Chapter 5.1 - Mathematical Models

> “You know, for a mathematician, he did not have enough imagination. But he has become a poet and now he is fine.” - D. Hilbert, talking about an ex-student.

It’s an injustice to students that they rarely learn about the applications of math until they’re already prepared for a quantitative career. Models are perhaps the most fascinating tool in mathematics, yet we avoid teaching about them until we absolutely have to. This chapter will attempt to fix that.

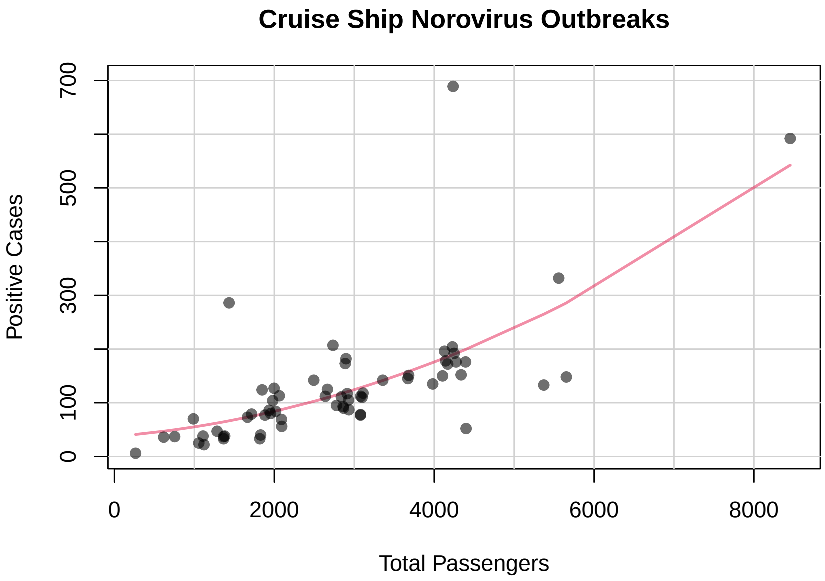
There is a problem with doing this, however. Mathematical models are far from a simple concept to grasp. When we’re observing a real process, (i.e., the flow of warm air around a room as it transitions to cold air), and attempting to convert this into pure mathematics we can’t escape the complexities of math. So while we’ll attempt to keep things simple enough to digest— this will still feel like a slightly out-of-reach concept. And that’s OK.

Phenomenological vs. Mechanistic

When discussing mathematical modeling we need to first establish a key difference between major “types” of models. If we were to build a model that simple described a relationship between variables by fitting them *into* the model, we refer to that model as **phenomenological**. Consider the example below where we take data from outbreaks of Norovirus on passenger cruise ships and place it into the model:

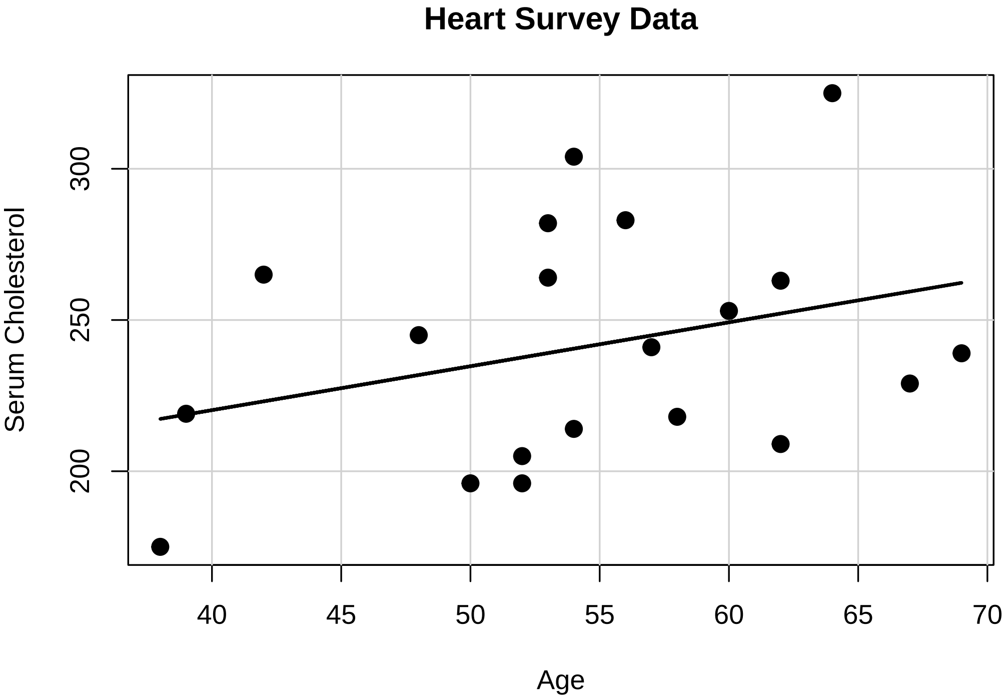
$$\text{Infected} = \text{Intercept} + \text{Total Passengers} + \text{Total Passengers}^2$$

This will show the trend of infected passengers relative to total passengers as a *smoothed* polynomial curve.



The model can be considered phenomenological because it's not developed from knowledge of the process but rather used to try and better understand the process. While many statistcians will consider any model that is complex enough to be uninterpretable to be phenomenological— this is not the truth. Consider the case of fitting a straight line through the heart survey data using the model:

$$\text{Cholesterol} = \text{Intercept} + \text{Age}$$

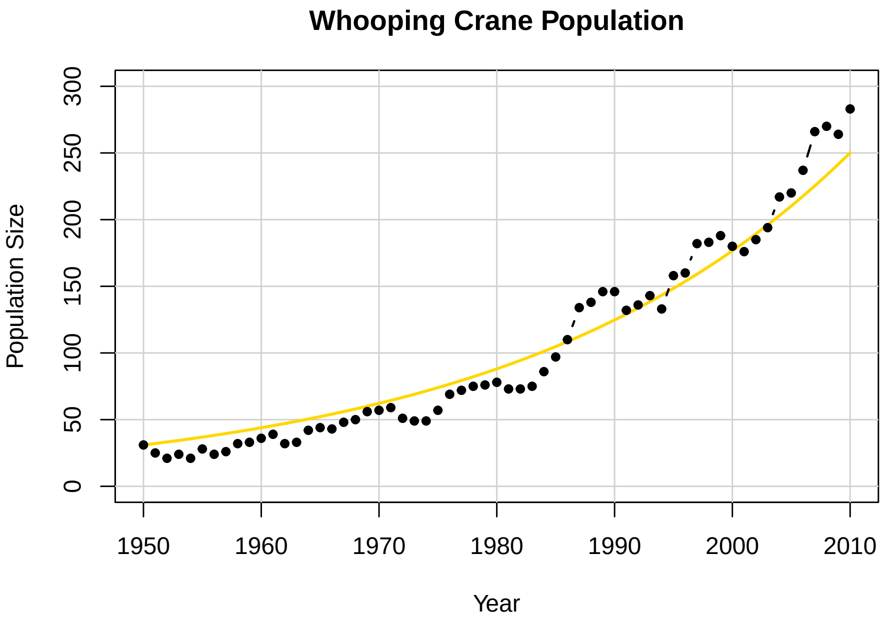


This is phenomenological if we're *uncertain* whether the relationship between age and cholesterol is strictly linear. We don't need to know whether the relationship between variables can be captured fully using just those variables, but we do need to be fairly certain of their isolated relationship.

If we were to develop a model from knowledge of a process or relationship and use that model to better understand our data then we would refer to it as **mechanistic**. A classic example of mechanistic modeling is fitting population data to the formula for exponential growth.

$$\lambda(t) = \lambda_0 e^{\gamma(t-t_0)}$$

Where $\lambda(t)$ is the population size at time t, t_0 is the initial time point, λ_0 is the initial population, and γ is the rate of population growth. This is actually a solution to an ordinary differential equation (ODE) and is well studied because of how clear the effect of population growth is over time.



Mechanistic models such as this one have the benefit of being highly interpretable; most biological scientists can comprehend a growth rate. We aren't gaining some new knowledge about the process— we're learning about the rate of population growth for this specific data.

Linear vs. Nonlinear

All models can be classified as **linear** or **non-linear**. For statisticians linearity is defined by the *parameters*; the following would be considered **linear**:

$$Y = aX + b$$

$$w = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$P = \mu + \alpha \log(r \times t)$$

$$\boldsymbol{\Delta} = \boldsymbol{\beta} \cos(\boldsymbol{X}) + 10$$

Nonlinear models are anything else where the parameters have some nonlinear component:

$$Y = a^2 X + b$$

$$w = \beta_0 + \log(\beta_1) x$$

$$\lambda(t) = \lambda_0 e^{\gamma(t - t_0)}$$

Deterministic vs. Probabilistic

The models we've covered so far perform the exact same way everytime we fit them to data, in that there's no "randomness" to them. Since these models operate under the assumption that the outcomes and relationships have been *determined* already, we refer to them as **deterministic**.

Showing randomness can be helpful in some scenarios since real life is seemingly filled with random outcomes. If we were modeling something like the movement of cattle in a pen we should expect a certain degree of chaotic movement since the decision making of cattle can sometimes be enigmatic. Statisticians refer to this chaos as *stochasticity*. Since we consider stochasticity to be dictated by probabilities when we develop a model that accounts for it we refer to that model as **probabilistic**.

We can take a deterministic model and place it inside of a probabilistic one; this is actually the core methodology that statistical models are developed through. Later on in this chapter we'll show how we can take the model for the effect of age on serum cholesterol levels, (a purely deterministic model as we've shown it), and transform it into a probabilistic model through the addition of one parameter:

$$\text{Cholesterol} = \text{Intercept} + \text{Age} + \text{Random Error}$$

Model Outputs

In many physics textbooks there's discussion about the difference between *explicit* and *implicit* models. Statisticians don't concern themselves with this distinction as much because we almost always use implicit models.

Explicit models are meant to calculate some finite value from the input of variable conditions. This would be something like the model for compound interest:

$$P(t) = P_0 e^{rt}$$

We generally use this formula by inputting the principal P_0, the rate r, and the time elapsed t, then calculating the final amount. Since we know all of the parameter values we're only interested in estimating $P(t)$.

Statistics aims to estimate parameters so our primary technique is implicit modeling. We don't know the effect of Age or Intercept in our cholesterol model so we're seeking to estimate them with the data:

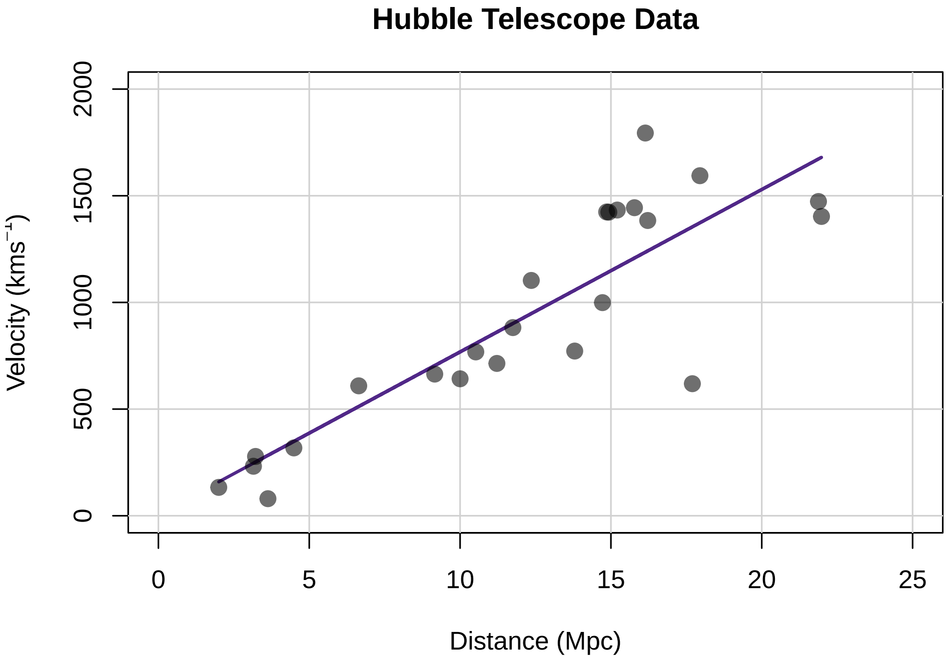
$$\text{Cholesterol} = \text{Intercept} + \text{Age}$$

$$\text{Cholesterol} = 162.127 + 1.452 \times \text{Age}$$

Once we calculate the value of these parameters we treat the model *as if* it's explicit, but our modeling technique was still implicit.

Sometimes the goal of a model isn't the parameters but rather a result of using the parameter values in an external formula. As an example: We can use data from the hubble telescope to estimate the parameters of the formula:

$$\text{Velocity} = \text{Distance}$$



Then plug the value of the Distance parameter into the equation:

$$\frac{979.708}{\text{Distance}}$$

To estimate the age of the universe. We refer to any such output as a **derived quantity** since it was derived from the parameters of the model.

As a final note, there are two more distinctions that statisticians make between models: Frequentist versus Bayesian. These are rarely discussed in introductory courses because of a major gap in scientific communication for bayesian methods. The result is that students have little or no knowledge of bayesian models and walk away from their education believing frequentist techniques to be the only valid ones.

The basic premise behind these differing modeling techniques is that they have competing philosophies on how inference should be informed. Frequentism dictates that inference should be purely the result of data with minimal assumptions. Bayesianism argues that inference can be supported with prior knowledge about the possible distributions of parameters.

In keeping with traditions I'll throw my hands up, cry out that the details are outside of the scope of the book, and leave you with a quote to either provide context or stretch the page count.

> "Bayesian theory requires a great deal of thought about the given situation to apply sensibly, and recommending that scientists use Bayes' theorem is like giving the neighborhood kids the key to your F-16." - Andrew Geldman