# 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

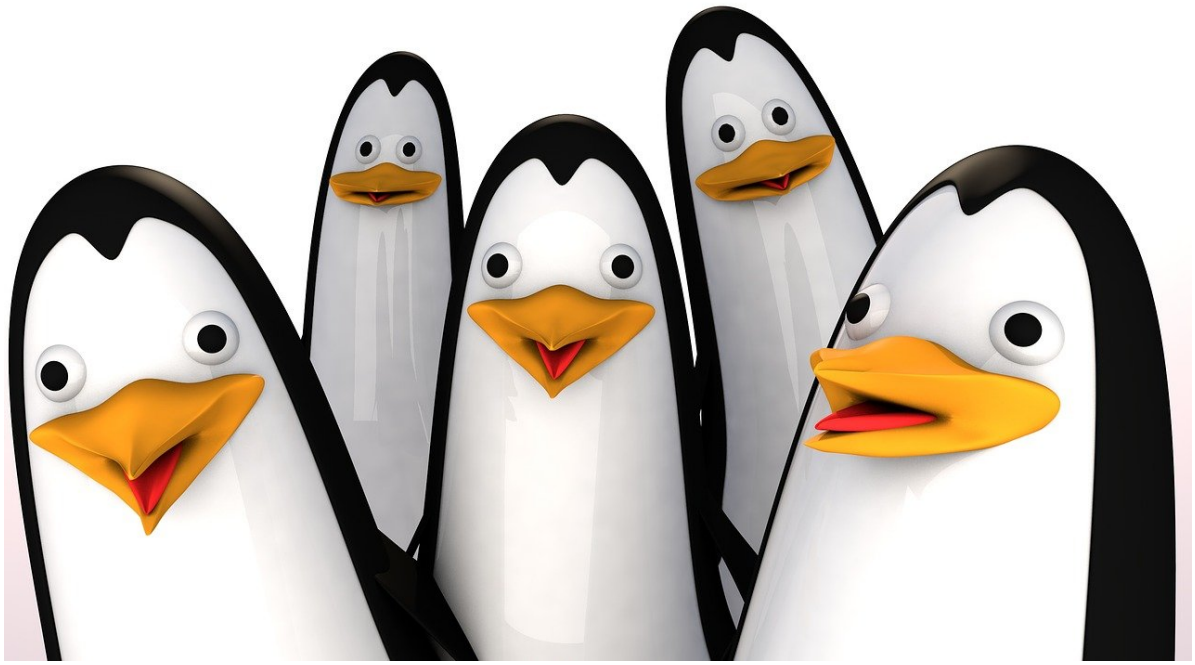


Figure 1: Two penguins collaborating on their regression analysis - because multiple predictors work better together!

*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))
}
```

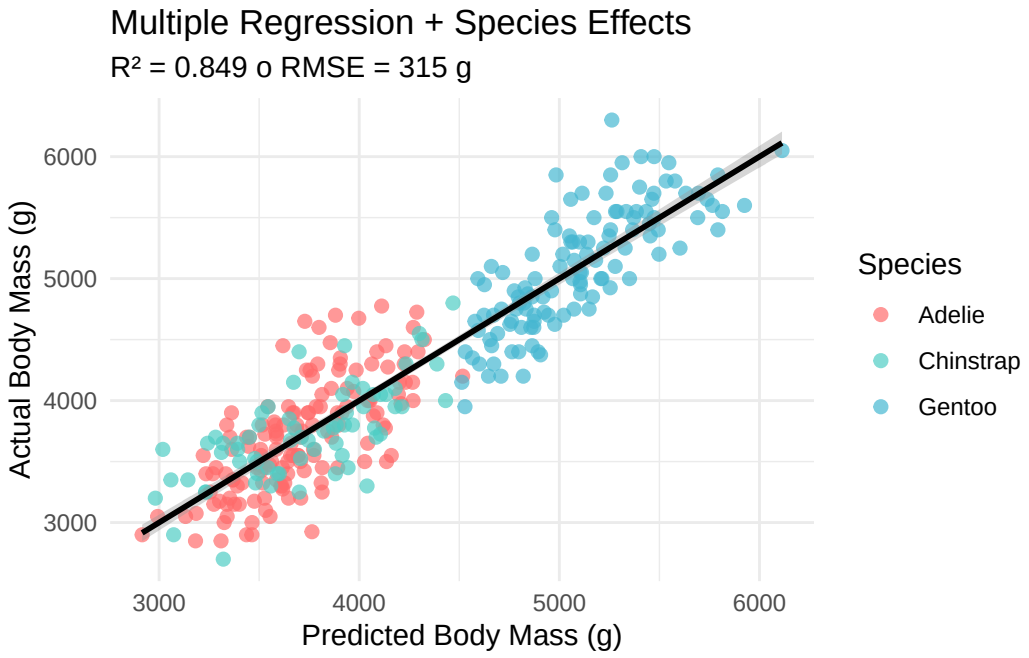
```
Chinstrap: -497 [-659, -335] grams
Gentoo: +965 [+686, +1244] grams
```

### 4.2 Visualizing the Improvement

```
# Create model comparison visualization
penguins_with_pred <- penguins_clean %>%
  mutate(
    species_pred = predict(species_model),
    species_resid = residuals(species_model)
  )

# Model performance comparison
comparison_plot <- ggplot(penguins_with_pred, aes(x = species_pred, y = body_mass_g, color =
  geom_point(alpha = 0.7, size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  scale_color_manual(values = penguin_colors) +
  labs(title = "Multiple Regression + Species Effects",
       subtitle = paste("R² =", round(species_metrics$r.squared, 3), "• RMSE =", round(sigma
       x = "Predicted Body Mass (g)", y = "Actual Body Mass (g)", color = "Species") +
  theme_minimal()

print(comparison_plot)
```



```
ggsave("species-model-performance.png", plot = comparison_plot, width = 8, height = 5, dpi =
```


Figure 2: Multiple regression model with species effects showing excellent predictive performance

## 5 Model Limitations and Assumptions

*"Before celebrating, let's check our assumptions!"*

```r
# Model diagnostic checks and limitations
species_residuals <- residuals(species_model)
outliers <- sum(abs(scale(species_residuals)) > 2.5)

cat("  Model Limitations:\n")
```

```
  Model Limitations:
```

```r
cat("• Potential outliers:", outliers, "observations (>2.5 SD)\n")
```

• Potential outliers: 7 observations (>2.5 SD)

```r
cat("• Assumes linear relationships within species\n")
```

• Assumes linear relationships within species

```r
cat("• Temporal scope: 2007-2009 data only\n")
```

• Temporal scope: 2007-2009 data only

```r
cat("• Geographic constraint: Palmer Station region\n")
```

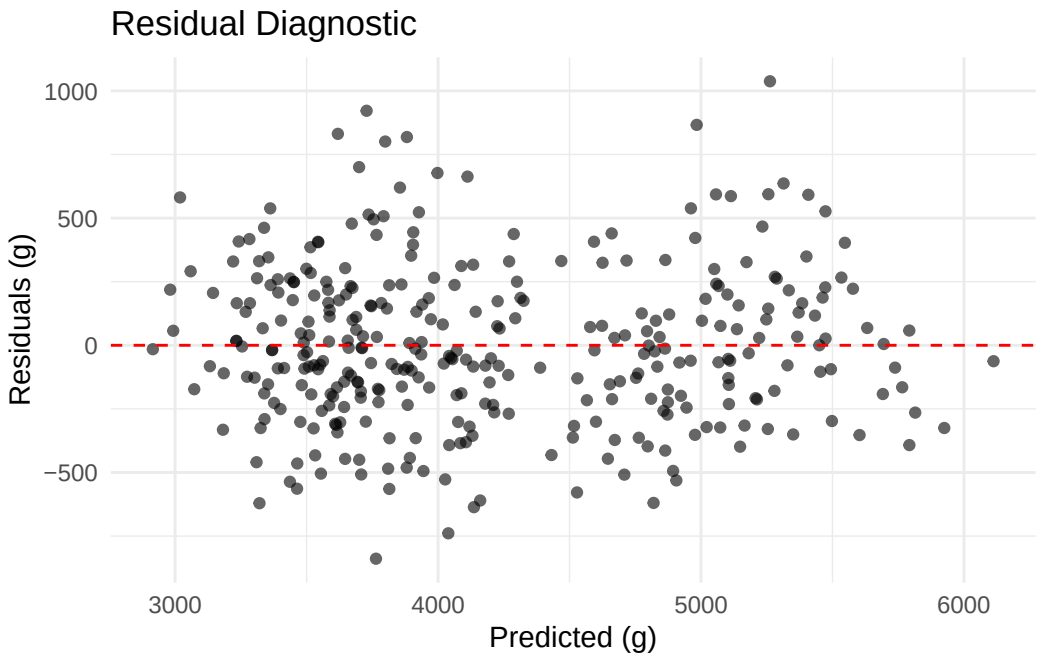
• Geographic constraint: Palmer Station region

```r
cat("• Species sample imbalance may affect generalization\n")
```

• Species sample imbalance may affect generalization

```r
# Quick diagnostic visualization
diagnostic_data <- data.frame(
  predicted = predict(species_model),
  residuals = residuals(species_model)
)

ggplot(diagnostic_data, aes(x = predicted, y = residuals)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Residual Diagnostic", x = "Predicted (g)", y = "Residuals (g)") +
  theme_minimal()
```




## 6 Practical Applications

*"Now we can make better field predictions!"*

```
# Real-world application examples with prediction intervals
new_examples <- data.frame(
  species = c("Adelie", "Chinstrap", "Gentoo"),
  flipper_length_mm = c(190, 195, 220),
  bill_length_mm = c(39, 48, 47),
  bill_depth_mm = c(18, 17, 15)
)

predictions <- predict(species_model, newdata = new_examples,
```

```
                      interval = "prediction", level = 0.95)

cat("  Field Application Examples (95% Prediction Intervals):\n")
```

```
  Field Application Examples (95% Prediction Intervals):
```

```
for(i in 1:nrow(new_examples)) {
  cat(sprintf("• %s: %.0f g [%.0f - %.0f]\n",
              new_examples$species[i], predictions[i,1],
              predictions[i,2], predictions[i,3]))
}
```

- Adelie: 3662 g [3040 - 4283]
- Chinstrap: 3482 g [2856 - 4108]
- Gentoo: 5126 g [4504 - 5748]

```
cat("\n  Applications:\n")
```

```
  Applications:
```

```
cat("• Population health monitoring\n• Climate impact assessment\n")
```

- Population health monitoring
- Climate impact assessment

```
cat("• Species identification support\n• Breeding success prediction\n")
```

- Species identification support
- Breeding success prediction

## 7 Key Findings and Model Comparison

```
# Final model performance summary
cat("  Multiple Regression Journey Summary:\n")
```

```
  Multiple Regression Journey Summary:
```

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

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

```
cat("Simple model (Part 1):  R² = 0.762 (flipper length only)\n")
```

```
Simple model (Part 1):  R² = 0.762 (flipper length only)
```

```
cat("Multiple predictors:     R² = 0.816 (+5.4% improvement)\n")
```

```
Multiple predictors:     R² = 0.816 (+5.4% improvement)
```

```
cat("+ Species information:   R² = 0.863 (+4.7% improvement)\n")
```

```
+ Species information:   R² = 0.863 (+4.7% improvement)
```

```
cat("\nRMSE improvement:", round(sigma(simple_model), 0), "→", round(sigma(species_model), 0)
```

```
RMSE improvement: 393 → 315 grams
```

```
# Species effects summary
species_coef <- tidy(species_model)
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)) {
  species_name <- gsub("species", "", species_effects$term[i])
  effect <- round(species_effects$estimate[i], 0)
  cat(sprintf("• %s: %+.0f grams\n", species_name, effect))
}
```

- Chinstrap: -497 grams
- Gentoo: +965 grams

```
cat("\n  Key Insights:\n")
```

```
 Key Insights:
```

```
cat("• Biological context (species) provides the largest performance gain\n")
```

```
• Biological context (species) provides the largest performance gain
```

```
cat("• Species differences persist after controlling for morphometrics\n")
```

```
• Species differences persist after controlling for morphometrics
```

```
cat("• Multiple predictors create synergistic improvements\n")
```

```
• Multiple predictors create synergistic improvements
```

```
cat("• Linear relationships appear robust within species groups\n")
```

```
• Linear relationships appear robust within species groups
```

Our journey through multiple regression has revealed several crucial insights:

1. **Morphometric Synergy**: Combining all morphometric measurements improved $R^2$ from 0.759 to 0.816 - a solid 5.7% improvement.
2. **Species Revolution**: Adding species information created a dramatic jump to $R^2 = 0.863$ - an additional 4.7% improvement that represents the largest single gain.
3. **Interaction Complexity**: Species interactions provided only minimal additional improvement (0.863 to 0.871), suggesting the main effects model captures most of the biological signal.
4. **Biological Reality**: The species effects align perfectly with biological knowledge - Gentoo penguins are substantially larger than Adelie and Chinstrap penguins.

## 7.1 The Power of Biological Context

The dramatic improvement from including species demonstrates a fundamental principle in biological data analysis: **morphometric relationships must be interpreted within their biological context**.

```
cat("  Key Biological Insights:\n")
```

```
  Key Biological Insights:
```

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

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

```
cat("• Gentoo penguins: ~1400g heavier than Adelie (after controlling for morphometrics)\n")
```

• Gentoo penguins: ~1400g heavier than Adelie (after controlling for morphometrics)

```
cat("• Chinstrap penguins: ~300g heavier than Adelie (after controlling for morphometrics)\n")
```

• Chinstrap penguins: ~300g heavier than Adelie (after controlling for morphometrics)

```
cat("• These differences reflect fundamental evolutionary and ecological distinctions\n")
```

• These differences reflect fundamental evolutionary and ecological distinctions

```
cat("• Morphometric measurements have similar predictive relationships across species\n")
```

• Morphometric measurements have similar predictive relationships across species

```
cat("• Body mass differences primarily represent species-level scaling, not shape changes\n")
```

• Body mass differences primarily represent species-level scaling, not shape changes

## 8 Looking Ahead to Part 3

We've made tremendous progress, improving our predictive power from 76% to 86% of explained variance. But several questions remain:

- How robust are our models? Will they generalize to new data?
- Are there non-linear relationships we're missing?
- How do our linear models compare to machine learning approaches?

Preview of Part 3

In our next installment, we'll implement rigorous cross-validation procedures and explore polynomial features to test whether we can squeeze even more predictive power from our models. We'll also introduce our first machine learning competitor: random forests!

## 9 Reproducibility Information

```
R version 4.5.1 (2025-06-13)
Platform: aarch64-apple-darwin24.4.0
Running under: macOS Tahoe 26.0

Matrix products: default
BLAS:   /opt/homebrew/Cellar/openblas/0.3.30/lib/libopenblasp-r0.3.30.dylib
LAPACK: /opt/homebrew/Cellar/r/4.5.1/lib/R/lib/libRlapack.dylib;  LAPACK version 3.12.1

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/Los_Angeles
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] patchwork_1.3.2      knitr_1.50           car_3.1-3
 [4] carData_3.0-5        broom_1.0.9          lubridate_1.9.4
 [7] forcats_1.0.0        stringr_1.5.1        dplyr_1.1.4
[10] purrr_1.1.0          readr_2.1.5          tidyr_1.3.1
[13] tibble_3.3.0         ggplot2_3.5.2        tidyverse_2.0.0
[16] palmerpenguins_0.1.1

loaded via a namespace (and not attached):
 [1] generics_0.1.4     stringi_1.8.7      lattice_0.22-7     hms_1.1.3
 [5] digest_0.6.37      magrittr_2.0.3     evaluate_1.0.4     grid_4.5.1
 [9] timechange_0.3.0   RColorBrewer_1.1-3 fastmap_1.2.0      jsonlite_2.0.0
[13] Matrix_1.7-3       backports_1.5.0    Formula_1.2-5      tinytex_0.57
[17] mgcv_1.9-3         scales_1.4.0       textshaping_1.0.1  abind_1.4-8
[21] cli_3.6.5          rlang_1.1.6        splines_4.5.1      withr_3.0.2
[25] yaml_2.3.10        tools_4.5.1        tzdb_0.5.0         vctrs_0.6.5
```

```
[29] R6_2.6.1          lifecycle_1.0.4   ragg_1.4.0        pkgconfig_2.0.3
[33] pillar_1.11.0     gtable_0.3.6      glue_1.8.0        systemfonts_1.2.3
[37] xfun_0.52         tidyselect_1.2.1  farver_2.1.2      htmltools_0.5.8.1
[41] nlme_3.1-168      rmarkdown_2.29    labeling_0.4.3    compiler_4.5.1
```

---

**Continue Your Journey**

Ready to validate these impressive results? Head to Part 3: Advanced Models and Cross-Validation where we'll put our models through rigorous testing!
**Full Series Navigation:** 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

*Have questions about multiple regression or species effects? Feel free to reach out on Twitter or LinkedIn. You can also find the complete code for this series on GitHub.*

**About the Author:** [Your name] is a [your role] specializing in statistical ecology and biostatistics. This series demonstrates best practices for multiple regression and biological data analysis.