A Brief Introduction to Modeling

Ryan Miller

What is Modeling?

How would you define modeling?

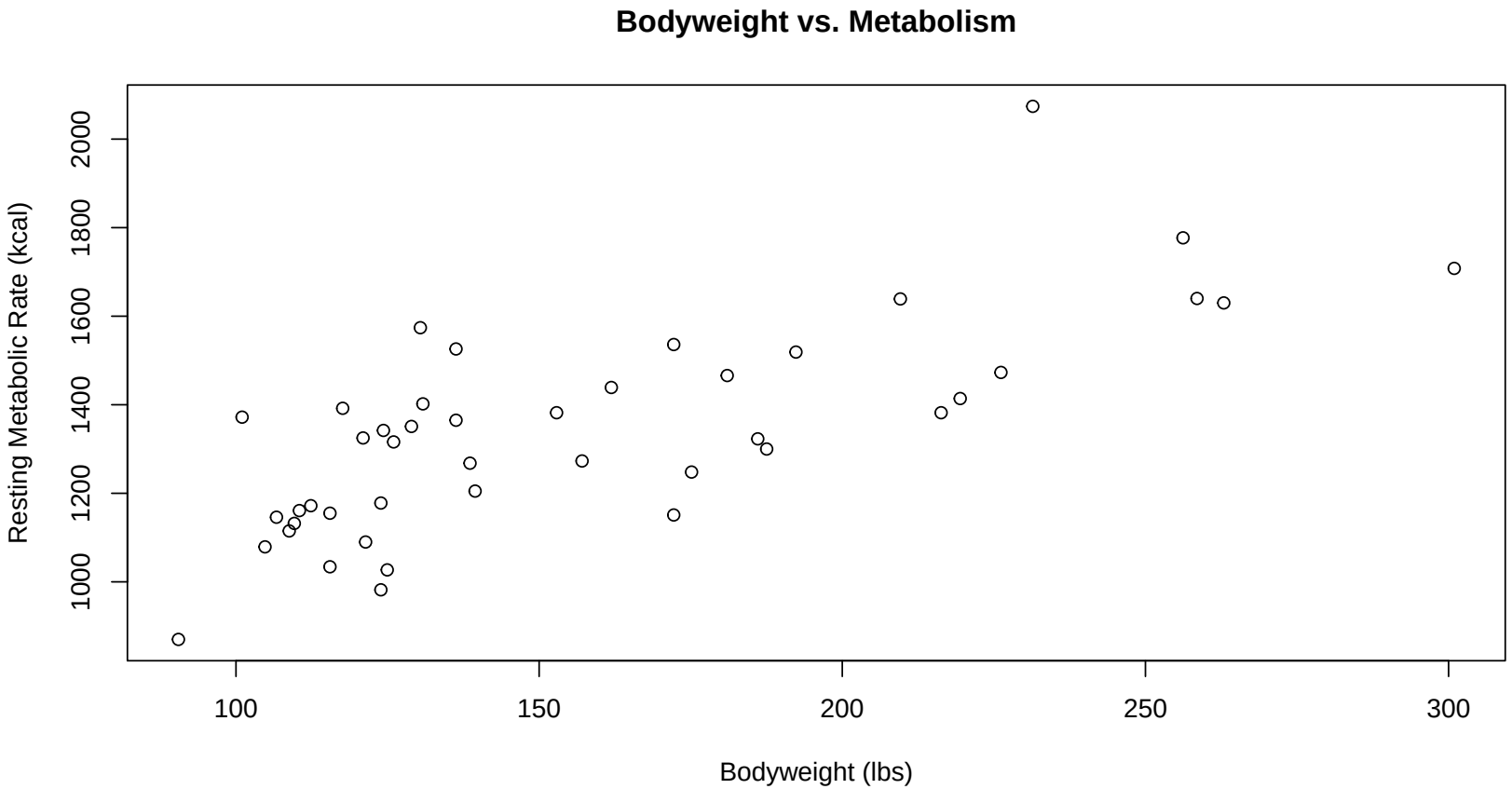
What is Modeling?

How would you define modeling?

- *"The goal of a model is to provide a low-dimensional summary of a dataset"* - Data Science for R (textbook)
- *"A system of postulates, data, and inferences presented as a mathematical description of an entity or state of affairs"* Marriam-Webster (dictionary)

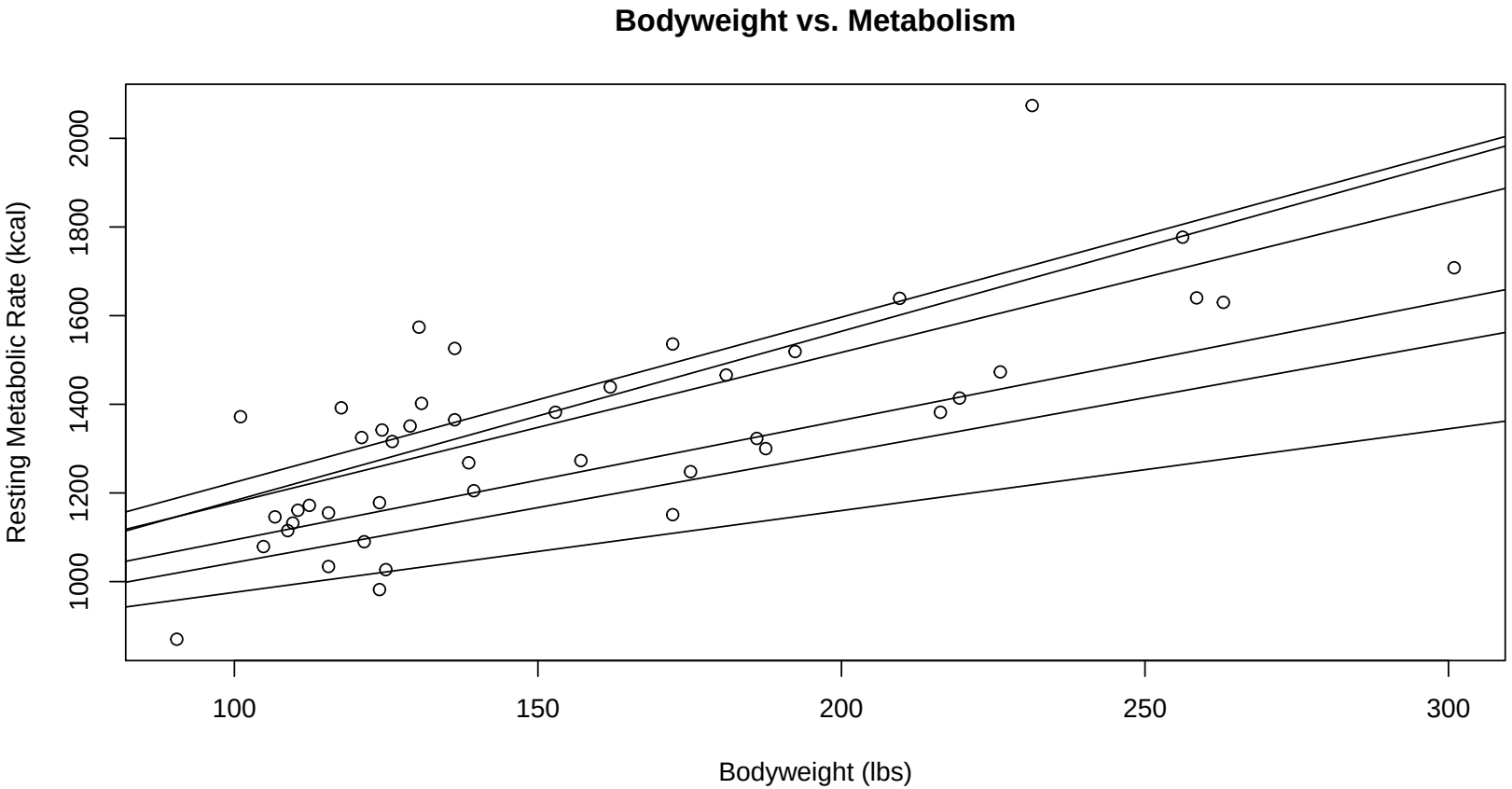
A Basic Example

Shown below are data from a random sample of 44 US adult women, how would you summarize the pattern?



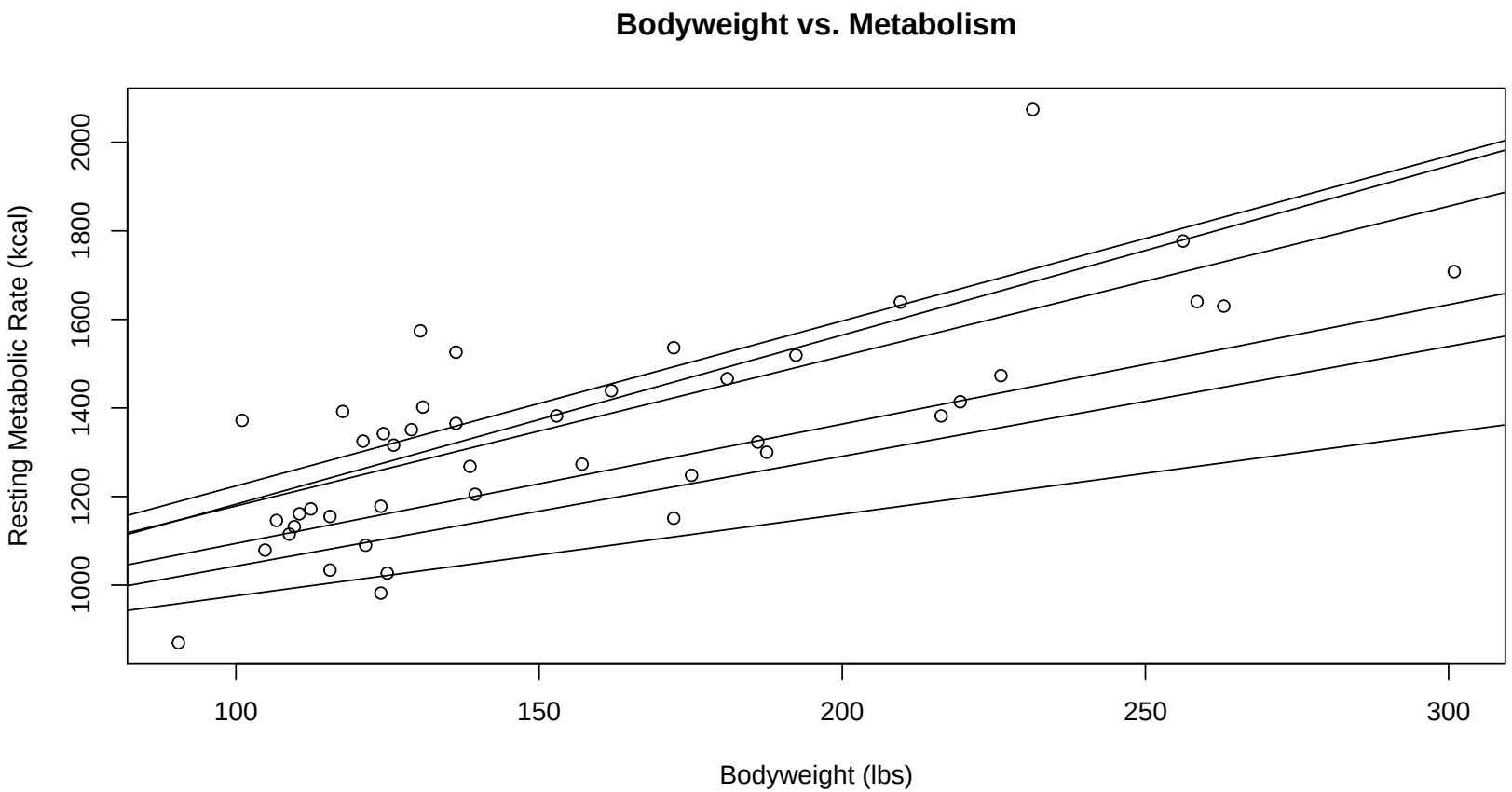
A Basic Example

You might consider a *family of models*, such as straight lines (ie: $Y = aX + b$), a few example models are depicted below:



A Basic Example

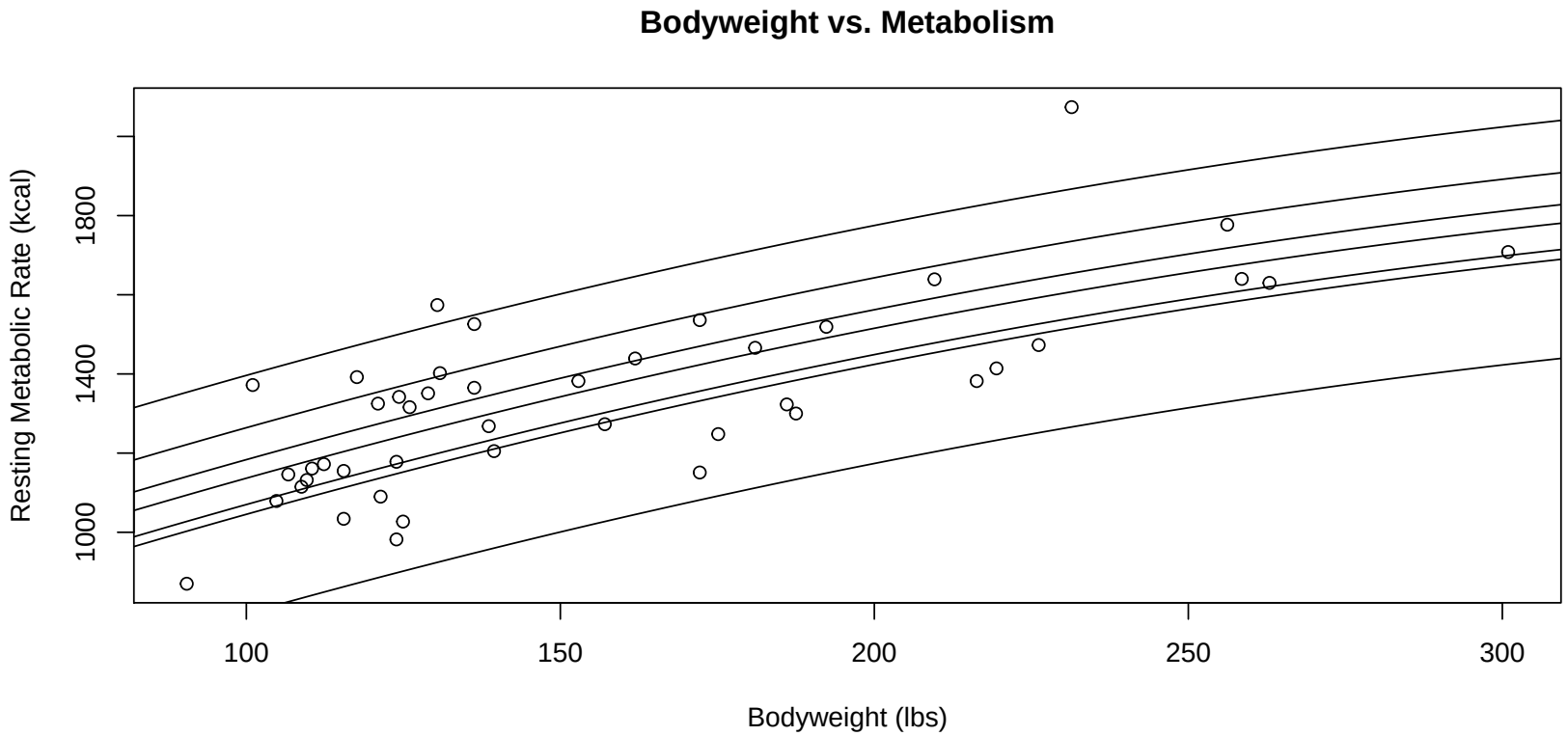
Each line represents a *candidate model* in this family, some look pretty good, while others do not.



In the coming weeks we'll learn more about choosing a good model...

Another Family of Models

Another *family of models* we might consider are quadratic polynomials (ie: $Y = aX^2 + bX + c$), below are some examples:



What advantages/disadvantages of this family relative to straight lines?

Parametric vs. Non-parametric Models

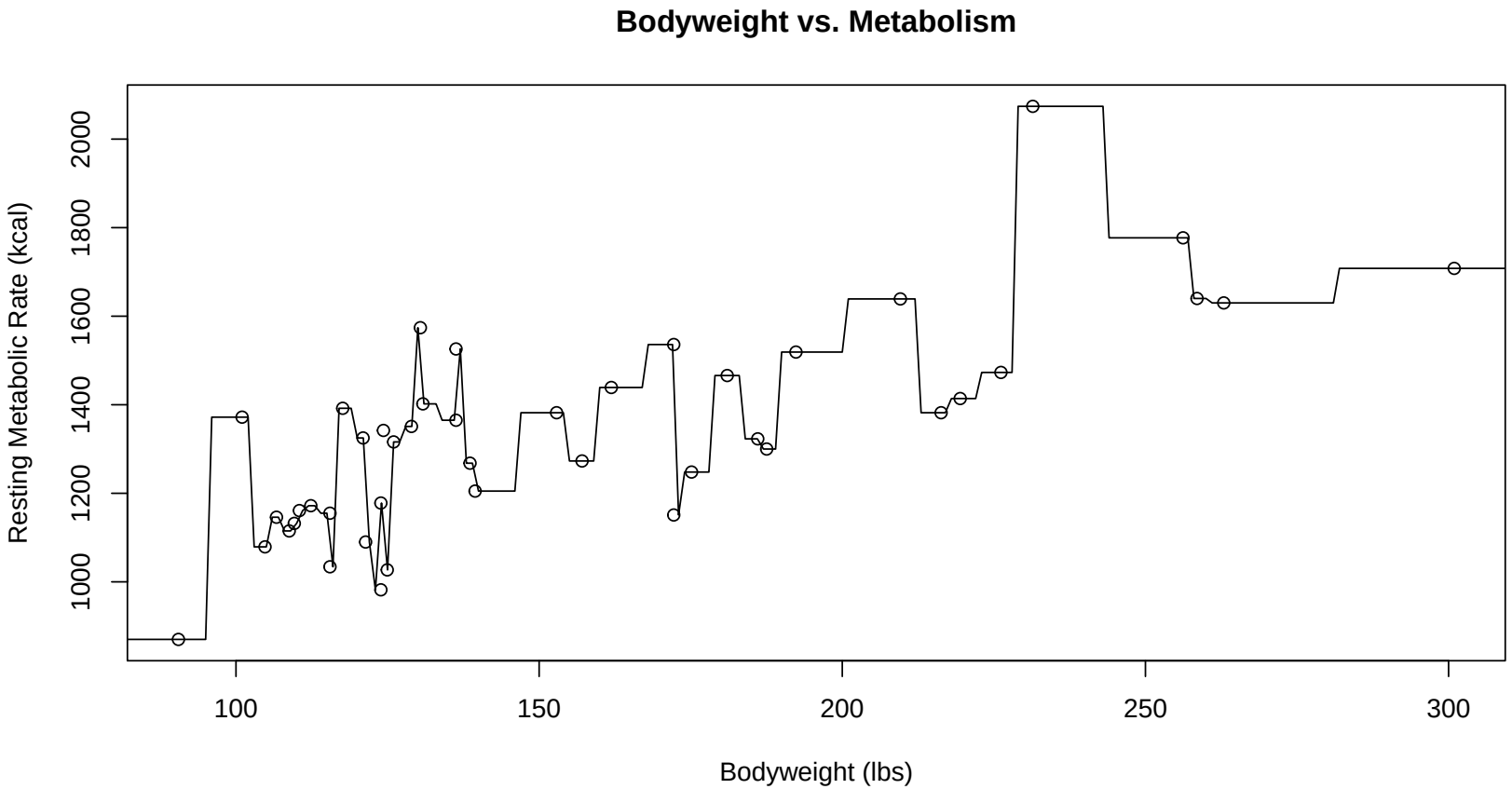
- The models we've seen so far are examples of **parametric models**
  - They can be entirely defined by a mathematical formula that involves a set of *parameters* (ie: a slope and intercept, or the coefficients $\{a, b, c\}$)

Parametric vs. Non-parametric Models

- The models we've seen so far are examples of **parametric models**
  - They can be entirely defined by a mathematical formula that involves a set of *parameters* (ie: a slope and intercept, or the coefficients $\{a, b, c\}$)
- An entirely different alternative are **non-parametric models**
  - You can think of these models as algorithms or sets of rules that do not conform to a rigid parametric structure

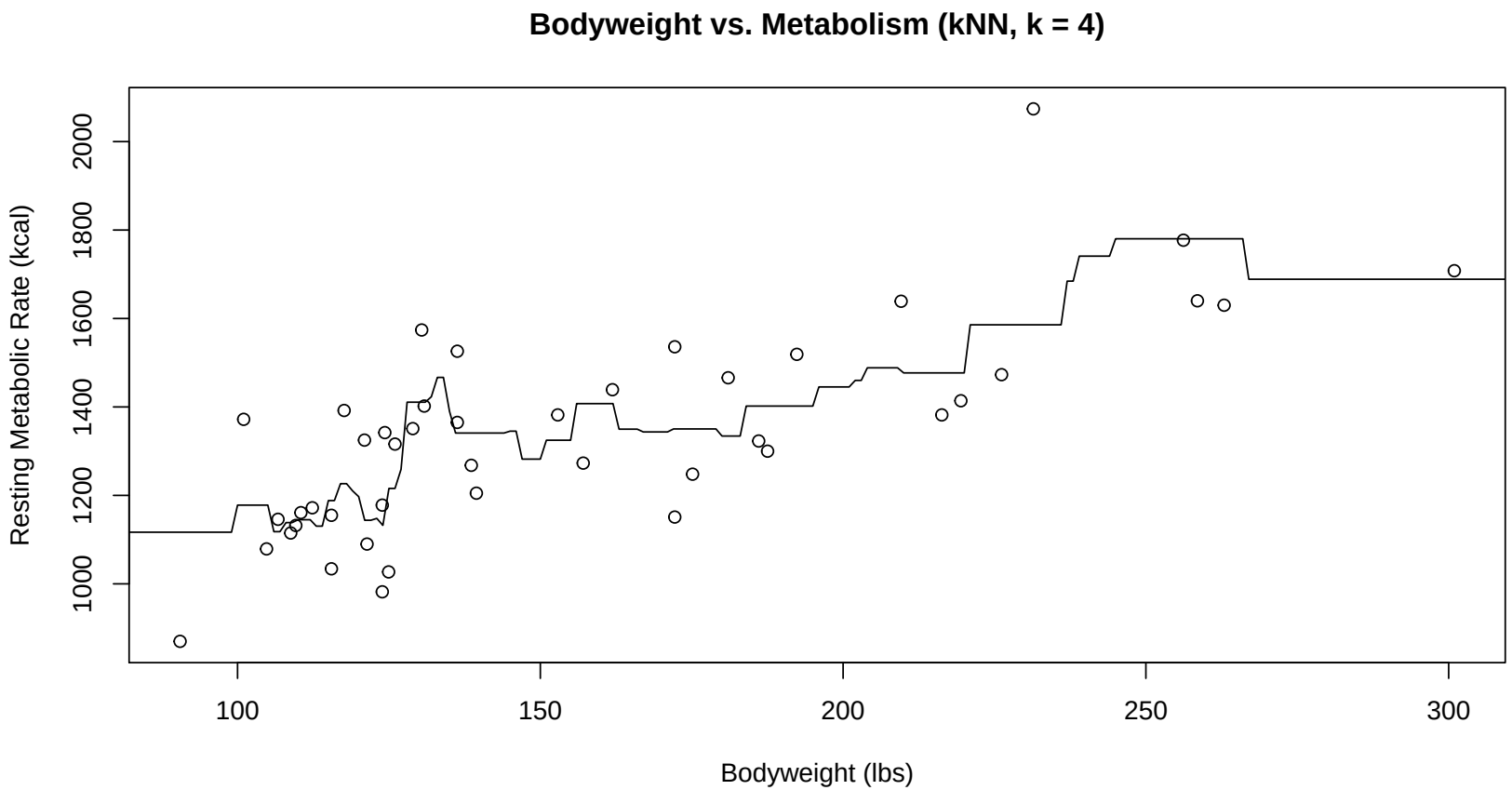
A Simple Non-Parametric Model

As an example, a non-parametric model might be the rule "the predicted RMR is the observed RMR of the nearest data-point"



K-Nearest Neighbors

We can generalize this to "the predicted RMR is the average observed RMR of the nearest k data-points", a type of non-parametric model known as *K-nearest Neighbors*



Closing Remarks

- Modeling involves a lot of decision making
  - Even with just two variables, there are tons of possible models we could use

Closing Remarks

- Modeling involves a lot of decision making
  - Even with just two variables, there are tons of possible models we could use
- Throughout the semester, we'll focus our attention on how to make modeling decisions
  - Choosing between non-parametric vs. parametric models
  - Choosing between different model families and algorithms
- We'll also spend time understanding our models
  - Statistical inference on model parameters, evaluating model fit, etc.