Model Comparison in the Introductory Physics Laboratory

Brian S. Blais, Bryant University, Smithfield, RI and Brown University, Providence, RI.

Model comparison is at the heart of all scientific methodologies. Progress is made in science by constructing many models possibly of different complexities, testing them against measurements, and determining which of them explain the data the best. It is my observation, however, that in many introductory physics labs we provide students with the materials and methods to verify the "correct" model of the experiment they are performing, e.g. measuring "g" or verifying the period of a pendulum. In this way, we do our students a disservice and don't allow them to experience the richness and creativity that constitutes the scientific enterprise. Limiting the lab to the "correct" model can have its uses — for example, getting the students to practice the proper methods to measure lengths and times or to support the specific theory covered in the lecture portion of the class. However, when students perform these labs they come to view these activities as repetitive and mechanical, reinforcing the notion that science concerns not the true exploration of nature but simply the verification of what we already know. By verifying what we already know, the laboratory experience does not improve overall understanding[1] and can mislead students about the methods of science overall.

This paper proposes to include model comparison — even in the introductory physics lab — to mimic the process of research as done by practicing scientists, and provides the students with a more enriching laboratory experience. We do this by making the outcome unknown or uncertain and to stress model building and comparison. Some of these steps are described by Holmes and Bonn[2], in the context of uncertainty in labs, but here we extend it to the entire laboratory activity. In the process, I introduce some elementary methods for doing model comparison in introductory physics laboratory settings which are not much more difficult than the alternatives, yet also expose students to a richer laboratory experience. One of the benefits of the methods proposed here is the introduction of the notion that increased model *complexity*, despite yielding a somewhat better fit to the data, should not be favored unless there is a *substantial* increase in the quality of the fit — the complexity comes at a cost. By thinking in terms of model complexity in addition to the quality of the fit we can explore models of many kinds and use them to help us design better laboratory experiences for the students.

Free fall

A common introductory laboratory is the so-called "free-fall" lab, often constructed as an exercise to estimate the acceleration due to gravity[3], g. Efforts are often taken to reduce air friction, a complexity typically ignored in any analysis in class. Other treatments make use of air-friction, but focus on the time-frame of the terminal velocity behavior[4] - simplifying the situation to constant speeds. As physicists we know that these are not the only two alternatives and further, even if they were the only two alternatives we would not want to

restrict ourselves to just one of them. To pave the way to a full model comparison treatment the instructor can make some simple adjustments: don't take efforts to reduce friction (perhaps even choose to *increase* it) and look at the *entire* time-frame. Specifically, in addition to small metal balls, I am proposing instructors include balls of paper or aluminum foil, even packing peanuts and coffee filters[5] — the added complexity is a *benefit*. By using a *variety* of objects, it becomes unclear if the theory described in class will be applicable in every case and the students have to explore multiple possible models. The students can be presented with models which are *approximately* correct, and the students can be required to justify when these approximations are useful. Two such models that are immediately available are the *air-free* model, where the object falls at a constant *acceleration*

Model Air-Free

$$y = y_o - \frac{1}{2} g t^2 \tag{1}$$

and the *air-dominant* model, where the object falls at a constant *speed*

Model Air-Dominant

$$y = y_o - v_f t \tag{2}$$

In order to determine which model is more correct we need data at two or more points in time. These measurements can be done with video or by hand with stopwatches/smart phones, and may be designed in a number of ways, e.g. time to pass half-way to the floor and to impact, many time points taken on a video, total time to fall from several different heights, etc… The particular experimental design does not matter for this paper, but the idea of *different model predictions* should be stressed. Let's presume we have the sample data shown in Figure 1of the height of two objects as they fall from 2 meters. These kinds of data can be easily obtained with modern range finders[6] or the analysis of movie[7]. However even data with far less temporal resolution can be approached in the fashion proposed here.

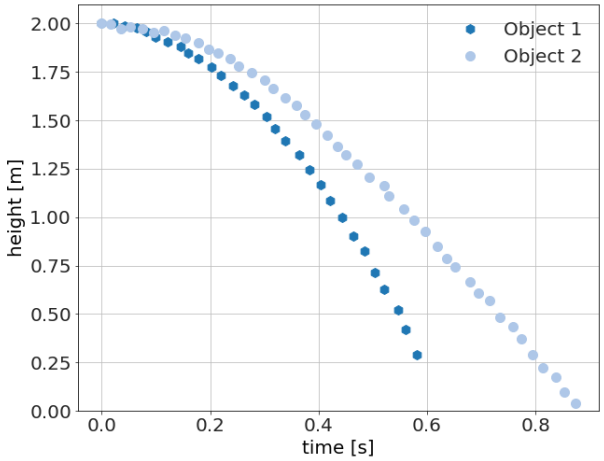

Fig 1. Data for two objects falling from 2 meters. Object 2 is experiencing a stronger effect of air friction, and can be seen nearing terminal velocity near the end of the time period.

We can then use these data to fit both models, to obtain best estimates and uncertainties for the unknown parameters, g and v_f, and to compare which model works best for this particular object. The analysis can be done in Excel[8], Python[9], or any other numerical curve fitting program. The result is shown in Figure 2. Although technically the Air-Free Model is "better" (i.e. has a lower MSE and thus is closer to the data) one can see even by eye that neither model is particularly good. Observing that both models are not acceptable motivates us to find a better one.

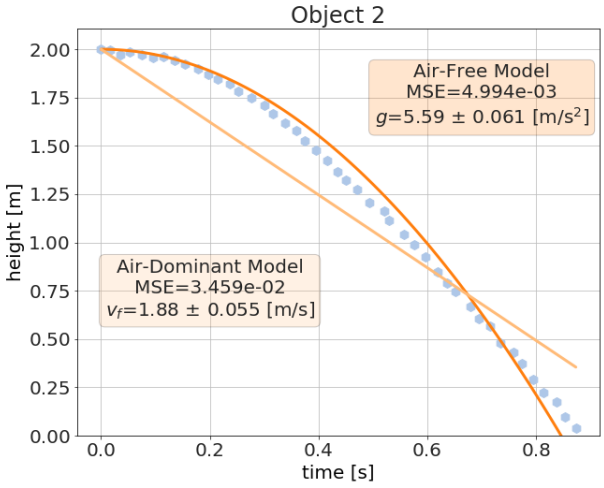

Fig 2. Fitting parameters to free-fall models, air-free and air-dominant, applied to data from Object 2. The air-free model is clearly "better" than the air-dominant model, but also clearly falls short of satisfactory - even by eye.

When all models fail

When faced with the situation where all of models do poorly, what do we do as scientists? Sometimes we have to tentatively proceed with the best of the poor alternatives, as we have done historically in the case of Newton's Law of Gravitation after Mercury's orbit was measured carefully[10]. Sometimes the failure of models motivates us to find an alternative which we pursue now. While some objects follow the *air-free* model and others the *air-dominant* model, students may observe that the rest seem to fall somewhere in between. Further, there aren't any objects observed which fall *faster* than the predictions of the *air-free* model and for many objects (like Object 2 above) the latter part of the trajectory appears to show a *constant* speed. This might make us propose a *transition to air-dominant* model — or several of them. A simple model might be,

Model Transition

$$y = \begin{cases} y_o - \frac{1}{2} g t^2 & \text{for } t \leq t_c \\ y_o - v_f t & \text{for } t > t_c \end{cases} \tag{3}$$

where t_c is an, unknown, *critical time* for transitioning from one model to another. Other models have been explored[5,11] and may have advantages to the one proposed here. To reiterate, it's not the *particular* model that is important but the fact that there are an infinite number of possible models of varying complexity, each with different assumptions and limitations. It's also valuable to point out to students that we may come up with models even when the ones we have actually work — that there could be more than one "correct" answer.

The process of model creation is the first step to improving our understanding of the system. One must follow this with the techniques to analyze and compare them. We often stress the minimization of the mean-squared error (MSE) as the measure of the ``goodness of fit'', however with models of varying complexity this measure needs to be improved. This we explore presently.

Beyond the Mean Squared Error - Enter the BIC

It is a numerical fact that a model with 3 free parameters (e.g. Model Transition) will *fit* better, i.e. have a lower mean squared error (MSE), than a subset of that model but with only 1 free parameter (e.g. Model Air-Free) *no matter what the data* (Figure 3). Because the latter model is a subset of the former, we have the freedom of two extra parameters in the Transition model to possibly "improve" the overall fit (and possibly overfit).

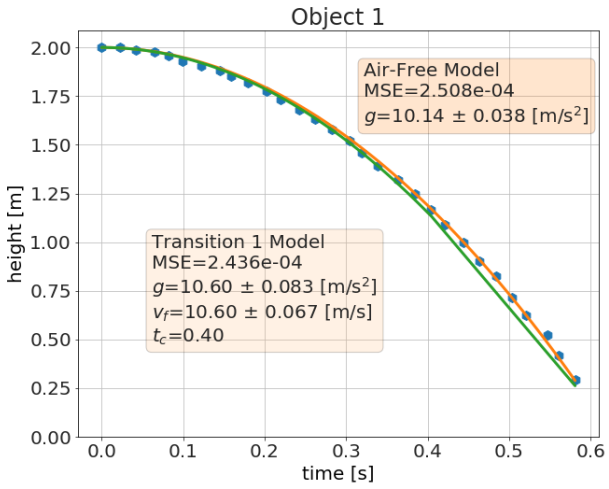

Fig 3. Fitting parameters to free-fall models, the air-free and simple Transition, applied to data from Object 1. The fits nearly the same, with the more complex model achieving a smaller mean squared error (MSE) due entirely to the added freedom of extra fit parameters.

As scientists, we typically favor the *simpler* model all things being equal — applying the so-called Occam's Razor[12]. However, we also admit that we should favor the *more complex* model in cases where it fits *substantially* better than the simple one - but how much is *substantial*? One procedure used in more advanced statistics is Bayesian model comparison, or the approximation with the Bayesian Information Criterion (BIC)[13]. In this procedure, the measure of the fit of the model is a combination of two terms — one including the mean squared error (MSE) and another which measures the *complexity* of the model. The complexity is measured by the number of *fitted* parameters, denoted by k. The mathematical form of BIC is simple enough to introduce in an introductory setting even if a full derivation is beyond such a class,

$$\mathrm{BIC} = k \cdot \log(N) + N \cdot \log\,(\mathrm{MSE}) \tag{4}$$

where N is the number of data points and k is the number of *fitted* parameters. Notice that since the BIC depends on the Mean Squared Error (MSE), a *larger BIC represents a poorer fit to the data*. For models with the same complexity (i.e. equal k), comparing BIC is identical to comparing MSE.

Applied to the data above (Figure 4), we can see that the *complex* Transition model applied to Object 1, which has a lower MSE and thus a better ``fit'', achieves a *higher BIC*, however, leading us to conclude that the extra complexity of the model is *not justified*. Applied to Object 2 the added complexity *is* warranted, as seen in Table 1. We note that, when applied to models of the same complexity, using the BIC measure is equivalent to using the MSE for goodness-of-fit. However, the BIC measure generalizes to models of varying complexity and provides students with a method of model comparison in these cases.

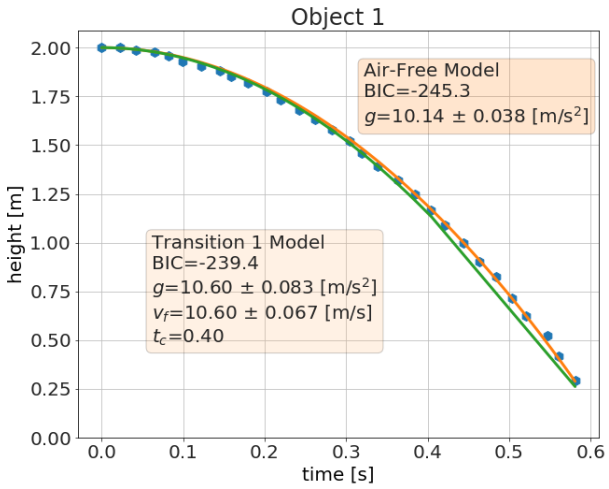

Fig 4. Fitting parameters to free-fall models, the air-free and simple Transition, applied to data from Object 1. The models fit nearly the same, with the simpler model achieving a smaller BIC due to penalty for the extra complexity of fit parameters in the Transition model.

Object	Model	MSE	BIC
Object 1	Air-Free	0.000251	**-245**
	Air-Dominant	0.034	-98.1
	Transition	**0.000244**	-239
Object 3	Air-Free	0.00499	-235
	Air-Dominant	0.0346	-148
	Transition	**0.000221**	**-367**

Table 1. MSE and BIC Results for Object 1 (an object with little air friction) and Object 2 (an object with more air friction) for several models. Using MSE as the measure of ”goodness of fit” leads to nearly always favoring the more complex model. Using the BIC as the measure of ”goodness of fit” leads one to favor the simpler model when the added complexities are not warranted.

Pennies

As another example that is both accessible and interesting to students, we look at the density of pennies across the years. These data are easily collected by students near the beginning of the semester, a sample shown in Figure 5. There seems to be at least one transition in these data, around 1980, where the density seems to drop. As before, we consider several models and apply them to the data.

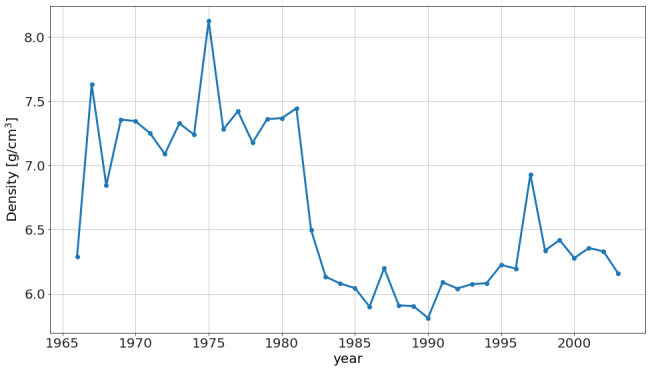

Fig 5. Density of pennies vs year.

Model 1: Constant

$$y = \text{constant} \tag{5}$$

Model 2: Linear

$$y = \text{slope} \cdot t + \text{intercept} \tag{6}$$

Model 3: Single-breakpoint Constant

$$y == \begin{cases} \text{constant}_1 & \text{for } t \leq t_c \\ \text{constant}_2 & \text{for } t > t_c \end{cases} \tag{7}$$

Calculating the MSE is straightforward for the constant models. The constant in Model 1 is just the average of the entire data set and the piecewise constants in Model 3 are just the average values in those time ranges, given the breakpoint time, t_c. The MSE for the linear model, as for the examples above, is provided by any software that performs linear fits. The results are shown in Figure 6. For a more elementary analysis, one can easily restrict the models to piecewise constant examples.

This is a good time to remind students about the meaning the ``number of parameters'' value, k, in the calculation for BIC (Equation) . In Model 1, there is only one parameter that can be fit or adjusted — the constant — so $k = 1$. For the linear Model 2 there are two such parameters, so $k = 2$. The Single-breakpoint Constant Model 3 has $k = 3$ because each of the two constants needs to be fit, but also the time of the breakpoint is a third adjustable parameter. It can be challenging for students to recognize the breakpoint time as an adjustable parameter, because our eye is naturally drawn to this value. However, the breakpoint of $t_c = 1980$ in this data is really the result of finding the optimum t_c (i.e. lowest MSE or equivalently BIC) for the breakpoint model across all possible values of t_c.

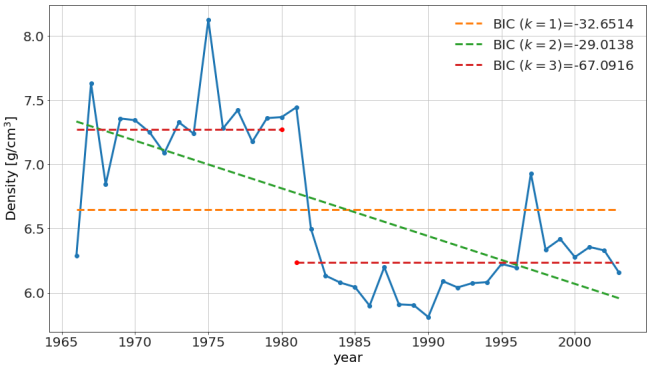

Figure 6. Density of pennies vs year with models of different complexity: single constant ($k = 1$), linear ($k = 2$), and single-breakpoint constant (k=3). Each case improves both MSE and BIC.

Matters get interesting when, once you see a single breakpoint around 1980, one might start seeing other transitions. Perhaps there is one around 1996? Perhaps others? If we introduce a double-breakpoint model, with breakpoints at $t_{c1} = 1980$ and $t_{c2} = 1996$, we can see (Figure 7) that the extra complexity is unwarranted even though our eye may think there is a transition there. A discussion of the idea that humans see patterns in random noise is good follow-up to this exercise. As a real world example using climate change, there are several examples of fitting the global temperature trends over the past 150 years using single linear models and piecewise-linear models with different numbers of breakpoints[14]. The criticism of the piecewise models is not that there are no transitions, but that the extra complexity introduced in the presumed transitions are not warranted without a correspondingly large decrease in MSE.

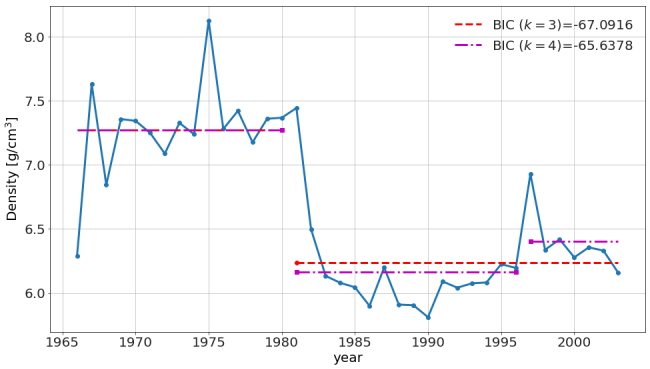

Figure 7. Density of pennies vs year with two piecewise models of different complexity: single-breakpoint constant ($k = 3$) and double-breakpoint constant ($k = 4$). Although the double-breakpoint model improves the fit (i.e. MSE is lower) it is not as favorable as the single-breakpoint model (i.e. BIC is higher) due to its extra complexity.

Further applications

One can generate a rich variety of examples once one introduces the notion of model comparison in introduction physics classes. The simple pendulum lab, for example, is typically presented to students to verify the relationship between the period of the pendulum and the length of the pendulum for small oscillations,

Small Angle

$$T = \sqrt{\frac{L}{g}} \tag{8}$$

Here, we can easily extend our comparison to the messier, large angle solution or one of its approximations[15,16],

Large Angle

$$T = \sqrt{\frac{L}{g}}\left(1 + \frac{1}{16}\theta_o^2 + \frac{11}{3072}\theta_o^4 + \cdots\right) \tag{9}$$

One could potentially introduce forms including the decay of the oscillation or any other interesting variations. Some of the questions which the students could answer using the model comparison techniques include,

- When is it appropriate to use the *Small Angle* model over the *Large Angle* model?
- How carefully need one measure the period and angle to distinguish three terms in the *Large Angle* model over two terms?
- If you don't, or somehow can't, measure the initial angle θ_o can you use it as a free parameter in the model comparison?

One can even introduce the model pendulum to explain walking speeds. For example, each leg can be seen as a pendulum during the swing phase of a normal walking step. How complex of a model is needed to understand the speed of walking quantitatively? Is a simple pendulum enough (i.e. we approximate the mass in the center of the leg)? Is a uniform solid pendulum a justified complexity? Is a two-part solid pendulum a further justified complexity?

This *process* of asking questions is more aligned with the way that scientists work in practice, is intrinsically more interesting to students, and provides a uniform framework for approaching *all* physics problems. Following this approach students are presented with some of the central points in all scientific endeavors:

1. All models are wrong
2. Some models are *good enough* and that's what we work with (for now)
3. The generation of models is a fundamentally creative human enterprise

Discussion and Conclusions

I have used this approach in a freshman physics lab, where to some students I have had to describe what mean-squared error is and when the last time they saw logarithms would have been Algebra II in high school. I'll admit, the calculation might appear a bit like a "blackbox" to these particular students, but in my experience it doesn't detract much from the application of the process, if one describes the process in stages:

1. MSE is a measure of the difference from the model to the data - larger MSE = larger difference
2. one term in the BIC is directly related to MSE (i.e. higher MSE = higher BIC) so it can be used in comparing differences between two models and the data in the same way
3. the other term in the BIC is a penalty for a model having adjustable parameters and thus being more complex

Neither the derivation of BIC, nor any advanced statistics, nor the detailed properties of logarithms are required to understand these stages and thus use the approach.

I have also used this approach in advanced freshman and sophomore physics labs. At first, the students find it a bit unusual — they haven't seen this approach even in their math classes. However, after doing it several times across the semester they become much more comfortable with it, especially with the idea of the complexity of different models.

Model comparison is the bread-and-butter of working scientists, yet it isn't stressed in introductory physics labs. Here we have presented some straight-forward examples,

extending traditional physics lab exercises to include the process of model comparison. I believe this approach makes these lab exercises both more interesting for the students and a better reflection of the core processes of science, without unduly complicating the analysis. The examples presented here can be modified to be as simple or as challenging for the needs of any particular class. The essential idea of this approach can be applied to nearly any lab activity, and generate an entire family of new and interesting student experiences. The original intention of this approach was to restrict it to only mathematical models — the BIC is a mathematical equation after all. However, the *idea* is much broader — those models with more adjustable parameters or pieces need to justify those parameters by fitting the data *even better* than models without those parameters. Philosophers have been using Occam's Razor for centuries but the approach here brings it into the introductory physics laboratory.

References

1 Natasha G. Holmes and Carl E. Wieman. "Introductory physics labs: We can do better." *Physics Today*, **71**(1):38–45, (Jan 2018).

2 NG Holmes and DA Bonn. "Quantitative comparisons to promote inquiry in the introductory physics lab." *The Physics Teacher*, **53**(6):352–355, (2015).

3 Patrik Vogt and Jochen Kuhn. "Analyzing free fall with a smartphone acceleration sensor." *The Physics Teacher*, **50**(3):182–183, (2012).

4 Paul Gluck. "Air resistance on falling balls and balloons." *The Physics Teacher*, **41**(3):178–180, (2003).

5 Norman F Derby, Robert G Fuller, and Phil W Gronseth. "The ubiquitous coffee filter." *The Physics Teacher*, **35**(3):168–171, (1997).

6 Derrick E Boucher. "A perspective on motion sensors and free-fall." *American Journal of Physics*, **83**(11):948–951, (2015).

7 Loo Kang Wee, Kim Kia Tan, Tze Kwang Leong, and Ching Tan. "Using tracker to understand 'toss up'and free fall motion: a case study." *Physics Education*, **50**(4):436, (2015).

8 Angus M Brown. "A step-by-step guide to non-linear regression analysis of experimental data using a Microsoft excel spreadsheet." *Computer Methods and Programs in Biomedicine*, 65(3): 191–200, (2001).

9 Matthew Newville, Till Stensitzki, Daniel B Allen, Michal Rawlik, Antonino Ingargiola, and Andrew Nelson. "Lmfit: non-linear least-square minimization and curve-fitting for python." *Astrophysics Source Code Library*, (2016).

10 William Harper. "Newton's methodology and Mercury's perihelion before and after Einstein." *Philosophy of Science*, **74**(5):932–942, (2007).

11 Richard A Young. “Improving the data analysis for falling coffee filters.” *The Physics Teacher*, **39**(7):398–400, (2001).

12 Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. “How to grow a mind: Statistics, structure, and abstraction.” *Science*, **331**(6022):1279–1285, (2011).

13 David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. “Bayesian measures of model complexity and fit.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4):583–639, (2002).

14 The global climate continues to warm rapidly, URL https://www.skepticalscience.com/ipcc-global-warming-pause.htm. (Retrieved April 2019).

15 Richard B Kidd and Stuart L Fogg. “A simple formula for the large-angle pendulum period.” *The Physics Teacher*, **40**(2):81–83, (2002).

16 Rajesh R Parwani. “An approximate expression for the large angle period of a simple pendulum.” *European Journal of Physics*, **25**(1):37, (2003).