# PSTAT 5LS Lab 7

YOUR NAME HERE

Winter 2025

Section 1

# Announcements & Recap

Section 2

# Learning Objectives

# R Learning Objectives

1. Creating a linear model and finding the relevant values
2. Creating a plot of (x,y) quantitative values with the least-squares regression line.

# Statistical Learning Objectives

1. Using a scatterplot to describe the relationship between two quantitative values.
2. Using the least-squares regression line to estimate the response variable.

# Functions covered in this lab

1. `cor()`
2. `plot()`
3. `lm()`
4. `abline()`

Section 3

# Lab Tutorial

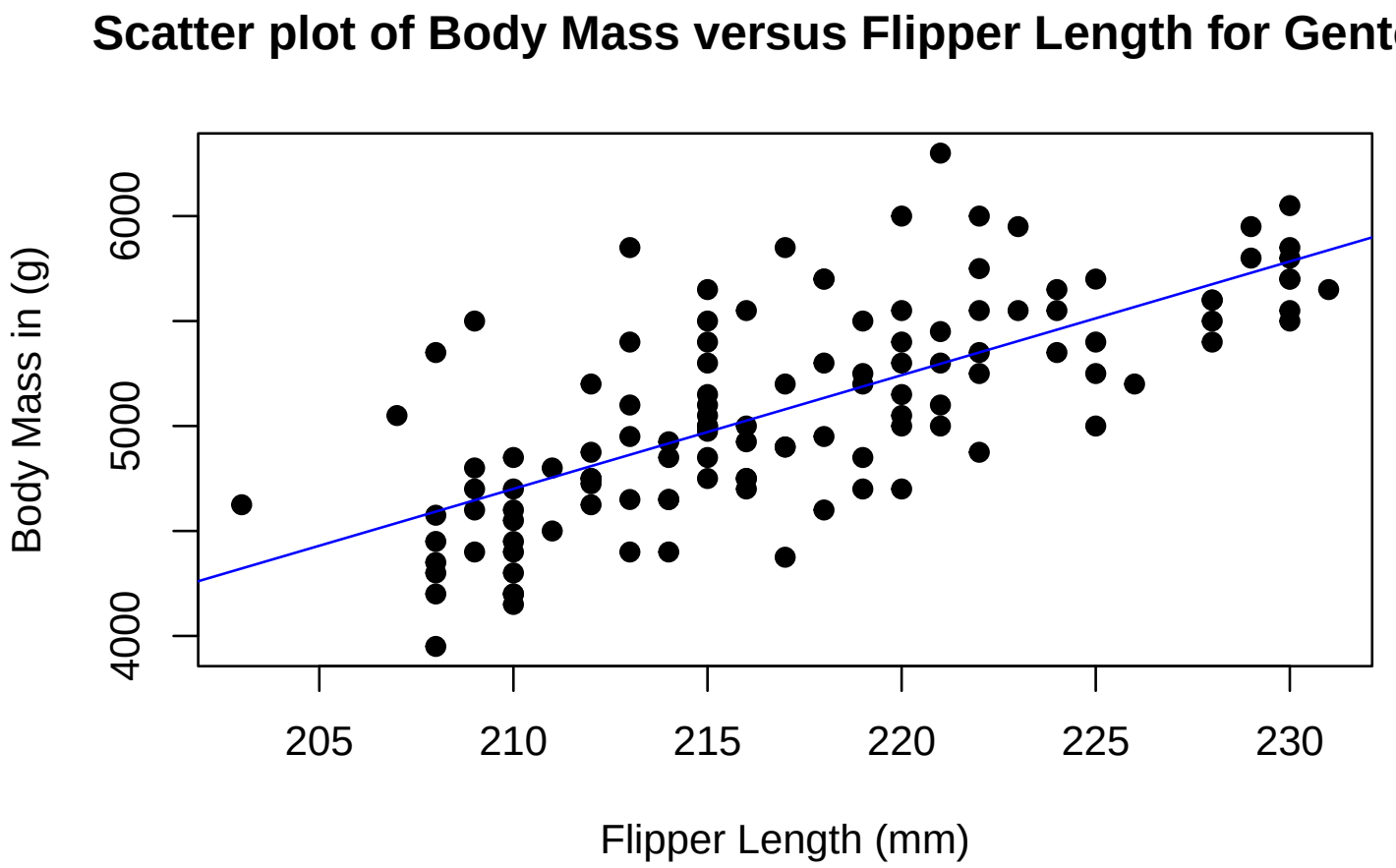
# Gentoo Penguins Data Set

We're back to hanging out with the Gentoo penguins that we worked with last time. To be sure you're working with the correct data set, run the `loadGentoo` chunk in the `lab7-notes.Rmd` markdown file, and verify that the `gentoo` data is in your environment in the top right corner of your project.

```
gentoo <- read.csv("gentoo.csv", stringsAsFactors = TRUE)
```

# Linear Regression

Last time, we found a strong positive linear relationship between body mass and flipper length for Gentoo penguins:



# Finding the Regression Equation

Now, we're going to perform a linear regression of body mass on flipper length to determine the equation of the line we saw superimposed on the scatterplot.

In this case, "linear regression of body mass on flipper length" means that we will use the flipper length to estimate body mass. This makes flipper length the explanatory variable ($x$) and body mass the response variable ($y$).

To perform the regression, we will use the `lm()` function (which stands for **l**inear **m**odel). Specifically, we'll provide:

- a formula (`y ~ x`)
- a `data` argument.

We will store the results (the model) as an object called `line1`, and then use the `summary()` function to obtain detailed results.

# Linear Regression Code

Let's try this code together in the `tryIt1` code chunk.

```
line1 <- lm(body_mass_g ~ flipper_length_mm, data = gentoo)
summary(line1)
```

# Linear Regression Output

Here's what the output looks like for our linear regression model.

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = gentoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -704.69 -244.29  -58.87  161.98 1003.65
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6674.204   1075.436  -6.206 8.51e-09 ***
## flipper_length_mm    54.165      4.948  10.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354 on 117 degrees of freedom
## Multiple R-squared:  0.506,  Adjusted R-squared:  0.5017
## F-statistic: 119.8 on 1 and 117 DF,  p-value: < 2.2e-16
```

# Understanding Linear Regression Output

**Interpreting the Output Table**

When reviewing the linear regression output, here's how to navigate the key components:

- The first two lines display the code that was run.
- The **Residuals** can be skipped over for now.
- The most important section if the We want the piece dealing with the **Coefficients** table, which contains two rows and four columns.
- The **Estimate** column is our main focus in this lab.

# Understanding Linear Regression Output

## Breaking Down the Coefficients Table

- The first row, labeled `(Intercept)`, represents the vertical ($y$) intercept of the regression line.
  - The `Estimate` value in this row gives the $y$-intercept of the least-squares regression line.
- The second row corresponds to our explanatory variable, `flipper_length_mm`.
  - This row's `Estimate` value provides the slope of the least-squares regression line.
  - This helps verify that we correctly specified the model in the formal `y ~ x`.

# Understanding Linear Regression Output

**Additional Key Values**

- **Residual standard error:** This value, denoted as $s$, measures the standard deviation of the residuals.
- **Multiple R-squared**: This value represents the *coefficient of determination* – a measure of how well the model explains the variation in the response variable.
  - We *ignore the Adjusted R-squared value* because it does not apply to simple linear regression.

# Summary: Essential Values to Pull from Regression Output

From the regression output, we need to identify:

1. the **y-intercept** of the least-squares regression line.
2. the **slope** of the least-squares regression line.
3. the **multiple R-squared value** (coefficient of determination)
4. the **residual standard error** ($s$)

# Putting It All Together: The Least-Squares Regression Equation

Now that we've identified the key values from the regression output—the intercept and the slope—it's time to put them together into the equation of the least-squares regression line.

This equation allows us to make predictions based on our explanatory variable. So, what is the equation for the least-squares regression line?

In general, the regression line for simple linear regression is

$$\hat{y} = b_0 + b_1 x$$

# Putting It All Together: The Least-Squares Regression Equation

What is the equation of the least-squares regression line for our example?

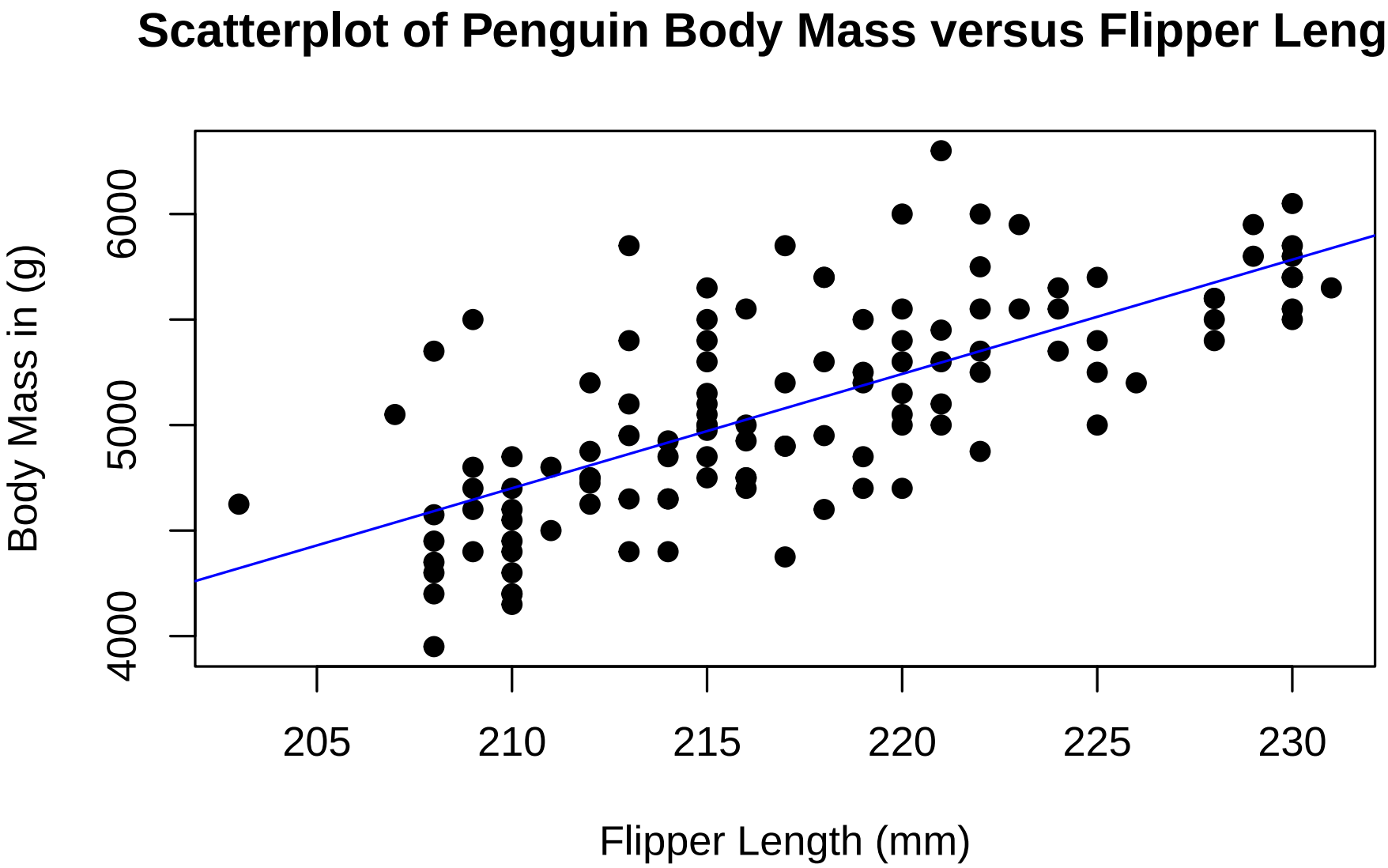
# Adding the Regression Line to the Scatterplot

Now that we have our regression equation, let's add the estimated least-squares regression line to our scatterplot. We can do this by passing our model object to the `abline()` function, which overlays the regression line on the plot.

Let's try this out in `tryIt2` in your notes:

```
plot(gentoo$body_mass_g ~ gentoo$flipper_length_mm,
     main = "Scatterplot of Body Mass versus Flipper Length fo
     xlab = "Flipper Length (mm)",
     ylab = "Body Mass in (g)",
     pch = 19)
abline(line1, col = "blue")
```

# Scatterplot and Least-Squares Regression Line



# Applying the Regression Equation: Estimating Body Mass

Now that we have our least-squares regression equation, let's use it to **predict** body mass for given flipper lengths.

Our regression equation is:

$$\hat{y} = -6674.20 + 54.165x$$

where:

- $x =$ **flipper length (mm)**
- $\hat{y} =$ **predicted body mass (g)**

# Applying the Regression Equation: Estimating Body Mass

Use the regression equation to estimate the body mass for the following flipper lengths:

1. **210 mm**

2. **230 mm**

# Calculating Residuals: How Good Is Our Prediction?

A **residual** measures how far off our prediction is from the actual data:

$$\text{Residual} = \text{Actual } y - \text{Predicted } \hat{y}$$

Using your predicted values from the regression equation, calculate the residuals for these observed values:

| Flipper Length (mm) | Obs. Body Mass (g) | Pred. Body Mass (g) | Residual |
|---|---|---|---|
| 210 | 4600 | ? | ? |
| 230 | 6050 | ? | ? |

Think about your residuals—are they large or small? What do they tell us about how well the regression model fits the data?