2.1 – Introduction to Predictive Analytics

1 Predictive Analytics

Predictive modeling uses data to make a prediction of what is likely to happen [3]. In our last reading, we discussed the K-Nearest Neighbors Algorithm as our first predictive model. Here we will explore predictive modeling with continuous functions, beginning by applying our modeling triangle to predictive modeling and then familiarizing ourselves with several families of functions.

2 Types of Questions

As we look at how we apply modeling to make predictions, we first need to understand what type of questions we can answer using it. This is because, when we remember the mathematical modeling triangle, understanding what we need to find is one of the key inputs to selecting a model. Some questions you can answer using predictive modeling are:

- Into which category is my response variable likely to fall based on past data?
- How will a future change in my explanatory variable affect my response variable based on past data?

3 Applying the Modeling Triangle

The mathematical modeling triangle we learned in our previous readings applies here, but we can tailor our general *transform*, *solve*, and *interpret* steps to be more specific to building a predictive model. The general triangle with the customizations for predictive modeling are found in Figure 1.

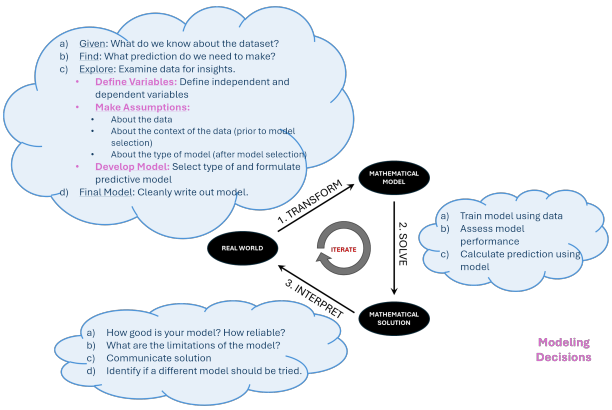

Figure 1: The mathematical modeling triangle as it is applied to predictive modeling. The pink writing identifies where the modeling decisions are made.

3.1 Transform

As shown in Figure 1, the flow through the transform step is the same. We start with what we are *given*, the data set. We also need to understand the question we are trying to answer; this is what we need to *find*. Once the inputs are understood, we *explore* the data to develop our model. Even the components of exploration are the same: define variables, make assumptions, and develop model. However, what is unique, is what we do within those components.

3.1.1 Define Variables

Understanding the dataset is instrumental to being able to build a predictive model. Recall the KNN exercise from our last reading using the following sample data:

Table 1: Sample Cadet Fitness Data

ID	Run Time	Deadlift	Age
A001	14.8	300	20
A002	13.5	275	19
A003	15.2	320	21
A004	13.9	290	20
A005	14.1	310	19
A010	14.0	295	20

As a part of defining our variables, we thought through and defined what each of these columns represent so that we didn't get confused later in the analysis. We then explored our data using a visualization and defined our *explanatory* and *response* variables. We will do this every time we build a predictive model.

3.1.2 Make Assumptions

We may make several assumptions throughout predictive modeling. When modeling with continuous functions, our assumptions will mainly address our conclusions about the shape of the data. After exploring our data with several visualizations, we may decide that it is best represented with a particular family of functions (linear, exponential, polynomial) [1]. This represents both an assumption about the data and a modeling decision. We must justify why this assumption is both reasonable and necessary.

3.1.3 Develop Model

What is a model in the context of predictive analytics? In this case, the model we are working toward includes a visualization and mathematical formulation that help answer the research question. Both of these items help translate the data into something simple to help us make sense of it. Your mathematical formulation is what allows you to calculate your prediction. A complete generalized model should include the explanatory and response variables, appropriate visualizations, and the mathematical formulation you plan to use.

3.2 Solve

The solve step of the mathematical modeling triangle includes calculating the parameters of your chosen math-

ematical formulation, generating the appropriate visualizations, making a prediction using the model you chose, and assessing model performance.

3.3 Interpret

During the interpret step you will communicate your prediction. Does your model allow you to sufficiently answer the research question? If not, do we need to iterate to try a different model?

4 Ethical Checklist

Remember that the ethical checklist we use requires three things: data validity, model validity, and clear communication. Below are some questions to think through as you consider whether your model is ethical.

- **Data validity**. As with descriptive statistics, you need look at where you data is coming from and how it was collected. Normally, when you are using data collected by someone else, there should be some report or description of how the data was collected. How much did you need to clean the data? What are the possible implications about this data set?
- **Model validity**. Examine your modeling decisions.
  - Did you make modeling decisions while cleaning or exploring your data? How did you identify your explanatory and response variables? Are they the best variables to choose? Does your model answer the question you are trying to find?
  - Is your predictive model accurate? How did you evaluate your model? How does your model perform on new data? Does your choice of model make sense in the context of the problem?
- **Communication**. Is your visualization clear? Is it misleading, or does it convey an honest representation of the data? Did you thoroughly communicate each of your assumptions and modeling decisions? Did you communicate the limitations of your model?

5 Families of Functions

We will now familiarize ourselves with linear, exponential, and polynomial functions.

5.1 Linear Functions

We will begin with **linear** functions of the form $y = mx+b$ where x in the independent variable, y is the dependent variable, m is the rate of change, and b is the y-intercept. Fixed values within the function, like m and b are called **parameters**. We will often estimate these parameters when formulating predictive models. By understanding how the parameters of a line affect its shape and location, we can begin to develop continuous linear models that represent the trends that may appear in discrete data sets [1]. While we will mainly use slope-intercept form in this block, you may see linear functions in any of the following forms.

Slope-intercept form of a line: $y = mx+b$ where m is the slope or rate of change and b is the y-intercept.

Point-slope form of a line: $y-y_0 = m(x-x_0)$ where m is the slope or rate of change and (x_0, y_0) is a point on the line.

General form of a line: $Ax + By + C = 0$ where A, B, and C are constants. [1]

5.2 Exponential Functions

The general form of the **exponential** function is $y = ab^x+d$ where the a parameter controls the vertical stretch and initial value, the b parameter controls the general shape, and the d parameter controls the vertical shift[1]. **Do not confuse exponential functions with power functions**! The exponential has the independent variable in the exponent. Take some time to pull up your favorite graphing tool to see what happens when you vary the parameters of an exponential function. The box below summarizes the behavior of exponential functions based on parameter values. By understanding how the parameters of an exponential function affect its shape and location, we can begin to develop continuous exponential models that represent the trends that may appear in discrete data sets [1].

Exponential Behavior Summary

- if $a > 0$, the graph of the function lies above the horizontal asymptote $y = d$
- if $a < 0$, the graph of the function lies below the horizontal asymptote $y = d$
- if $b > 1$, the function will diverge away from the horizontal asymptote $y = d$ as x grows
- if $0 < b < 1$, the function will converge toward the horizontal asymptote $y = d$ as x grows
- if $b = 1$, the function will remain a constant distance away from the horizontal asymptote $y = d$ as x grows [1]

In this course we will also use exponential functions of the form $y = a(1+r)^x + d$ where $b = 1+r$. This form draws attention to the parameter r as the *growth rate* or *decay rate* of the function. Earlier we said that b controls the general shape of an exponential function, now we can say it controls how quickly the function grows or decays.

5.3 Polynomial Functions

A **polynomial** function takes the form $y = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$ where n is a nonnegative integer and the a_i is the constant coefficient of the i^{th} term. The **degree** of the polynomial is n, the largest exponent of the independent variable [2]. The box below summarizes behaviors of polynomials based on the degree of the function.

Polynomial Behavior Summary

- the graph of a polynomial function of degree 2 or greater is an unbroken smooth curve (for degrees 1 and 0, the graph is a line)
- the graph of a polynomial function of degree n has at most $n - 1$ local extrema
- for the graph of any polynomial function (other than a constant function), as $|x|$ gets very large, $|y|$ grows very large [2]

6 Modeling Approaches: First Principles and Empirical

You may be wondering about how to select a family of functions when presented with a modeling problem. In some cases, the context of the problem provides us with clues about the family we might select, and our knowledge of the scenario may help us make meaningful estimations of our parameters. This approach of building a model based on what we expect to see if called a **first principles** approach.

In some cases, we may not have context clues, and our modeling approach must be based on what we observe. We must collect and explore data to determine what models are appropriate and use the data to estimate the parameters. This type of approach, where the model type and parameters are driven by the data, is called an **empirical** approach.

7 Assessing a Model

We will go into more specifics on how to assess a model in the next block, but for now there is a useful tool in Excel that can us understand how to assess a model. If we graph data and then add a trendline in Excel, we have the option to display the equation of the trendline on the chart. We also have the ability to add something called R^2 to our chart. If we add this, we should see a number displayed that is between 0 and 1. As the value for R^2 increases, we should notice that our trendline appears that to match our data better, and, as the value lowers, the trendline seems to match our data less.

This value, R^2, gives us a general sense of how well our model fits the data, it tells us how much of the variation in the response variable is explained by the model. While we will dive into how it's calculated later, for now, we can use R^2 as a useful, visual indicator of model fit when comparing different trendlines.

References

[1] US Military Academy. *Modeling in a Real and Complex World*. West Point, New York: Department of Mathematical Sciences, 2022.

[2] David Cohen, Theodore Lee, and David Sklar. *Precalculus*. Brooks/Cole, 2012.

[3] Frederick Hillier et al. *MA103 Mathematical Modeling: Introduction to Management, Science, & Business Analytics with Connect*. McGraw-Hill, 2024.