2.5 – Model Selection, Assessment, and Use

1 Introduction

Recall from our first reading in this block that predictive models use data to make a prediction of what is likely to happen [1]. In this reading, we will draw on the math modeling triangle to select, build, assess, and apply several models using both a first principles and an empirical approach. This reading should serve as a general guide for how to approach a predictive modeling problem.

2 Problem: Engine Cooling

You notice that your platoon's vehicles have a tendency to overheat. You want to create a model that will help you predict the temperature of the engine after a cooling period. You collect the following data at the site of several engine failures.

Time (min)	Average Engine Temp (°F)
0	250.50
2	207.61
4	152.29
6	112.36
8	93.24
10	86.84
12	85.30
14	85.04
16	85.00
18	85.00
20	85.00

2.1 Transform

1. Given: We are given some context in this problem that may be useful for a first principles approach. After reading through your old math textbook, you find that objects cool to ambient temperature over time according to an exponential model at a rate that is proportional to the difference between an object's temperature and ambient temperature. This is called Newton's Law of Cooling and is given by the following equation

$$T(t) = T_{ambient} + (T_0 - T_{ambient})e^{-kt}$$

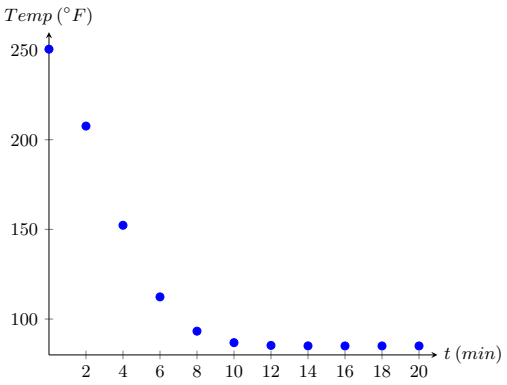
where $T_{ambient}$ is the ambient temperature, T_0 is the initial temperature of the object, k is a cooling constant based on the characteristics of the object's materials and environment, and t is time [2].

We are also given data that we can use to make a model using an empirical approach. This problem is an example where it may be good to use our contextual knowledge as a sanity check for an empirical model that we expect to be a decaying exponential. Notice that the given engine temperatures are averaged, indicating that data was collected over multiple incidents, but we do not know how many or how the data was collected.

2. Find: engine temperature after a specified time period.

3. Explore:

   - Define Variables:
     - t, time elapsed since engine overheat in minutes
     - $T(t)$, engine temperature in $^\circ F$ at time t
   - Make Assumptions:
     - We might be tempted to make an assumption here based on our contextual knowledge of the problem, but let's explore the data with a scatter plot first.



     The data appears to follow the shape of a decaying exponential function, so we will assume an exponential model. This is reasonable both because we have explored the data and because we have some contextual knowledge about the problem serving as our sanity check. It is necessary because we must assume some underlying trend to be able to make predictions and because a simpler linear model would not capture the shape of the data.
     - We will explore several different methods to solve for parameters in the Solve step, one of which will involve the following assumption. We know that the engine will cool down to ambient temperature based on the context of the problem, and we will need to find a value for ambient temperature as a parameter for our exponential function, so we must decide on a method for determining that value. It is reasonable to make this estimation by looking at the plot of our data to see where it appears to approach a horizontal asymptote.
   - Develop Model: We assumed an exponential model so our model will be of the form $y = ab^x + d$

4. Final Model: $T(t) = ab^t + d$

   - t, time elapsed since engine overheat in minutes

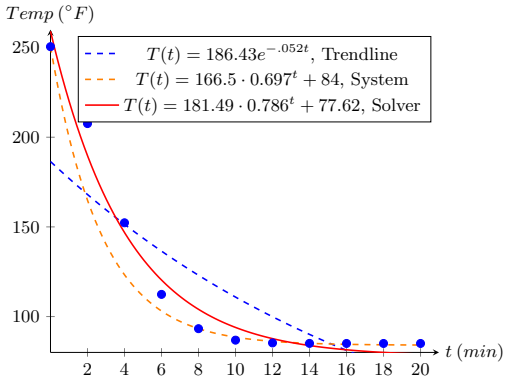
- $T(t)$, engine temperature in $^\circ F$
- a, scaling factor which affects the vertical stretch
- b, decay rate and the base of our exponential function
- d, horizontal asymptote, in this case ambient temperature

Note that these expressions complete the transform step. We define the general structure of the model symbolically, without yet substituting the numerical values we will find in the Solve step.

2.2 Solve

We have several options for finding the parameters of our exponential model. We will explore finding parameters using a system of equations, Excel trendline, and Excel Solver. We will discuss Excel Solver in detail in our prescriptive analytics block. For now, we need only know that it is an optimization tool that we can use to determine the parameter values that minimize SSE and that it will allow us to incorporate a vertical shift into our exponential model.

1. First let's find our parameters by solving a system of equations and using our assumed method for finding the ambient temperature, d.
   - If I look at the plot of my data, the temperate values appear to approach 84 (°F), so $d = 84$
   - I can now solve the following system of equations for a and b using two of my data points, $(0, 250.5)$ and $(8, 93.24)$,
   $$250.5 = ab^0 + 84$$
   $$93.24 = ab^8 + 84$$
   which results in $a = 166.5$ and $b = 0.697$.
   - Our model is now: $T(t) = 166.5 \cdot 0.697^t + 84$
   - We calculate that $SSE = 2756$ and $R^2 = 0.98$
2. Second let's try adding a trendline to our plot in Excel
   - Using this method, we find that $T(t) = 186.43e^{-0.052t}$
   - We calculate that $SSE = 8473$ and $R^2 = 0.79$
3. Lastly let's use Excel Solver to fine the parameter values that minimize SSE
   - Using this method, we find that $T(t) = 181.49 \cdot 0.78^t + 77.6$
   - We found the minimized value of $SSE = 733$ and $R^2 = 0.99$



Let's assess and compare our models using a chart to compare SSE and R^2 for each model.

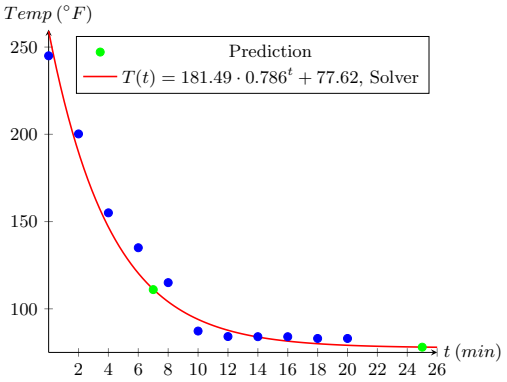
	System of Equations	Trendline	**Solver**
SSE	2756	8473	733
R^2	0.98	0.79	0.99

We have an interesting result here. Solving our system of equations and using Excel Solver both resulted in a lower SSE and higher R^2 value than the Excel trendline. This highlights both the usefulness of exploring more than one method for parameter estimation and the necessity of using a quantitative tool to assess and compare models. Here we will select the Solver model because it resulted in the least error and highest value of R^2.

Before proceeding, we should also consider whether or not our model could be overfit. From a visual inspection of the plot, we do not appear to be fitting noise. Let's test our model against new data. You collect the following data at the scene of another engine failure.

Time (min)	Engine Temp (°F)
0	245.00
2	200.23
4	155.00
6	135.00
8	115.00
10	87.25
12	84.10
14	84.02
16	84.00
18	83.00
20	83.00

Let's plot this data with our Solver model and calculate SSE to see how well it would have predicted this data.



The Solver model seems to fit the overall shape of the new data and results in $SSE = 772$, about a 5% increase in error. This is not a large jump in error or decline in performance when compared to the data with which we built our model. Compare this increase in error to the performance of our other models against the same new data using the table below. The large increase in error in the Excel trendline indicates that it is not representative of the overal trend of the data.

	System of Equations	Trendline	**Solver**
SSE	3785	109765	772
Percent Increase SSE	37	1195	5

We can now use our model to make predictions. For example, let's predict the temperature of an engine after 7 minutes of cool down time. We find $T(7) = 111.55$ as plotted above.

Seven minutes is between our observed values, so this prediction is an *interpolation*. If we were to predict the engine temperature at 25 minutes (78°F), that would be outside the range of our observed values and would be an *extrapolation*. Although we do have some context for this problem, we have no data about what happens to the engine temperature after 20 minutes, and would be less certain when making predictions outside of the observed window.

2.3 Interpret

We can be reasonably confident that our model is reliable. The shape of the data appears to be exponential, which fits with what we knew about the context of the problem. We used our first principles knowledge to check that our empirical model followed our understanding of cooling to ambient temperature.

Our first principles knowledge can also point us to the possible limitations of this model. We know that objects will cool to ambient temperature, which is explicitly defined as the horizontal asymptote in our model. This means that our model is only valid near a certain ambient temperature. If the ambient temperature is very different from this value, our model is no longer valid. In future readings, we will discuss how to determine a range of parameter values for which your model remains valid.

Given that we are near the ambient temperature observed in our data, our final model is,

$$T(t) = 181.49 \cdot 0.786^t + 77.62,$$

where $T(t)$ is the engine temperature in °F, at a given time, t. This model allows us to predict the engine temperature after a given cooling period. For example, we predict the engine temperature will be 111°F after 7 minutes of cooling.

3 Reflect - a Different Approach

Let's imagine that we do not know ahead of time that cooling can be modeled with an exponential function and had to go through our entire transform step with no first principles knowledge as a sanity check for our model. Do we end up with the same result?

Beginning again from the data you collected, let's build a purely empirical model. Our model selection will now focus on deciding between a linear, exponential, or polynomial function instead of different solving techniques for parameters.

Time (min)	Average Engine Temp (°F)
0	250.50
2	207.61
4	152.29
6	112.36
8	93.24
10	86.84
12	85.30
14	85.04
16	85.00
18	85.00
20	85.00

3.1 Transform (again)

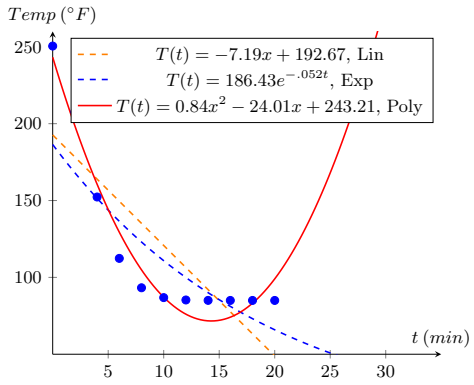
1. Given: We are given data that we can use to make an empirical model. The data consists of eleven average temperature observations at two minute intervals, starting at the time the engine overheats. We are not given any information on the total number of observations that make up the average temperature readings.

2. Find: engine temperature after a specified time period

3. Explore:

   - Define Variables:
     - t, time elapsed since engine overheat in minutes
     - $T(t)$, engine temperature in $^\circ F$ at time t
   - Make Assumptions:

- Before we make assumptions, let's explore the data using a scatter plot
- The data appears to be exponential in shape, but we should explore a variety of possible models before assuming a trend. We can use Excel's trendline tool to explore linear, exponential, and polynomial curves, each of which is shown in the following plot.



- We will use a chart to compare the SSE and R^2 of each function type. Let's assess and compare our models using a chart to compare SSE and R^2 for each model.

	Linear	Exponential	**Quadratic Polynomial**
SSE	10954	8473	1214
R^2	0.67	0.79	0.96

We select the polynomial model because it has the lowest error and highest R^2 value.

We've just made a modeling decision, so we should stop and ask ourselves if it's valid and address any concerns with our model selection. It appears that the polynomial model is valid within the range of observations, but what happens beyond that range? The data appears to level out at some ambient temperature, but our model is a quadratic and thus displays parabolic behavior, indicating that the temperature would begin to increase again. This is not reflected in the data.

We must be extremely cautious with our use of this model and warn decision makers against extrapolating outside the observed range.

- These observations demonstrate the importance of context and critical thinking in the modeling process. If we select a model solely on the metrics, without understanding context or appropriate usage, we may end up with invalid results. Ethical modeling must account for data validity and model validity, and modelers have a duty to communicate these validity conditions.

- Develop Model: We assumed a polynomial model, so our model will be of 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 a_i is the i^{th} constant coefficient.

4. Final Model: $T(t) = a_n t^n + a_{n-1}t^{n-1} + ... + a_1 t + a_0$ where n is a nonnegative integer and a_i is the i^{th} constant coefficient.

   - t, time elapsed since engine overheat in minutes
   - $T(t)$, engine temperature in $^\circ F$ at time t

3.2 Solve (again)

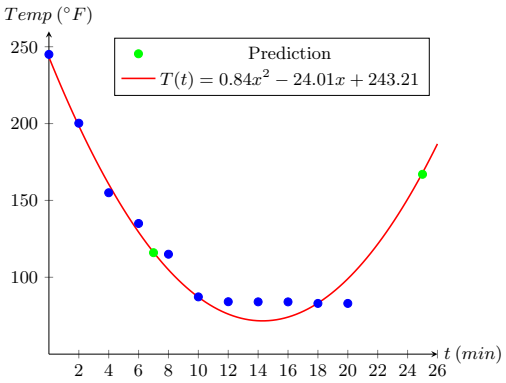
From here we build our model using the degree and coefficients determined using the Excel trendline tool. Our polynomial model becomes

$$T(t) = 0.84x^2 - 24.01x + 243.21.$$

We will use the same test data to assess under or overfit in our polynomial model.

Time (min)	Engine Temp (°F)
0	245.00
2	200.23
4	155.00
6	135.00
8	115.00
10	87.25
12	84.10
14	84.02
16	84.00
18	83.00
20	83.00

Let's plot this data with our model and calculate SSE to see how well it would have predicted this data.



The polynomial model seems to fit the overall shape of the data inside of the range of observations and results in $SSE = 1242$. This is not a large jump in error or decline in performance when compared to the data with which we built our model.

We can now use our model to make predictions. We will use the same data points as we did when we built our

blended first principles model. We find $T(7) = 116.09$ and $T(25) = 167.21$.

We see again the importance of critical thinking and context when using using a model. While this model works well for interpolation, it fails dramatically when we extrapolate outside of the range of observed data. While we do not have the knowledge of Newton's Law of Cooling for this exercise, we know that an engine does not begin increasing in temperature again while it is turned off for a cooling period.

3.3 Interpret (again)

We can be reasonably confident that our model is reliable within the observed range of data. Within this range, and with very little context, the shape of the data could reasonably be modeled with a polynomial function.

We identified a significant limitation to our model at every step in the process: we should not use this model to extrapolate. We used critical thinking to determine that the model is not valid outside the range of observed data.

We should think about other instances where our model may not be valid. Ambient temperature is implicitly built into our model as the vertical shift of our quadratic function, meaning that our model is only valid near the observed ambient temperature. Given that we are at or near this observed ambient temperature, our final model is,

$$T(t) = 0.84x^2 - 24.01x + 243.21,$$

where $T(t)$ is the engine temperature in °F, at a given time, t. We can write this in its equivalent form $T(t) = 0.84(t - 14.29)^2 + 71.6$ to explicitly show the vertical shift to ambient temperature. This model allows us to predict the engine temperature after a given cooling period. For example, we predict the engine temperature will be 116°F after 7 minutes of cooling.

4 Conclusion

This reading was meant to remind ourselves of the math modeling triangle and the importance of critical thinking when modeling. We built and assessed several different models and selected the best based on our understanding of the problem. This exercise should serve as a general guide for how to approach predictive modeling problems and highlight the nuance required when using mathematical models.

References

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

[2] James Stewart, Daniel Clegg, and Saleem Watson. *Calculus: Early Transcendentals*. Cengage, 2020.