Palmer Penguins Data Analysis Series (Part 1): Exploratory Data Analysis and Simple Regression

Getting acquainted with our Antarctic friends and their morphometric relationships

Your Name

2025-01-01

Table of contents

Figure 1: Curious Adelie penguins beginning their data science journey - because every great analysis starts with getting to know your data!

Photo: African penguins at Boulders Beach, South Africa. Licensed under CC BY 2.0 via Wikimedia Commons

Palmer Penguins Data Analysis Series

This is **Part 1** of a 5-part series exploring penguin morphometrics:

1. **Part 1: EDA and Simple Regression** (This post)
2. Part 2: Multiple Regression and Species Effects
3. Part 3: Advanced Models and Cross-Validation
4. Part 4: Model Diagnostics and Interpretation
5. Part 5: Random Forest vs Linear Models

1 Introduction

Welcome to our comprehensive exploration of the Palmer penguins dataset! In this 5-part series, we'll journey through the complete data science workflow, from initial data exploration to advanced modeling techniques. The Palmer penguins dataset has become a beloved alternative to the iris dataset, providing real-world biological data that's both engaging and educationally valuable.

Collected by Dr. Kristen Gorman at Palmer Station Antarctica, this dataset contains morphometric measurements for three penguin species: Adelie (*Pygoscelis adeliae*), Chinstrap (*Pygoscelis antarcticus*), and Gentoo (*Pygoscelis papua*). Understanding these relationships is crucial for Antarctic ecology research, as body mass serves as a key indicator of penguin health and reproductive success.

In this first part, we'll focus on:

- Getting familiar with the Palmer penguins dataset
- Conducting thorough exploratory data analysis
- Understanding the relationships between morphometric variables
- Building our first simple regression model
- Establishing the foundation for more complex analyses in subsequent parts

By the end of this post, you'll have a solid understanding of the data structure and the strongest individual predictors of penguin body mass.

2 Prerequisites and Setup

Before we begin our Antarctic adventure, let's ensure we have the right tools:

Required Packages:

```
# Install required packages if not already installed
install.packages(c("palmerpenguins", "tidyverse", "broom", "corrplot",
                   "GGally", "patchwork", "knitr"))
```

Load Libraries:

```
library(palmerpenguins)
library(tidyverse)
library(broom)
library(corrplot)
library(GGally)
```

```r
library(patchwork)
library(knitr)

# Set theme for consistent plotting
theme_set(theme_minimal(base_size = 12))

# Set penguin-friendly colors
penguin_colors <- c("Adelie" = "#FF6B6B", "Chinstrap" = "#4ECDC4", "Gentoo" = "#45B7D1")
```

3 Meet the Penguins: Dataset Overview

Let's start by getting acquainted with our Antarctic research subjects:

```r
# Load the Palmer penguins data
data(penguins)

# Basic dataset information
cat("  Palmer Penguins Dataset Overview  \n")
```

```
  Palmer Penguins Dataset Overview
```

```r
cat("=====================================\n")
```

```
=====================================
```

```r
cat("Dimensions:", nrow(penguins), "observations ×", ncol(penguins), "variables\n\n")
```

```
Dimensions: 344 observations × 8 variables
```

```r
# Display variable information
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
```

```
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

3.1 Data Structure and Variables

Our dataset contains the following key measurements:

```
# Create a summary table of variables
variable_info <- tibble(
  Variable = names(penguins),
  Description = c(
    "Penguin species (Adelie, Chinstrap, Gentoo)",
    "Island location (Biscoe, Dream, Torgersen)",
    "Bill length in millimeters",
    "Bill depth in millimeters",
    "Flipper length in millimeters",
    "Body mass in grams",
    "Penguin sex (female, male)",
    "Study year (2007, 2008, 2009)"
  ),
  Type = map_chr(penguins, class)
)

kable(variable_info, caption = "Palmer Penguins Dataset Variables")
```

Table 1: Palmer Penguins Dataset Variables

Variable	Description	Type
species	Penguin species (Adelie, Chinstrap, Gentoo)	factor
island	Island location (Biscoe, Dream, Torgersen)	factor
bill_length_mm	Bill length in millimeters	numeric
bill_depth_mm	Bill depth in millimeters	numeric
flipper_length_mm	Flipper length in millimeters	integer
body_mass_g	Body mass in grams	integer
sex	Penguin sex (female, male)	factor
year	Study year (2007, 2008, 2009)	integer

3.2 Missing Data Assessment

Before diving into analysis, let's check for missing values:

```
# Check for missing values
missing_summary <- penguins %>%
  summarise_all(~sum(is.na(.))) %>%
  pivot_longer(everything(), names_to = "Variable", values_to = "Missing_Count") %>%
  mutate(Percentage = round(Missing_Count / nrow(penguins) * 100, 1)) %>%
  filter(Missing_Count > 0)

if(nrow(missing_summary) > 0) {
  kable(missing_summary, caption = "Missing Values Summary")
} else {
  cat(" No missing values found!")
}
```

Table 2: Missing Values Summary

Variable	Missing_Count	Percentage
bill_length_mm	2	0.6
bill_depth_mm	2	0.6
flipper_length_mm	2	0.6
body_mass_g	2	0.6
sex	11	3.2

```
# Create clean dataset for analysis
penguins_clean <- penguins %>%
  drop_na()

cat("\n After removing missing values:")
```

```
 After removing missing values:
```

```
cat("\n   Original dataset:", nrow(penguins), "rows")
```

```
   Original dataset: 344 rows
```