Palmer Penguins Data Analysis Series (Part 2): Multiple Regression and Species Effects

Discovering the power of combining predictors and biological groupings

Your Name

2025-01-02

Table of contents

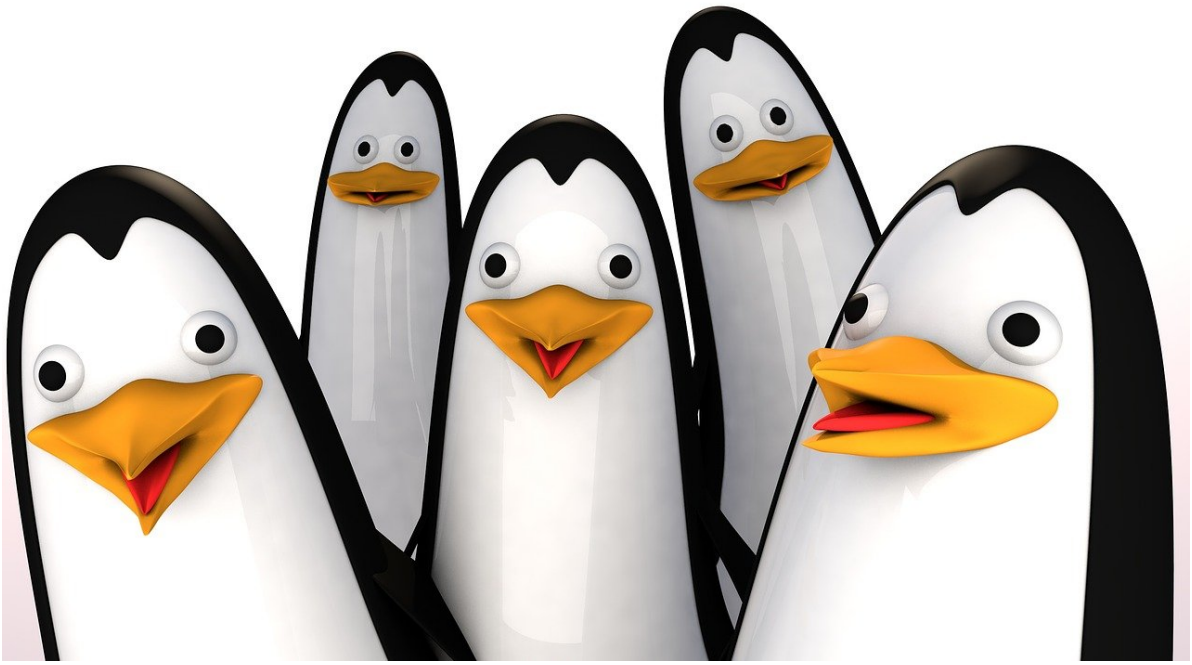

Figure 1: Two penguins collaborating on their regression analysis - because multiple predictors work better together!

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 2** of a 5-part series exploring penguin morphometrics:

1. Part 1: EDA and Simple Regression
2. **Part 2: Multiple Regression and Species Effects** (This post)
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 back to our Antarctic data science adventure! In Part 1, we discovered that flipper length alone can explain 76% of the variance in penguin body mass. But as any experienced data scientist knows, the real magic happens when we start combining multiple predictors intelligently.

In this second installment, we'll explore how incorporating additional morphometric measurements and, most importantly, species information can dramatically improve our predictive power. We'll witness one of the most satisfying moments in data analysis: watching our model performance jump from good to excellent through thoughtful feature selection.

In this post, we'll cover:

- Building multiple regression models with all morphometric predictors
- Understanding multicollinearity and variance inflation factors
- The dramatic impact of including species information
- Interaction effects between species and morphometric measurements
- Model comparison techniques to quantify improvements

By the end of this post, you'll see how our R^2 improves from 0.759 to over 0.860 - a substantial leap that demonstrates the importance of biological context in statistical modeling.

2 Quick Recap and Setup

"We learned that flipper length predicts body mass, but what if we combine forces?"

```
library(palmerpenguins)
library(tidyverse)
library(broom)
library(car)  # for VIF calculations
```

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

# Set theme and colors
theme_set(theme_minimal(base_size = 12))
penguin_colors <- c("Adelie" = "#FF6B6B", "Chinstrap" = "#4ECDC4", "Gentoo" = "#45B7D1")

# Load clean data and Part 1 baseline
data(penguins)
penguins_clean <- penguins %>% drop_na()
simple_model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_clean)

cat("  From Part 1: Flipper length alone R² =", round(glance(simple_model)$r.squared, 3))
```

```
  From Part 1: Flipper length alone R² = 0.762
```

3 Multiple Linear Regression

"Let's combine all our measurements and see what happens!"

3.1 Building Multiple Predictor Models

```r
# Build multiple regression with all morphometric variables
multiple_model <- lm(body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
                     data = penguins_clean)

# Extract key metrics with confidence intervals
multiple_metrics <- glance(multiple_model)
```

```r
multiple_coef <- tidy(multiple_model, conf.int = TRUE)

cat("  Multiple Regression Results:\n")
```

```
  Multiple Regression Results:
```

```r
cat("R² improvement:", round(multiple_metrics$r.squared, 3), "vs", round(glance(simple_model)
```

```
R² improvement: 0.764 vs 0.762 (simple)
```

```r
cat("RMSE:", round(sigma(multiple_model), 1), "grams\n")
```

```
RMSE: 393 grams
```

```r
# Check multicollinearity
vif_values <- vif(multiple_model)
cat("\n  Multicollinearity Check (VIF):\n")
```

```
  Multicollinearity Check (VIF):
```

```r
for(i in 1:length(vif_values)) {
  cat(sprintf("%s: %.1f %s\n", names(vif_values)[i], vif_values[i],
              ifelse(vif_values[i] < 5, " ", " ")))
}
```

```
bill_length_mm: 1.9
bill_depth_mm: 1.6
flipper_length_mm: 2.6
```

```r
# Key coefficients with confidence intervals
cat("\n  Key Effects (95% CI):\n")
```

```
  Key Effects (95% CI):
```

```r
for(i in 2:nrow(multiple_coef)) {
  term <- multiple_coef$term[i]
  est <- multiple_coef$estimate[i]
  ci_low <- multiple_coef$conf.low[i]
  ci_high <- multiple_coef$conf.high[i]
  cat(sprintf("%s: %.1f [%.1f, %.1f] g/mm\n", term, est, ci_low, ci_high))
}
```

```
bill_length_mm: 3.3 [-7.3, 13.8] g/mm
bill_depth_mm: 17.8 [-9.4, 45.0] g/mm
flipper_length_mm: 50.8 [45.8, 55.7] g/mm
```

4 The Species Revolution

"Wait... what if we account for the fact that we're different species?"

4.1 Adding Species Information

```
# Model with species as predictor
species_model <- lm(body_mass_g ~ bill_length_mm + bill_depth_mm +
                    flipper_length_mm + species, data = penguins_clean)

species_metrics <- glance(species_model)
species_coef <- tidy(species_model, conf.int = TRUE)

cat("  Species Model - Dramatic Improvement:\n")
```

```
  Species Model - Dramatic Improvement:
```

```
cat("R² jump:", round(multiple_metrics$r.squared, 3), "→", round(species_metrics$r.squared, 3
```

```
R² jump: 0.764 → 0.849
```

```
cat(" (+", round((species_metrics$r.squared - multiple_metrics$r.squared) * 100, 1), "%)\n")
```

```
 (+ 8.6 %)
```

```
cat("RMSE reduction:", round(sigma(multiple_model), 1), "→", round(sigma(species_model), 1),
```

```
RMSE reduction: 393 → 314.8 grams
```

```
# Species effects with confidence intervals
species_effects <- species_coef[grepl("species", species_coef$term), ]
cat("\n  Species Effects (vs Adelie baseline):\n")
```

```
  Species Effects (vs Adelie baseline):
```

```
for(i in 1:nrow(species_effects)) {
  term <- gsub("species", "", species_effects$term[i])
  est <- species_effects$estimate[i]
  ci_low <- species_effects$conf.low[i]
  ci_high <- species_effects$conf.high[i]
  cat(sprintf("%s: %+.0f [%+.0f, %+.0f] grams\n", term, est, ci_low, ci_high))
}
```

```
Chinstrap: -497 [-659, -335] grams
Gentoo: +965 [+686, +1244] grams
```

```
knit_exit()
```