WHOSE MODEL IS IT?: BRIDGING THE GAP BETWEEN ENGINEERING AND STATISTICS

As a statistician who occasionally consults with engineers, I find that we often get our wires crossed when discussing a project. We often use the same words, but they can mean very different things. This situation is even worse than using jargon. When jargon is used, the listener knows what he or she doesn't know and has the opportunity to ask clarifying questions. When common words represent different concepts, each party believes they are clearly communicating, when the opposite is true. I have found that two terms that generate considerable confusion are "model" and "parameter."

For engineers a model is derived from physical principles to describe a phenomenon. For example, consider measuring the temperature within a fuel tank over time while a vehicle is driven. A physical model can be developed from the heat transfer equation[1,2] that relates changes in temperature to changes in heat:

$$\Delta Q_t = m_t c_P \frac{dZ_t}{dt}$$

where

- t is time with $t \geq 0$;
- Q_t is the heat in the system at time t;
- m_t is the mass of the fuel and tank at time t;
- c_P is the coefficient of specific heat of the fuel; and
- Z_t is the temperature within the fuel tank at time t.

Changes in heat in the fuel tank are from three primary sources:

1. Convection heat from under the tank: $Q_{ct} = UA(T_U - Z_t)$ where U is the heat transfer coefficient for the fuel tank; A is the surface area of the bottom of the fuel tank; and T_U is the temperature below the fuel tank, taken to be constant.
2. The heat in the fuel that is returned to the tank from the engine: $Q_{Rt} = m_R c_P (T_R - Z_t)$ where m_R is the mass of the returned fuel, and T_R is the temperature of the returned fuel, assumed to be constant. If the engine does not return fuel, then Q_{Rt} is zero.
3. Heat from the fuel pump, if it is adjacent to the fuel tank: Q_P.

The change in heat within the fuel tank is: $\Delta Q_t = Q_{Ct} + Q_{Rt} + Q_P$.

Assume that the vehicle uses fuel at a constant rate. The mass of the fuel and the tank at time t is $m_t = m_0 - gt$ where m_0 is the initial mass, and g is the constant rate of fuel consumption. Then the solution of the heat transfer equation is:

$$Z_t = \frac{D}{B}\left[1 - \left(1 - \frac{gt}{m_0}\right)^{B/c_P g}\right] + Z_0\left(1 - \frac{gt}{m_0}\right)^{B/c_P g} \quad (1)$$

where Z_0 is initial temperature, and

$$D = Q_P + c_P m_R T_R + UAT_U$$
$$B = c_P + m_R + UA; \quad (2)$$

The solution is a weighted average of the initial temperature Z_0 and the asymptotic or equilibrium temperature D/B. The weights are between zero and one, sum to one, and vary monotonically with time. If g, the rate of fuel usage, is small, then an approximation is:

$$Z_t = \frac{D}{B}\left[1 - \exp\left(-\frac{Bt}{c_P m_0}\right)\right] + Z_0 \exp\left(-\frac{Bt}{c_P m_0}\right). \quad (3)$$

The physical model given in Equations 1 and 2 relates the dynamics of temperature to design parameters, such as surface area and heat transfer coefficient, and to operating conditions, such as initial temperature and the temperature under the fuel tank. The challenge for the design engineer is to select the design parameters to obtain desired temperature profiles for a variety of field conditions or for test protocols mandated by government agencies.

The model was derived under assumptions, such as constant fuel usage and temperature under the tank, that are certainly not true, and the solution depends on parameters that are measured with different levels of accuracy or may be unknown. The utility of the model depends on the model's sensitivity to these assumptions and parameter uncertainty.

In contrast to engineers, statisticians do not think in terms of differential equations and Newtonian physics. In fact, most graduate programs in statistics do not require a course in differential equations or physics. How do statisticians think? ("If they think at all," my engineering friends would all too readily add.) Specifying a statistical model begins with the goals of the study and understanding how the observations are collected. Consider the experiment where the temperature within the tank is measured on n vehicles at m points in time. Most statistical models have the general form:

$$Y_{it} = f(t, X_i, \theta) + \varepsilon_{it} \quad (4)$$

where the components of the model are:

- Y_{it} the observed temperature at time t on the i^{th} experimental run or vehicle;
- f is a known function with three arguments;
- X_i is a vector of observations or known constants for the i^{th} experimental unit;
- θ is a vector of unknown statistical parameters; and
- ε_{it} is "measurement error."

P. Lenk is Associate Professor of Statistics with The University of Michigan Business School, Ann Arbor, MI.

Measurement error is not observed and is included to balance the left-hand and right-hand sides of eq (4). Measurement errors are assumed to have a specified distribution. Most frequently, they are mutually independent and have a normal distribution with mean 0 and unknown variance. Their inclusion in the model explicitly recognizes that the observations will deviate from f. Statistical inference quantifies the variation of the observations from f.

The function f may be related to the physical model, but it need not be. The vector of observations X_i may include variables such as ambient temperature or return fuel temperature if they are measured during the experiment. This vector may also include physical parameters such as the coefficient of specific heat. Statisticians frequently treat physical parameters as data in statistical models. On the other hand, the statistical parameters θ often are not directly related to physical parameters.

Statistical inference estimates the unknown statistical parameters θ. The statistician will report, "The estimated (statistical) parameters of the (statistical) model are...." Meanwhile, the engineer is thinking, "Hold on. We selected the (physical) parameters when we designed the fuel tank. Why would you estimate them?"

The choice of the components of the statistical model depends on information at the time of the experiment, the variables being measured during the experiment, the measurement system, and the goal of the experiment. The simplest case of eq (4) states that the observed temperature for the i^{th} experimental run at time t is due to a mean temperature at time t and a random deviation from the mean:

$$Y_{it} = \theta_t + \varepsilon_{it}. \qquad (5)$$

It is not surprising that an engineer would be dismayed at calling eq (5) a model for fuel tank temperatures: it is divorced from the underlying physics. This statistical model does not attempt to explain the temperature dynamics in the fuel tank. It is designed for very limited purposes and can be used to validate the engineering design or to test the adequacy of the physical model.

Consider the experiment where one vehicle is repeatedly tested under nearly identical conditions. After each test run, the fuel tank is returned to its initial temperature, and the test is repeated n times on the same vehicle. Based on this experiment, the statistician could test the hypothesis that the mean temperature θ_t at time t is equal to the theoretical temperature Z_t. Of course, the sample average $1/n\ \Sigma_{i=1}^{n} Y_{it}$ will be different from the theoretical values Z_t. If the discrepancies were within the range that would be anticipated due to random sampling, the statistician would conclude that the data supports the physical model.

The means $\{\theta_t\}$ in eq (5) are allowed to vary freely: θ_{t+s} may be larger or smaller than θ_t, while the solution of the heat transfer equation is monotonically increasing in time. A more sophisticated statistical model could be motivated from the theoretical solution in eq (3):

$$Y_{it} = \theta_1[1 - \exp(-\theta_2 t)] + Y_{i0} \exp(-\theta_2 t) + \varepsilon_{it} \qquad (6)$$

This model forces f to be monotonically increasing as a function of time. The statistical parameters are θ_1, the equilibrium temperature, and θ_2, the exponential rate at which the temperature increases from its initial temperature to θ_1. This model is appropriate under the following conditions:

- The same vehicle or vehicles with nearly identical physical parameters are used in each test run;
- The test conditions on each run are nearly identical; and
- Only the interior temperature of the fuel tank is measured.

This model is used to predict the equilibrium temperate and the rate that it is reached for a given configuration and testing conditions. The statistician uses the form of Z_t as a function of time to motivate eq (6) but is not interested in the fine detail of the physical model.

After estimating the statistical model in eq (6), the statistician and engineer have the following conversation.

Statistician: I have the results of the experiment and the estimated equilibrium temperature is 110 degrees.
Engineer: Great! That temperature is lower than we expected, but it is still too high. What happens to the equilibrium temperature if we decreased the bottom area of the fuel tank?
Statistician: Well, we will not know that until we perform additional experiments with the new fuel tanks.
Engineer: What? We just spent our budget! You said you had a model for fuel tank temperatures. Why can't we just use it to predict the temperatures for different parameters?
Statistician: Well, we didn't run an experiment that varied the area of the fuel tank. There is no way we use our data to predict the effect of area on the equilibrium temperature.

To answer the engineer's question, the statistical model in eq (6) could be expanded to include physical parameters, such as bottom area, and observed variables or covariates, such as road temperature:

$$Y_{it} = \frac{\theta_1 + \theta_2 c_P + \theta_3 A T_U}{c_P + \theta_4 A}\left[1 - \exp\left(-\frac{c_P + \theta_4 A}{c_p m_0} t\right)\right] + Y_{0t} \exp\left(-\frac{c_P + \theta_4 A}{c_p m_0} t\right) + \varepsilon_{it} \qquad (7)$$

The statistical parameters can be viewed as functions of physical parameters. However, their estimated values may differ substantially from the physical parameters due to a variety of reasons including sampling and measurement errors, missing sources of heat in the physical model, and violations of the assumptions of the physical model. Equation (7) is appropriate if vehicles with different fuel tank configurations were tested under different ambient temperatures. Then the model could be used to predict the impact of changing design parameters such as fuel tank area or of operating the vehicle in different temperatures. Of the three statistical models eq (7) is closest in spirit to the theoretical model.

Statistical models and experiments are used to answer limited questions and not to investigate fully the dynamics of the heat transfer equation for all possible configurations. As a last example, the engineer may be interested in the effects

of three different design factors on the temperature within the fuel tank after thirty minutes of operating the vehicle. The factors are the thickness for the fuel tank walls, the geometry of the fuel tank, and the type of return fuel line. Fuel tank walls have one of three widths. The tanks have two levels of geometry: tall and narrow versus short and wide. The return lines have two levels: insulated versus uninsulated. Vehicles with different configurations are operated for thirty minutes before taking the temperature within the tank. Ambient temperature is recorded and is used as a covariate. The traditional analysis of variance or response surface methods[3] provide useful design information without recourse to physical models.

The information being sought by this experiment is relatively crude: which combination of experimental factors results in the lowest temperature over a range of ambient temperatures. In contrast, physical model attempts to describe temperature dynamics for every possible configuration and operating condition. Accurate physical models supercede statistical models. However, most physical models are derived from assumptions that often are violated, and they use approximations that may effect their accuracy. In addition, they usually contain parameters that are imperfectly measured or are unknown. Statistical models are designed to provide "fast and dirty" answers to well posed and limited questions. Statistical models may be motivated by physical models, but, more often than not, they are not.

If an engineer is asked for a model for fuel tank temperatures, he or she will immediately think, "heat transfer equation." The statistician does not have a fixed model. He or she cannot begin to formulate a model until knowing the goal of the experiment, the experimental design, the measurement system, the variables being measured, and the experimental protocol. The continuing challenge is to combine physical and statistical models to reduce development times and increase design effectiveness. A needed step in this direction is to understand each other's models and parameters.

ACKNOWLEDGMENTS

I wish to recognize Chester Szczotka and Todd Crompton at the Ford Motor Company and Andy Andrews at the University of Michigan for their collaboration in developing the physical and statistical models for fuel tank temperatures. The events in the article have been fictionalized to illustrate the differences in models and parameters.

References

1. Kreider, D. L., Kuller, R. G., Ostberg, D. R., and Perkins, F. W., *An Introduction to Linear Analysis*, Addison-Wesley, Reading, Massachusetts (1966).

2. Whalley, P. B., *Basic Engineering Thermodynamics*, Oxford University Press, Oxford (1992).

3. Box, G. E. P., and Draper, N. R., *Empirical Model-Building and Response Surfaces*, John Wiley & Sons, New York (1987). ■