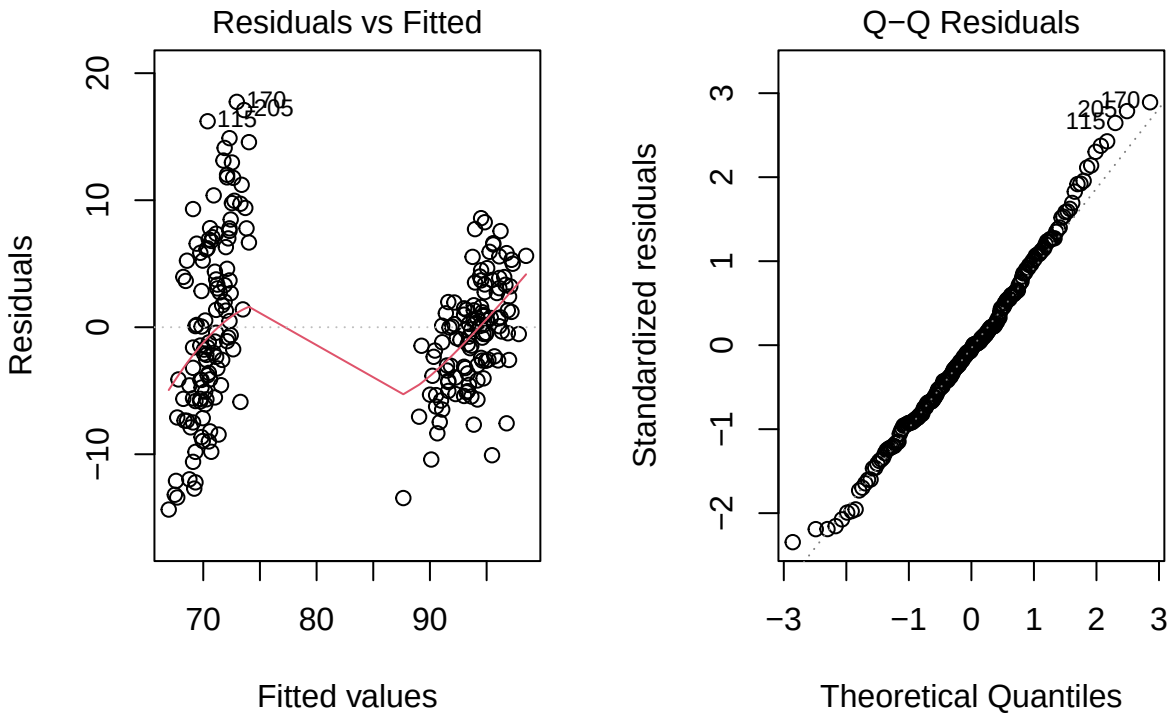
```
##
##  Two Sample t-test
##
## data:  cockatoo$bodylength[cockatoo$colour == "white"] and cockatoo$bodylength[cockatoo$colour == "bl
## t = 93.852, df = 232, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  21.15743 22.06479
## sample estimates:
## mean of x mean of y
##  61.91282  40.30171
```

Adjusted r-squared is 0.7848, meaning that nearly 78% of data can be explained by this linear model (lm.1). Even though 78% is a great goodness-of-fit, dividing the dataset into 2 would be recommended to capture detailed relationship between wingspan and bodylength for white cockatoos and black cockatoos separately.

```
summary(lm.1)
```

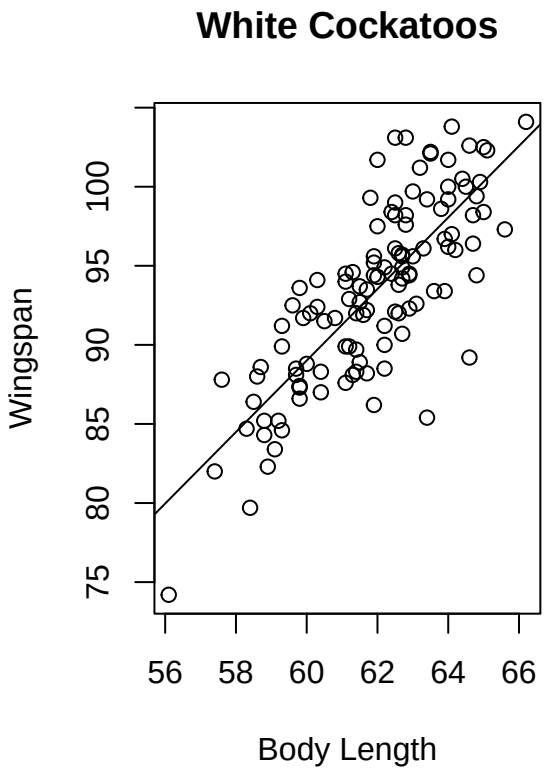
```
##
## Call:
## lm(formula = wingspan ~ bodylength, data = cockatoo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3537  -4.1472  -0.2777   3.7130  17.7411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.49105    1.92151   14.31   <2e-16 ***
## bodylength   1.07235    0.03676   29.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.156 on 232 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7848
## F-statistic: 850.8 on 1 and 232 DF,  p-value: < 2.2e-16
```

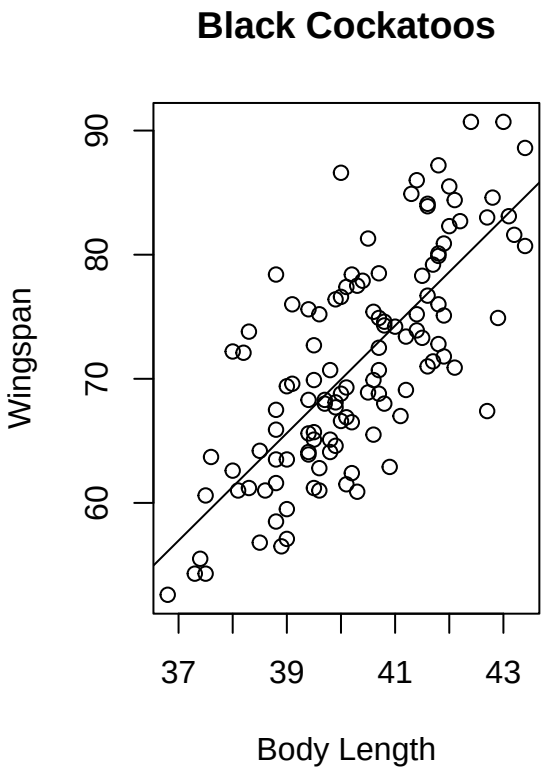
```
par(mfrow=c(1,2))
plot(lm.1, which=1:2)
```



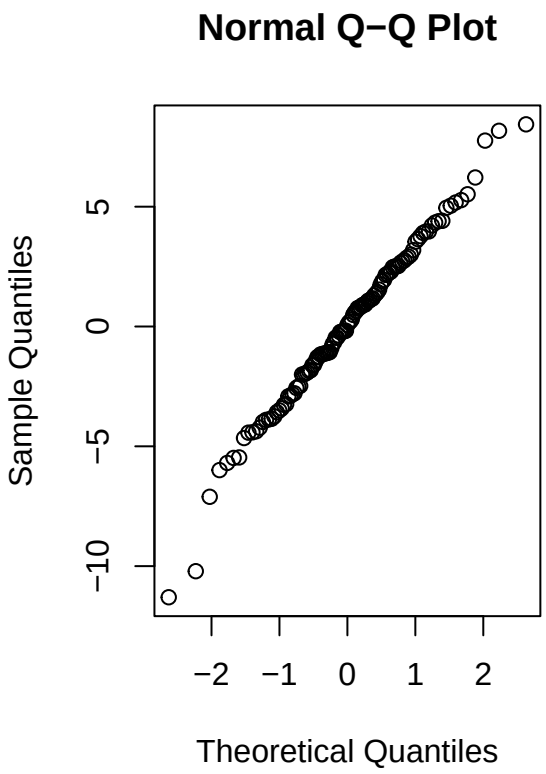
Split the Dataset Based on Colour

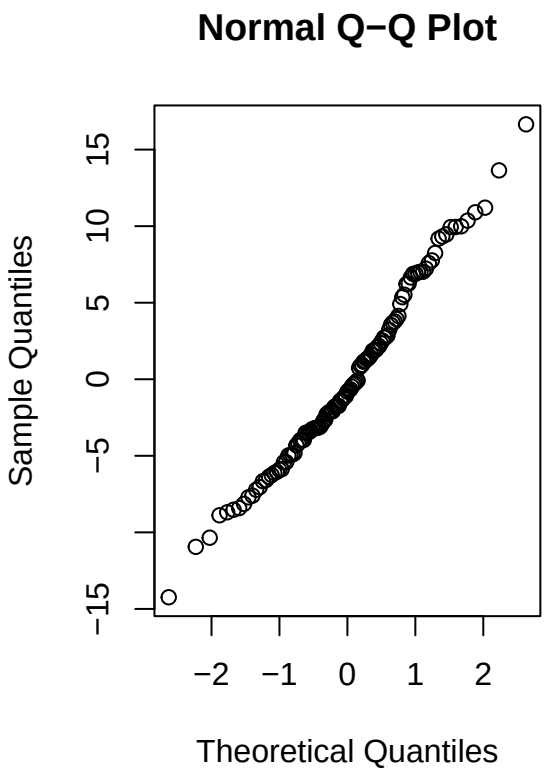
```
white = cockatoo[cockatoo$colour == 'white',]
black = cockatoo[cockatoo$colour == 'black',]

lm.white = lm(wingspan~bodylength, data=white)
lm.black = lm(wingspan~bodylength, data=black)
```






```
par(mfrow=c(1,2))
qqnorm(lm.white$residuals)
qqnorm(lm.black$residuals)
```





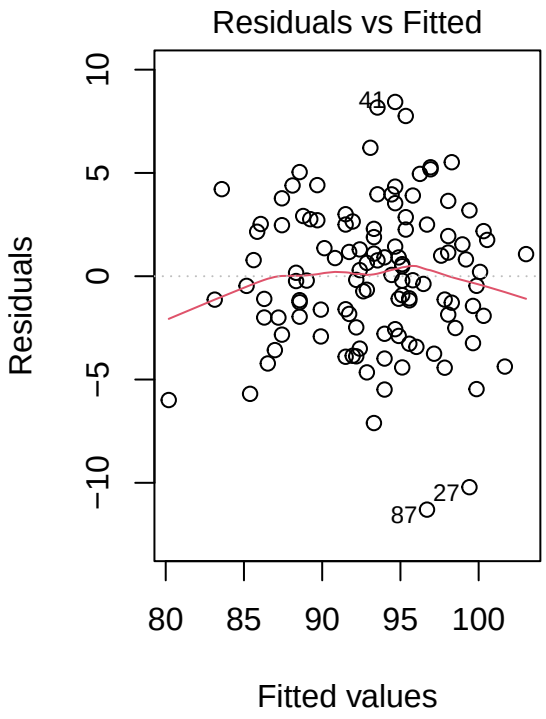
Before conducting the test, we need to check **assumptions**:

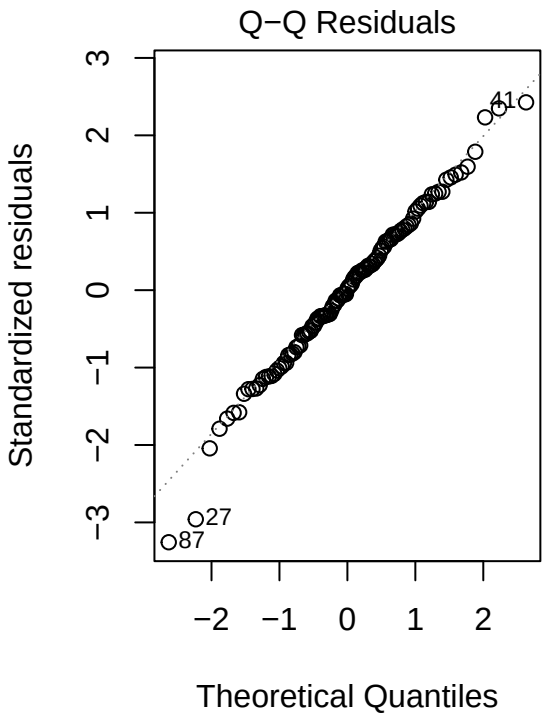
1. **Linearity:** From the scatter plot, we can see some linear relationship between 'wingspan' and 'bodylength' for both groups.
2. **Residuals close to normally distributed:** QQ-plot for both groups show normality in their distribution. There seems a few outliers in the white cockatoo dataset, however it still shows patterns close to normality.

Simple Linear Regression for White Cockatoos:

There seem to be some outliers, we will remove them and fit the linear model.

```
par(mfrow=c(1,2))
plot(lm.white, which=1:2)
```






```
(badpoint = which(lm.white$residuals < -9))
```

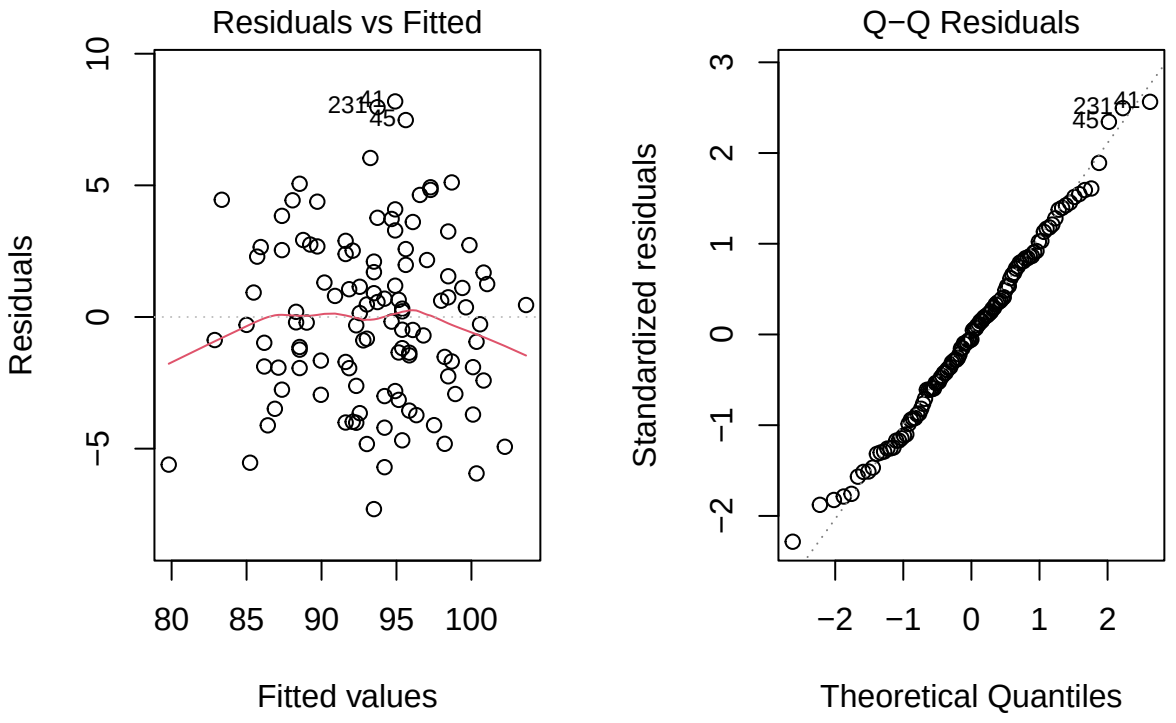
```
## 27 87
## 15 47
```

```
clean.white = white[-badpoint,]

lm.clean.white = lm(wingspan~bodylength, data=clean.white)

summary(lm.clean.white)
```

```
##
## Call:
## lm(formula = wingspan ~ bodylength, data = clean.white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2963 -2.0984 -0.1764  2.3422  8.1876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.6010     9.3210  -5.643 1.25e-07 ***
## bodylength    2.3602     0.1506  15.676  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.208 on 113 degrees of freedom
## Multiple R-squared:  0.685,  Adjusted R-squared:  0.6822
## F-statistic: 245.7 on 1 and 113 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
plot(lm.clean.white, which=1:2)
```



Even though the adjusted r-squared score is not good compared to the previous model, we now have β_0 and β_1 specifically for the White Cockatoos whose bodylength ranging from 55 to 67. That is:

$$\hat{y}_{white} = 2.3602 * X - 52.6010 | X \in (55, 67)$$

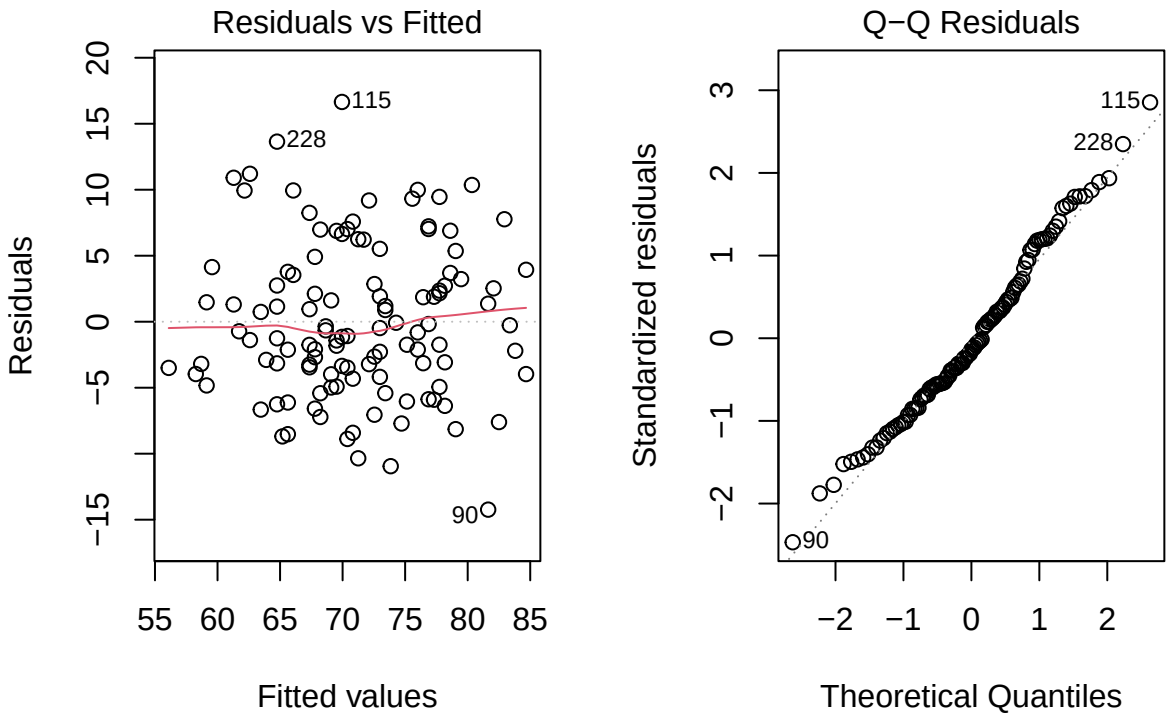
Which means that if there is a unit increase in bodylength, the wingspan of white cockatoo increases by about 2.3

Now we check black cockatoo group.

Simple Linear Regression for Black Cockatoos:

There is no. extreme outlier, we do not need to clean the dataset.

```
par(mfrow=c(1,2))
plot(lm.black, which=1:2)
```




```
summary(lm.black)
```

```
##
## Call:
## lm(formula = wingspan ~ bodylength, data = black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2373  -3.9672  -0.8103   3.6926  16.6495
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -103.1876    14.7711  -6.986 1.96e-10 ***
## bodylength     4.3285     0.3663  11.818  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.861 on 115 degrees of freedom
## Multiple R-squared:  0.5484, Adjusted R-squared:  0.5445
## F-statistic: 139.7 on 1 and 115 DF,  p-value: < 2.2e-16
```

We now have β_0 and β_1 specifically for the Black Cockatoos whose bodylength ranging from 36 to 44.

$$\hat{y}_{black} = 4.3285 * X - 103.1876 | X \in (36, 44)$$

Which means that if there is a unit increase in black cockatoo's bodylength, their wingspan increases by about 4.3, which is nearly double of white cockatoo's unit increase.

Conclusion:

As suspected, these two groups have different slopes as well as bodylength ranges. Even though the goodness-of-fit drops when dividing the dataset by colour, it is more appropriate to do the analysis in this manner. Now we have more detailed relationship between wingspan and bodylength, that is:

- $\hat{y}_{white} = 2.3602 * X - 52.6010 | X \in (55, 67)$
- $\hat{y}_{black} = 4.3285 * X - 103.1876 | X \in (36, 44)$
- ***Wingspan increases about 2.3*** units when Bodylength increases 1 unit for ***White Cockatoos***
- ***Wingspan increases about 4.3*** units when Bodylength increases 1 unit for ***Black Cockatoos***
- There is nearly double unit increase difference between black and white cockatoos.
- ***White Cockatoos' body length*** is raging from ***55 to 67*** .
- Meanwhile, ***Balck Cockatoos' body length*** is raging from ***36 to 44*** .