Section 1.D: Build the Model

# The Second Stage

After we have defined our problem, the primary tools we use to build our model will be mathematical. The mathematics needed could range from simple arithmetic to a complicated differential equation. Sometimes simulation or data analysis software will be needed. These mathematical tools will be discussed in more detail in Chapter 2. However, there are some general ideas that will be applicable for any modeling problem. We discuss these in this section.

# Continuous versus Discrete

Many modeling problems that are a function of time can be categorized as “continuous” or “discrete.” After defining the problem, we can classify our variables as changing continuously or not. For example, the length of a shadow changes continuously as the sun moves across the sky. No matter how often we wish to measure the length of the shadow, we can expect the value to change. Another example is the concentration of carbon dioxide in a room. At any moment, this value could be measured. Tools of calculus, such as differential equations, are ideal for analyzing continuous variables.

On the other hand, modeling the number of students enrolled at a university each semester is a discrete problem. This situation must be discrete as intermediate enrollment numbers are not reported. Even if a university reported weekly enrollment numbers, the variable would still be discrete as no intermediate change is being measured between weeks. Simulation and discrete dynamical systems are among the tools that can help solve these types of problems.

Notice that the line between continuous and discrete problems is blurry. Even though a shadow moves continuously, we could collect data in ten-minute increments and treat shadow length as changing discretely. Likewise, we could approximate a discrete situation with a continuous function, acknowledging that the function will only be valid for certain input values.

# Interpolation and Extrapolation

In data-based models, we typically fit a model using known data points. The domain of data points is often incomplete, and the model is needed to predict the dependent variable for values not included in the data set. It is called **interpolation** when we use the model to predict the dependent variable *within* the domain of the independent variable. It is called **extrapolation** when we use the model to make predictions *outside* the domain of the independent variable. Note that predicting the future will always involve extrapolation. However, extrapolation is generally considered unreliable as we have no mathematical assurance our model will continue to hold true outside of the domain of the data set.

# Data-driven versus Theory-driven

When we build our model, we often have the choice of using a data-driven or theory-driven approach. A **data-driven model** (also called an **empirical model**) blindly follows the data, with no intuition about the type of model that would be appropriate. We merely try to fit a model to data points. In this case, interpolation would be appropriate, but not extrapolation.

**Theory-driven modeling** is based on using mathematical reasoning to deduce the most appropriate type of model. Data may then be used to calculate specific values used in the model. In this case, the model is trying to explain any trends in the data. Thus, extrapolation may be possible, although one needs to check if the theoretical basis for the model is still valid.

To summarize, a data-driven model starts with data and attempts to fit the best possible model to the data. A theory-driven model starts with a type of model, and then fits it to available data.

# Accuracy versus Precision

It is tempting to use these words interchangeably, but they have different meanings. Consider two scales used to measure the weight of a 50 lb. pumpkin. On one scale, the pumpkin repeatedly weighs between 54.2 and 54.6 pounds. On the other scale, the pumpkin weighs 48, 51, 51, 52, and 47 pounds. The first scale is more **precise** than the second, because all values are close to each other. Meanwhile, the second scale is more **accurate** than the first because the average of the values is closer to the true weight. Of course, we would prefer a scale to be both accurate and precise.

# Sources of Error

It is helpful to identify sources of error and quantify the amount of error that is affecting the model. One source is **formulation** **error** and is caused by faulty assumptions and the impreciseness of model fitting. This may be hard or impossible to quantify, but it is important to acknowledge its existence. We can confront this error more carefully in the next stage of the modeling process.

A second source is **round-off error**. We can handle this error appropriately by maintaining a similar level of precision in all our calculations. This is particularly important when dealing with exponential functions. Suppose we are modeling population with census data that rounds to the nearest 1000 people. Then any values we use in the model should preserve that level of precision and accuracy. If we are using a model of the form, $P\left(t\right)=P\_{0}e^{kt}$, it will be important that $k$ is accurate to a precision of at least four decimal places. We should also be careful to not report results that are overly precise.

A third source is **measurement** or **data collection error**. This occurs when our method of collecting data introduces error. This can occur several ways:

* Imprecision of our measurement device,
* Inaccuracy of our measurement device,
* Inaccuracy of the method of measurement or data collection, and
* Recording error.

It is hard to account for recording error, but we can quantify measurement error by taking repeated measurements and tracking the variation in our results.

# Absolute and Relative Error

There are multiple ways to describe error. Assume $x\_{actual}$ represents the actual value for $x$ and $x\_{estimate}$ represents an estimation of $x.$ Then

$$absolute error=\left|x\_{actual}-x\_{estimate}\right|$$

and

$relative error=\frac{\left|x\_{actual}-x\_{estimate}\right|}{x\_{actual}}$.

Note that **percent error** is simply relative error expressed as a percentage.

A more complicated situation arises when trying to quantify error for a set of data, rather than a single point. We will discuss this in more detail in Section 2.H.

# Example #1:

Suppose a [video](https://emporia-my.sharepoint.com/personal/bhollenb_emporia_edu/Documents/Desktop/Modeling/Tools%20of%20MM/CNFreeFall.mov) shows a ball dropping frame-by-frame. In Frame 13, the height of the ball is 5.1 feet. At Frame 16 its height is 4.9 feet. At Frame 19, its height is 4.4 feet. Assume the heights are accurate to within 0.2 feet. Work through the “Build the Model” stage for the real-world question: *Describe the height of the ball as a function of frame number. Use both data-driven and theory-driven approaches.*

## Answer:

Normally we would need to work through the “define the problem” stage before thinking about building a model. However, this problem is more straightforward than most, especially if we are going to take a data-driven approach. We first observe that this problem could be considered continuous or discrete. On one hand, the frame-by-frame nature of the data is certainly discrete. On the other hand, we know the ball is moving continuously through time. We will take a function-based approach which is flexible enough to accommodate this type of situation. We let $t$ represent the independent variable, frame number, and $h$ represent height, the dependent variable. Figure 1 shows a plot of the three data points.

Based on this plot, we might first assume the relationship is linear. Using the first and third data points, we find the linear equation:

$$h(t)=-\frac{0.7}{6}\left(t-13\right)+5.1.$$

Figure 1: The height of the ball (in ft) as a function of frame number.

Figure 2 shows plots this line with the data. Although the line does not perfectly fit the middle data point, it is within the accuracy of our measurements. We could interpolate using the function of the graph to see that we expect the ball to be at a height between 4.5 and 4.6 at Frame 18. Although we could use the linear function to extrapolate the height of the ball for other frame numbers, we do not have much confidence in such predictions. (See Exercise #1.)

Figure 2: A linear model plotted with the data.

Let us next create a theory-driven model. In this case, we need to make some simplifying assumptions. For instance, we assume air resistance is negligible. We also assume the only force acting on the ball is gravity and that it is only moving in a vertical direction. Thus, we can start to build our model by writing the equation,

$$a\left(t\right)=h^{''}\left(t\right)=-g,$$

where $t$ represents time in number of frames as before. Integrating twice will yield a quadratic function of time. How do we deduce the coefficients of the quadratic function? We have two choices: continue using a theoretical approach or solve for them using the provided data. Because we do not know the initial velocity or height of the ball, we will use the data. Three data points is enough to provide a perfect fit through the data, using the function:

$$h\left(t\right)=-0.0167t^{2}+0.4167t+2.5.$$

Before we proceed, we should check that we have used enough decimal places to maintain the level of precision desired for the domain of interest. (See Exercise #4.)

Figure 3 shows the function plotted with the data. Since we have used a theory-driven approach it is possible to extrapolate as is shown in the plot. Compare this plot to the video to judge the reliability of extrapolation in this case. You will see that the predicted values are much better than the data-driven linear model.

Figure 3: A quadratic model plotted with the data.

However, we should still be skeptical. We assumed our measurements are within 0.2 feet of the actual value. If the data point for Frame 19 is off by 0.2 feet, how much would our model change if this single value was changed? As Figure 4 shows, the resulting quadratic functions differ by several feet for Frames 1 and 28. Without a better measurement, we cannot be certain which model is correct. Notice this effect would be less dramatic if we did not extrapolate so far from the domain of data. This effect can also be minimized by building a model from more data and/or a more theoretical approach. Regardless, we need to be aware that the usefulness of our model is sensitive to such considerations.

Figure 4: Two alternative quadratic models

# Question #1:

When is extrapolation considered reliable?

# Question #2:

Suppose I am throwing darts at a dartboard and my throws are precise, but not accurate. Describe what the distribution of my throws would look like.

# Question #3:

My model predicts the height of a falling ball after 3 seconds will be 37 feet. Its height is 33 feet. Calculate the absolute and percent error of my estimate.

# Exercise #1:

Extrapolate the linear model above to predict the height of the ball at Frame 1 and Frame 28 and compare to the video. How does the model do? Calculate the absolute and relative error for the predictions for each of these frames.

# Exercise #2:

Repeat Exercise #1 for the quadratic model discussed above.

# Exercise #3:

Express the potential measurement error in Example #1 as relative error.

# Exercise #4:

The exact function for the quadratic model above should be $h\left(t\right)=-\frac{1}{60}t^{2}+\frac{5}{12}t+2.5$. Calculate $h(1)$ and $h(28)$ using approximations for the coefficients with 2, 3, 4, and 5 decimal places of precision. What is the fewest number of decimal places needed to guarantee that these values are within 0.2 feet of the exact value (from the formula, not the video)?

# Exercise #5:

Assume there are 30 frames per second and the falling ball in the example above reaches its peak of 5.1 feet between 12 and 13 frames. Using only this information, find a purely theory-driven model for the ball’s height. How does it compare to the previous models?

# Problem #1:

Given two meter sticks and a sunny day, estimate the height of a nearby tall object. Identify the various sources of error and try to quantify them. Which one is likely to impact your model the most?

# Problem #2:

We have seen that a quadratic function can model the height of a falling object well in at least one situation. Use this model to estimate the speed of a baseball as it hits the ground after it is dropped from the top of the Space Needle in Seattle. Do you think this is realistic? Do some research to find out if the quadratic model is appropriate for this situation.

# Problem #3:

Repeat Problem #2 for the situation of dropping a penny from the observation deck of the Empire State Building.

# Problem #4:

Repeat Problem #2 for the situation of Felix Baumgartner’s record-breaking skydive in 2012. (Ignore the part of the jump after he opens his parachute.)

# Problem #5:

Work through the “Build the Model” stage for the staircase problem discussed in Section 1.C.

# Project #1:

Find a mathematical model that predicts an object will eventually reach a **terminal velocity** when falling. What are the parameters of the model? Use it to analyze one of the situations in Problems #2, 3, or 4. Does this model provide a better fit than the quadratic model?