Relatório de Regressão Linear Múltipla

Luis Gustavo Lopes Leal Silva (251363) Matheus Queiroz Mota (251495)
Pascual Matheo Mazolini Soto (251557)

2024-06-24

Introdução

Descrição do Problema

Este relatório apresenta uma análise do conjunto de dados de peixes, com o objetivo principal de estimar o peso dos peixes com base em suas espécies e medidas físicas.

Objetivo Principal do Estudo

Estimar o peso de um peixe utilizando um modelo de regressão linear múltipla, considerando as variáveis de espécie e medidas físicas como preditoras.

Objetivos Específicos

1. Explorar e descrever as variáveis presentes no conjunto de dados.
2. Ajustar um modelo de regressão linear múltipla.
3. Interpretar os parâmetros e resultados do modelo de maneira clara e compreensível.

Análise Exploratória

Descrição do Conjunto de Dados

O conjunto de dados contém as seguintes variáveis:

- **Species**: A espécie do peixe.
- **Weight**: O peso do peixe em gramas.
- **Length1**: Comprimento vertical em centímetros.
- **Length2**: Comprimento diagonal em centímetros.
- **Length3**: Comprimento cruzado em centímetros.
- **Height**: Altura em centímetros.
- **Width**: Largura em centímetros.

```
library(GGally)
library(ggplot2)
library(tidyverse)
```

```r
# Carregar os dados
data <- read.csv("Fish.csv")

# Visualizar as primeiras linhas do dataset
head(data)
```

```
##   Species Weight Length1 Length2 Length3  Height  Width
## 1   Bream    242    23.2    25.4    30.0 11.5200 4.0200
## 2   Bream    290    24.0    26.3    31.2 12.4800 4.3056
## 3   Bream    340    23.9    26.5    31.1 12.3778 4.6961
## 4   Bream    363    26.3    29.0    33.5 12.7300 4.4555
## 5   Bream    430    26.5    29.0    34.0 12.4440 5.1340
## 6   Bream    450    26.8    29.7    34.7 13.6024 4.9274
```

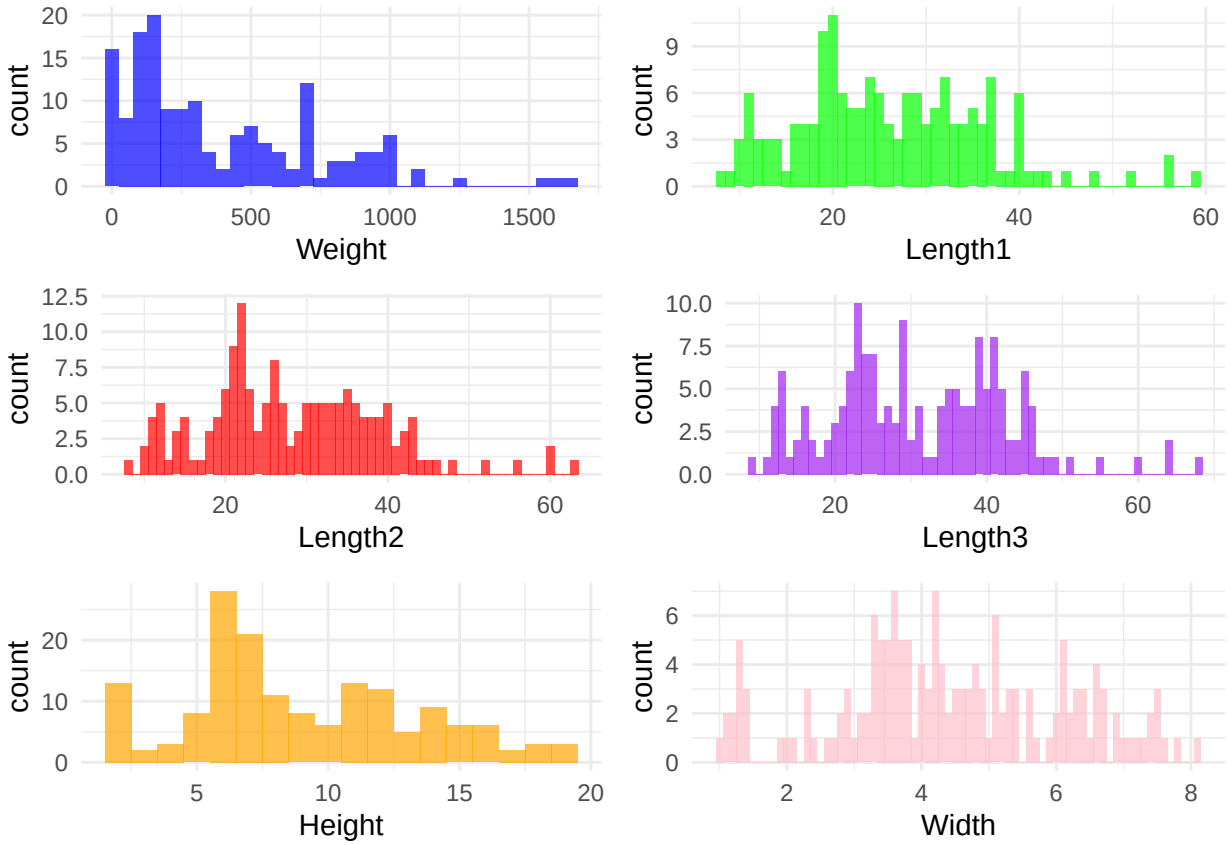
Análise Descritiva

Vamos visualizar estatísticas descritivas e gráficos das variáveis.

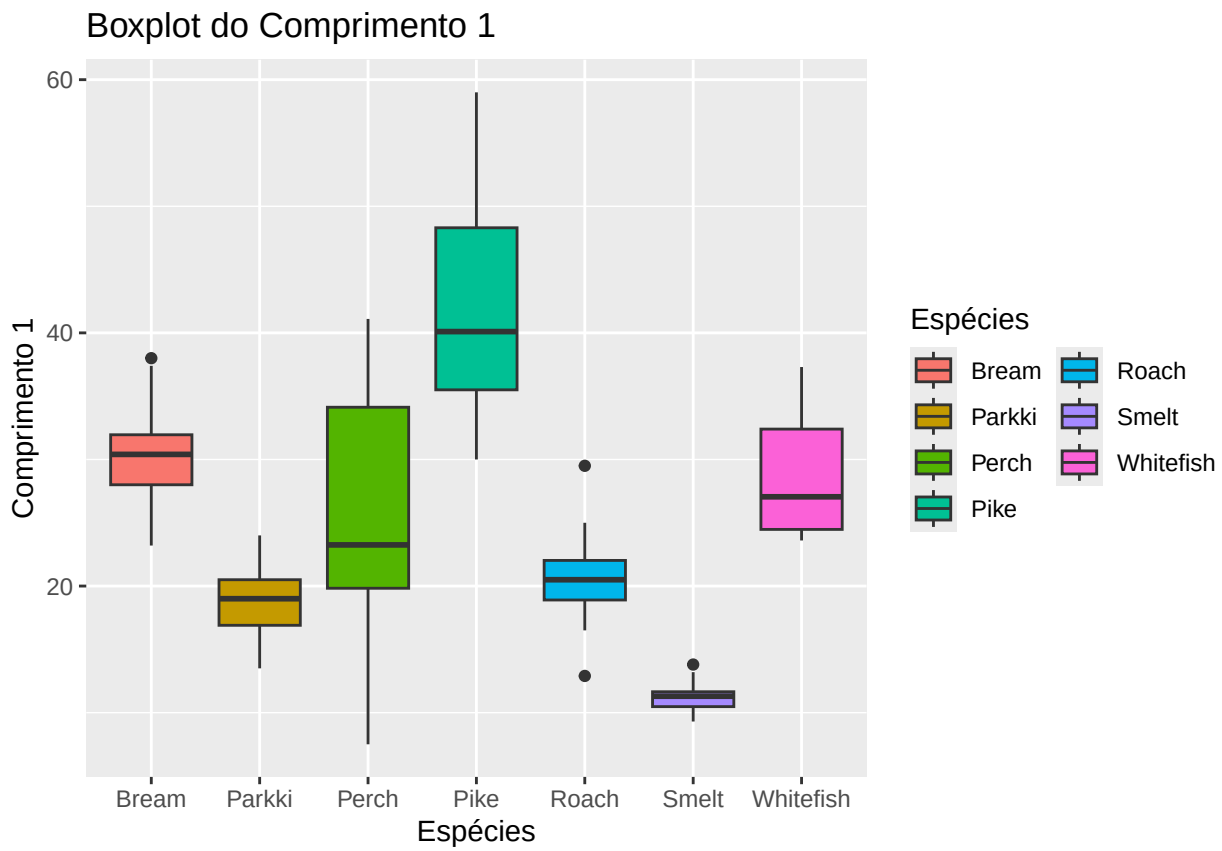
```r
# Estatísticas descritivas
summary(data)
```

```
##    Species              Weight          Length1         Length2
##  Length:159         Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Class :character   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
##  Mode  :character   Median : 273.0   Median :25.20   Median :27.30
##                     Mean   : 398.3   Mean   :26.25   Mean   :28.42
##                     3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##                     Max.   :1650.0   Max.   :59.00   Max.   :63.40
##     Length3          Height           Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386
##  Median :29.40   Median : 7.786   Median :4.248
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
```

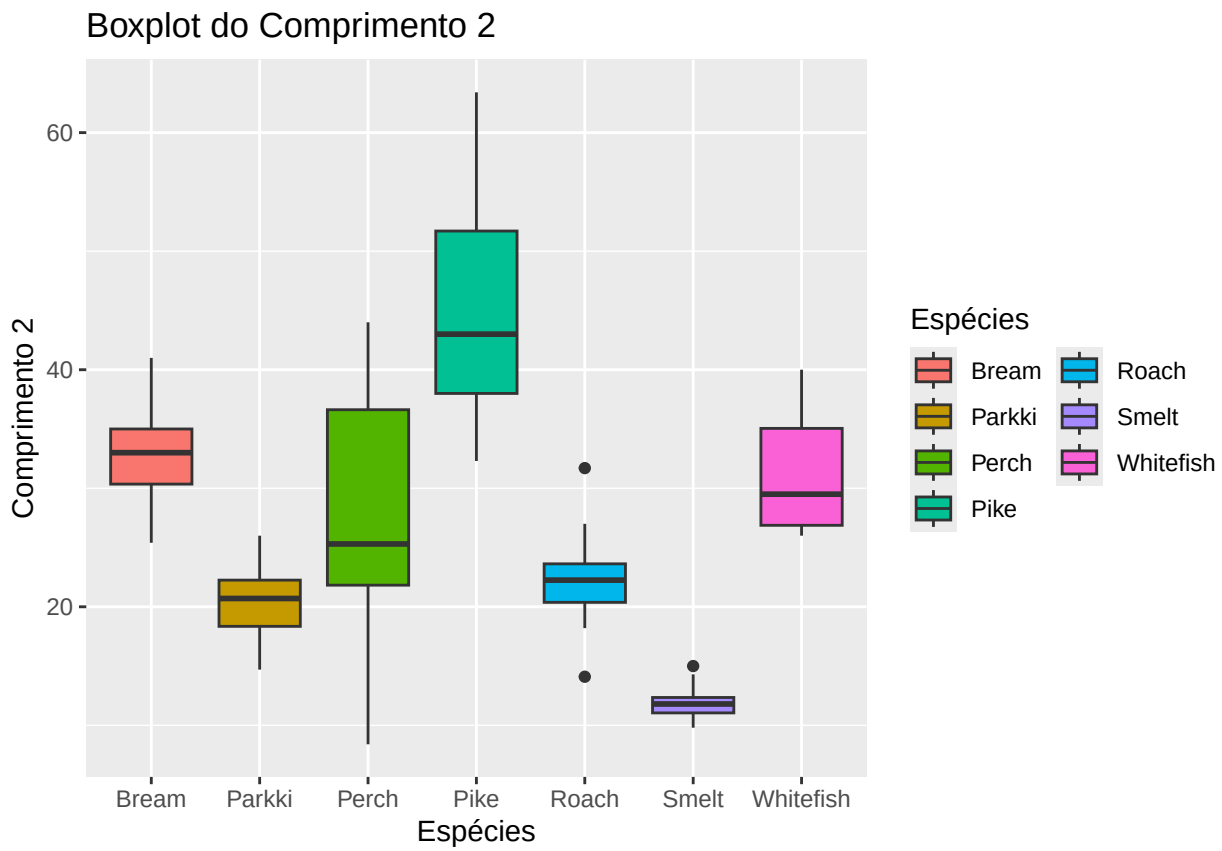
```r
# Histogramas das variáveis numéricas
library(ggplot2)
library(gridExtra)

p1 <- ggplot(data, aes(x = Weight)) + geom_histogram(binwidth = 50, fill = "blue", alpha = 0.7) + theme
p2 <- ggplot(data, aes(x = Length1)) + geom_histogram(binwidth = 1, fill = "green", alpha = 0.7) + theme
p3 <- ggplot(data, aes(x = Length2)) + geom_histogram(binwidth = 1, fill = "red", alpha = 0.7) + theme_
p4 <- ggplot(data, aes(x = Length3)) + geom_histogram(binwidth = 1, fill = "purple", alpha = 0.7) + ther
p5 <- ggplot(data, aes(x = Height)) + geom_histogram(binwidth = 1, fill = "orange", alpha = 0.7) + theme
p6 <- ggplot(data, aes(x = Width)) + geom_histogram(binwidth = 0.1, fill = "pink", alpha = 0.7) + theme

grid.arrange(p1, p2, p3, p4, p5, p6, ncol = 2)
```



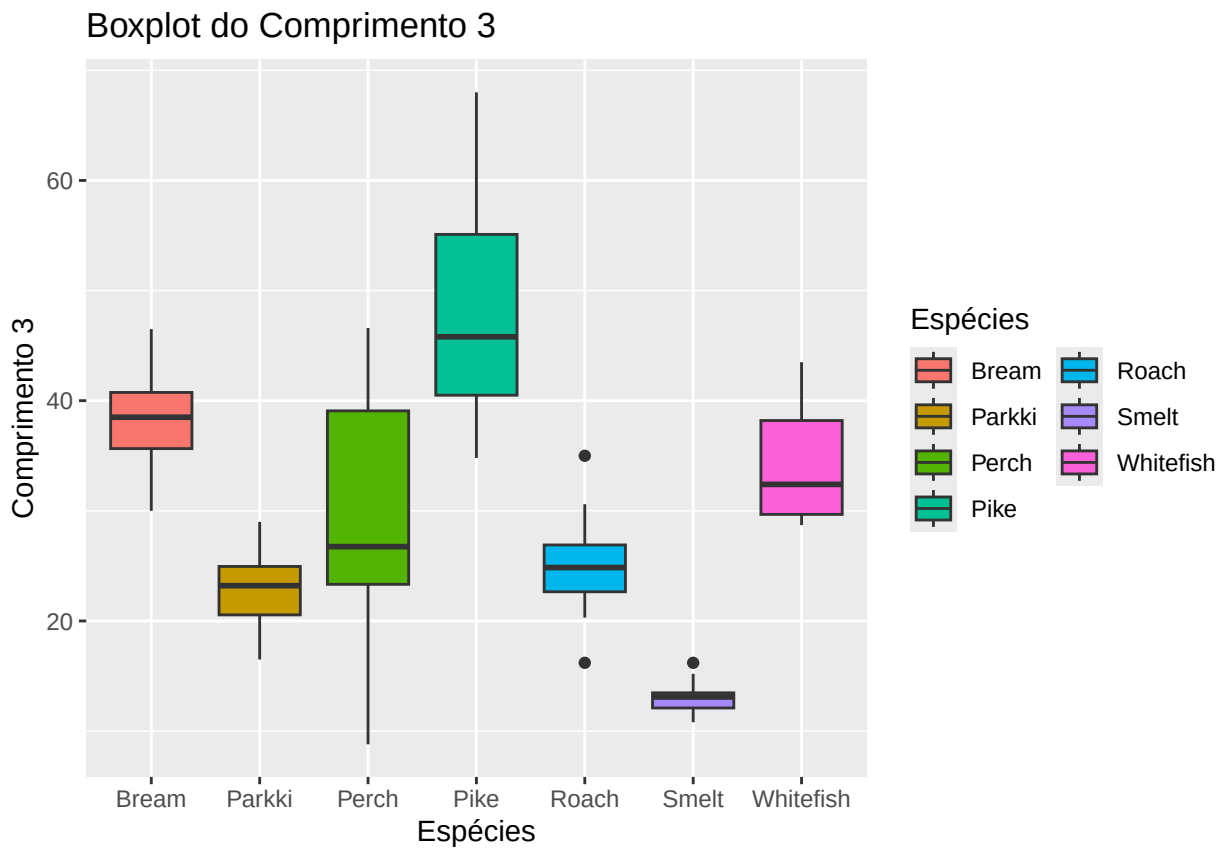

```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Length1, fill = Species)) +
  labs(title = "Boxplot do Comprimento 1", x = "Espécies", y = "Comprimento 1") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```



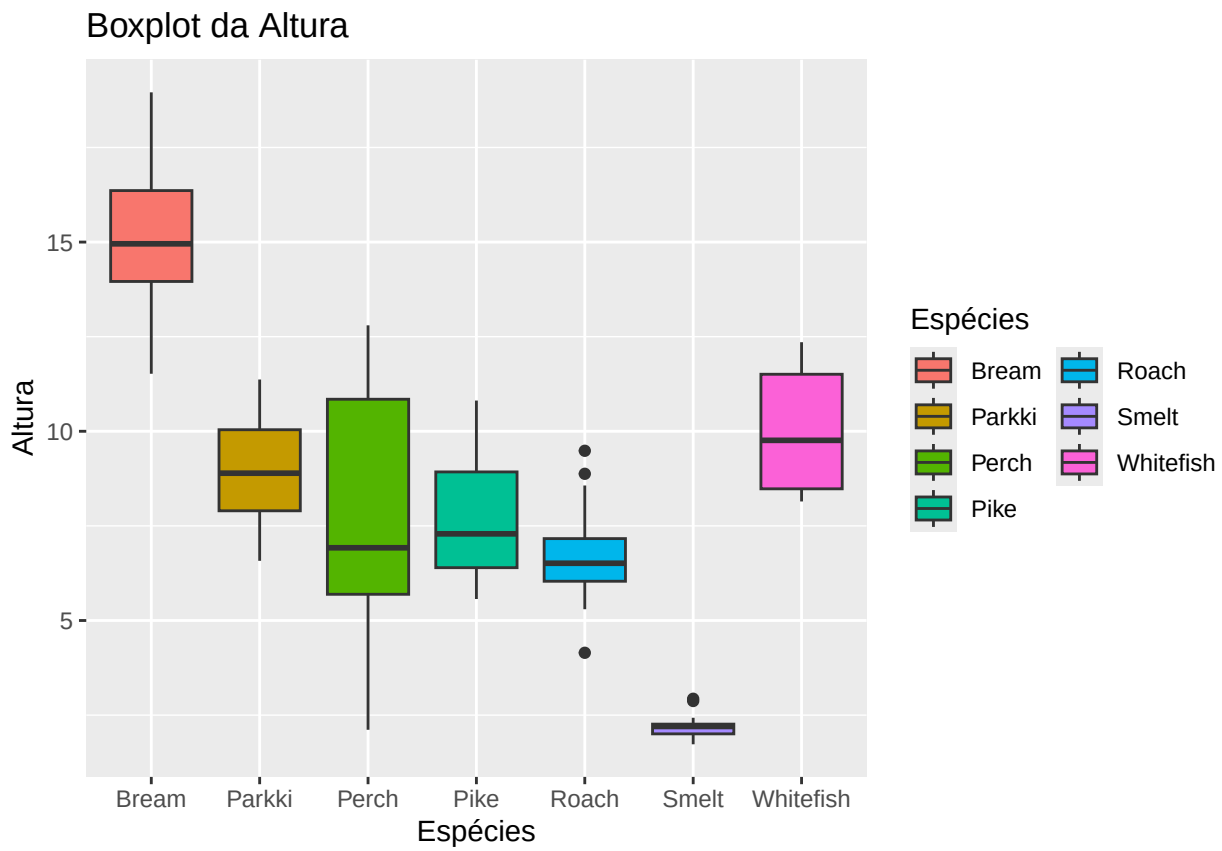

```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Length2, fill = Species)) +
  labs(title = "Boxplot do Comprimento 2", x = "Espécies", y = "Comprimento 2") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```



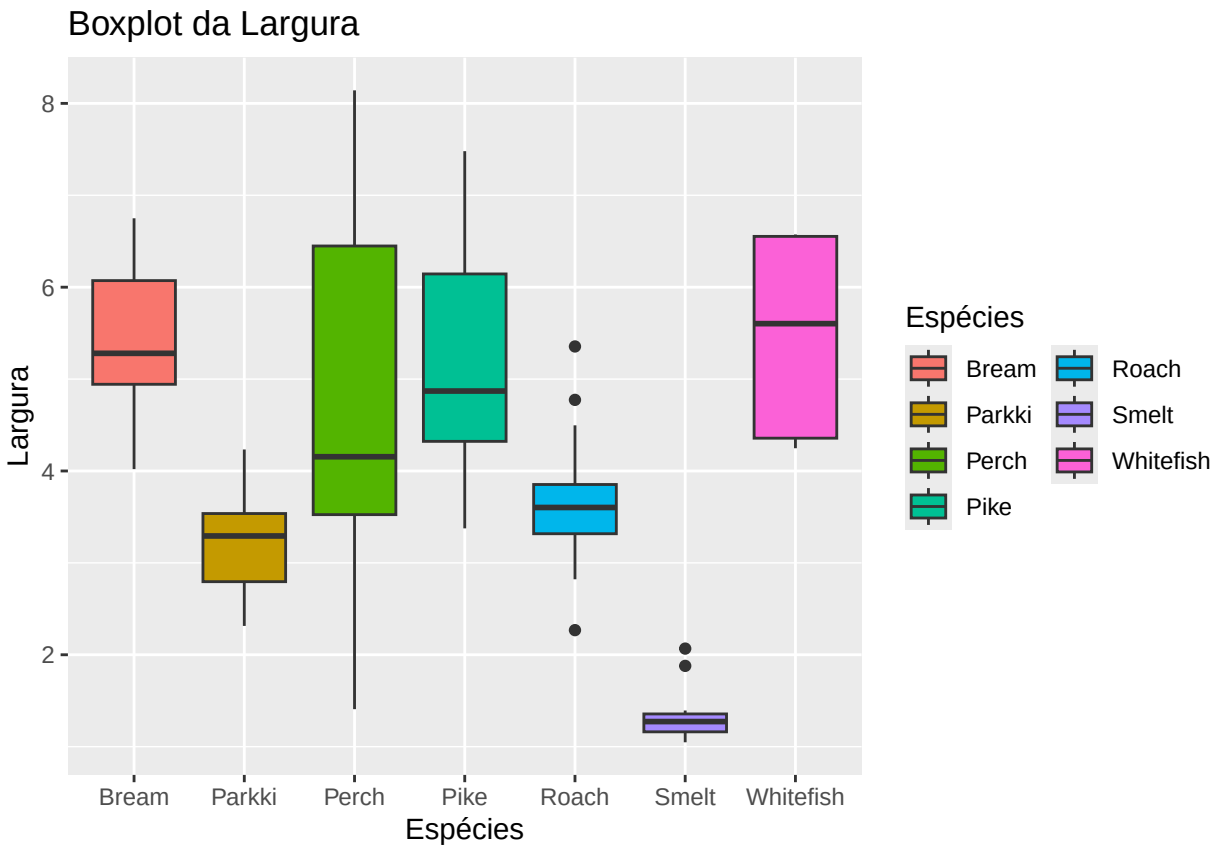

```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Length3, fill = Species)) +
  labs(title = "Boxplot do Comprimento 3", x = "Espécies", y = "Comprimento 3") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```



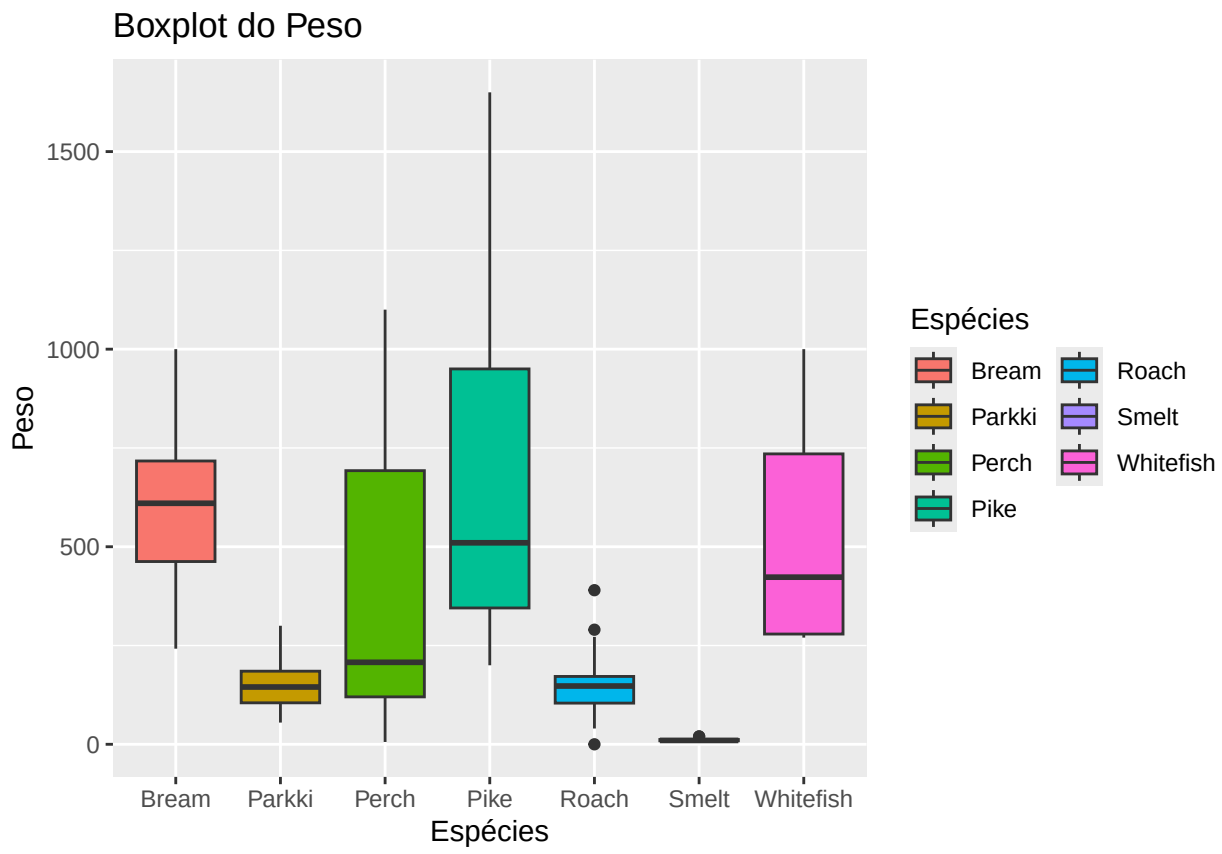

```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Height, fill = Species)) +
  labs(title = "Boxplot da Altura", x = "Espécies", y = "Altura") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```



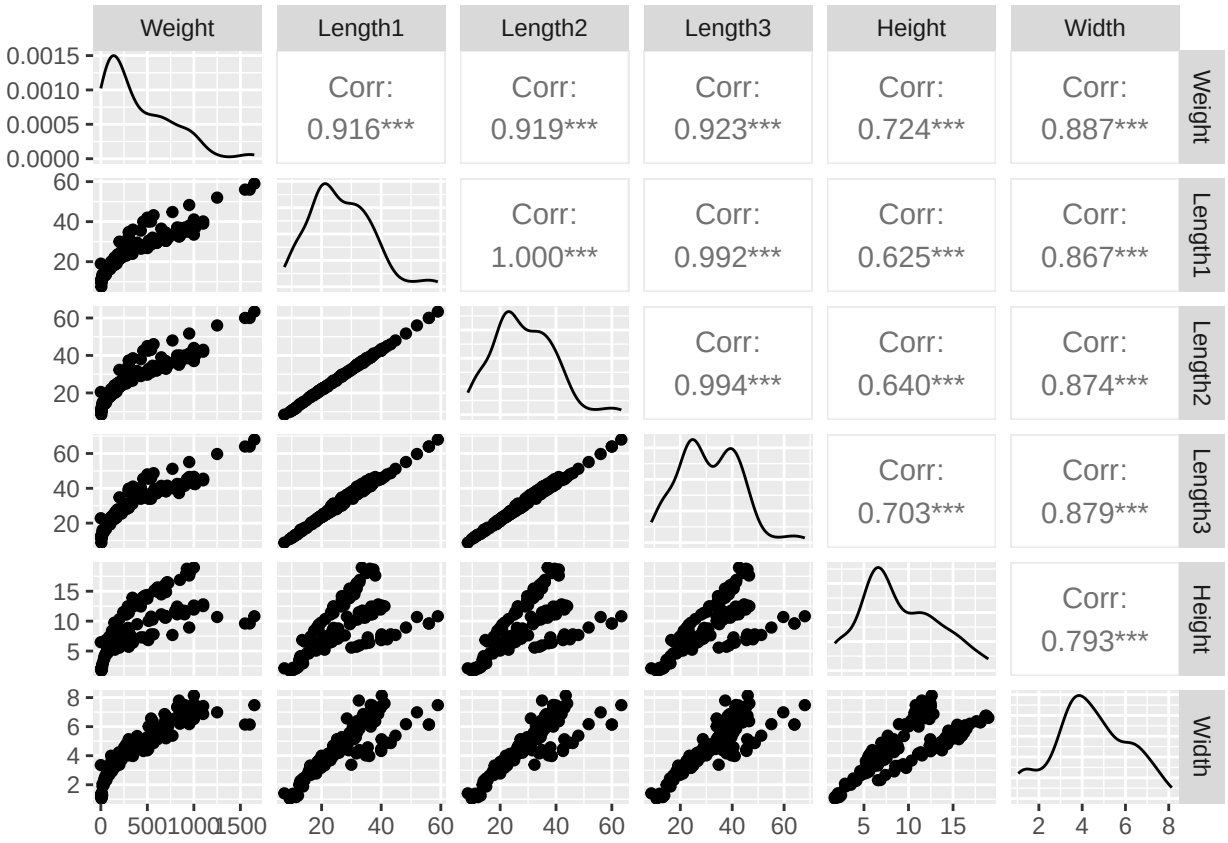

```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Width, fill = Species)) +
  labs(title = "Boxplot da Largura", x = "Espécies", y = "Largura") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```




```
data %>%
  ggplot() +
  geom_boxplot(aes(x = Species, y = Weight, fill = Species)) +
  labs(title = "Boxplot do Peso", x = "Espécies", y = "Peso") +
  guides(fill = guide_legend(title = "Espécies", ncol = 2))
```




```
# gráfico de dispersão entre todas as variáveis exceto "Species"
ggpairs(data[, -which(names(data) == "Species")],
        lower = list(continuous = "points"),
        diag = list(continuous = "densityDiag"),
        upper = list(continuous = "cor"))
```



Tratamento dos Dados

Primeiramente, renomeamos as colunas para nomes mais descritivos e de fácil compreensão: 'Especie', 'Peso', 'Comprimento_Vertical', 'Comprimento_Diagonal', 'Comprimento_Cruzado', 'Altura' e 'Largura'.

Em seguida, removemos os registros de peixes com peso nulo, pois esses dados não seriam úteis para o modelo de predição do peso dos peixes.

Transformamos a variável Especie em um fator, permitindo uma melhor análise categórica dessa variável.

Para evitar problemas de multicolinearidade devido a alta correlação das variáveis de comprimento, criamos uma nova variável chamada Comprimento_Geral, que é a média das três medidas de comprimento (Comprimento_Vertical, Comprimento_Diagonal e Comprimento_Cruzado).

Após a criação da variável Comprimento_Geral, as colunas individuais de comprimento foram removidas, deixando o conjunto de dados mais enxuto e focado nas variáveis realmente necessárias para a análise.

```
colnames(data) <- c("Especie", "Peso", "Comprimento_Vertical",
"Comprimento_Diagonal",
  "Comprimento_Cruzado", "Altura", "Largura")

# Remover peixes com peso nulo
data<- data %>% filter(Peso>0)

data$Especie <- as.factor(data$Especie)

data<-data %>% mutate(Comprimento_Geral = (Comprimento_Vertical + Comprimento_Diagonal+Comprimento_Cruza
```

```r
data<-data %>% select(-c("Comprimento_Vertical", "Comprimento_Diagonal","Comprimento_Cruzado"))
```

Ajuste do Modelo de Regressão Linear Múltipla

Ajustamos um modelo de regressão linear múltipla para estimar o peso com todas as variáveis.

```r
modelo <- lm(Peso ~ Especie + Comprimento_Geral + Altura + Largura, data = data)
summary(modelo)
```

```
##
## Call:
## lm(formula = Peso ~ Especie + Comprimento_Geral + Altura + Largura,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228.03  -56.48   -8.36   32.25  404.69
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -797.894     84.372  -9.457  < 2e-16 ***
## EspecieParkki       94.827     47.728   1.987 0.048788 *
## EspeciePerch        72.749     80.885   0.899 0.369889
## EspeciePike       -252.511    122.799  -2.056 0.041511 *
## EspecieRoach        47.066     76.376   0.616 0.538679
## EspecieSmelt       338.512     89.614   3.777 0.000229 ***
## EspecieWhitefish    60.238     78.003   0.772 0.441197
## Comprimento_Geral   37.257      3.793   9.822  < 2e-16 ***
## Altura              10.835     13.141   0.824 0.410984
## Largura             -2.330     24.152  -0.096 0.923263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95.47 on 148 degrees of freedom
## Multiple R-squared:  0.9329, Adjusted R-squared:  0.9288
## F-statistic: 228.5 on 9 and 148 DF,  p-value: < 2.2e-16
```

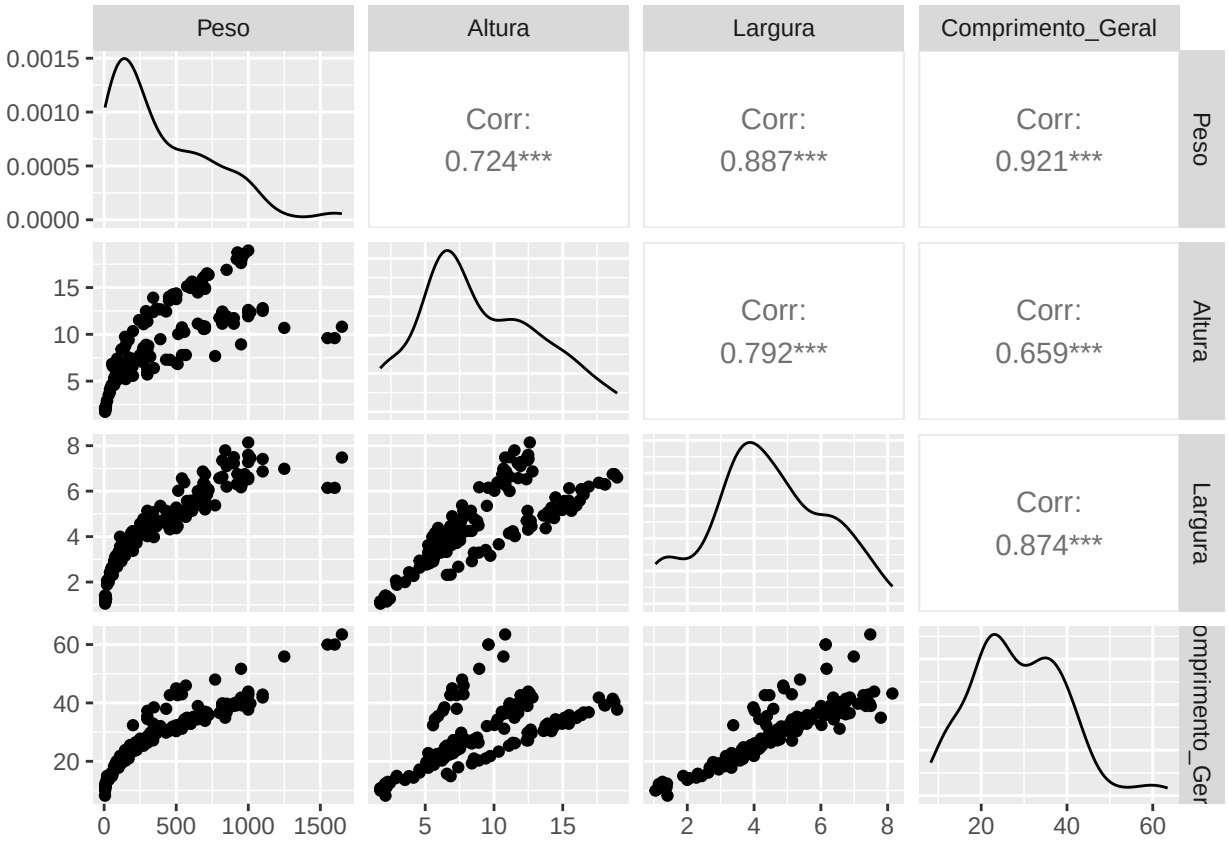
Note que neste caso o intercepto não possui interpretação já que representaria o valor esperado para o peso de um peixe da espécie de referência (Bream) com altura, largura e comprimento iguais a 0. Para tornar o intercepto interpretável, nós centralizamos as variáveis.

```r
# Ajustando o modelo com as variáveis centralizadas diretamente na fórmula (Linear)
modelo <- lm(Peso ~ Especie + I(Comprimento_Geral - mean(Comprimento_Geral)) +
                I(Altura - mean(Altura)) +
                I(Largura - mean(Largura)), data = data)

# Resumo do modelo centralizado
summary(modelo)
```

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

```
## lm(formula = Peso ~ Especie + I(Comprimento_Geral - mean(Comprimento_Geral)) +
##     I(Altura - mean(Altura)) + I(Largura - mean(Largura)), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228.03  -56.48   -8.36   32.25  404.69
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  357.688     61.950   5.774
## EspecieParkki                                 94.827     47.728   1.987
## EspeciePerch                                  72.749     80.885   0.899
## EspeciePike                                 -252.511    122.799  -2.056
## EspecieRoach                                  47.066     76.376   0.616
## EspecieSmelt                                 338.512     89.614   3.777
## EspecieWhitefish                              60.238     78.003   0.772
## I(Comprimento_Geral - mean(Comprimento_Geral))  37.257      3.793   9.822
## I(Altura - mean(Altura))                      10.835     13.141   0.824
## I(Largura - mean(Largura))                    -2.330     24.152  -0.096
##                                             Pr(>|t|)
## (Intercept)                                  4.4e-08 ***
## EspecieParkki                               0.048788 *
## EspeciePerch                                0.369889
## EspeciePike                                 0.041511 *
## EspecieRoach                                0.538679
## EspecieSmelt                                0.000229 ***
## EspecieWhitefish                            0.441197
## I(Comprimento_Geral - mean(Comprimento_Geral))  < 2e-16 ***
## I(Altura - mean(Altura))                    0.410984
## I(Largura - mean(Largura))                  0.923263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95.47 on 148 degrees of freedom
## Multiple R-squared:  0.9329, Adjusted R-squared:  0.9288
## F-statistic: 228.5 on 9 and 148 DF,  p-value: < 2.2e-16
```

```
ggpairs(data[, -which(names(data) == "Especie")],
        lower = list(continuous = "points"),
        diag = list(continuous = "densityDiag"),
        upper = list(continuous = "cor"))
```



Veja que no novo modelo, a correlação de algumas variáveis ainda é alta, vamos calcular o VIF e verificar se a multicolinearidade está causando efeitos nas estimativas dos coeficientes de regressão.

```
library(car)
vif(modelo)
```

```
##                                                 GVIF Df GVIF^(1/(2*Df))
## Especie                                    225.27675  6        1.570579
## I(Comprimento_Geral - mean(Comprimento_Geral))  28.76168  1        5.362991
## I(Altura - mean(Altura))                    54.87994  1        7.408099
## I(Largura - mean(Largura))                  28.66588  1        5.354053
```

Como o VIF está abaixo de 10 para todas as variáveis, podemos assumir que o nosso modelo não será afetado pela multicolinearidade.

Teste com seleção de modelos

Como é possível ver nos gráficos de dispersão acima, as variáveis comprimento, largura e altura aparentam ter uma relação quadrática com o peso. Deste modo, nosso objetivo agora é criar um dataframe considerando contribuição linear e quadrática das variáveis e submetê-lo a teste de seleção de modelos.

```
xy <- data %>%
  mutate(
    Comprimento_Geral_c = Comprimento_Geral - mean(Comprimento_Geral, na.rm = TRUE),
```

```
    Altura_c = Altura - mean(Altura, na.rm = TRUE),
    Largura_c = Largura - mean(Largura, na.rm = TRUE)
  )

# Criando o dataframe xy2 com termos lineares e quadráticos
xy2 <- xy %>%
  mutate(
    Comprimento_Geral_c2 = Comprimento_Geral_c^2,
    Altura_c2 = Altura_c^2,
    Largura_c2 = Largura_c^2
  ) %>%
  select(Peso, Especie,
         Comprimento_Geral_c, Comprimento_Geral_c2,
         Altura_c, Altura_c2, Largura_c, Largura_c2) %>%
  rename(y = Peso)

# Visualizando as primeiras linhas do dataframe xy2
head(xy2)
```

```
##     y Especie Comprimento_Geral_c Comprimento_Geral_c2 Altura_c Altura_c2
## 1 242   Bream          -2.4797468            6.1491444  2.53321  6.417154
## 2 290   Bream          -1.5130802            2.2894116  3.49321 12.202517
## 3 340   Bream          -1.5130802            2.2894116  3.39101 11.498950
## 4 363   Bream           0.9202532            0.8468659  3.74321 14.011622
## 5 430   Bream           1.1535865            1.3307618  3.45721 11.952302
## 6 450   Bream           1.7202532            2.9592710  4.61561 21.303857
##     Largura_c Largura_c2
## 1 -0.40423165 0.16340322
## 2 -0.11863165 0.01407347
## 3  0.27186835 0.07391240
## 4  0.03126835 0.00097771
## 5  0.70976835 0.50377112
## 6  0.50316835 0.25317839
```

Utilizando AIC como critério, vamos identificar o melhor modelo

```
library(bestglm)
library(knitr)
modelos <- bestglm(xy2, IC = "AIC")
modelos$Subsets
```

```
##    Intercept     y Especie Comprimento_Geral_c Comprimento_Geral_c2 Altura_c
## 0       TRUE FALSE   FALSE               FALSE                FALSE    FALSE
## 1       TRUE FALSE    TRUE               FALSE                FALSE    FALSE
## 2       TRUE FALSE    TRUE               FALSE                FALSE    FALSE
## 3       TRUE FALSE    TRUE               FALSE                 TRUE    FALSE
## 4       TRUE FALSE    TRUE                TRUE                 TRUE    FALSE
## 5*      TRUE FALSE    TRUE                TRUE                 TRUE    FALSE
## 6       TRUE FALSE    TRUE                TRUE                 TRUE     TRUE
## 7       TRUE  TRUE    TRUE                TRUE                 TRUE     TRUE
##    Altura_c2 Largura_c logLikelihood       AIC
## 0      FALSE     FALSE    -193.46218  386.9244
```

```
## 1      FALSE     FALSE    -147.63381  307.2676
## 2      FALSE      TRUE    -108.60015  231.2003
## 3      FALSE      TRUE     -85.24871  186.4974
## 4      FALSE      TRUE     -55.48111  128.9622
## 5*      TRUE      TRUE     -44.28972  108.5794
## 6       TRUE      TRUE     -44.24544  110.4909
## 7       TRUE      TRUE    -607.91014 1239.8203
```

```r
# Identificando o melhor modelo
melhor <- which(modelos$Subsets$AIC == min(modelos$Subsets$AIC))
numvar <- ncol(xy2) - 1 # total de variáveis consideradas inicialmente
varincluidas <- modelos$Subsets[melhor, 2:(numvar + 1)] # variáveis escolhidas

# Exibindo as variáveis escolhidas
varincluidas %>% kable(booktabs = TRUE)
```

	y	Especie	Comprimento_Geral_c	Comprimento_Geral_c2	Altura_c	Altura_c2	Largura_c
5*	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE

```r
# Ajustando o modelo com as variáveis selecionadas
modeloescolhidoAIC <- lm(y ~ ., data = xy2[, c(which(varincluidas == TRUE), which(names(xy2) == "y"))])

# Resumo do modelo escolhido
summary(modeloescolhidoAIC)
```

```
##
## Call:
## lm(formula = y ~ ., data = xy2[, c(which(varincluidas == TRUE),
##     which(names(xy2) == "y"))])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -174.012  -23.060    0.683   16.478  186.456
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           309.97441   16.08498  19.271  < 2e-16 ***
## EspecieParkki          56.90613   24.59879   2.313   0.0221 *
## EspeciePerch           -6.72907   17.96003  -0.375   0.7084
## EspeciePike          -217.39658   30.95585  -7.023 7.49e-11 ***
## EspecieRoach          -11.63432   20.82780  -0.559   0.5773
## EspecieSmelt           29.02564   25.26918   1.149   0.2526
## EspecieWhitefish       44.41274   26.63875   1.667   0.0976 .
## Comprimento_Geral_c    19.73265    2.24015   8.809 3.26e-15 ***
## Comprimento_Geral_c2    0.60770    0.03839  15.828  < 2e-16 ***
## Altura_c2               2.16844    0.32075   6.760 3.02e-10 ***
## Largura_c              87.09525   11.58397   7.519 5.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.54 on 147 degrees of freedom
```

```
## Multiple R-squared:  0.9813, Adjusted R-squared:   0.98
## F-statistic: 771.6 on 10 and 147 DF,  p-value: < 2.2e-16
```

O melhor modelo encontrado inclui a contribuição quadrática da altura mas não a linear, porém é crucial manter as contribuições de ordem mais baixa se as de ordem mais alta estão incluídas. Sendo assim, temos o seguinte modelo:

```
xy3 <- xy2 %>% select(-Largura_c2)
modelo2<-lm(y~.,xy3)
summary(modelo2)
```

```
##
## Call:
## lm(formula = y ~ ., data = xy3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.887  -25.805   -0.456   17.487  159.300
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          192.61254   37.37428   5.154 8.16e-07 ***
## EspecieParkki         80.49031   24.69731   3.259 0.001391 **
## EspeciePerch         156.80333   50.44513   3.108 0.002262 **
## EspeciePike           58.15048   85.23074   0.682 0.496148
## EspecieRoach         130.54974   45.83127   2.848 0.005027 **
## EspecieSmelt         225.85578   62.01533   3.642 0.000375 ***
## EspecieWhitefish     171.54965   44.91270   3.820 0.000197 ***
## Comprimento_Geral_c   12.27272    3.05647   4.015 9.46e-05 ***
## Comprimento_Geral_c2   0.70577    0.04668  15.119  < 2e-16 ***
## Altura_c              36.98804   10.71549   3.452 0.000729 ***
## Altura_c2              0.93763    0.47213   1.986 0.048913 *
## Largura_c             62.20087   13.30133   4.676 6.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.77 on 146 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9814
## F-statistic: 754.6 on 11 and 146 DF,  p-value: < 2.2e-16
```

Resultados

Vamos inicialmente definir os parâmetros para o modelo linear, com todas as variáveis centralizadas:

Intercepto (β_0)

- **Intercepto (357.688):** Este é o valor esperado do peso do peixe (variável dependente) quando todas as variáveis preditoras são iguais a zero (ou seja, para a espécie de referência (Bream) e para valores médios das variáveis contínuas centralizadas).

Variáveis Categóricas (Espécie)

Os coeficientes para as espécies representam a diferença média no peso dos peixes dessa espécie em comparação com a espécie de referência (Bream).

- **EspecieParkki (94.827):** Em média, peixes da espécie Parkki pesam 94.827 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspeciePerch (72.749):** Em média, peixes da espécie Perch pesam 72.749 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspeciePike (-252.511):** Em média, peixes da espécie Pike pesam 252.511 unidades a menos do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieRoach (47.066):** Em média, peixes da espécie Roach pesam 47.066 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieSmelt (338.512):** Em média, peixes da espécie Smelt pesam 338.512 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieWhitefish (60.238):** Em média, peixes da espécie Whitefish pesam 60.238 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.

Variáveis Contínuas (Centralizadas)

Os coeficientes para as variáveis contínuas representam a mudança no peso do peixe associada a um aumento de uma unidade nessa variável, com todas as outras variáveis mantidas constantes.

- **Comprimento Geral (37.257):** Para cada aumento de uma unidade no comprimento geral (centralizado), espera-se que o peso do peixe aumente em 37.257 unidades, mantendo todas as outras variáveis constantes.
- **Altura (10.835):** Para cada aumento de uma unidade na altura (centralizada), espera-se que o peso do peixe aumente em 10.835 unidades, mantendo todas as outras variáveis constantes.
- **Largura (-2.330):** Para cada aumento de uma unidade na largura (centralizada), espera-se que o peso do peixe diminua em 2.330 unidades, mantendo todas as outras variáveis constantes.

Combinação Linear do Modelo

A equação do modelo de regressão linear é dada por:

$$Y = \beta_0 + \beta_1 \times \text{EspecieParkki} + \beta_2 \times \text{EspeciePerch} + \beta_3 \times \text{EspeciePike} + \beta_4 \times \text{EspecieRoach} + \beta_5 \times \text{EspecieSmelt} + \beta_6 \times \text{Esp}$$
$$+ \beta_7 \times \text{ComprimentoGeral} + \beta_8 \times \text{Altura} + \beta_9 \times \text{Largura}$$

onde:

$$\beta_0 = 357.688$$
$$\beta_1 = 94.827$$
$$\beta_2 = 72.749$$
$$\beta_3 = -252.511$$
$$\beta_4 = 47.066$$
$$\beta_5 = 338.512$$
$$\beta_6 = 60.238$$
$$\beta_7 = 37.257$$
$$\beta_8 = 10.835$$
$$\beta_9 = -2.330$$

Para o modelo 2 com termos polinomiais, temos a seguinte interpretação dos parâmetros:

Intercepto (β_0)

- **Intercepto (192.61254):** Este é o valor esperado do peso do peixe (variável dependente) quando todas as variáveis preditoras são iguais a zero (ou seja, para a espécie de referência (Bream) e para valores médios das variáveis contínuas centralizadas).

Variáveis Categóricas (Espécie)

Os coeficientes para as espécies representam a diferença média no peso dos peixes dessa espécie em comparação com a espécie de referência (Bream).

- **EspecieParkki (80.49031):** Em média, peixes da espécie Parkki pesam 80.49031 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspeciePerch (156.80333):** Em média, peixes da espécie Perch pesam 156.80333 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspeciePike (58.15048):** Em média, peixes da espécie Pike pesam 58.15048 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieRoach (130.54974):** Em média, peixes da espécie Roach pesam 130.54974 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieSmelt (225.85578):** Em média, peixes da espécie Smelt pesam 225.85578 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.
- **EspecieWhitefish (171.54965):** Em média, peixes da espécie Whitefish pesam 171.54965 unidades a mais do que a espécie de referência, mantendo todas as outras variáveis constantes.

Variáveis Contínuas (Centralizadas)

Quando existem termos polinomiais no modelo de regressão múltipla, a interpretação dos parâmetros do modelo é diferente. Enquanto que para o caso linear o aumento de uma unidade de (β_j) representa a diferença no valor esperado da variável resposta ao se adicionar uma unidade na variável (X_j), para o modelo polinomial isso não é verdade, já que temos no modelo outro termo que depende de (X_j), logo todos os parâmetros referentes a (X_j) devem ser considerados simultaneamente. Logo, vamos interpretar conjuntamente os parâmetros de variáveis com contribuição linear e quadrática.

Vamos considerar o seguinte modelo de regressão polinomial como exemplo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

E o modelo após aumentar x_2 em uma unidade:

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1) + \beta_3 (x_2 + 1)^2$$

Vamos calcular a diferença $y^* - y$ para entender o impacto de aumentar x_2 em uma unidade.

Expandindo y^*:

$$\begin{aligned} y^* &= \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + 1) + \beta_3 (x_2 + 1)^2 \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 + \beta_3 (x_2^2 + 2x_2 + 1) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 + \beta_3 x_2^2 + 2\beta_3 x_2 + \beta_3 \end{aligned}$$

Então a diferença $y^* - y$ é:

$$\begin{aligned} y^* - y &= (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 + \beta_3 x_2^2 + 2\beta_3 x_2 + \beta_3) \\ &\quad - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2) \\ &= \beta_2 + 2\beta_3 x_2 + \beta_3 \\ &= \beta_2 + \beta_3 + 2\beta_3 x_2 \end{aligned}$$

Portanto, a diferença ao aumentar x_2 em uma unidade é:

$$y^* - y = (\beta_2 + \beta_3) + 2\beta_3 x_2$$

Com isso em mente, podemos interpretar os parâmetros da seguinte forma:

- **Comprimento Geral (12.27272) e Comprimento Geral ao Quadrado (0.70577):**
  - Para cada aumento de uma unidade no comprimento geral, o impacto no peso do peixe é dado pela soma dos coeficientes dos termos linear e quadrático do comprimento. Portanto, o efeito total é $(12.27272 + 0.70577) + 2 \times 0.70577 \times$ Comprimento unidades, mantendo todas as outras variáveis constantes.
  - Isso significa que o impacto do comprimento no peso do peixe depende do valor atual do comprimento.
  - O termo quadrático sugere uma relação não-linear, onde o efeito do comprimento aumenta com o comprimento.
- **Altura (36.98804) e Altura ao Quadrado (0.93763):** Para cada aumento de uma unidade na altura, o impacto no peso do peixe é dado por $(36.98804 + 0.93763) + 2 \times 0.93763 \times$ Altura unidades, mantendo todas as outras variáveis constantes.
- **Largura (62.20087):** Para cada aumento de uma unidade na largura (centralizada), espera-se que o peso do peixe aumente em 62.20087 unidades, mantendo todas as outras variáveis constantes.

Combinação Linear do Modelo

A equação do modelo de regressão linear é dada por:

$$y = \beta_0 + \beta_1 \times \text{EspecieParkki} + \beta_2 \times \text{EspeciePerch} + \beta_3 \times \text{EspeciePike} + \beta_4 \times \text{EspecieRoach} + \beta_5 \times \text{EspecieSmelt} + \beta_6 \times \text{Esp}$$
$$+ \beta_7 \times \text{ComprimentoGeral} + \beta_8 \times \text{ComprimentoGeral}^2 + \beta_9 \times \text{Altura} + \beta_{10} \times \text{Altura}^2 + \beta_{11} \times \text{Largura}$$

onde:

$$\begin{aligned}
\beta_0 &= 192.61254 \\
\beta_1 &= 80.49031 \\
\beta_2 &= 156.80333 \\
\beta_3 &= 58.15048 \\
\beta_4 &= 130.54974 \\
\beta_5 &= 225.85578 \\
\beta_6 &= 171.54965 \\
\beta_7 &= 12.27272 \\
\beta_8 &= 0.70577 \\
\beta_9 &= 36.98804 \\
\beta_{10} &= 0.93763 \\
\beta_{11} &= 62.20087
\end{aligned}$$

Apêndice

Para verificar a significância dos coeficientes do modelo, calculamos a estatística F.

```
# summary para o modelo
summary(modelo2)
```

```
##
## Call:
## lm(formula = y ~ ., data = xy3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.887  -25.805   -0.456   17.487  159.300
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           192.61254   37.37428   5.154 8.16e-07 ***
## EspecieParkki          80.49031   24.69731   3.259 0.001391 **
## EspeciePerch          156.80333   50.44513   3.108 0.002262 **
## EspeciePike            58.15048   85.23074   0.682 0.496148
## EspecieRoach          130.54974   45.83127   2.848 0.005027 **
## EspecieSmelt          225.85578   62.01533   3.642 0.000375 ***
## EspecieWhitefish      171.54965   44.91270   3.820 0.000197 ***
## Comprimento_Geral_c    12.27272    3.05647   4.015 9.46e-05 ***
## Comprimento_Geral_c2    0.70577    0.04668  15.119  < 2e-16 ***
## Altura_c               36.98804   10.71549   3.452 0.000729 ***
## Altura_c2               0.93763    0.47213   1.986 0.048913 *
```

```
## Largura_c             62.20087   13.30133   4.676 6.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.77 on 146 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9814
## F-statistic: 754.6 on 11 and 146 DF,  p-value: < 2.2e-16
```

Teste de Hipóteses

Teste F: As hipóteses para este teste são:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

$$H_A : \text{pelo menos um } \beta_j \neq 0,\ j = 1, 2, \ldots, p-1$$

A estatística F calculada é 754.64 com 11 e 146 graus de liberdade, e o p-valor correspondente é ≈ 0.

Como o p-valor é pequeno ($<< 0.05$), temos evidência para rejeitar H_0 e concluir que pelo menos um coeficiente é diferente de zero. // //

Teste t: Depois de verificar que o teste F é significativo, tem-se o interesse em testar se

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0, \quad \text{para algum} \quad j = 1, \ldots, p$$

Vamos realizar o teste t para cada coeficiente β_j para determinar se algum deles é significativamente diferente de zero.

Por meio da tabela:

```
##
## Call:
## lm(formula = y ~ ., data = xy3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -162.887  -25.805   -0.456   17.487  159.300
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          192.61254   37.37428   5.154 8.16e-07 ***
## EspecieParkki         80.49031   24.69731   3.259 0.001391 **
## EspeciePerch         156.80333   50.44513   3.108 0.002262 **
## EspeciePike           58.15048   85.23074   0.682 0.496148
## EspecieRoach         130.54974   45.83127   2.848 0.005027 **
## EspecieSmelt         225.85578   62.01533   3.642 0.000375 ***
## EspecieWhitefish     171.54965   44.91270   3.820 0.000197 ***
## Comprimento_Geral_c   12.27272    3.05647   4.015 9.46e-05 ***
## Comprimento_Geral_c2   0.70577    0.04668  15.119  < 2e-16 ***
## Altura_c              36.98804   10.71549   3.452 0.000729 ***
## Altura_c2              0.93763    0.47213   1.986 0.048913 *
## Largura_c             62.20087   13.30133   4.676 6.59e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.77 on 146 degrees of freedom
## Multiple R-squared:  0.9827, Adjusted R-squared:  0.9814
## F-statistic: 754.6 on 11 and 146 DF,  p-value: < 2.2e-16
```

Os resultados do modelo de regressão linear indicam que vários coeficientes têm p-valores menores que 0.05, fornecendo evidências para rejeitar a hipótese nula de que esses coeficientes são iguais a zero. Isso significa que essas variáveis têm um impacto significativo no peso dos peixes. Por outro lado, a variável 'EspeciePike' apresenta p-valor maior que 0.05, logo não possuímos evidências para rejeitar a hipótese de que seu coeficiente seja nulo.

Validação de modelo

```r
# Configurar uma semente aleatória para reprodutibilidade
set.seed(251495)

# Dividir os dados em treino (70%) e teste (30%)
n <- nrow(xy3)
indices_treino <- sample(seq_len(n), size = 0.7 * n)

# Criar os conjuntos de treino e teste
treino <- xy3[indices_treino, ]
teste <- xy3[-indices_treino, ]

# Ajustar o modelo de regressão linear com o conjunto de treino
modelo2 <- lm(y ~ ., data = treino)

# Resumo do modelo ajustado
summary(modelo2)
```

```
##
## Call:
## lm(formula = y ~ ., data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -106.984  -28.414   -1.838   20.699  148.505
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          140.47641   46.87821   2.997 0.003458 **
## EspecieParkki        110.33229   29.93998   3.685 0.000375 ***
## EspeciePerch         231.57684   61.50820   3.765 0.000284 ***
## EspeciePike          153.65845  102.19023   1.504 0.135887
## EspecieRoach         190.69429   57.03120   3.344 0.001172 **
## EspecieSmelt         312.54314   72.43724   4.315 3.82e-05 ***
## EspecieWhitefish     244.00684   53.94187   4.524 1.71e-05 ***
## Comprimento_Geral_c   10.92335    3.40525   3.208 0.001807 **
## Comprimento_Geral_c2   0.69958    0.05153  13.575  < 2e-16 ***
## Altura_c              50.01170   12.57652   3.977 0.000134 ***
```

```
## Altura_c2              0.59385    0.53937   1.101 0.273594
## Largura_c             46.80196   15.12548   3.094 0.002572 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.13 on 98 degrees of freedom
## Multiple R-squared:  0.985,  Adjusted R-squared:  0.9833
## F-statistic: 584.3 on 11 and 98 DF,  p-value: < 2.2e-16
```

```r
# Fazer previsões no conjunto de teste
predicoes <- predict(modelo2, newdata = teste)

# Calcular a raiz do erro médio quadrático (RMSE) para avaliar o modelo
mse <- mean((teste$y - predicoes)^2)
rmse <- sqrt(mse)
cat("Raíz do Erro Médio Quadrático (MSE) no conjunto de teste:", rmse, "\n")
```

```
## Raíz do Erro Médio Quadrático (MSE) no conjunto de teste: 57.80239
```

Referências

- Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*. 5th Edition.
- Seber, G. A. F., Lee, A. J. (2003). *Linear Regression Analysis*. 2nd Edition.
- Outros materiais e fontes utilizadas durante o curso.

Este relatório foi gerado automaticamente utilizando RMarkdown. Ele apresenta uma análise clara e detalhada do modelo de regressão linear múltipla aplicado ao dataset de peixes.