Exploring multivariable relationships

Prof. Maria Tackett

Click for PDF of slides

Carbohydrates in Starbucks food

- Starbucks often displays the total calories in their food items but not the other nutritional information.
- Our goal is to analyze the relationship between the calories and total carbohydrates (carbs) in Starbucks food items, and assess if it differs based on the type of food item (bakery, salad, sandwich, etc.)
- We can use our analysis to estimate the total carbs using information about the total calories and type for a given food time

Starbucks data

- **Observations**: 77 Starbucks food items
- **Variables**:
  - **`carb`**: Total carbohydrates (in grams)
  - **`calories`**: Total calories
  - **`bakery`**: 1: bakery food item, 0: other food type

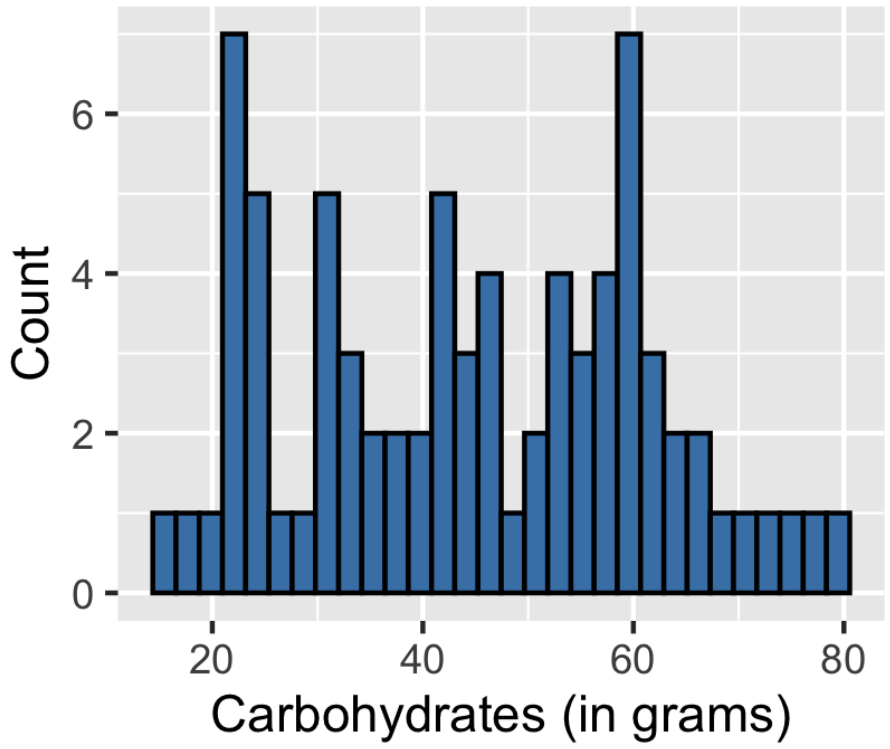
Terminology

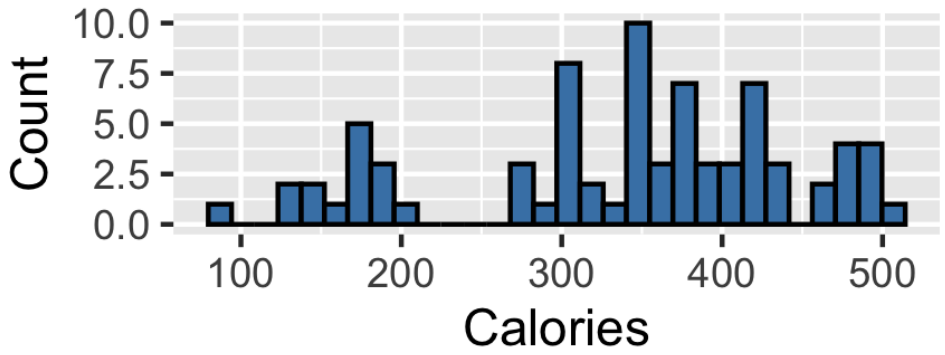
- **`carb`** is the **response variable**
  - variable whose variation we want to understand / variable we wish to predict
  - also known as *outcome* or *dependent* variable

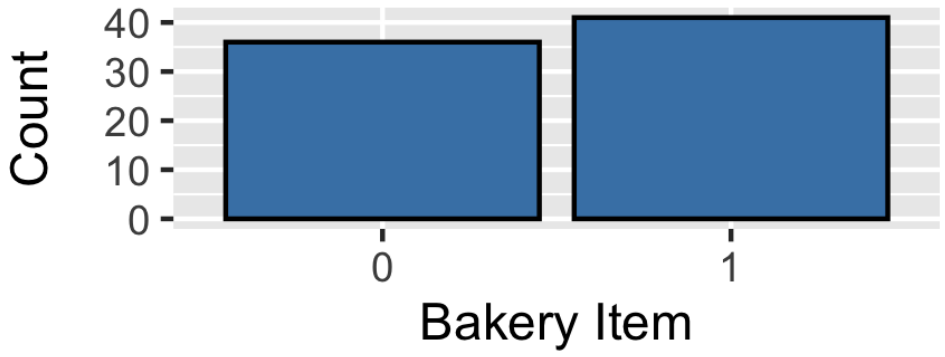
- **`calories`**, **`bakery`** are the **predictor variables**
  - variables used to account for variation in the outcome
  - also known as *explanatory*, *independent*, or *input* variables

Let's look at the data

Plot Code

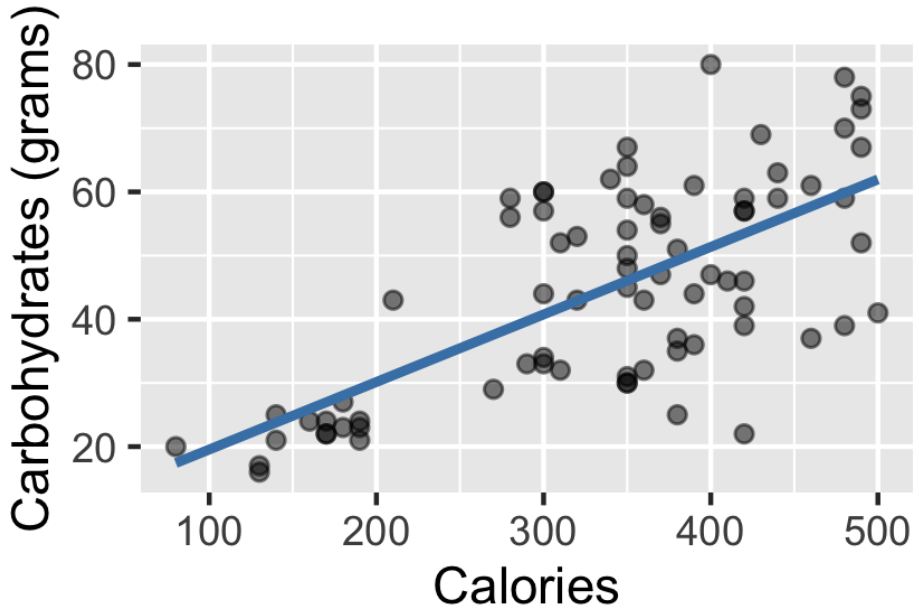


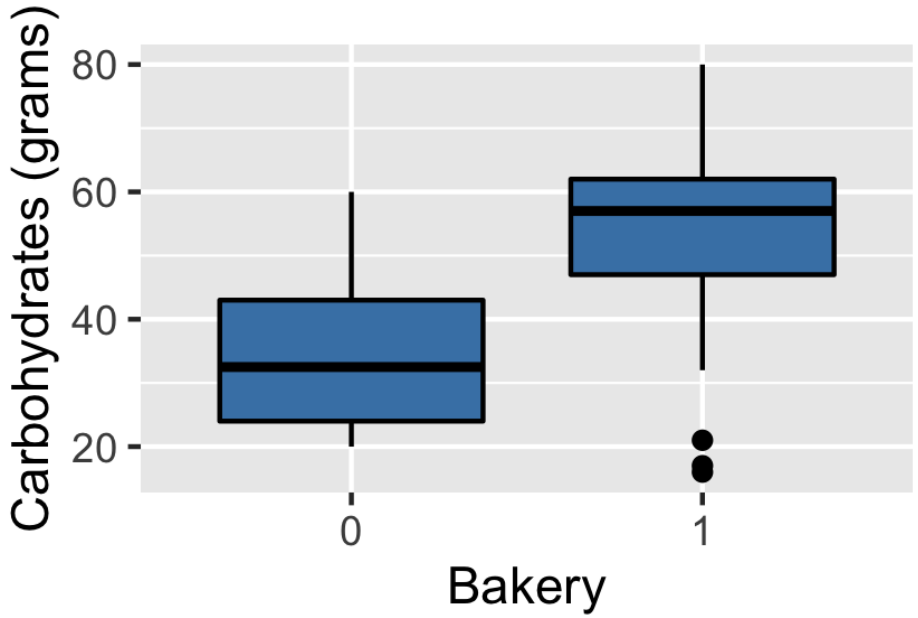




Response vs. Predictors

Plot Code





$$\text{carbs} = f(\text{calories}, \text{bakery}) + \epsilon$$

Model

$$\text{carbs} = f(\text{calories}, \text{bakery}) + \epsilon$$

- **Goal**: Determine f
- How do we determine f?
  - Make an assumption about the functional form f
  - Use the data to fit a model based on that form

Determine f

In general,

1) Choose the functional form of f, i.e. **choose the appropriate model given the response variable**

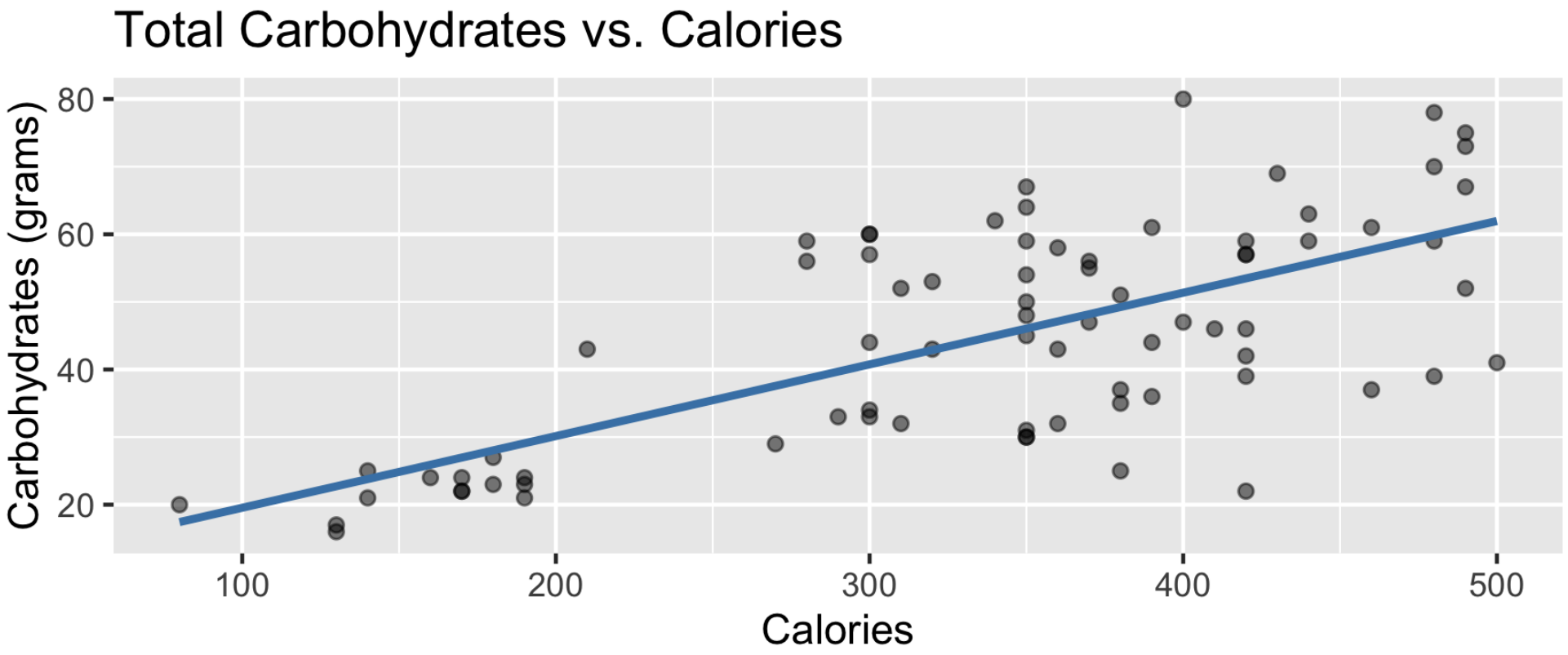
- Suppose f is a linear model

$$y = f(\mathbf{X}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

2) Use the data to fit (or train) the model, i.e **estimate the model parameters**

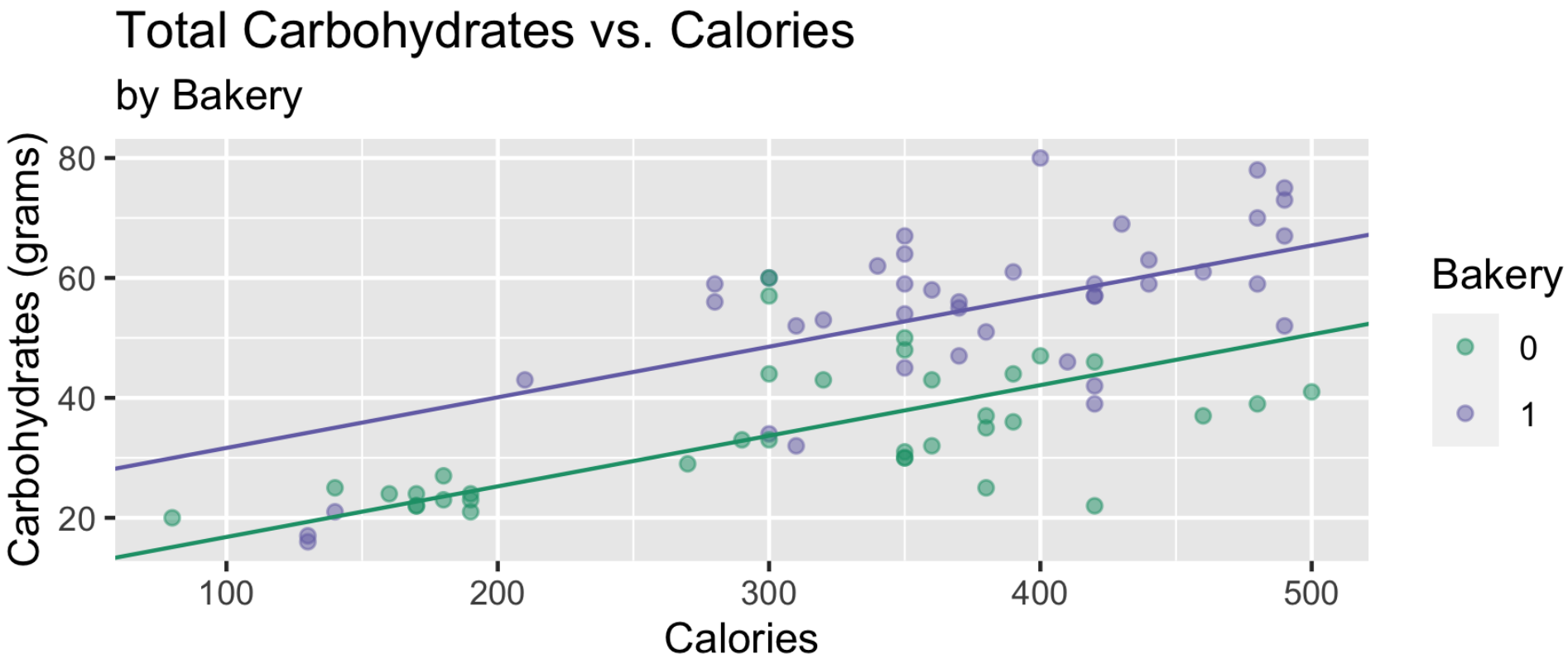
- Estimate $\beta_0, \beta_1, \ldots, \beta_p$

Carbs vs. Calories



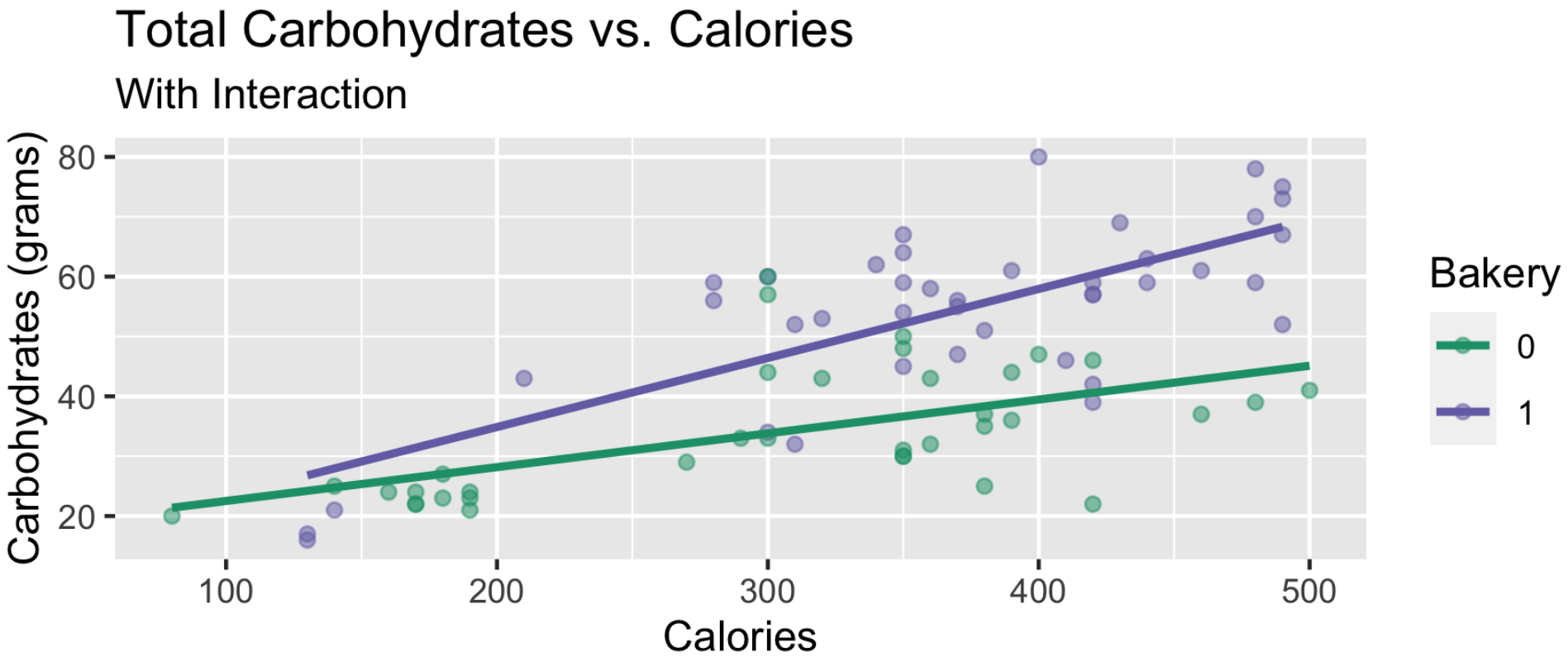
$$\text{carbs} = \beta_0 + \beta_1 \text{ calories} + \epsilon$$

Carbs vs. Calories + Bakery



$$\text{carbs} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \epsilon$$

Carbs vs. Calories + Bakery (with interaction)



$$\text{carbs} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \beta_3 \text{ calories} \times \text{bakery} + \epsilon$$

Code for plot on previous slide

```
ggplot(data = starbucks, aes(x = calories, y = carb, color = bakery)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Calories",
       y  = "Carbohydrates (grams)",
       color = "Bakery",
       title = "Total Carbohydrates vs. Calories",
       subtitle = "With Interaction") +
  scale_color_manual(values=c("#1B9E77", "#7570B3"))
```

Why?

$$\text{carbs} = \beta_0 + \beta_1 \text{ calories} + \beta_2 \text{ bakery} + \beta_3 \text{ calories} \times \text{bakery} + \epsilon$$

Prediction:

What do we expect the total carbohydrates to be in a piece of Starbucks pumpkin bread, a bakery item that is 410 calories?

Inference:

What is the relationship between the calories and total carbohydrates for bakery items at Starbucks? For non-bakery items?

Course Outline

Unit 1: Quantitative Response Variables

- Simple Linear Regression
- Multiple Linear Regression

Unit 2: Categorical Response Variable

- Logistic Regression
- Multinomial Logistic Regression

Unit 3: Looking Ahead

- Log-linear Regression
- Weighted Least Squares
- Presenting statistical results