# Statistical Modelling Tech. HW-4

Ceren Yıldız

2025-11-20

As is known here, the basic principles of experimentation are a frequently used expression and concept in experimental design. It is based on the rules that must be followed to conduct an experiment in a correct,valid and repeatable manner. Therefore, it is very important.Therefore, I will include these principles in my homework.

**Principles of the Experiment:**

1. **Blocking**

2. **Randomization**

3. **Replication**

- **Randomization:** It's clear that in experimental design, experimental units are desired to be as homogeneous as possible. Furthermore, inherent differences between experimental units always exist. It's also clear that these differences are random. To fully define randomness, we can define it as the condition of randomly assigning experimental units to groups. If trials are not randomly assigned to experimental units, the differences between trial effects and the estimates of error variance will be biased. This is undesirable. Furthermore, this rule of randomization ultimately ensures that experimental units have equal probabilities of being assigned to trials.
- **Blocking:**In order to increase the sensitivity of the experiment, experimental units with systematic differences between them can be divided into groups called blocks, which are homogeneous within themselves and heterogeneous among themselves.
- **Replication:**It is the number of experimental units to which each possible trial is applied.

**After making the basic explanations, we can start ANOVA.**

**ANOVA(ANalysis of VAriance)** is a statistical test to determine whether two or more population means are different.In other words, it is used to compare two or more groups to see if they are significantly different.Also,if you want to understand how many ways ANOVA is used, you must determine how many independent (factor) variables are included in the model. For example, if there is only one independent variable, it is possible to say that it is a one-way ANOVA. The important point here is not to include interaction terms.

### Assumptions of ANOVA

- **1. Normality**
- **2. Homogeneity of variance**
- **3. Independence**

```
library(ggplot2)
library(tidyr)
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
```

```
set.seed(123)
group1 <- rnorm(30, mean = 15, sd = 3)
group2 <- rnorm(30, mean = 18, sd = 3)
group3 <- rnorm(30, mean = 16, sd = 3)

data <- data.frame(
value = c(group1, group2, group3),
group = factor(rep(c("Group A", "Group B", "Group C"), each = 30))
)
```

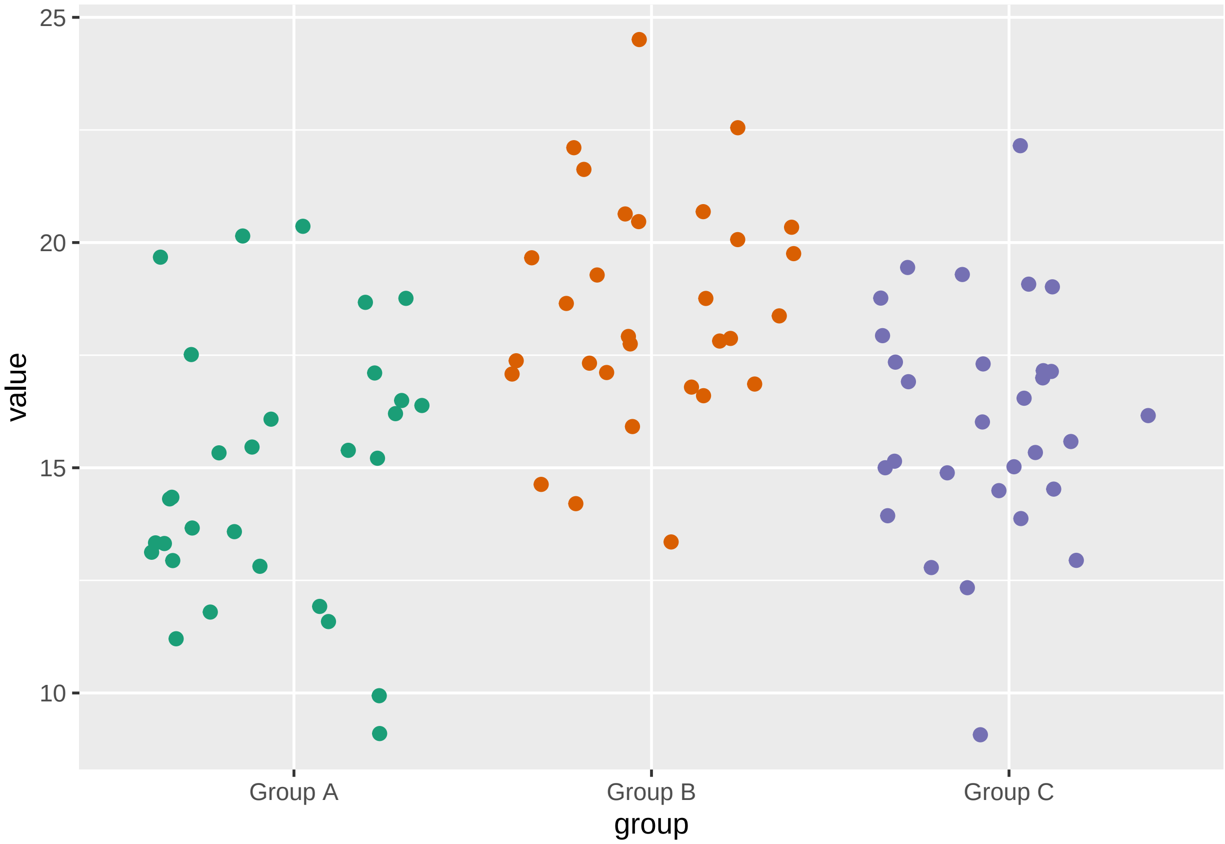
```
# EXPLORATORY DATA ANALYSIS
summary(data)
```

```
##      value            group
##  Min.   : 9.072   Group A:30
##  1st Qu.:14.383   Group B:30
##  Median :16.696   Group C:30
##  Mean   :16.489
##  3rd Qu.:18.738
##  Max.   :24.507
```

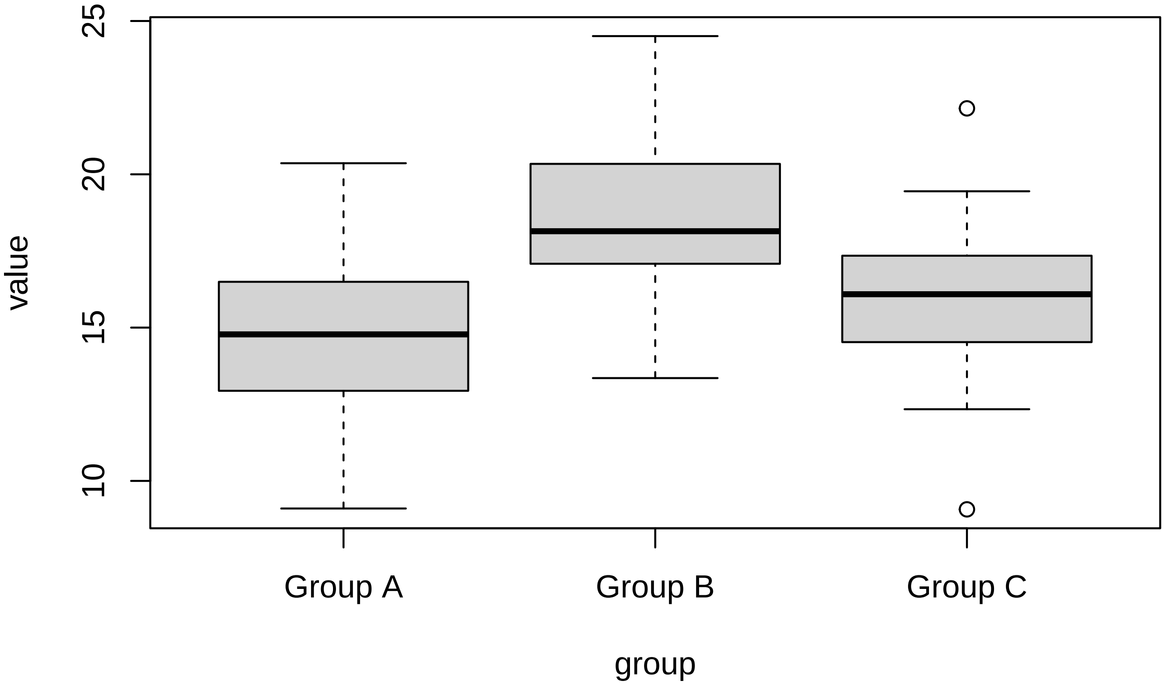
```
str(data)
```

```
## 'data.frame':    90 obs. of  2 variables:
##  $ value: num  13.3 14.3 19.7 15.2 15.4 ...
##  $ group: Factor w/ 3 levels "Group A","Group B",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
ggplot(data) +
  aes(x = group, y = value, color = group) +
  geom_jitter(size = 2) +scale_color_manual(values = c("#1B9E77", "#D95F02", "#7570B3")) +
  theme(legend.position = "none")
```



```
boxplot(value ~ group,
data = data
)
```



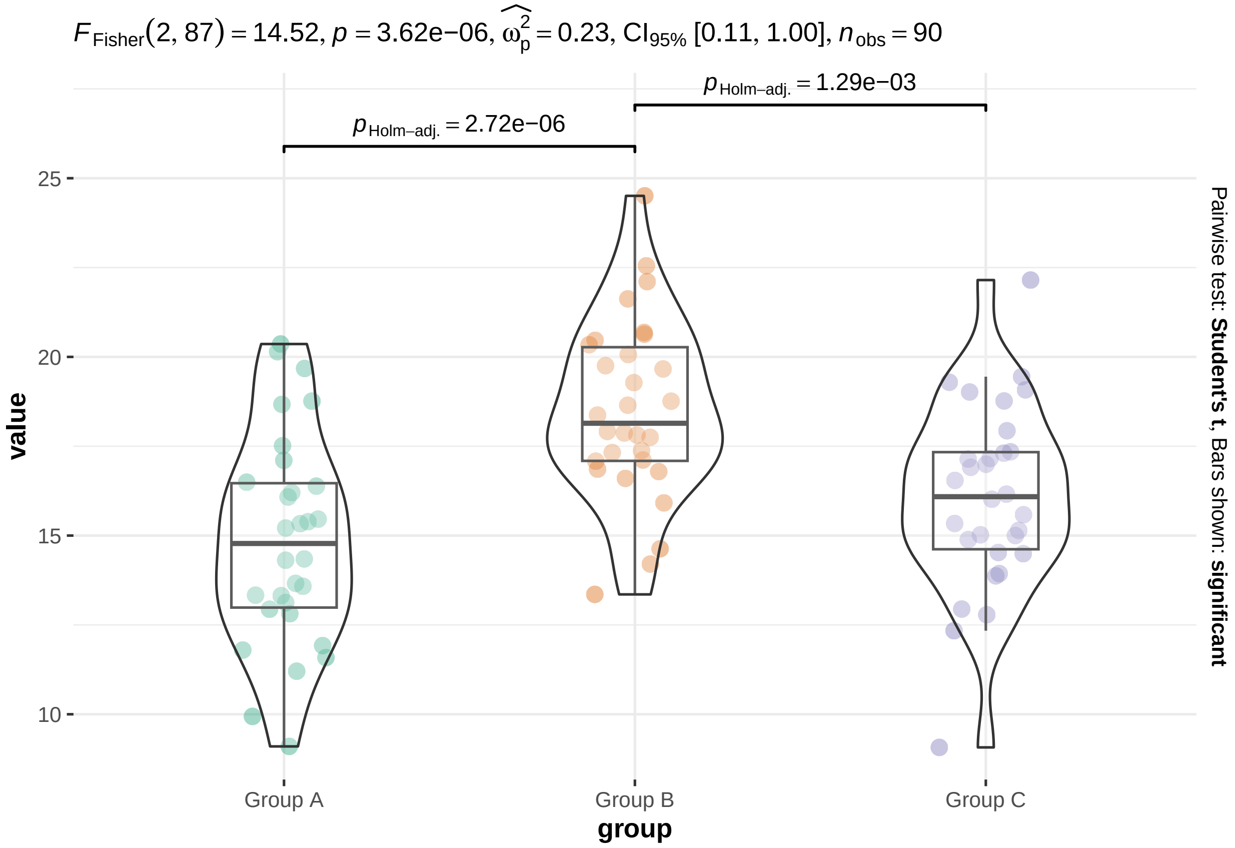
```
install.packages("ggstatsplot")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

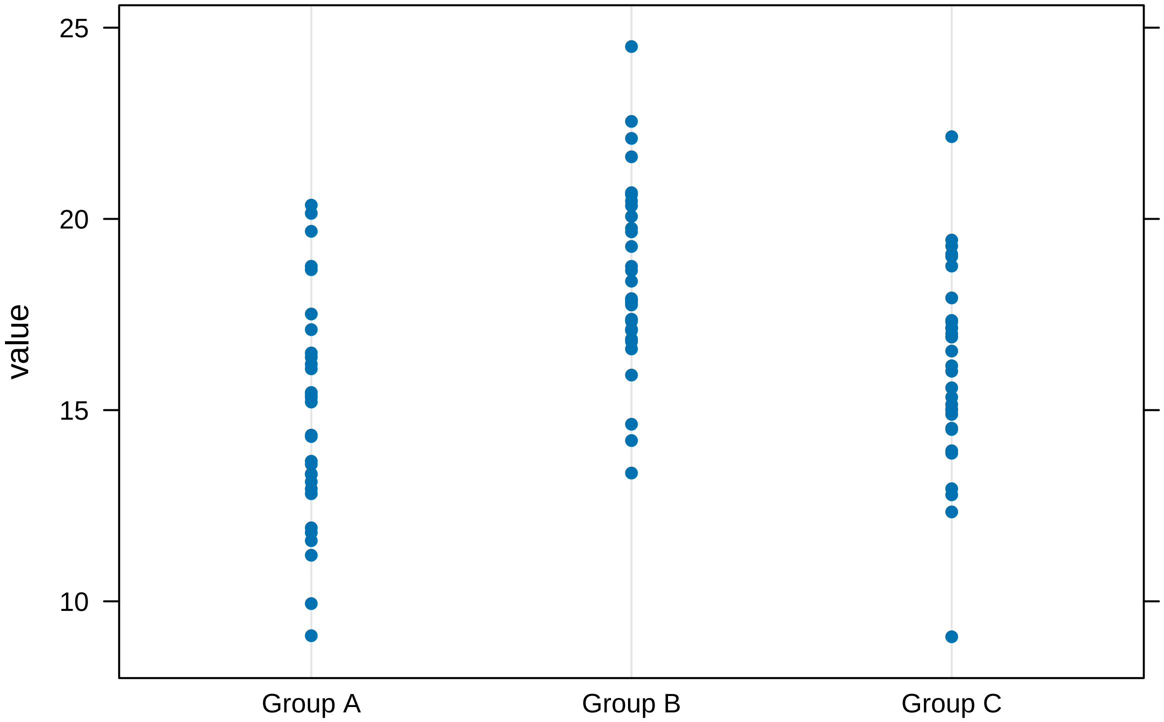
```
library(ggstatsplot)
```

```
## You can cite this package as:
##      Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.
##      Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167
```

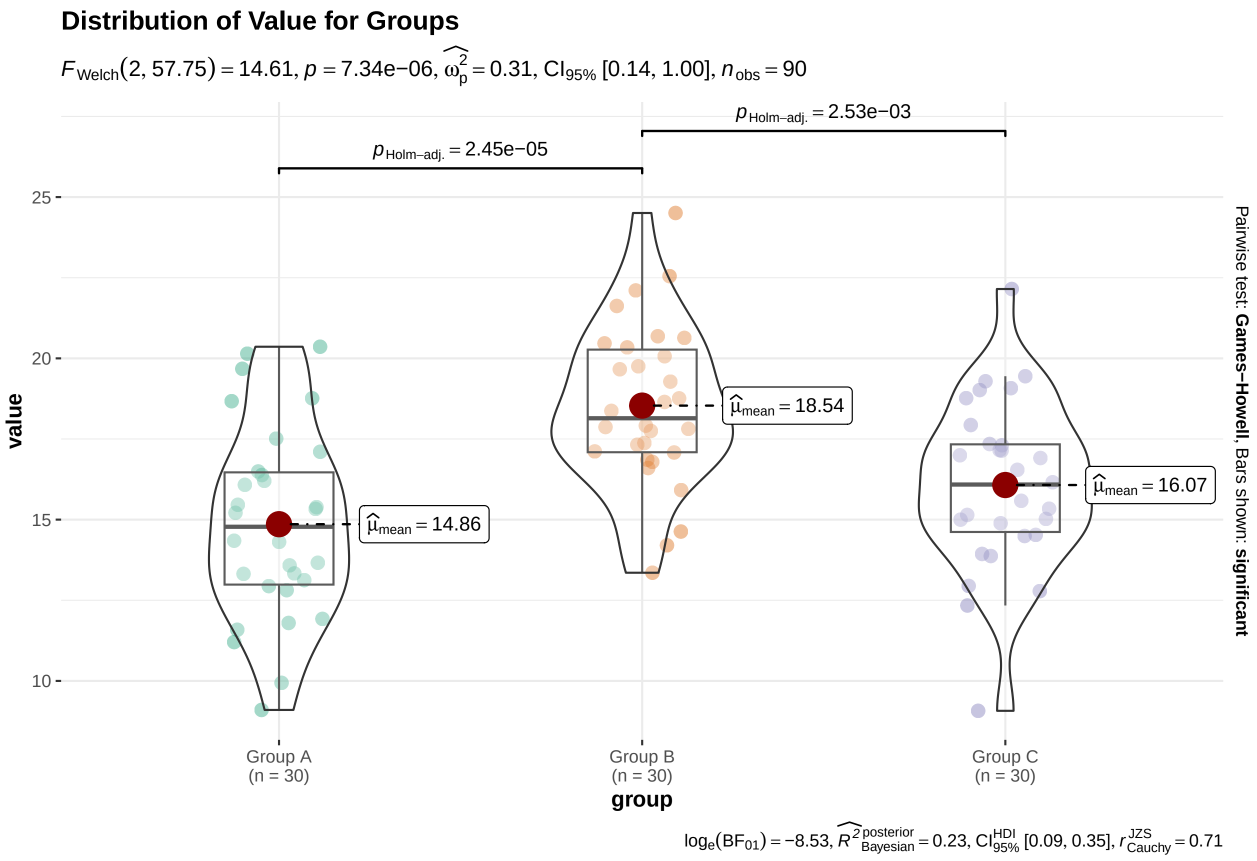
```
ggbetweenstats(
data = data,
x = group,
y = value,
type = "parametric", # ANOVA or Kruskal-Wallis
#The non-parametric equivalent of one-way ANOVA is Kruskal-Wallis, which is what I wanted to explain he
var.equal = TRUE, # ANOVA or Welch ANOVA
plot.type = "box",
pairwise.comparisons = TRUE,
pairwise.display = "significant",
centrality.plotting = FALSE,
bf.message = FALSE
)
```



```
# Dotplot
library("lattice")
dotplot(value ~ group,
data = data
)
```



```
set.seed(123)
ggbetweenstats(
data = data,
x = group,
y =        value,
title = "Distribution of Value for Groups"
)
```



**NOTE :The Welch ANOVA test is a type of ANOVA used when the variance between groups is unequal. It is used in conditions where equal variance, one of the classical ANOVA assumptions, is not met. I wanted to use this test to see how much of a difference it made even though the classical assumptions were met. It's safe to say that the Welch test is more cautious and has slightly less power.**

**Analysis of Covariance (ANCOVA)** is a technique that blends ANOVA and regression analysis. ANCOVA is named according to the number of covariates and whether the relationship between the dependent variable and the dependent variable is linear. For example, if there is only one covariate and the relationship between this variable and the dependent variable is linear, the analysis is called simple ANCOVA. However,it is possible to say that there are also types of curvilinear and multiple ANCOVA. To avoid digressing, I will not discuss curvilinear and multiple ANCOVA. Furthermore, as is known, there are certain assumptions in the ANCOVA section, just as there are in the ANOVA section.Let's start by generating data through simulation. Our teacher has asked us to perform regression and create data suitable for ANOVA and ANCOVA. To better understand the topic, I will first create data suitable for ANOVA. Then, I will perform the same steps for ANCOVA. After completing all of these, I will simulate only one data set and analyze the regression, ANOVA, and ANCOVA with a single data set. I believe it will be explanatory because it's a detailed assignment.

**Step 1: Simulating data in R via set.seed(123).**

```r
# Simulated data with set.seed(123)
set.seed(123)
group1 <- rnorm(30, mean = 15, sd = 3)
group2 <- rnorm(30, mean = 18, sd = 3)
group3 <- rnorm(30, mean = 16, sd = 3)
data <- data.frame(
value = c(group1, group2, group3),
group = factor(rep(c("Group A", "Group B", "Group C"), each = 30))
)
# One-way ANOVA
anova_result <- aov(value ~ group, data = data)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## group        2  210.5  105.25   14.52 3.62e-06 ***
## Residuals   87  630.7    7.25                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anova_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = value ~ group, data = data)
## 
## $group
##                       diff        lwr        upr     p adj
## Group B-Group A  3.676326  2.0186752  5.3339774 0.0000027
## Group C-Group A  1.214572 -0.4430786  2.8722235 0.1937346
## Group C-Group B -2.461754 -4.1194049 -0.8041028 0.0018428
```

First of all,we can do the Shapiro-Wilk Normality Test for normality, which is one of the assumptions of ANOVA. Let's start by hypothesizing.

- $H_0$:The data set has a normal distribution
- $H_1$: The data set is not normal distribution.

**Step 2:Application of normality test.**

```
shapiro.test(data$value)
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  data$value
## W = 0.99505, p-value = 0.9849
```

Since $p = 0.9849 > 0.05$, we fail to reject the null hypothesis of normality. This indicates that the data do not significantly differ from a normal distribution. In other words, the distribution can be considered approximately normal.

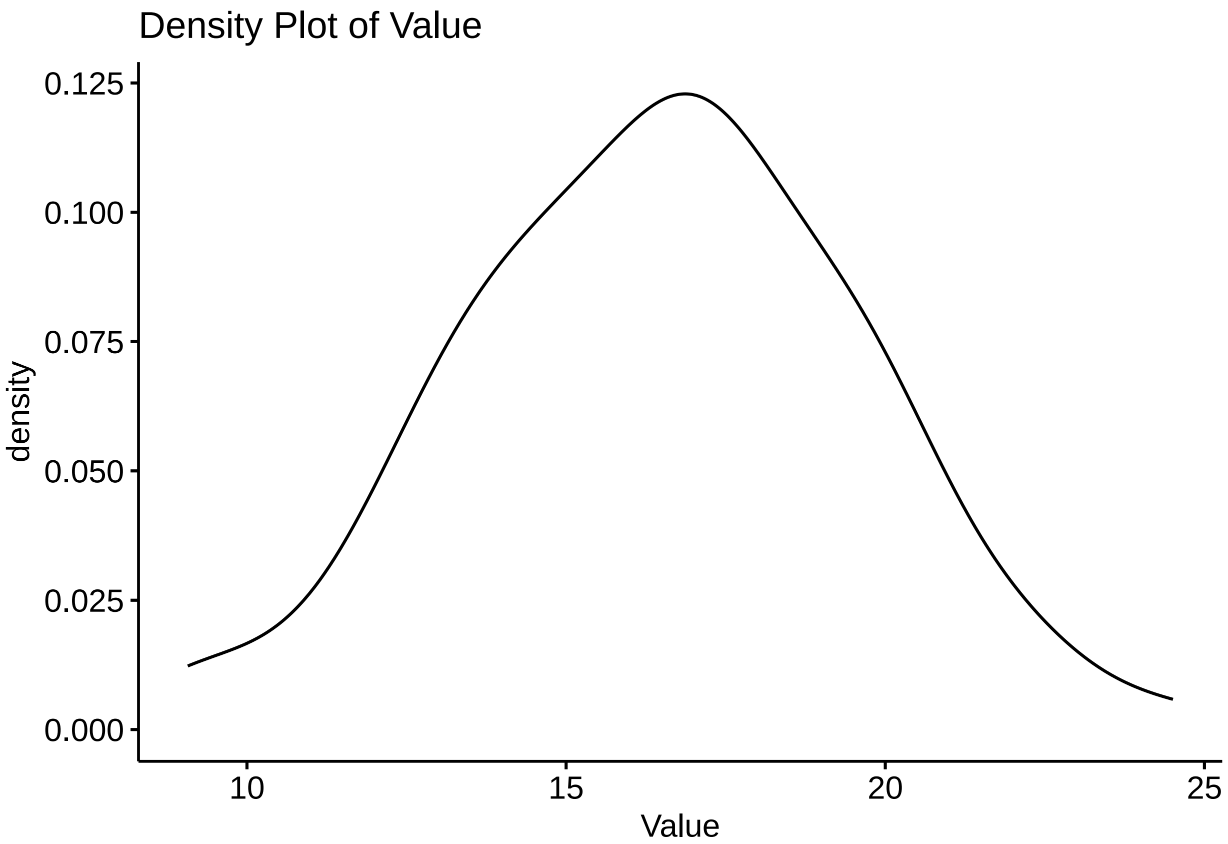
Also, we can look at visual methods such as q-q plot and density plot.

**Step 3:Additionally, application of visual methods with Q-Q Plot and density graph.**

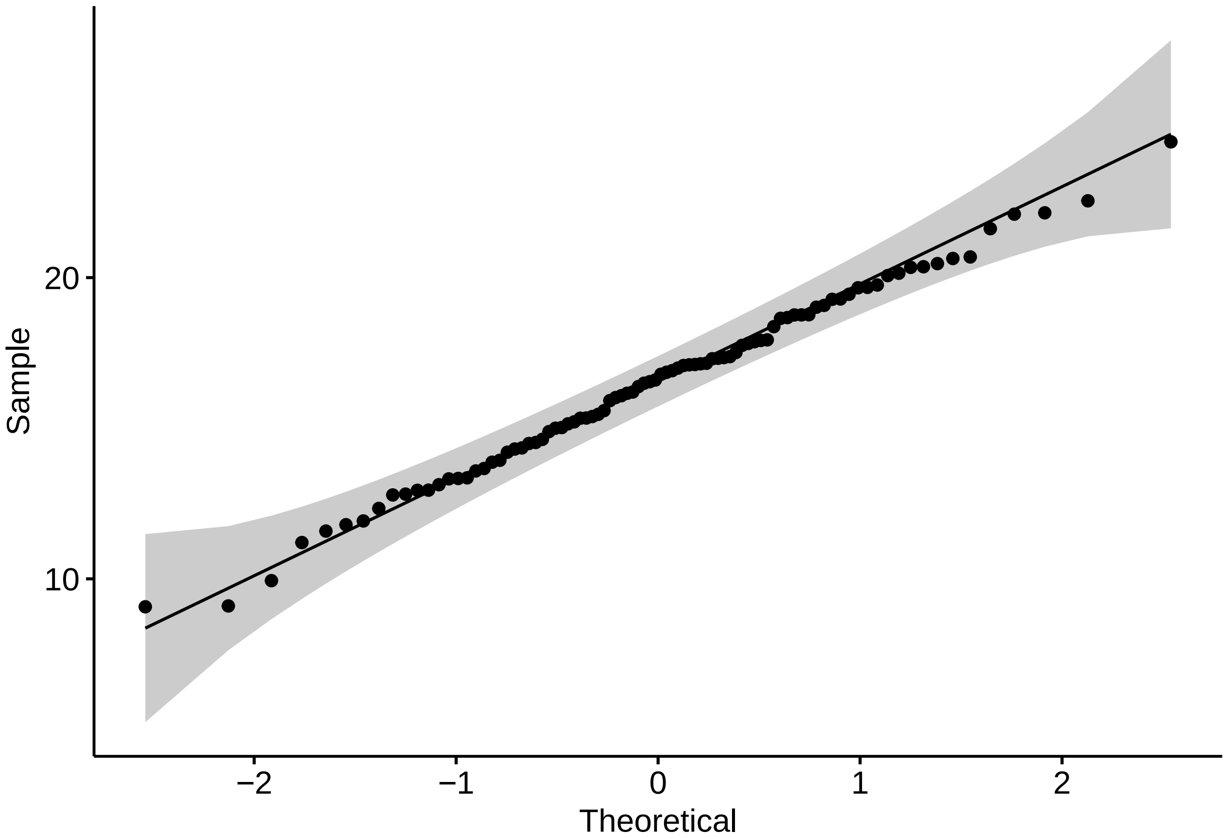
```
install.packages("ggpubr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```
library("ggpubr")
ggdensity(data$value,
main = "Density Plot of Value",
xlab = "Value")
```



```r
ggqqplot(data$value)
```



Now that we have examined the establishment of normality, we can proceed with the homogeneity of variances.In an experimental design, in hypothesis tests regarding the equality of trial means, it is assumed that the variances are homogeneous. The hypotheses for the test of equality of variances are given as follows.

**Hypothesis:**

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

$$H_1 : \sigma_i^2 \neq \sigma_j^2 \quad \text{for at least one pair}$$

There are several statistics for testing this hypothesis. Some examples are: i) Bartlett, ii) Cochran, iii)Levenes, and iv) Hartley. I will proceed with Bartlett here because it is the most commonly used test. In this testing process, random samples are assumed to come from independent populations. Furthermore,the sampling distribution of the test statistic used here has a chi-square distribution with (k-1) degrees of freedom.

**NOTE**:In this test, the number of observations in the trials does not need to be equal.( for Bartlett)

**Step 4:Application of the homogeneity of variances test.**

```r
bartlett.test(data$value, data$group)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data$value and data$group
## Bartlett's K-squared = 0.81811, df = 2, p-value = 0.6643
```

As a result of the test, we looked at the Bartlett result and saw that p = 0.6643>0.05, so we cannot reject $H_0$ and take one more step for ANOVA.We have looked and examined all the other conditions.Therefore, we can move on to the ANOVA test.

**Step 5:After proving the 3 assumptions, proceed to the ANOVA step.**

```
# One-way ANOVA
anova_result <- aov(value ~ group, data = data)
summary(anova_result)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## group        2  210.5  105.25   14.52 3.62e-06 ***
## Residuals   87  630.7    7.25                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We know from the Statistics 1 course that the p value of our tests is important and we check the p value. Approximately the p value was found to be 3.62e-06 . Therefore, since we work according to the 0.05(5%) significance level, we reach a result of p>alpha and reject the $H_0$(null) hypothesis.

- $\boldsymbol{H_0 : \mu_1 = \mu_2 = \mu_3 = \cdot \quad \cdot \quad \cdot = \mu_k}$
- ***$H_1$ : Means are not all equal.***

### Conclusion(Post Hoc. Comparison Test)

If we need to comment on the ratios in the foods as a result of the ANOVA test, we can say that the ratios are different between at least two treatments (when significance is 5%).We tested that there was a significant difference as a result of ANOVA. But we do not have enough information about which groups it is. Therefore, Post Hoc. We can apply one of the Comparison Tests. I will apply Tukey.

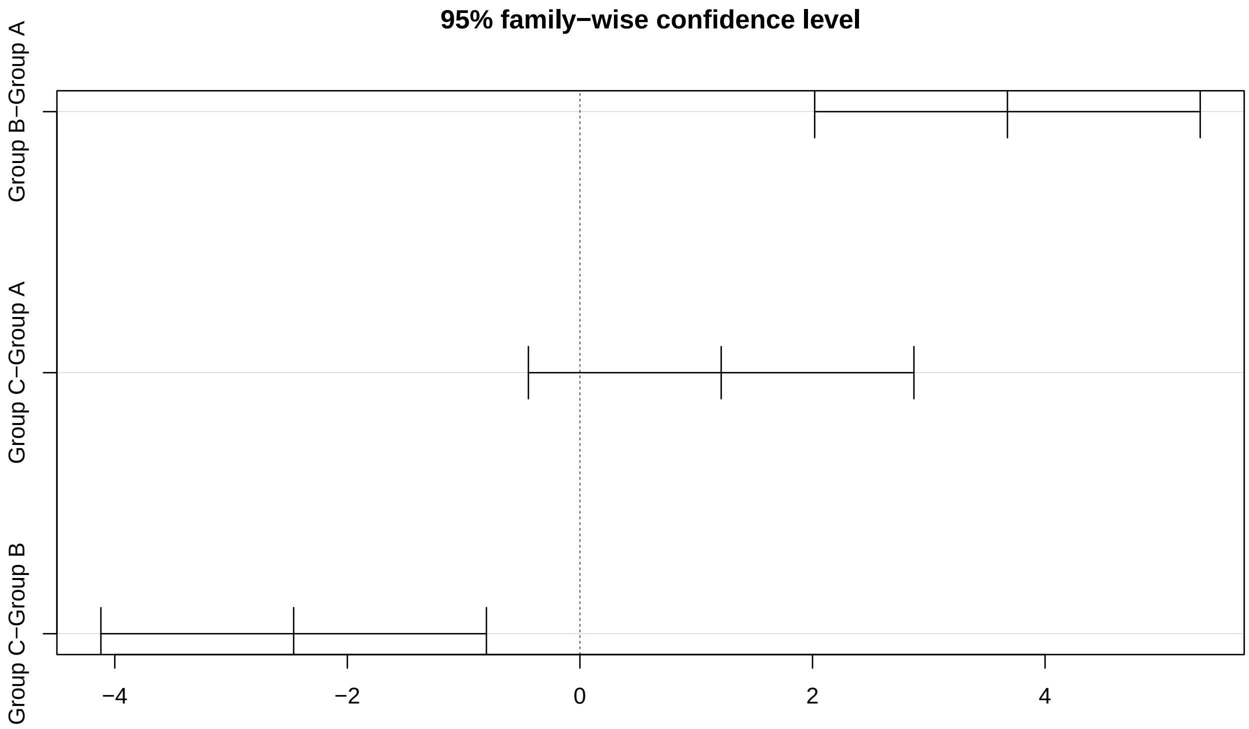
**Step 6: Application of Post Hoc. paired comparison tests.**

The reason for doing a Post Hoc test after an ANOVA is that an ANOVA only tells us that there is at least one difference between groups; it does not tell us which groups the difference is between.Post hoc tests evaluate all groups in pairwise comparisons.

```
TukeyHSD(anova_result)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = value ~ group, data = data)
##
## $group
##                       diff        lwr        upr     p adj
## Group B-Group A  3.676326  2.0186752  5.3339774 0.0000027
## Group C-Group A  1.214572 -0.4430786  2.8722235 0.1937346
## Group C-Group B -2.461754 -4.1194049 -0.8041028 0.0018428
```

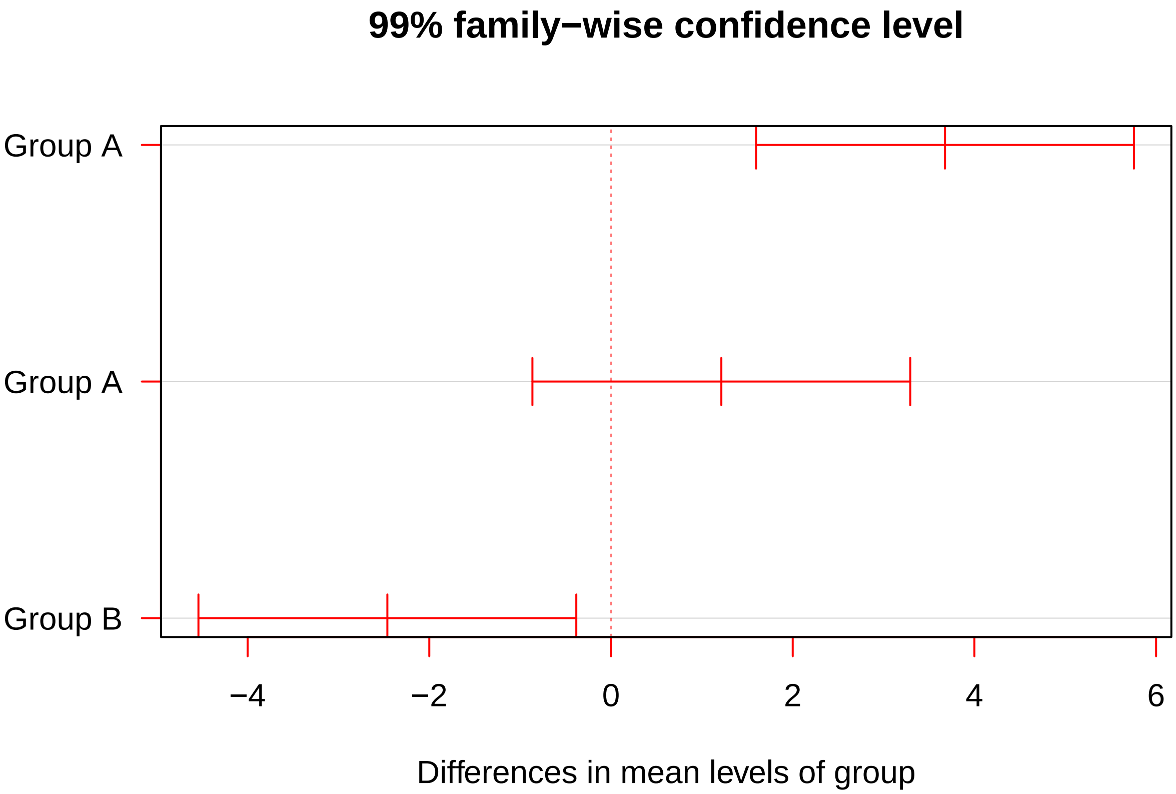
```
plot(TukeyHSD(anova_result))
```



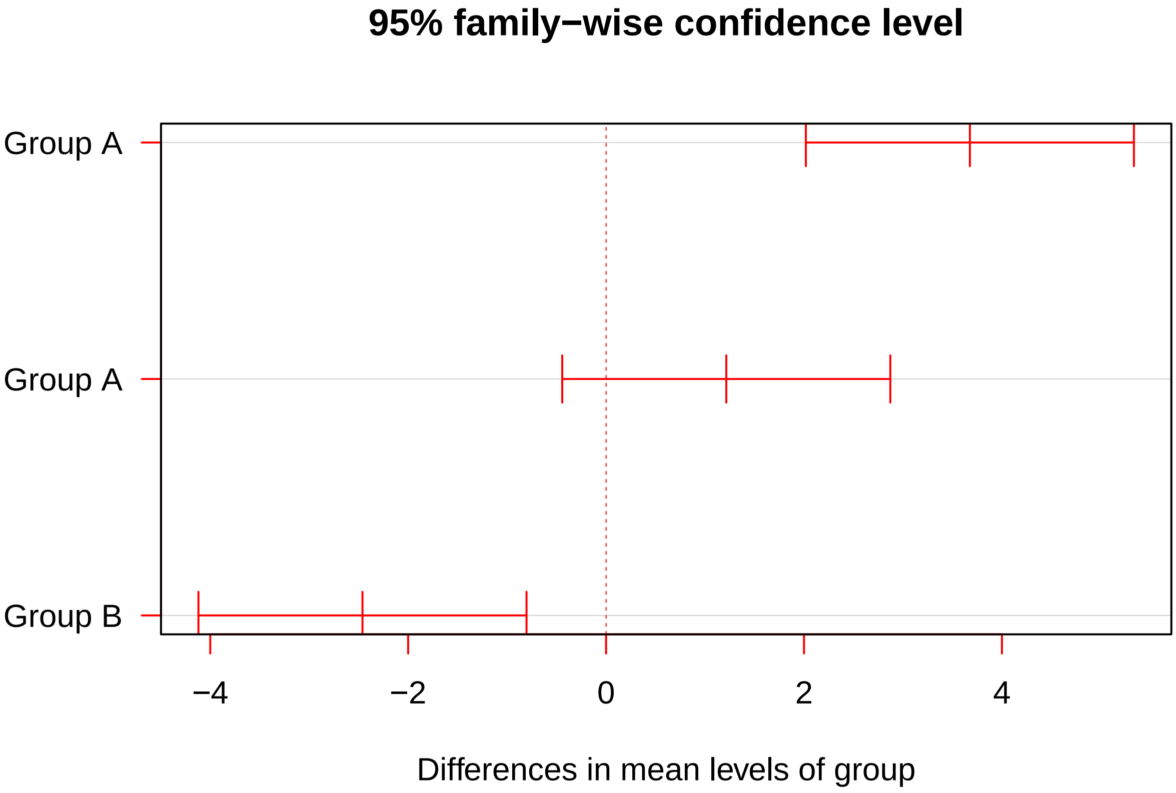
```
install.packages("agricolae")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```
library(agricolae)
plot(TukeyHSD(anova_result, conf.level = 0.99),las=1, col = "red")
```



```
plot(TukeyHSD(anova_result, conf.level = 0.95),las=1, col = "red")
```



```r
# We can look at different confidence level.
```

**NOTE:One might question why we didn't use a t-test instead of a Post Hoc test. This is because making too many pairwise comparisons increases the likelihood of a Type I error.While the likelihood of finding a false difference with a t-test increases exponentially, it's clear that we found a solution to this problem with a Post Hoc test. Post Hoc is more convenient than a t-test because it controls for Type I error when making multiple comparisons.**

**It is not possible to perform ANCOVA with the data I generated through simulation above. Only statistical analyses such as ANOVA can be performed.**

The ANCOVA has similar assumptions to the ANOVA.

### Assumptions of ANCOVA

- The **dependent variable and the covariate(s) are continuous** while the **independent variable is categorical.**
- The observations are **independent** and **randomly sampled.**
- There are no **outliers** or **points with high leverage.**
- The errors (residuals) are **approximately normally distributed** and have **equal variances.**
- If using more than one covariate, **no covariates should be highly correlated with another**

### Real Data Set

`Penguins Data` for 344 penguins, with size measurements, clutch observations and blood isotope ratios from three penguin species observed near Palmer Station, Antarctica.First of all, I would like to begin by applying the principles of the experiment to this data set.

- **Blocking:Penguins create blocks according to their species. (Adelie, Gentoo, Chinstrap)**
- **Randomization:Randomness makes the sample more representative of the population.Palmer Penguins is an observational study, there is no full experimental randomization, but there is randomness in the sampling process.**
- **Replication:Measuring hundreds of penguins of each species is an example of the repeatability step.**

```r
data("penguins")
head(penguins)
```

```
##   species    island bill_len bill_dep flipper_len body_mass    sex year
## 1  Adelie Torgersen     39.1     18.7         181      3750   male 2007
## 2  Adelie Torgersen     39.5     17.4         186      3800 female 2007
## 3  Adelie Torgersen     40.3     18.0         195      3250 female 2007
## 4  Adelie Torgersen       NA       NA          NA        NA   <NA> 2007
## 5  Adelie Torgersen     36.7     19.3         193      3450 female 2007
## 6  Adelie Torgersen     39.3     20.6         190      3650   male 2007
```

```r
colSums(is.na(penguins))
```

```
##     species      island    bill_len    bill_dep flipper_len   body_mass 
##           0           0           2           2           2           2 
##         sex        year 
##          11           0
```

```r
penguins_clean<-na.omit(penguins)
summary(penguins_clean)
```

```
##       species          island       bill_len        bill_dep      flipper_len
##  Adelie   :146   Biscoe   :163   Min.   :32.10   Min.   :13.10   Min.   :172
##  Chinstrap: 68   Dream    :123   1st Qu.:39.50   1st Qu.:15.60   1st Qu.:190
##  Gentoo   :119   Torgersen: 47   Median :44.50   Median :17.30   Median :197
##                                  Mean   :43.99   Mean   :17.16   Mean   :201
##                                  3rd Qu.:48.60   3rd Qu.:18.70   3rd Qu.:213
##                                  Max.   :59.60   Max.   :21.50   Max.   :231
##    body_mass        sex           year
##  Min.   :2700   female:165   Min.   :2007
##  1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :4050                Median :2008
##  Mean   :4207                Mean   :2008
##  3rd Qu.:4775                3rd Qu.:2009
##  Max.   :6300                Max.   :2009
```

```
head(penguins_clean)
```

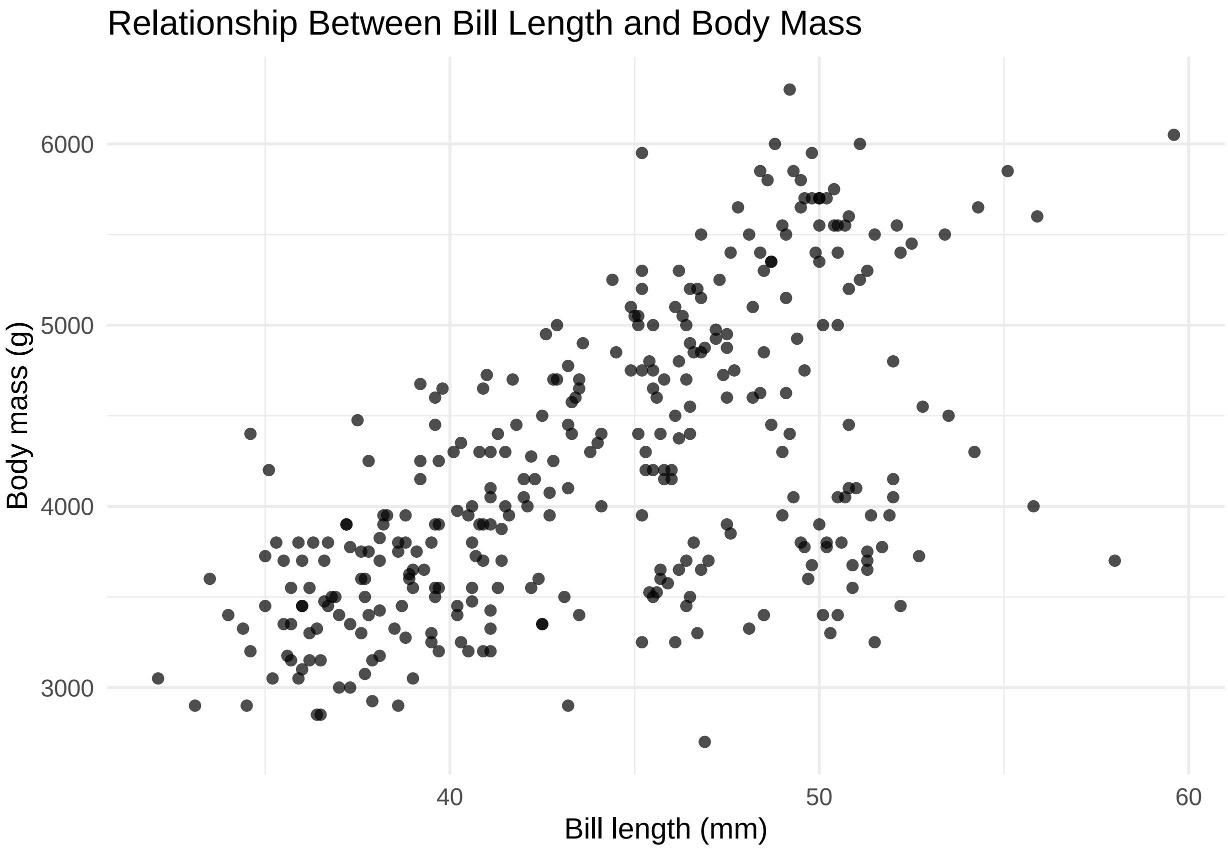
```
##   species    island bill_len bill_dep flipper_len body_mass    sex year
## 1  Adelie Torgersen     39.1     18.7         181      3750   male 2007
## 2  Adelie Torgersen     39.5     17.4         186      3800 female 2007
## 3  Adelie Torgersen     40.3     18.0         195      3250 female 2007
## 5  Adelie Torgersen     36.7     19.3         193      3450 female 2007
## 6  Adelie Torgersen     39.3     20.6         190      3650   male 2007
## 7  Adelie Torgersen     38.9     17.8         181      3625 female 2007
```

```
str(penguins_clean)
```

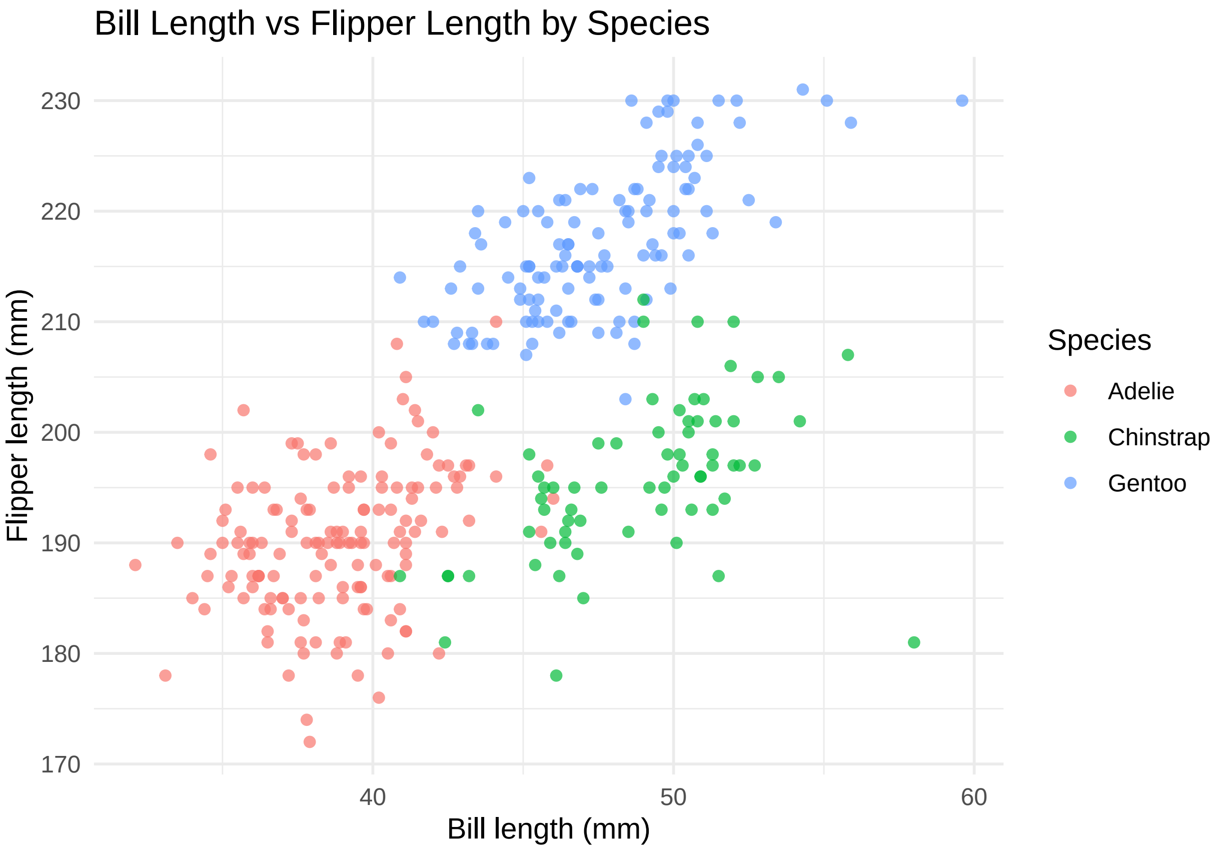
```
## 'data.frame':    333 obs. of  8 variables:
##  $ species    : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island     : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_len   : num  39.1 39.5 40.3 36.7 39.3 38.9 39.2 41.1 38.6 34.6 ...
##  $ bill_dep   : num  18.7 17.4 18 19.3 20.6 17.8 19.6 17.6 21.2 21.1 ...
##  $ flipper_len: int  181 186 195 193 190 181 195 182 191 198 ...
##  $ body_mass  : int  3750 3800 3250 3450 3650 3625 4675 3200 3800 4400 ...
##  $ sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 2 1 2 2 ...
##  $ year       : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
##  - attr(*, "na.action")= 'omit' Named int [1:11] 4 9 10 11 12 48 179 219 257 269 ...
##   ..- attr(*, "names")= chr [1:11] "4" "9" "10" "11" ...
```

## EXPLORATORY DATA ANALYSIS

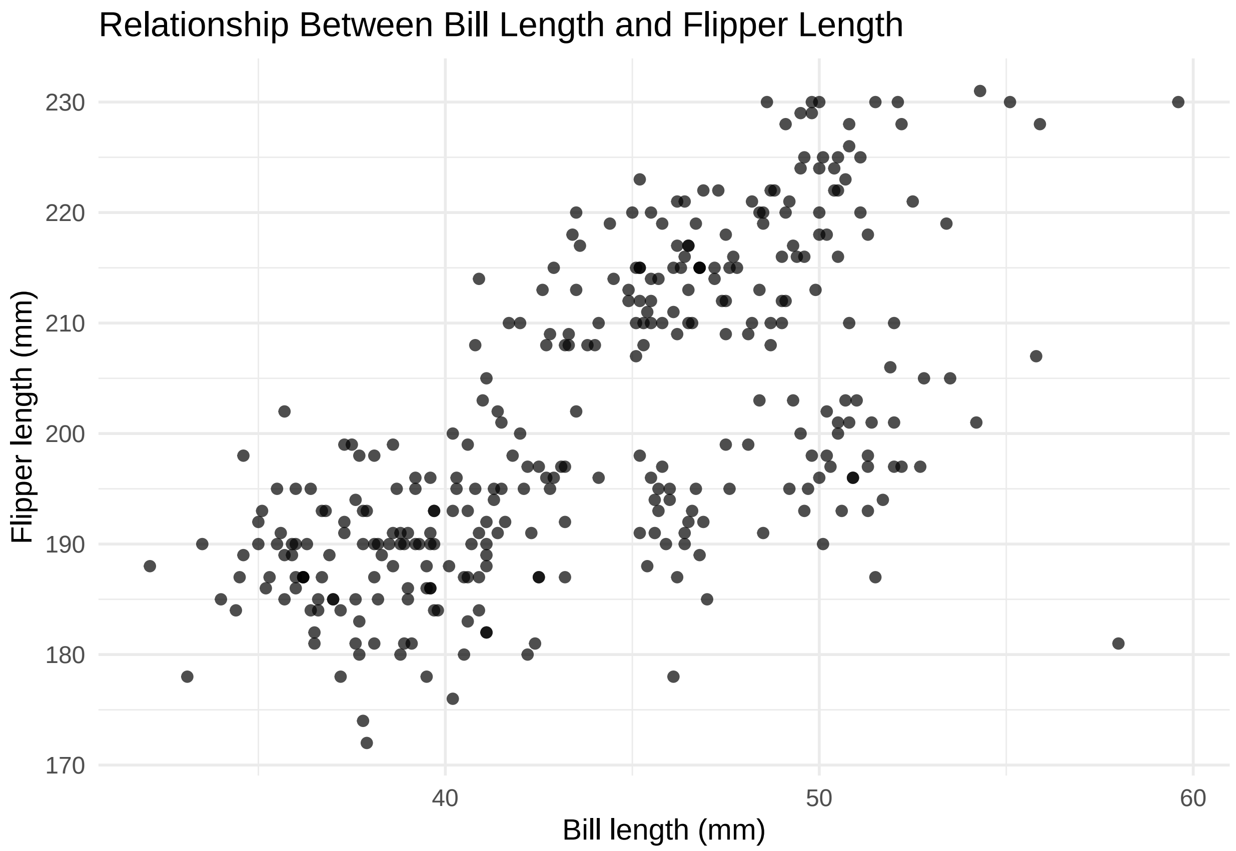
```
ggplot(penguins_clean,
       aes(x = bill_len, y = body_mass)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Bill length (mm)",
    y = "Body mass (g)",
    title = "Relationship Between Bill Length and Body Mass"
  ) +
  theme_minimal()
```



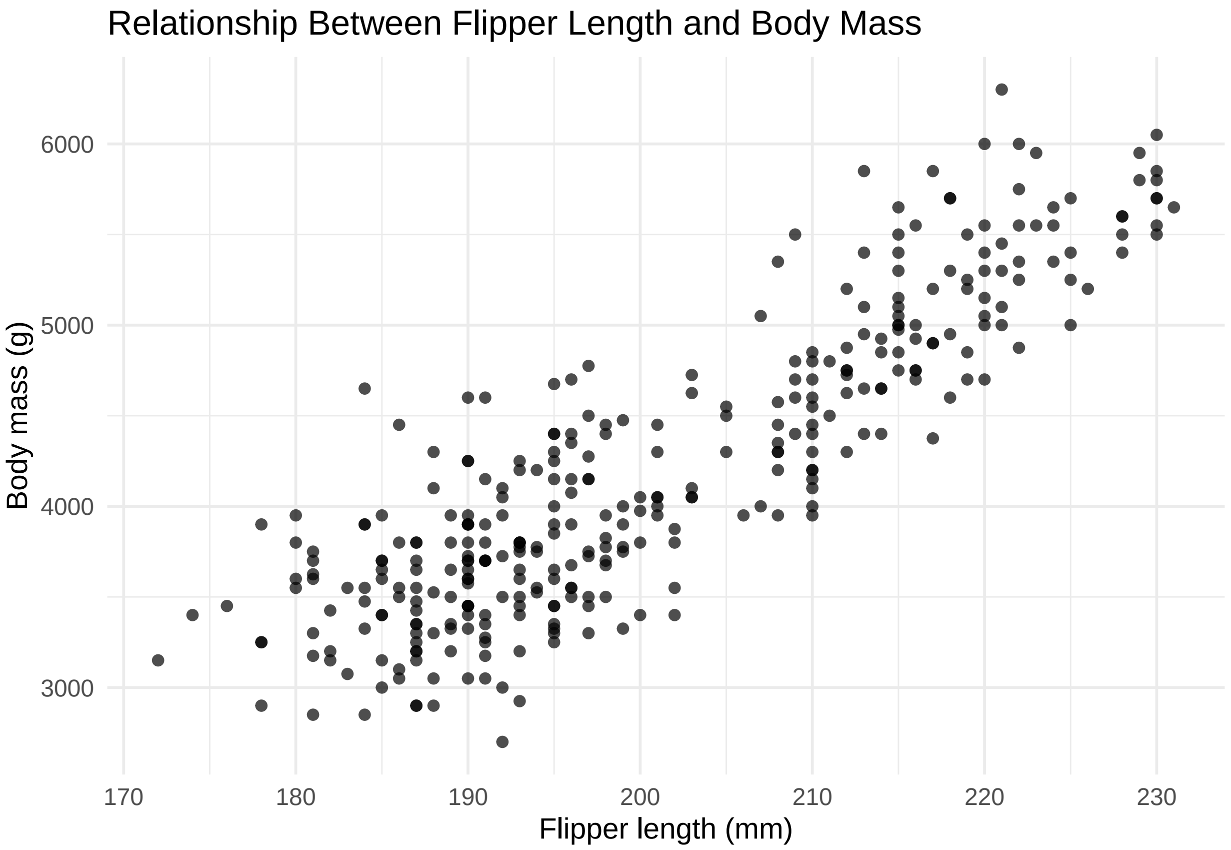
```r
ggplot(penguins_clean,
       aes(x = bill_len, y = flipper_len, color = species)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Bill length (mm)",
    y = "Flipper length (mm)",
    color = "Species",
    title = "Bill Length vs Flipper Length by Species"
  ) +
  theme_minimal()
```



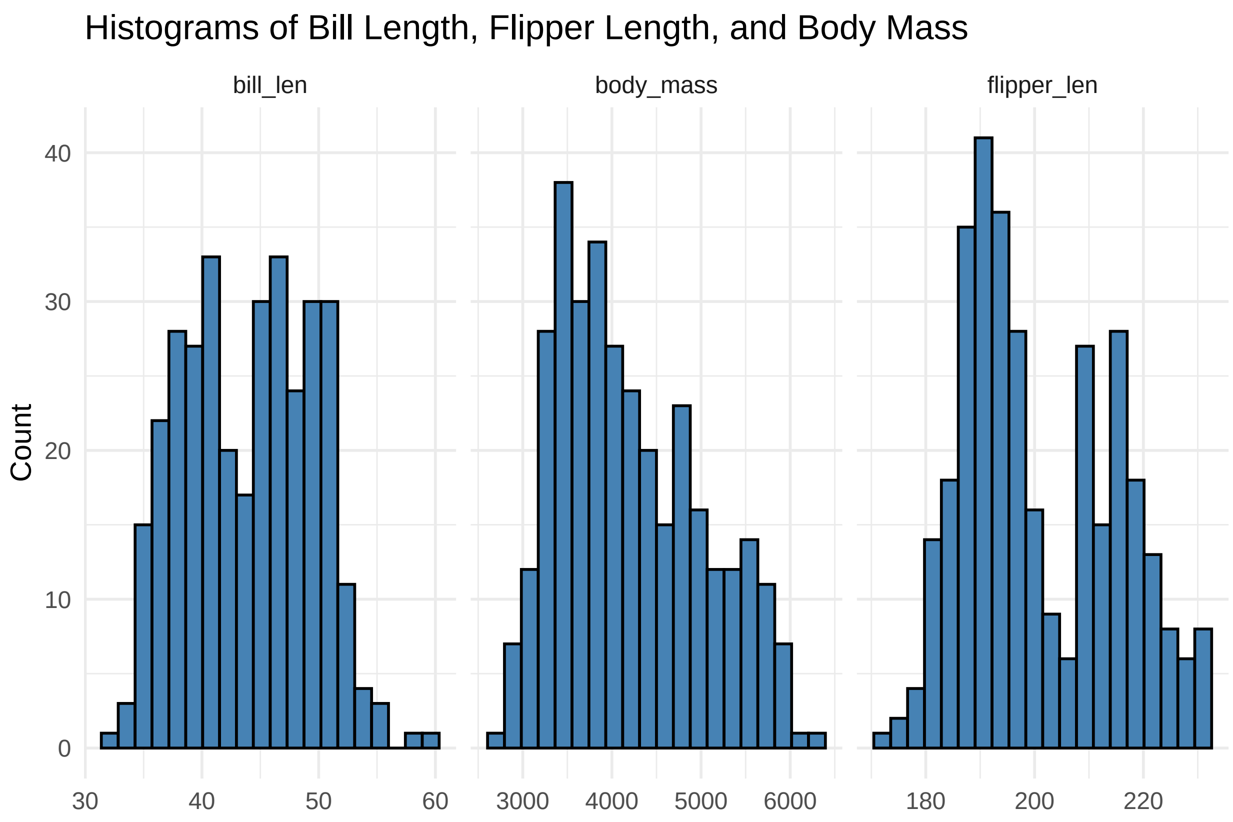
```r
ggplot(penguins_clean,
       aes(x = bill_len, y = flipper_len)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Bill length (mm)",
    y = "Flipper length (mm)",
    title = "Relationship Between Bill Length and Flipper Length"
  ) +
  theme_minimal()
```



```
ggplot(penguins_clean,
       aes(x = flipper_len, y = body_mass)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Flipper length (mm)",
    y = "Body mass (g)",
    title = "Relationship Between Flipper Length and Body Mass"
  ) +
  theme_minimal()
```



```
penguins_clean %>%
  select(bill_len, flipper_len, body_mass) %>%
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20, color = "black", fill = "steelblue") +
  facet_wrap(~ variable, scales = "free_x") +
  labs(x = "", y = "Count",
       title = "Histograms of Bill Length, Flipper Length, and Body Mass") +
  theme_minimal()
```



### Defining the Regression Model

```
model1<- lm(body_mass~bill_len,data=penguins)
summary(model1)
```

```
##
## Call:
## lm(formula = body_mass ~ bill_len, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1762.08  -446.98    32.59   462.31  1636.86
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  362.307    283.345   1.279    0.202
## bill_len      87.415      6.402  13.654   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.4 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3523
## F-statistic: 186.4 on 1 and 340 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(body_mass ~ bill_len, data = penguins_clean)

summary(model2)
```

```
##
```

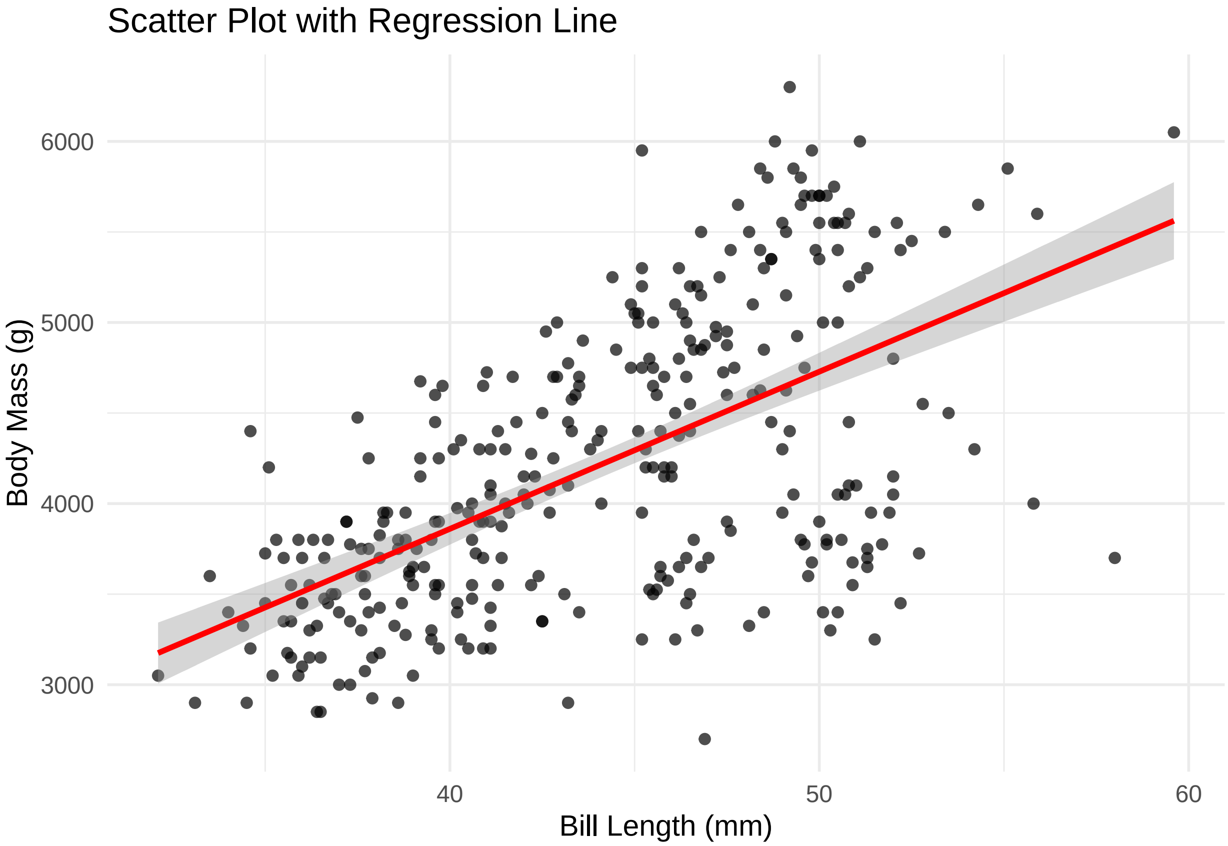
```
## Call:
## lm(formula = body_mass ~ bill_len, data = penguins_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1759.38  -468.82    27.79   464.20  1641.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  388.845    289.817   1.342    0.181
## bill_len      86.792      6.538  13.276   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 651.4 on 331 degrees of freedom
## Multiple R-squared:  0.3475, Adjusted R-squared:  0.3455
## F-statistic: 176.2 on 1 and 331 DF,  p-value: < 2.2e-16
```

```r
model3<-lm(body_mass~bill_len+flipper_len+bill_dep+species,data=penguins_clean)
summary(model3)
```

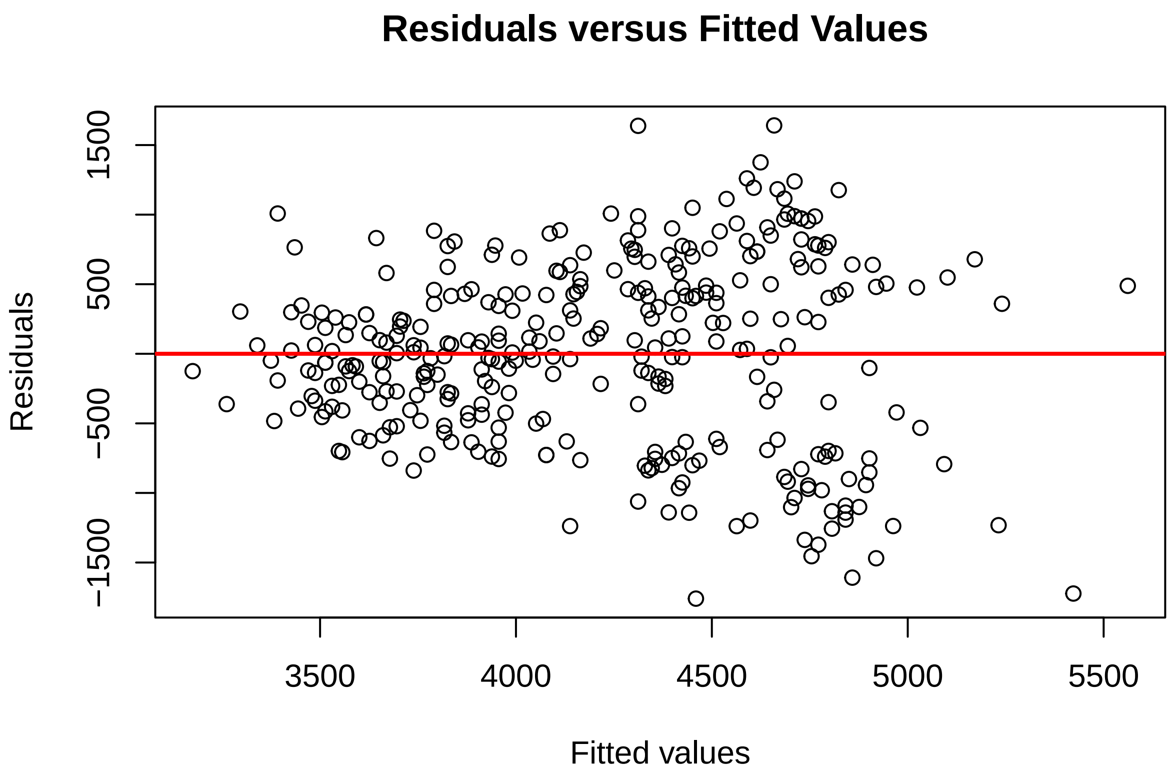
```
##
## Call:
## lm(formula = body_mass ~ bill_len + flipper_len + bill_dep +
##     species, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -838.90 -210.22  -21.17  199.67 1037.77
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4282.080    497.832  -8.601 3.33e-16 ***
## bill_len            39.718      7.227   5.496 7.85e-08 ***
## flipper_len         20.226      3.135   6.452 3.98e-10 ***
## bill_dep           141.771     19.163   7.398 1.17e-12 ***
## speciesChinstrap  -496.758     82.469  -6.024 4.59e-09 ***
## speciesGentoo      965.198    141.770   6.808 4.74e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 314.8 on 327 degrees of freedom
## Multiple R-squared:  0.8495, Adjusted R-squared:  0.8472
## F-statistic: 369.1 on 5 and 327 DF,  p-value: < 2.2e-16
```

```r
ggplot(penguins_clean, aes(x = bill_len, y = body_mass)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(
    x = "Bill Length (mm)",
    y = "Body Mass (g)",
    title = "Scatter Plot with Regression Line"
  ) +
  theme_minimal()
```

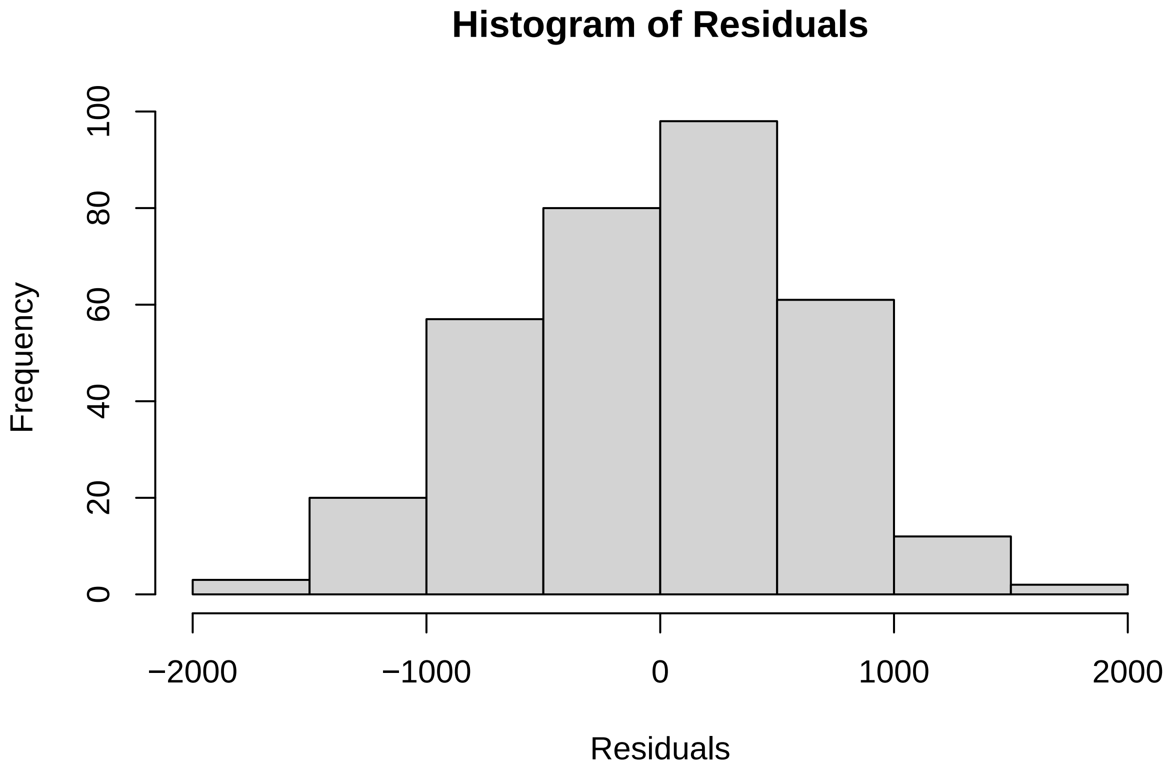
```
## `geom_smooth()` using formula = 'y ~ x'
```



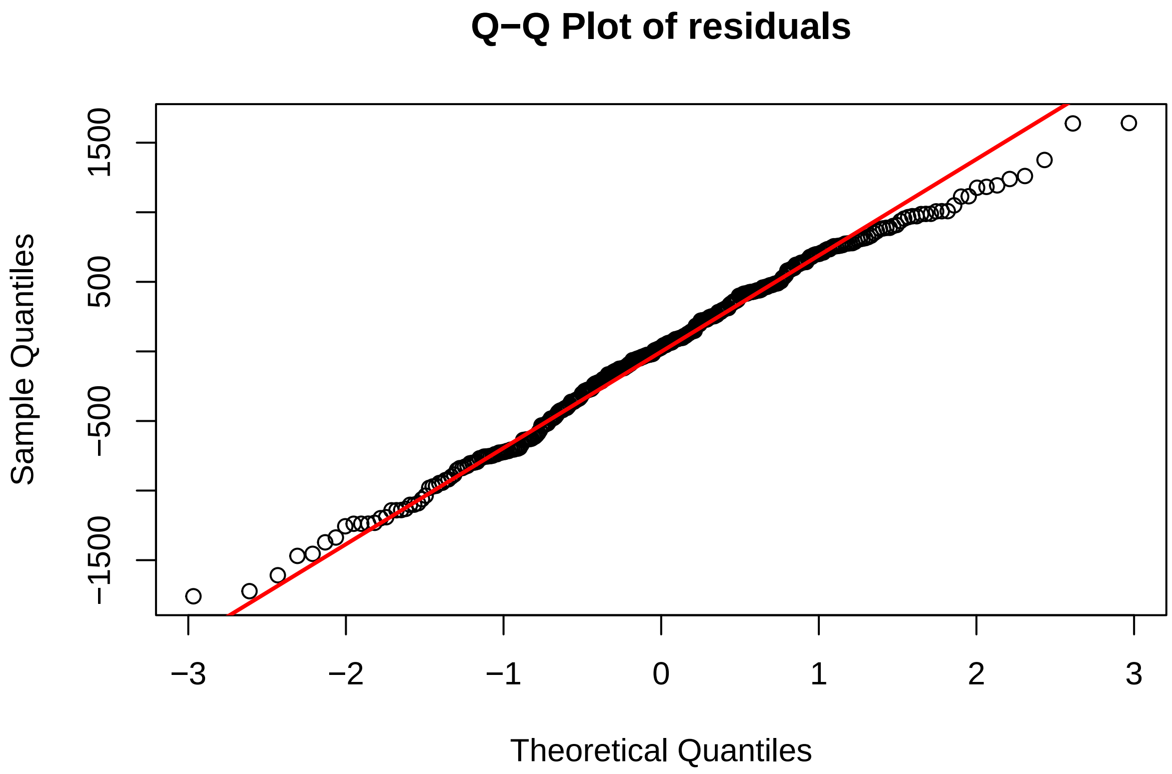
```
model <- lm(body_mass ~ bill_len, data = penguins_clean)

# Fitted Values and Residuals
fitted_vals <- fitted(model)
residuals_vals <- resid(model)

plot(fitted_vals, residuals_vals,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals versus Fitted Values ")
abline(h = 0, col = "red", lwd = 2)
```



```
#For normality
hist(residuals_vals,
     main = "Histogram of Residuals",
     xlab = "Residuals")
```



```r
qqnorm(residuals_vals,
       main = "Q-Q Plot of residuals")
qqline(residuals_vals, col = "red", lwd = 2)
```

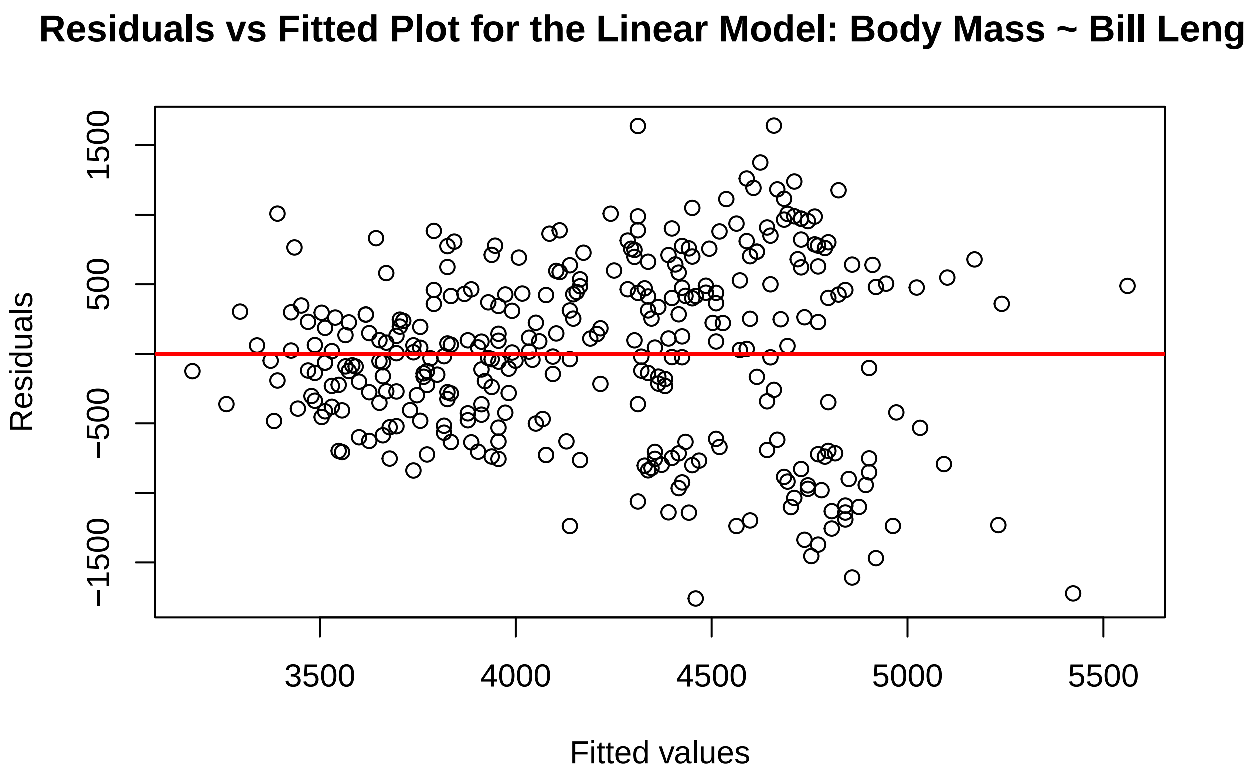


```r
shapiro.test(residuals_vals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_vals
## W = 0.99122, p-value = 0.04492
```

**For Homoscedasticity**

```r
# I draw again
plot(fitted_vals, residuals_vals,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Plot for the Linear Model: Body Mass ~ Bill Length")
abline(h = 0, col = "red", lwd = 2)
```



```
install.packages("lmtest")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 76.51, df = 1, p-value < 2.2e-16
```

Since p-value $<$ 2.2e-16, we can say that there is heteroscedasticity here and the regression coefficient is different from zero.

**NOTE: Here, classical assumptions are not met at some point. The variance is not constant.**

**ANOVA**

```
anova_result2<- aov(flipper_len~species,data=penguins_clean)
summary(anova_result2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)    
## species       2  50526   25263   567.4 <2e-16 ***
## Residuals   330  14693      45                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Post Hoc.Comparison Test**

```
install.packages("multcomp")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
```

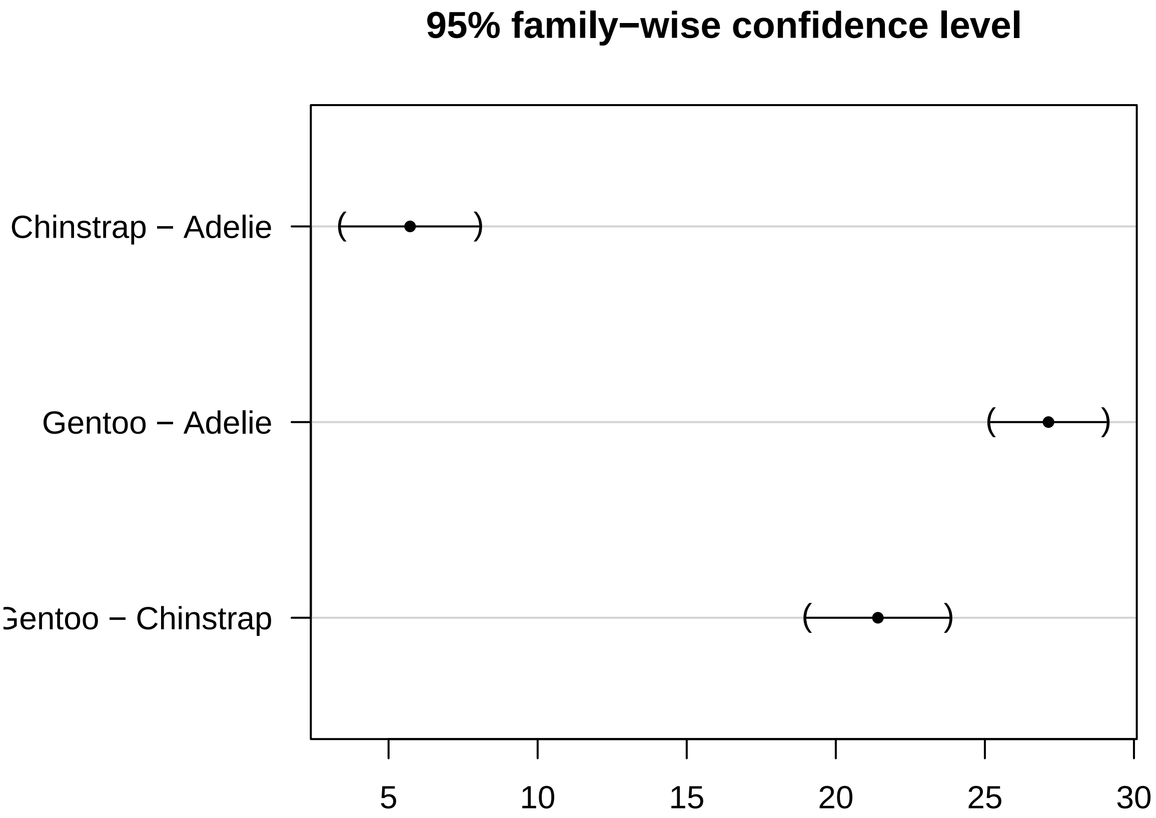
```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

## 
## Attaching package: 'MASS'

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

## 
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
## 
##     geyser
```

```
#Tukey HSD Test:
penguins_post_test<-glht(anova_result2,linfct=mcp(species="Tukey"))
summary(penguins_post_test)
```

```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Multiple Comparisons of Means: Tukey Contrasts
## 
## 
## Fit: aov(formula = flipper_len ~ species, data = penguins_clean)
## 
## Linear Hypotheses:
##                           Estimate Std. Error t value Pr(>|t|)    
## Chinstrap - Adelie == 0     5.7208     0.9796    5.84 2.65e-08 ***
## Gentoo - Adelie == 0       27.1326     0.8241   32.92  < 1e-08 ***
## Gentoo - Chinstrap == 0    21.4118     1.0143   21.11  < 1e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
par(mar=c(3,8,3,3))
    plot(penguins_post_test)
```



```
TukeyHSD(anova_result2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = flipper_len ~ species, data = penguins_clean)
##
## $species
##                        diff       lwr       upr p adj
## Chinstrap-Adelie  5.72079  3.414364  8.027215     0
## Gentoo-Adelie    27.13255 25.192399 29.072709     0
## Gentoo-Chinstrap 21.41176 19.023644 23.799885     0
```

### ANCOVA

There are two different models here.

```
ancova_model1<-lm(body_mass~species+sex+bill_len,data=penguins_clean)
anova(ancova_model1)
```

```
## Analysis of Variance Table
##
## Response: body_mass
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 145190219 72595110 765.705 < 2.2e-16 ***
## sex         1  37090262 37090262 391.214 < 2.2e-16 ***
## bill_len    1   1882101  1882101  19.852  1.15e-05 ***
## Residuals 328  31097084    94808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ancova_model2 <- lm(flipper_len ~ species + bill_len,
                    data = penguins_clean)

anova(ancova_model2)
```

```
## Analysis of Variance Table
##
## Response: flipper_len
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2  50526 25262.9  742.44 < 2.2e-16 ***
## bill_len    1   3498  3497.8  102.80 < 2.2e-16 ***
## Residuals 329  11195    34.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**It is clear that this table tells us three different things for *`ancova_model1`*.**

- **Question 1: Does species affect body mass?**
- **Question 2:Does sex affect body mass?**
- **Question 3:Does beak length (bill_len, covariate) still affect it when controlled for?**

**Answer 1:** Since $p < 0.05$ was found, it can be considered significant. This means there is a strong difference in body mass between species.

**Answer 2:**Since $p < 0.05$, it is still significant. It is possible to say that it has a very strong effect on body mass, even when sex, species, and beak length are controlled.

**Answer 3:**It's significant because $p < 0.05$. So, even if beak length is short or long, body mass still changes significantly. But there's something we need to be careful about here: the effect is much weaker than species and gender.

**CONCLUSION**:Species, sex, and bill length all significantly affect body weight. The effects of species and sex are very strong; the effect of bill length is weaker but still statistically significant.

**NOTE:Here I examined several different models through ANOVA and ANCOVA.**

- ***ancova_model1<-lm(body_mass~species+sex+bill_len,data=penguins_clean)***
- ***ancova_model2 <- lm(flipper_len ~ species + bill_len, data = penguins_clean)***