Components of Modeling
Part 4

Robustness

A model is *robust* if the conclusions you reach don't change dramatically if some of the assumptions are only approximately true. Of course, it can be very hard to determine how robust a model is. One particularly important measure of robustness is the sensitivity of your conclusions to variation in the parameters discussed above but robustness can also refer to the sensitivity of the conclusions to other kinds of assumptions that are made.

Example: An electronics manufacturer produces a variety of diodes. Quality control engineers attempt to insure that faulty diodes will be detected in the factory before they are shipped. It is estimated that 0.3% of the diodes produced will be faulty. It is possible to test each diode individually or to place a number of diodes in series and test the entire group. If the test of an entire group fails then one or more of the diodes in the group is faulty. The estimated testing cost is 5 cents per diode and $4+n$ cents for a group of $n \geq 1$ diodes. If a group test fails then each diode must be retested individually to find the bad one(s). Find the most cost-effective quality control procedure for detecting bad diodes.

Let

$n =$ number of diodes per test group

$C =$ testing cost for one group (cents)

$A =$ average testing cost (cents/diode)

If $n = 1$ then $C = 5$. If $n > 1$ then

$$C = \begin{cases} 4 + n & \text{all diodes in group are good} \\ (4 + n) + 5n & \text{at least one diode is bad} \end{cases}$$

$A =$ average value of C$/n$.

We are free to choose the value of n, the value of C is random and depends on n and the particular group of diodes that we test, the value of A is non-random; it is the average over many many groups.

Objective: Find the value of n that minimizes A.

The random variable C is equal to $4+n$ if all the diodes are good. The probability of this is 0.997^n. On the other hand, it is equal to $4+n+5n = 4+6n$ with probability $1-0.997^n$.

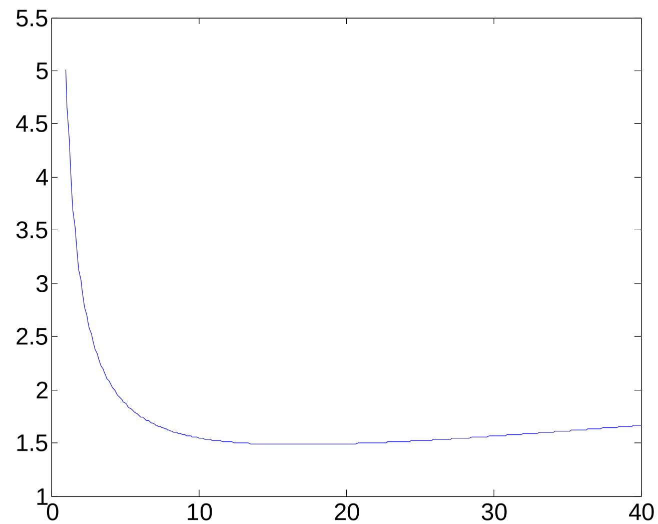
Thus, the average or expected value of C is

$$0.997^n(4+n)+(4+6n)(1-0.997^n) = 4+6n-5n(0.997^n)$$

so the value of A is

$$A = \frac{4}{n}+6-5(0.997^n).$$

Here is a Matlab plot of A; it looks very well-behaved in terms of finding the minimum.



The minimum returned by `fminsearch` is attained at $n = 16.7330$ which makes sense with the plot. Since the number of diodes must be an integer and the function decreases to its minimum and then increases, the minimum on the integers must be at either $n = 16$ or $n = 17$.

$$A/_{n=16} = 1.4847$$

$$A/_{n=17} = 1.4843$$

They should test 17 diodes at a time.

Now let's address the issue of practicality, sensitivity to parameter values, and robustness.

It may not be practical to test the diodes in bunches of 17; it may be more practical to test them in bunches of 10 or 20. The graph of A as a function of n is quite flat for n ranging between about 10 and 35, so testing in bunches of 20 won't make much difference to the cost.

The parameter $q = 0.003$ is the failure rate of the manufacturing process. It is likely to have been estimated roughly and could quite well change over time. It is natural to ask, therefore, 'How sensitive is A to the value of q?' In particular, if q changed by, say 10%, by what percentage would A change?

The sensitivity of A to changes in q is

$$S(A, q) = \left(\frac{dA}{dq}\right)\left(\frac{q}{A}\right)$$

Since

$$A = \frac{4}{n} + 6 - 5(1-q)^n$$

it follows that

$$S(A, q) = \frac{5nq(1-q)^{n-1}}{\frac{4}{n} + 6 - 5(1-q)^n}.$$

So when $n = 17$, $S(A, q) = 0.16$. In other words, if we test groups of 17 diodes, then the average cost A will increase by approximately 0.16% for every 1% increase in q. This means that A isn't very sensitive to q which is good!

More generally, our model and analysis of the situation relied on independence of the diodes; when we said that the probability that all of the diodes in a group of n are good is 0.997^n we assumed that one diode being good is *independent* of another being good.

This would definitely be violated if the manufacturing process was going awry (maybe some machine got maladjusted) since then once one diode was bad it would be very likely for all the next ones to be bad too.

However, we aren't trying to model the case when the manufacturing process has gone awry. Instead, we are modeling the case when the manufacturing process is in control, but there are minor things that cause diodes to be bad. These things might be a person on the assembly line tying their shoe or how the diode falls into the machines.

These are probably roughly independent; if a person ties their shoe they might miss a few diodes but how the diode falls into a machine probably doesn't depend on how the previous diode fell into the machine, so most of the factors that affect whether a diode is bad or not only affect one diode.

Simulation results tend to indicate that expected values are relatively robust to the assumption of independence, so our conclusion is probably not far off the optimum.

Model Validation

A model has been *validated* when predictions based on the model agree with data.

When a model is developed based solely on qualitative information about how the variables behave (and not using any data) and then parameters are found that produce a good match with the data, then this is considered a validation of the model.

If the model is developed based on observations of data and not on qualitative information about how the variables behave, then another set of data needs to be collected in order to validate the model. If you don't collect another set of data to validate the model you can be accused of *over fitting* the data.

Example: In the warm-up problem we modeled the concentration of drug in the body c as decaying exponentially. This model was based on our understanding of how the kidneys work and was develop without reference to any data set.

When we then checked the model against data, and found a decay rate that matched the concentrations predicted by the model to the concentrations in the data, this provided a validation of the model.

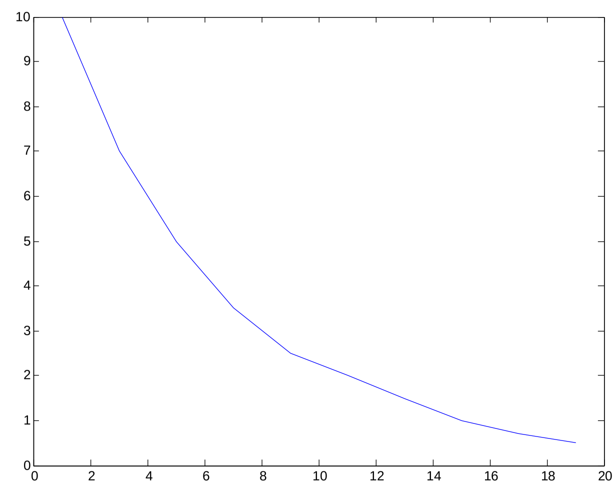
Suppose on the other hand that we had started with the data.

Data from a person that weighs 50 kg with an initial injection of 300 mg and no further injections.

t hrs	1	3	5	7	9
c mg/l	10.0	7.0	5.0	3.5	2.5

t hrs	11	13	15	17	19
c mg/l	2.0	1.5	1.0	0.7	0.5

Here's a plot of the data:



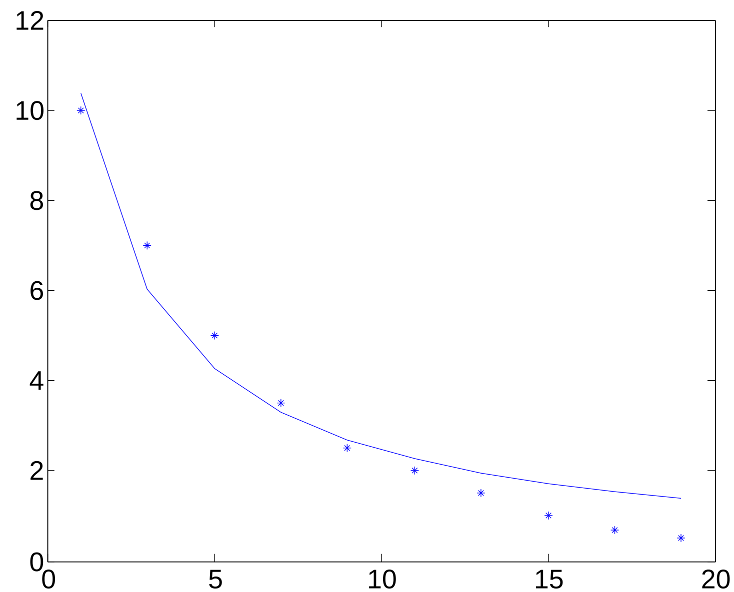
Based on this data we might conjecture that the concentration decays according to the form:

$$c(t) = \frac{a}{kt + 1}$$

where a and k are parameters.

To fit this model to the data we find the parameters that make this curve match the data most closely.

```
>> c = [10, 7, 5, 3.5, 2.5, 2, 1.5, 1, 0.7, 0.5];
>> t = 1:2:19;
>> e = e = @(par) sum((par(1)./(par(2)*t+1) - c).^2);
>> best = fminsearch(e, [12, 2])
best =
16.1854 0.5599
>> plot(t, best(1)./(best(2)*t+1))
>> hold on
>> plot(t, c, '*')
```



Doesn't look bad.

However, since the model was developed *using* the data and not independently of the data, this good fit does not validate the model. To validate the model one should collect other data (probably from another person and probably with a different initial dose) and see if you can get a good fit there as well.

(Actually to a trained eye, the consistent overestimation at the tail is a suggestion that the wrong functional form has been used.)

Use of the computer and simulation

Hopefully through all the examples we have discussed and through your own modeling you've seen what an important role the computer has to play in the modeling process. Here's a partial list of how you might use a computer.

- To estimate values of the parameters.
- To do a sensitivity analysis and explore the effect of changing the parameters.
- To form hypotheses about how the variables evolve. If you can then *prove* these hypotheses (or at least give a partial proof) this will give you greater understanding of the model and may give you information on how to modify the model for other circumstances or improve the model.
- To validate the model.
- To make predictions.

A good modeler will mover freely between simulating on the computer and doing analysis of their model.

Simulation is incredibly powerful and simulating the model as it is developed will give you confidence and help you to understand your model.

However, once you've seen some simulations and made conjectures on how the variables will behave when you change the parameters, try to go back to the model and understand why you are seeing that behavior.