Lesson 5

Federico Reali

Introduction

Linear Regression

Simple Linear Regression

- **Model**:

- **Goal**: Predict the dependent variable (y) using one independent variable (x)
- **Example**: Predicting bill length from body mass

Multiple Linear Regression

- **Model**:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p$$

- **Goal**: Predict the dependent variable using multiple independent variables
- **Example**: Predicting bill length from body mass and sex

Model Selection

AIC (Akaike Information Criterion)

- **Purpose**: Measure the quality of a model, balancing goodness of fit and complexity
- **Formula**:

$$\mathrm{AIC} = 2k - 2\ln(L)$$

  - k: Number of parameters
  - L: Likelihood of the model

Stepwise Selection

- **Purpose**: Select the best model by adding or removing predictors based on AIC
- **Types**: Forward selection, backward elimination, and stepwise selection

ANOVA (Analysis of Variance)

Purpose

- Compare means across multiple groups
- Test if there are significant differences between group means

Interpretation

- **F-statistic**: Ratio of variance between groups to variance within groups
- **p-value**: Probability of observing the data if the null hypothesis is true

Exercise

Overview

- Analysis of the Palmer Penguins dataset
- Focus on various statistical techniques and visualizations

Dataset

- The dataset contains measurements for penguins from three different species
- Variables include bill length, bill depth, flipper length, body mass, sex, and island

Data Loading

```r
# Load penguins dataset from the specified URL
df_peng <- read.csv("https://raw.githubusercontent.com/rfordatascience/tidy
```

Initial Exploration

```r
# Display the structure of the dataset
# str(df_peng)

# Provide a summary of the dataset
summary(df_peng)
```

```
  species             island          bill_length_mm  bill_depth_mm
Length:344         Length:344         Min.   :32.10   Min.   :13.10
Class :character   Class :character   1st Qu.:39.23   1st Qu.:15.60
Mode  :character   Mode  :character   Median :44.45   Median :17.30
                                      Mean   :43.92   Mean   :17.15
                                      3rd Qu.:48.50   3rd Qu.:18.70
                                      Max.   :59.60   Max.   :21.50
                                      NA's   :2       NA's   :2
flipper_length_mm  body_mass_g       sex                 year
Min.   :172.0     Min.   :2700   Length:344         Min.   :2007
1st Qu.:190.0     1st Qu.:3550   Class :character   1st Qu.:2007
Median :197.0     Median :4050   Mode  :character   Median :2008
Mean   :200.9     Mean   :4202                      Mean   :2008
3rd Qu.:213.0     3rd Qu.:4750                      3rd Qu.:2009
Max.   :231.0     Max.   :6300                      Max.   :2009
```

Handling Missing Values

Identify Missing Values

```r
# Identify rows with missing values in the 'bill_length_mm' column
df_peng[is.na(df_peng$bill_length_mm),]
```

```
    species    island bill_length_mm bill_depth_mm flipper_length_mm
4    Adelie Torgersen             NA            NA                NA
272  Gentoo    Biscoe             NA            NA                NA
    body_mass_g  sex year
4            NA <NA> 2007
272          NA <NA> 2009
```

Remove Missing Values

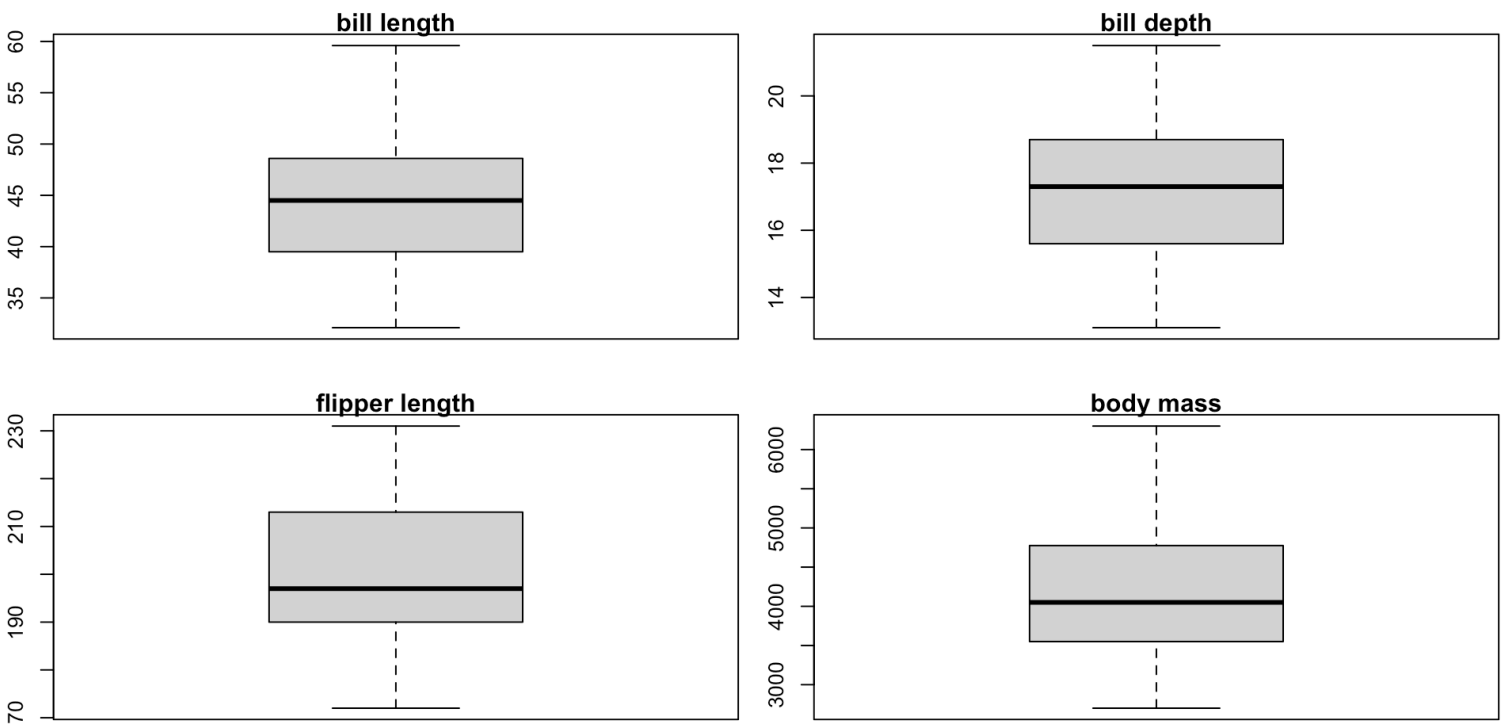
```r
# Remove rows with missing values
df_peng_1 <- df_peng
df_peng <- na.omit(df_peng)

# Display the summary after removing missing values
summary(df_peng)
```

```
  species             island          bill_length_mm  bill_depth_mm
Length:333         Length:333         Min.   :32.10   Min.   :13.10
Class :character   Class :character   1st Qu.:39.50   1st Qu.:15.60
Mode  :character   Mode  :character   Median :44.50   Median :17.30
                                      Mean   :43.99   Mean   :17.16
                                      3rd Qu.:48.60   3rd Qu.:18.70
                                      Max.   :59.60   Max.   :21.50
flipper_length_mm  body_mass_g       sex                 year
Min.   :172       Min.   :2700   Length:333         Min.   :2007
1st Qu.:190       1st Qu.:3550   Class :character   1st Qu.:2007
Median :197       Median :4050   Mode  :character   Median :2008
Mean   :201       Mean   :4207                      Mean   :2008
3rd Qu.:213       3rd Qu.:4775                      3rd Qu.:2009
Max.   :231       Max.   :6300                      Max.   :2009
```

Data Visualization

Box Plots

```r
# Visualize numerical variables using box plots
par(mfrow = c(2,2), mar = c(2,2,1,1))
boxplot(df_peng$bill_length_mm, main = "bill length")
boxplot(df_peng$bill_depth_mm, main = "bill depth")
boxplot(df_peng$flipper_length_mm, main = "flipper length")
boxplot(df_peng$body_mass_g, main = "body mass")
```

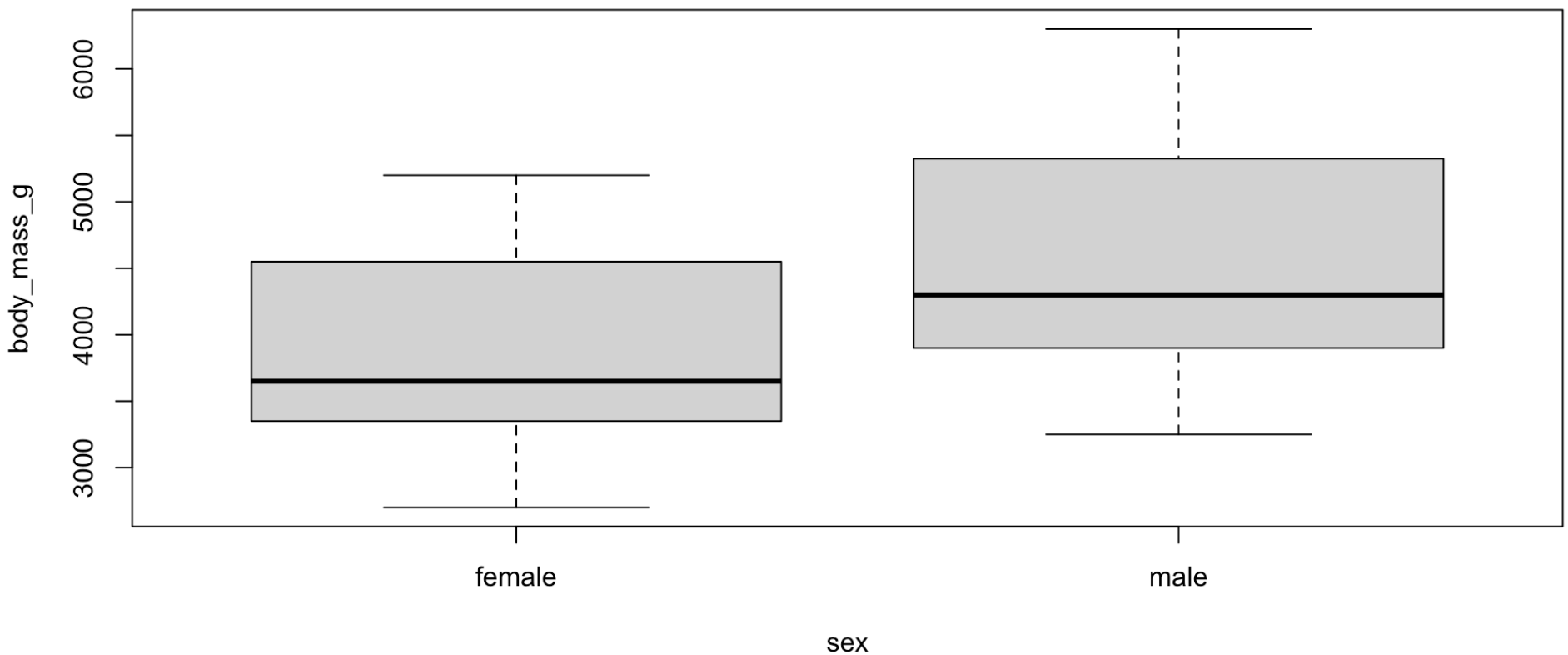


Histograms

```
# Visualize body mass distribution
hist(df_peng$body_mass_g)
```

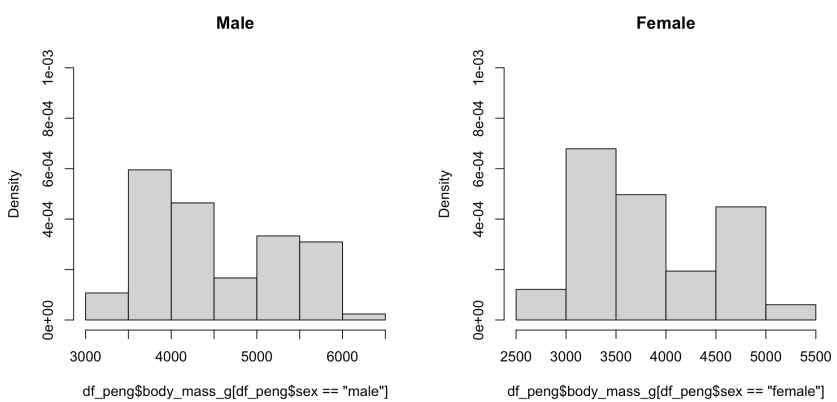
Box Plots by Sex

```r
# Analyze body mass according to sex using box plots
boxplot(body_mass_g ~ sex, data = df_peng)
```



Side-by-Side Histograms

```r
# Create side-by-side histograms for male and female body mass
par(mfrow = c(1,2))
hist(df_peng$body_mass_g[df_peng$sex == "male"],
     main = "Male",
     freq = F,
     breaks = 5,
     ylim = c(0, 1e-03))
hist(df_peng$body_mass_g[df_peng$sex == "female"],
     main = "Female",
     freq = F,
     breaks = 5,
     ylim = c(0, 1e-03))
```



Statistical Analysis

T-Test by Sex

```r
# Conduct t-test to compare male and female body mass
t.test(body_mass_g ~ sex, data = df_peng)
```

```
    Welch Two Sample t-test

data:  body_mass_g by sex
t = -8.5545, df = 323.9, p-value = 4.794e-16
alternative hypothesis: true difference in means between group female and
group male is not equal to 0
95 percent confidence interval:
 -840.5783 -526.2453
sample estimates:
mean in group female   mean in group male
            3862.273             4545.685
```

Unique Values

```r
# Explore unique values of the 'island' and 'species' columns
unique(df_peng$island)
```

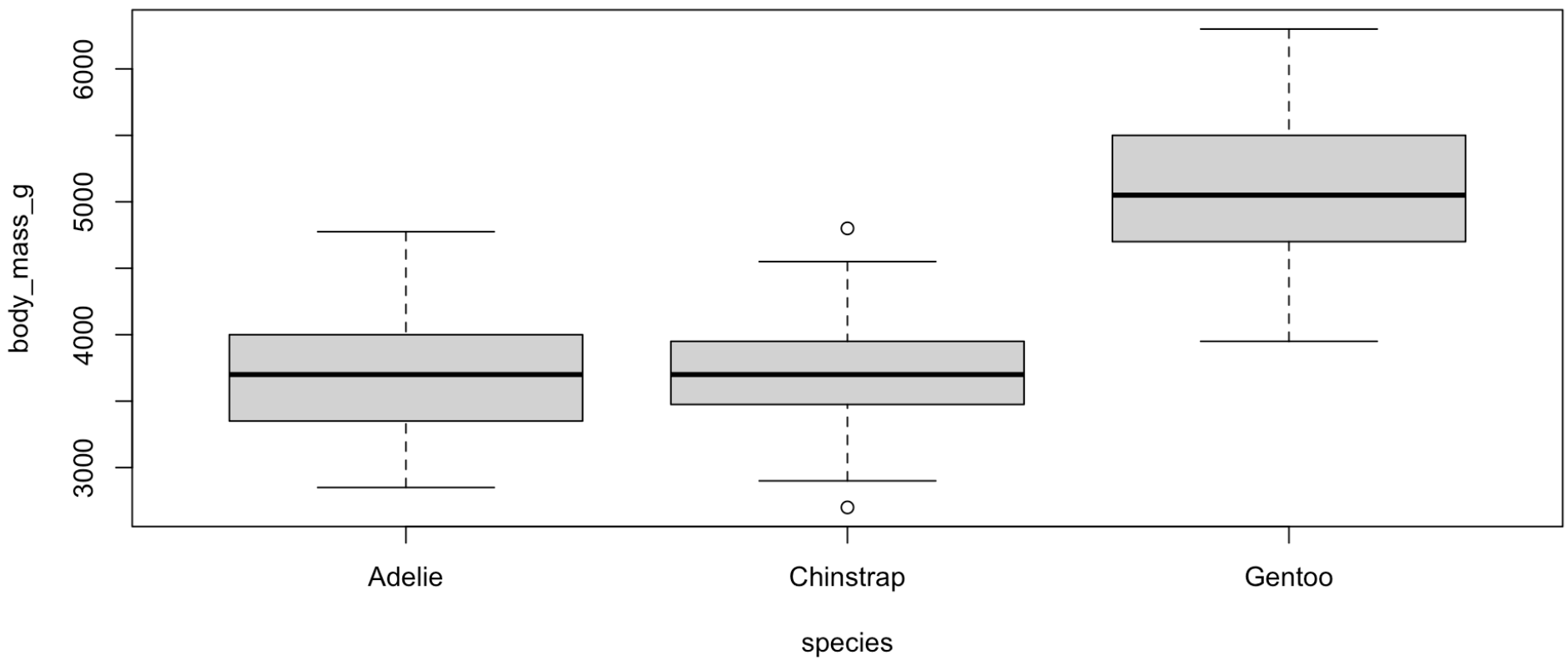
```
[1] "Torgersen" "Biscoe"    "Dream"
```

```r
unique(df_peng$species)
```

```
[1] "Adelie"    "Gentoo"    "Chinstrap"
```

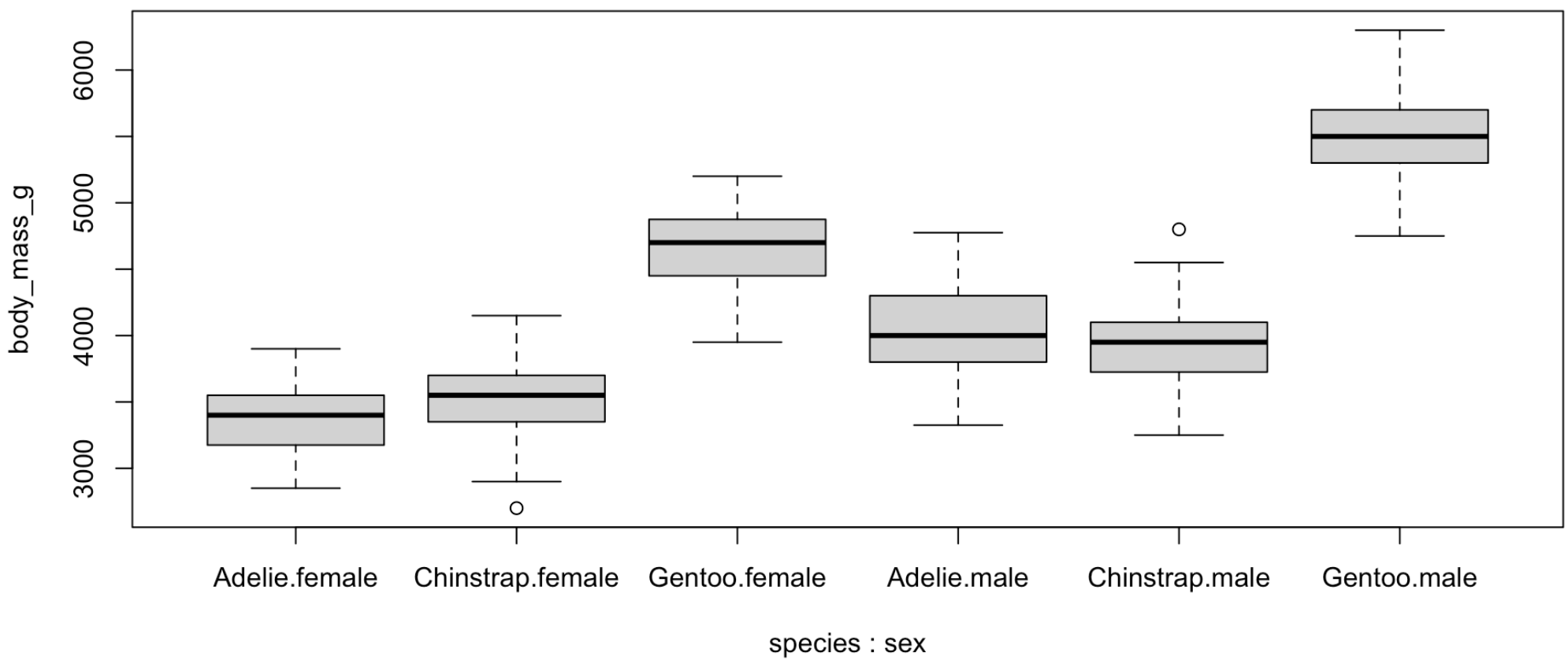
Box Plots by Species

```
1 # Analyze body mass across penguin species using box plots
2 boxplot(body_mass_g ~ species, data = df_peng)
```



Box Plots by Species and Sex

```
1 boxplot(body_mass_g ~ species + sex, data = df_peng)
```



T-Tests by Species Adelie

```r
# Conduct t-tests for body mass comparison within each penguin species
t.test(body_mass_g ~ sex, data = df_peng[df_peng$species == "Adelie",])
```

```
    Welch Two Sample t-test

data:  body_mass_g by sex
t = -13.126, df = 135.69, p-value < 2.2e-16
alternative hypothesis: true difference in means between group female and
group male is not equal to 0
95 percent confidence interval:
 -776.3012 -573.0139
sample estimates:
mean in group female   mean in group male
            3368.836             4043.493
```

Gentoo

```r
t.test(body_mass_g ~ sex, data = df_peng[df_peng$species == "Gentoo",])
```

```
    Welch Two Sample t-test

data:  body_mass_g by sex
t = -14.761, df = 116.64, p-value < 2.2e-16
alternative hypothesis: true difference in means between group female and
group male is not equal to 0
95 percent confidence interval:
 -913.1130 -697.0763
sample estimates:
mean in group female   mean in group male
            4679.741             5484.836
```

Chinstrap

```r
t.test(body_mass_g ~ sex, data = df_peng[df_peng$species == "Chinstrap",])
```

```
    Welch Two Sample t-test

data:  body_mass_g by sex
t = -5.2077, df = 62.575, p-value = 2.264e-06
alternative hypothesis: true difference in means between group female and
group male is not equal to 0
95 percent confidence interval:
 -569.7903 -253.7391
sample estimates:
mean in group female   mean in group male
            3527.206             3938.971
```

T-Test by Species

```
# Conduct a t-test for body mass comparison across penguin species
# t.test(body_mass_g ~ species, data = df_peng)
#t-test only works for pairs!
```

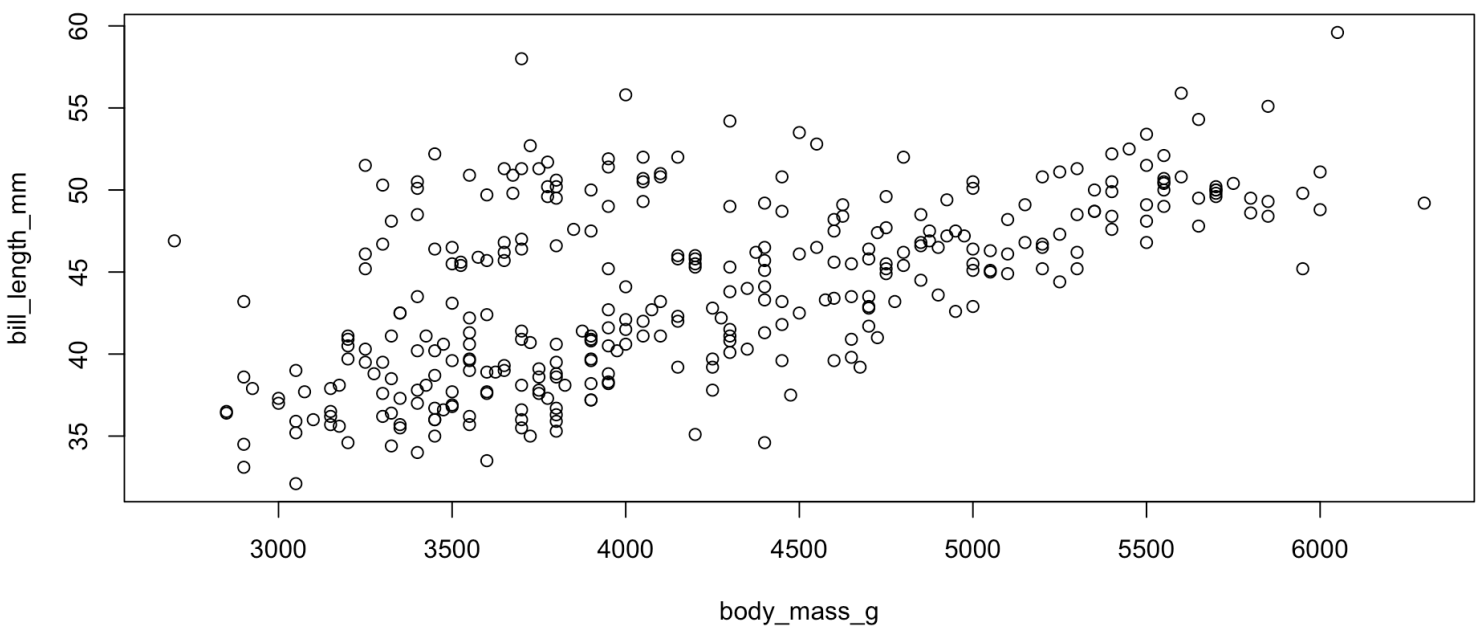
Linear Regression Analysis

Correlation

```r
# Visualize the relationship between bill length and body mass
plot(bill_length_mm ~ body_mass_g, data = df_peng)
```

```r
# Calculate and display the correlation between bill length and body mass
cor(df_peng$bill_length_mm, df_peng$body_mass_g)
```

```
[1] 0.5894511
```



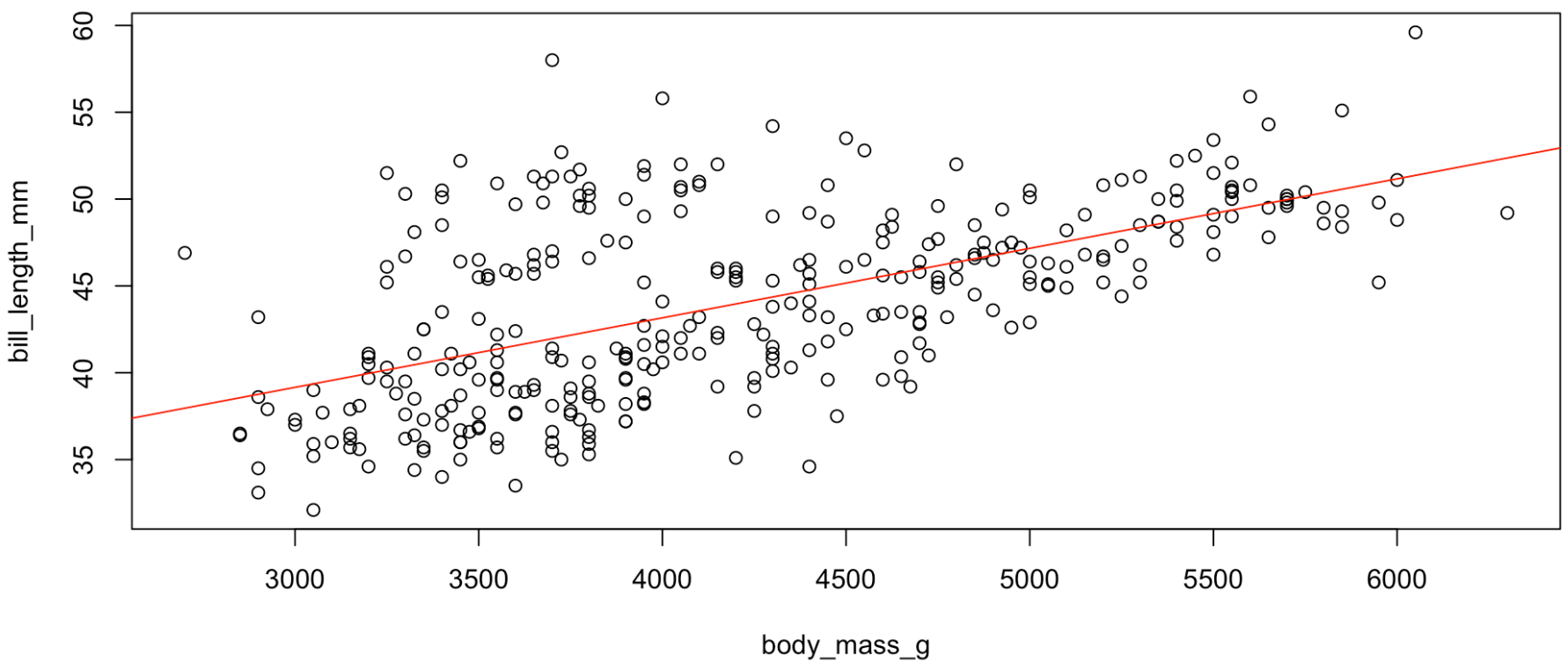
Simple Linear Regression

```r
my_lm1 <- lm(bill_length_mm ~ body_mass_g, data = df_peng)
summary(my_lm1)
```

```
Call:
lm(formula = bill_length_mm ~ body_mass_g, data = df_peng)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1652  -3.0664  -0.7672   2.2356  16.0371 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.715e+01  1.292e+00   21.02   <2e-16 ***
body_mass_g 4.003e-03  3.016e-04   13.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Plot the Linear Regression

```r
# Plot the regression line on the scatter plot
plot(bill_length_mm ~ body_mass_g, data = df_peng)
abline(my_lm1$coefficients, col = 'red')
```



Multiple Linear Regression - Sex

```r
# Fit linear regression models with additional predictors (sex and species)
my_lm2 <- lm(bill_length_mm ~ body_mass_g + sex, data = df_peng)
summary(my_lm2)
```

```
Call:
lm(formula = bill_length_mm ~ body_mass_g + sex, data = df_peng)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.7196  -3.2501  -0.7724   2.5415  16.4992 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.791e+01  1.323e+00  21.095   <2e-16 ***
body_mass_g 3.674e-03  3.309e-04  11.102   <2e-16 ***
sexmale     1.247e+00  5.321e-01   2.344   0.0197 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

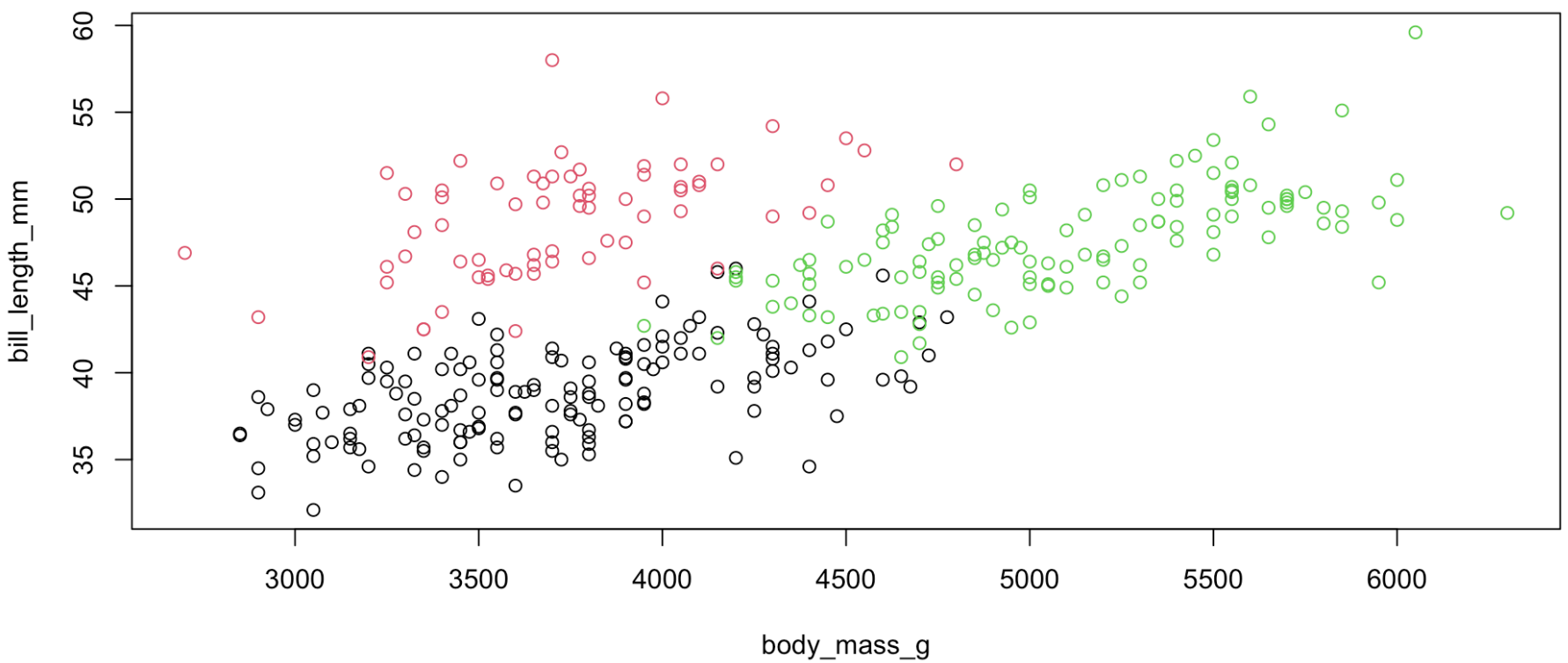
Multiple Linear Regression - Species

```r
my_lm3 <- lm(bill_length_mm ~ body_mass_g + species, data = df_peng)
summary(my_lm3)
```

```
Call:
lm(formula = bill_length_mm ~ body_mass_g + species, data = df_peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8291 -1.6728  0.1244  1.5318  9.2904 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      24.908763   1.089730  22.858  < 2e-16 ***
body_mass_g       0.003755   0.000289  12.991  < 2e-16 ***
speciesChinstrap  9.908762   0.355289  27.889  < 2e-16 ***
speciesGentoo     3.539179   0.499814   7.081 8.71e-12 ***
---
```

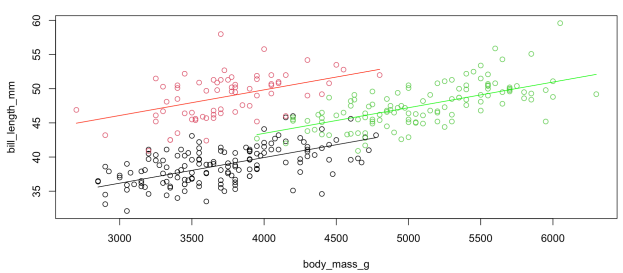
Visualization by Species

```
# Visualize the relationship between bill length and body mass by species
plot(bill_length_mm ~ body_mass_g, data = df_peng, col =
as.factor(df_peng$species))
```



Predictions by species

```r
data_A <- df_peng[df_peng$species == "Adelie",]
data_C <- df_peng[df_peng$species == "Chinstrap",]
data_G <- df_peng[df_peng$species == "Gentoo",]

# Predicted values for each species
data_A_y <- predict(my_lm3, data_A)
data_C_y <- predict(my_lm3, data_C)
data_G_y <- predict(my_lm3, data_G)

# Overlay predicted values on the scatter plot by species
plot(bill_length_mm ~ body_mass_g, data = df_peng, col = as.factor(df_peng$
lines(data_A$body_mass_g, data_A_y)
lines(data_C$body_mass_g, data_C_y, col = 'red')
lines(data_G$body_mass_g, data_G_y, col = 'green')
```



Multiple Linear Regression - Body Mass and Flipper Length

```r
# Fit a linear regression model for bill length, body mass, and flipper len
my_lm4 <- lm(bill_length_mm ~ body_mass_g + flipper_length_mm, data = df_pe
summary(my_lm4)
```

```
Call:
lm(formula = bill_length_mm ~ body_mass_g + flipper_length_mm,
    data = df_peng)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8251 -2.6432 -0.7281  2.0229 18.8227

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.9812079  4.7215540  -0.843    0.400
body_mass_g        0.0005513  0.0005797   0.951    0.342
flipper_length_mm  0.2271747  0.0333014   6.822 4.31e-11 ***
---
```

Model Analysis

Prediction - Data preparation

```r
# Create new data for prediction and predict bill length
new_data_for_pred <- df_peng[1:3,]
new_data_for_pred$species <- c("Gentoo", "Adelie", "Chinstrap")
new_data_for_pred$island <- rep("Torgersen", 3)
new_data_for_pred$bill_length_mm <- c(60, 50, 42)
new_data_for_pred$bill_depth_mm <- rep(12,3)
new_data_for_pred$flipper_length_mm <- rep(179, 3)
new_data_for_pred$body_mass_g <- c(4000, 3300, 4500)
```

Prediction - Data preparation

```
# Display predicted values
predict(my_lm3, new_data_for_pred)
```

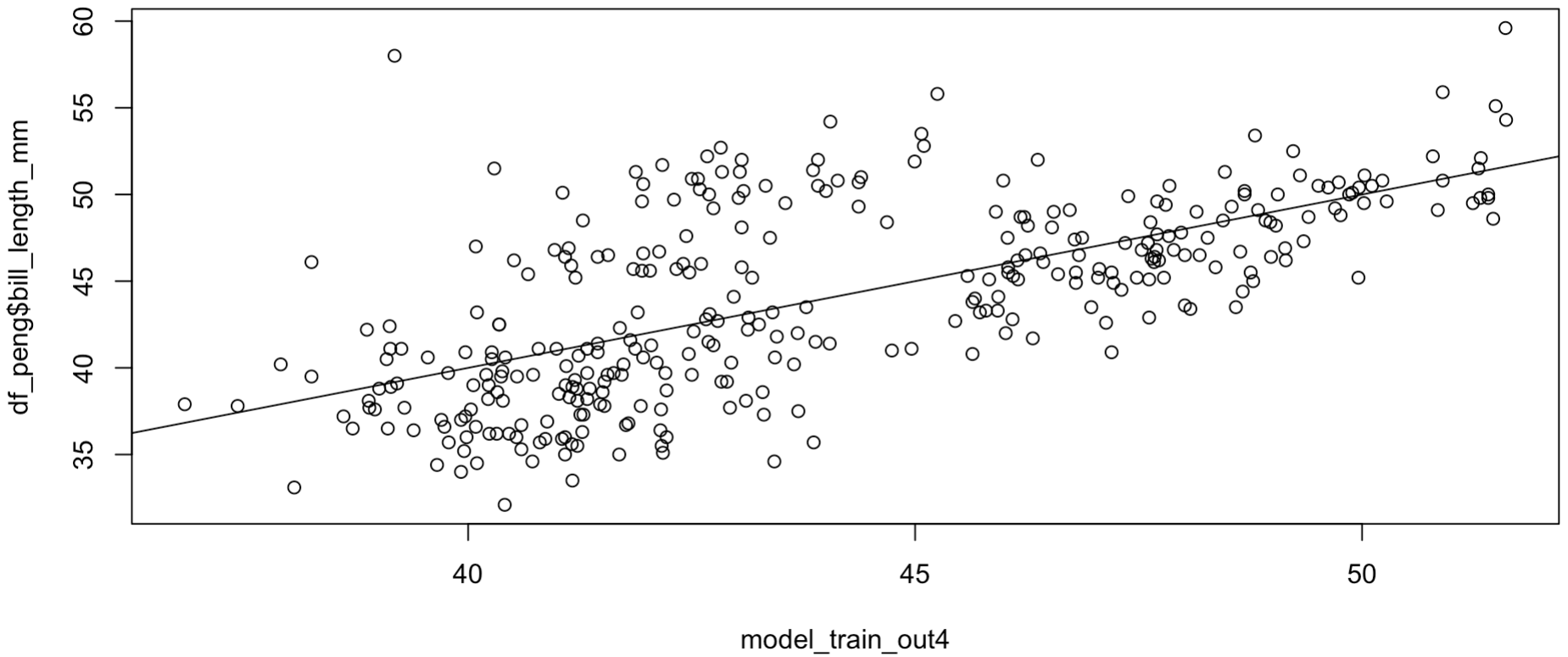
```
       1        2        3 
43.46639 37.29898 51.71328 
```

```
rbind(predict(my_lm3, new_data_for_pred), new_data_for_pred$bill_length_mm)
```

```
             1        2        3
[1,] 43.46639 37.29898 51.71328
[2,] 60.00000 50.00000 42.00000
```

Model Performance - Plot

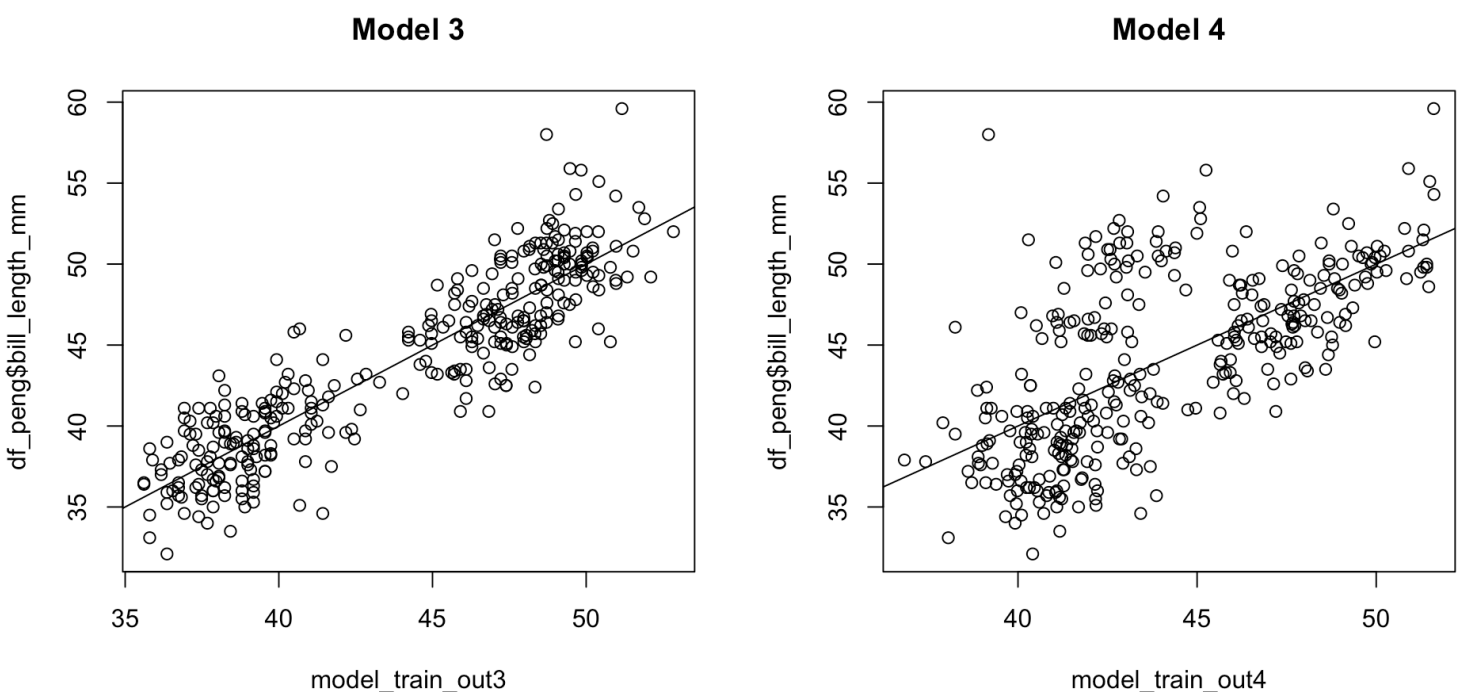
```
# Train and test the models on the same dataset
model_train_out4 <- predict(my_lm4, df_peng)

# Plot the predicted values against the actual bill length
plot(model_train_out4, df_peng$bill_length_mm)
abline(0,1)
```




```r
# Compare the predictions of two models
model_train_out3 <- predict(my_lm3, df_peng)
plot(model_train_out3, df_peng$bill_length_mm)
abline(0,1)
```

Model Performance - Plots

```r
# Side-by-side plots comparing model 3 and model 4 predictions
par(mfrow = c(1,2))
plot(model_train_out3, df_peng$bill_length_mm, main = "Model 3")
abline(0,1)
plot(model_train_out4, df_peng$bill_length_mm, main = "Model 4")
abline(0,1)
```



R-Squared Values

```r
# Calculate and display the R-squared values for the models
cor(model_train_out3, df_peng$bill_length_mm)^2
```

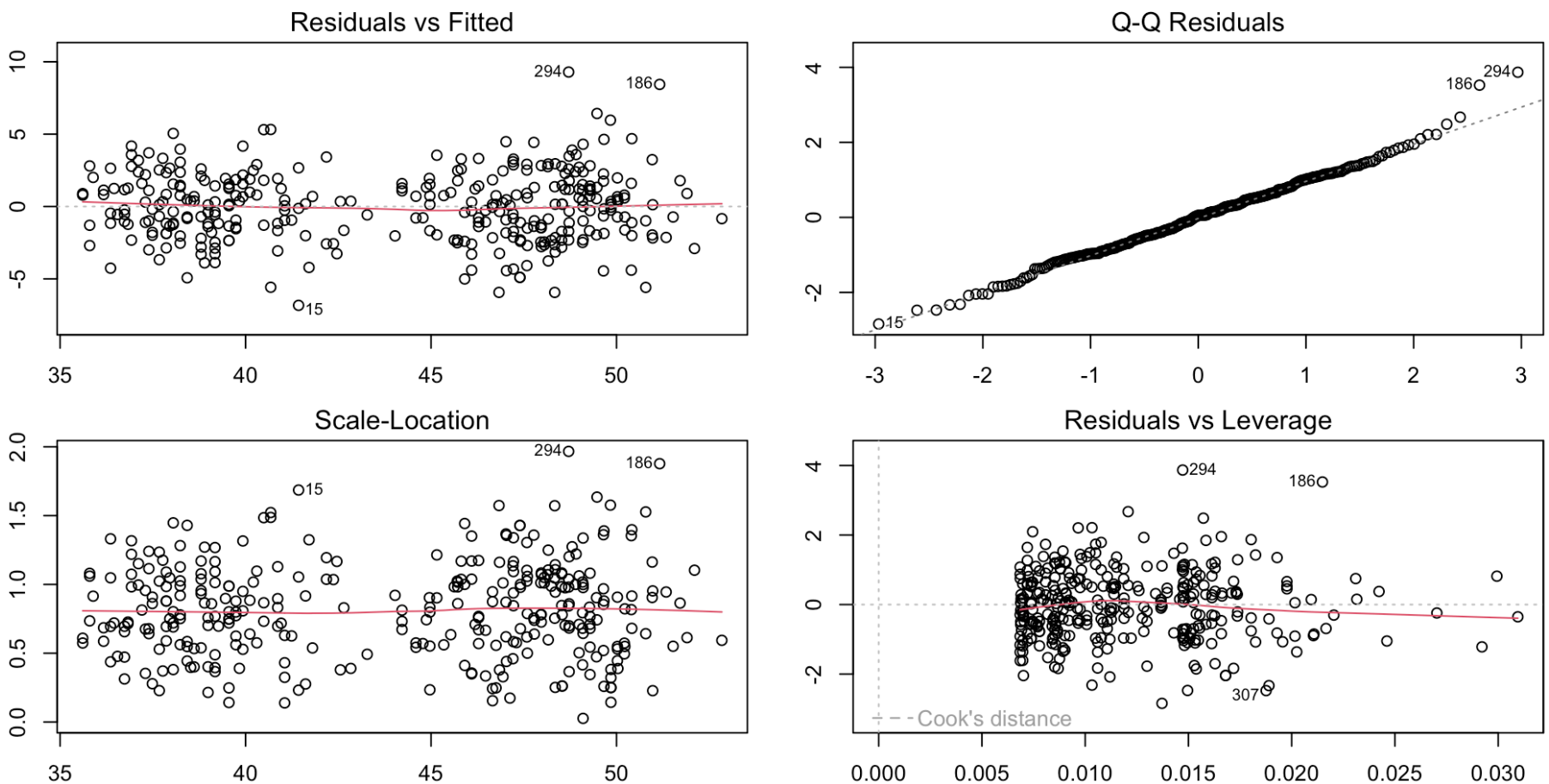
```
[1] 0.806047
```

```r
cor(model_train_out4, df_peng$bill_length_mm)^2
```

```
[1] 0.4281016
```

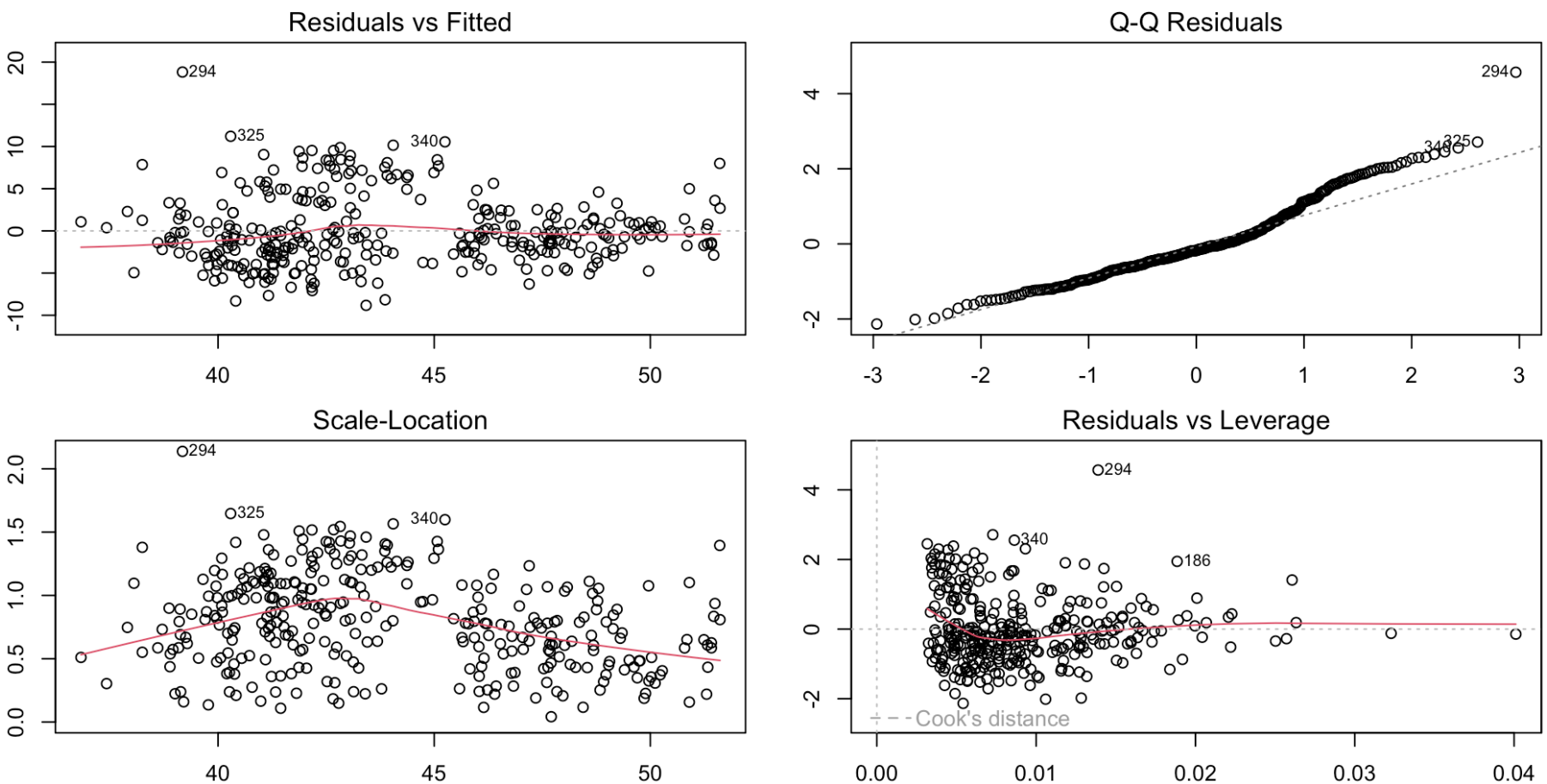
Diagnostics

```r
## Diagnostics plots for model 3
par(mfrow = c(2,2), mar = c(2,2,2,2))
plot(my_lm3)
```



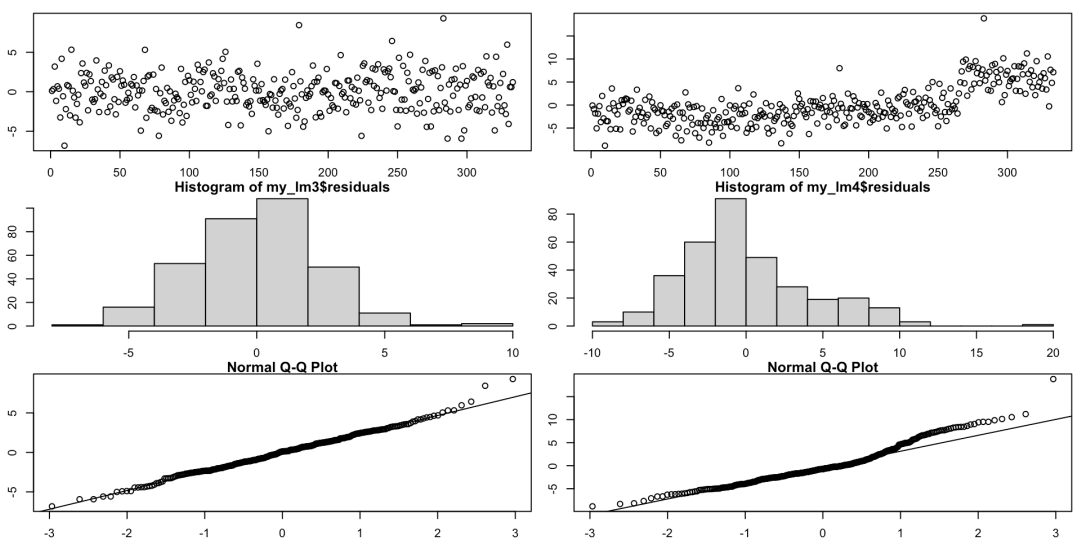
Diagnostics

```r
## Diagnostics plots for model 3
par(mfrow = c(2,2), mar = c(2,2,2,2))
plot(my_lm4)
```

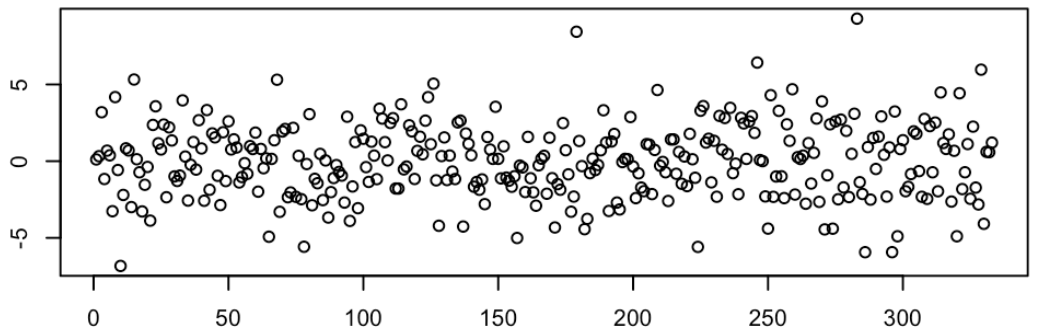


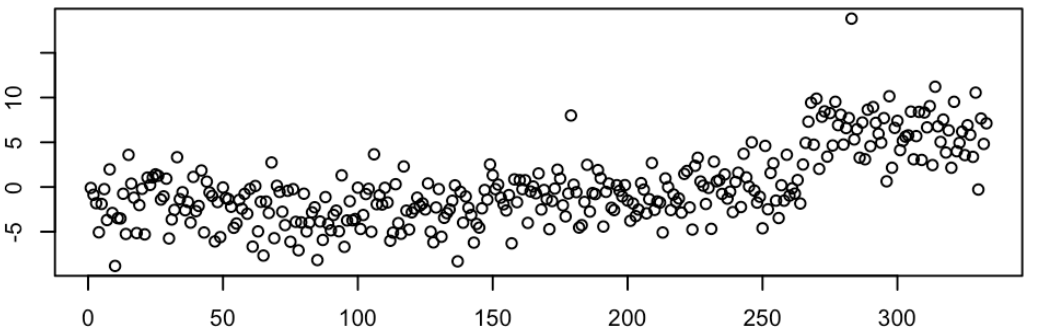
Plot Regression Residuals

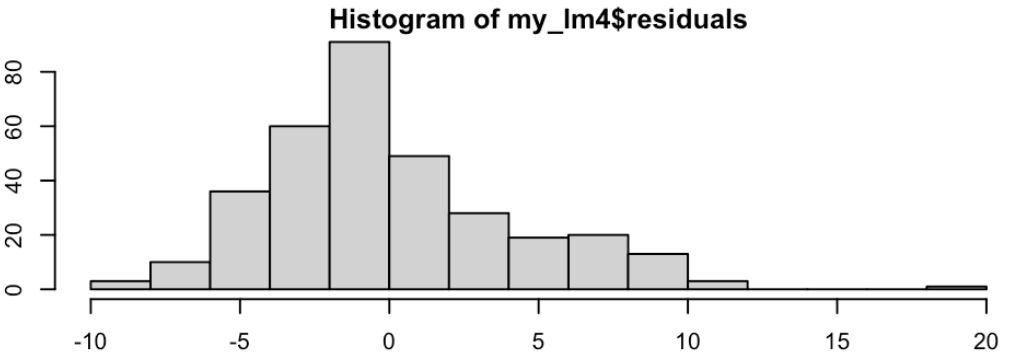
```
# Plot the regression residuals
par(mfrow = c(3,2), mar = c(2,2,1,1))
plot(my_lm3$residuals)
plot(my_lm4$residuals)
hist(my_lm3$residuals)
hist(my_lm4$residuals)
qqnorm(my_lm3$residuals)
qqline(my_lm3$residuals)
qqnorm(my_lm4$residuals)
qqline(my_lm4$residuals)
```

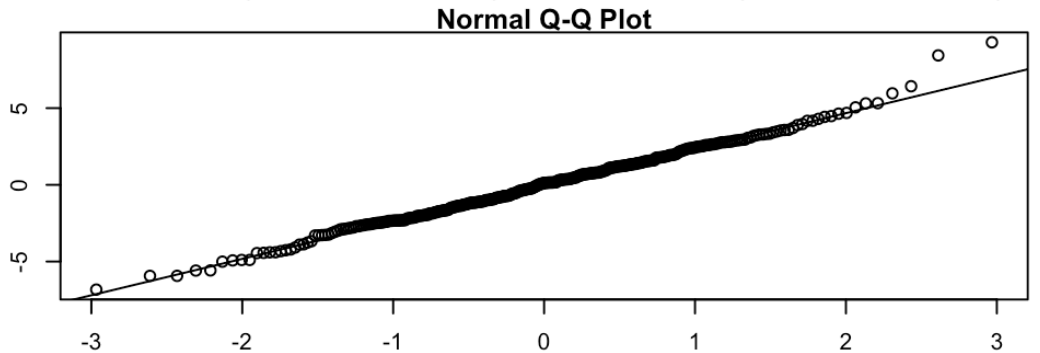


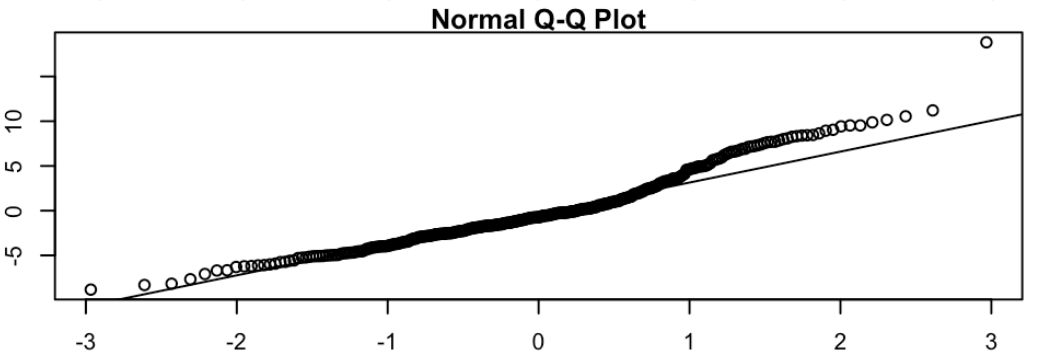
Plot Regression Residuals











Full Model

```r
my_lm_all <- lm(bill_length_mm ~ . , data = df_peng)
summary(my_lm_all)
```

```
Call:
lm(formula = bill_length_mm ~ ., data = df_peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3028 -1.2877 -0.0806  1.2938 11.4785 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -3.893e+02  3.257e+02  -1.196  0.23275    
speciesChinstrap   9.910e+00  4.277e-01  23.167  < 2e-16 ***
speciesGentoo      6.487e+00  1.126e+00   5.759 1.97e-08 ***
islandDream       -4.624e-01  4.512e-01  -1.025  0.30623    
islandTorgersen   -7.327e-02  4.716e-01  -0.155  0.87663    
```

Confidence Intervals

```r
confint(my_lm3)
```

```
                     2.5 %       97.5 %
(Intercept)      22.76504606 27.052480987
body_mass_g       0.00318604  0.004323184
speciesChinstrap  9.20983862 10.607686003
speciesGentoo     2.55594473  4.522413244
```

```r
confint(my_lm_all, level = 0.99)
```

```
                         0.5 %       99.5 %
(Intercept)       -1.233148e+03 4.544835e+02
speciesChinstrap   8.801177e+00 1.101785e+01
speciesGentoo      3.568297e+00 9.405191e+00
islandDream       -1.631675e+00 7.068039e-01
islandTorgersen   -1.295251e+00 1.148710e+00
bill_depth_mm     -7.684177e-02 7.312687e-01
flipper_length_mm -8.289657e-03 1.227639e-01
body_mass_g        3.211030e-05 2.239499e-03
sexmale            1.043576e+00 3.063959e+00
year              -2.194927e-01 6.240964e-01
```

Model Selection

R2 Comparison

```r
# Selection criteria (Higher is better)
summary(my_lm_all)$r.squared
```

```
[1] 0.8400442
```

```r
summary(my_lm3)$r.squared
```

```
[1] 0.806047
```

AIC Comparison

```
# Selection criteria (Smaller is better)
AIC(my_lm_all)
```

```
[1] 1487.227
```

```
AIC(my_lm3)
```

```
[1] 1539.402
```

```
#OR
AIC(my_lm_all,my_lm3)
```

```
           df      AIC
my_lm_all  11 1487.227
my_lm3      5 1539.402
```

Stepwise Model

```r
# STEP
my_ml_step <- step(my_lm_all)
```

```
Start:  AIC=540.21
bill_length_mm ~ species + island + bill_depth_mm + flipper_length_mm +
    body_mass_g + sex + year

                   Df Sum of Sq    RSS    AIC
- island            2      6.26 1594.4 537.52
- year              1      7.59 1595.8 539.80
<none>                         1588.2 540.21
- bill_depth_mm     1     21.65 1609.8 542.72
- flipper_length_mm 1     25.19 1613.4 543.45
- body_mass_g       1     34.96 1623.1 545.47
- sex               1    136.45 1724.6 565.66
- species           2   2657.85 4246.0 863.68

Step:  AIC=537.52
```

Stepwise Model

```r
# STEP summary
summary(my_ml_step)
```

```
Call:
lm(formula = bill_length_mm ~ species + bill_depth_mm + flipper_length_mm + 
    body_mass_g + sex, data = df_peng)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3939 -1.3424 -0.0421  1.2695 11.4274 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.502e+01  4.374e+00   3.433 0.000674 ***
speciesChinstrap  9.566e+00  3.497e-01  27.351  < 2e-16 ***
speciesGentoo     6.404e+00  1.030e+00   6.215 1.56e-09 ***
bill_depth_mm     3.130e-01  1.541e-01   2.032 0.043000 *  
```

```r
AIC(my_lm_all, my_ml_step)
```

```
           df      AIC
my_lm_all  11 1487.227
my_ml_step  8 1484.255
```

Conclusion

- Successfully applied various statistical techniques to analyze and model the penguin dataset.
- Demonstrated the importance of data cleaning, visualization, and model selection in statistical analysis.
- Highlighted the use of transformations and ANOVA for deeper insights.