Quantitative Understanding in Biology Short Course
Session 7
Model Fitting and F-Tests

Jason Banfelder
Luce Skrabanek

April 29th, 2025

1 Fitting a Michaelis-Menten Model to Myoglobin Binding Data

A classic mathematical model for enzyme kinetics is the Michaelis-Menten equation:

$$V = \frac{V_{max}[S]}{K_m + [S]} \tag{1}$$

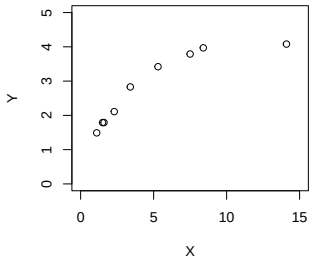
Given data of V versus $[S]$ for a particular enzyme and substrate, we can determine the Michaelis-Menten parameters V_{max} and K_m using a regression procedure.

While the Michaelis-Menten equation is formulated in terms of a reaction rate as a function of enzyme concentration, its mathematical form is often used (sometimes empirically) to describe any phenomenon that saturates.

Consider, for example, the following data for the association of myoglobin and oxygen.

P_{O_2} (torr)	1.1	1.5	1.6	2.3	3.4	5.3	7.5	8.4	14.1
$[O_2]$ (mL/dL)	1.49	1.79	1.79	2.11	2.83	3.42	3.79	3.97	4.08

These data are available in the supplementary file `myoglobin.txt`, and are plotted here.

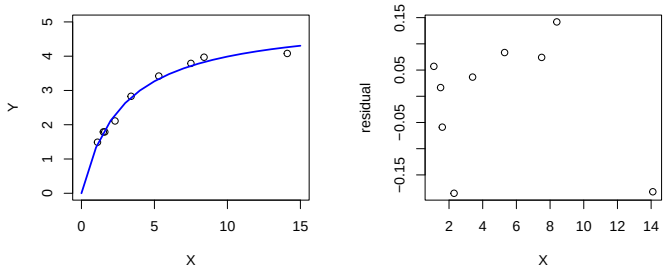


We can fit these data to this equation by asking Prism to perform a non-linear regression, and choosing "One site – Specific binding" from the "Binding - Saturation" section of equations. Prism will fit to the equation:

$$Y = \frac{B_{max} \cdot X}{K_d + X} \tag{2}$$

If you choose "Michaelis-Menten" from the "Enzyme Kinetics - Substrate vs. Velocity" section of equations, you'll perform an equivalent fit; just the terminology is different.

The best fit value of B_{max} is 5.117 (95% CI: 4.747-5.535), and that of K_d is 2.828 (95% CI: 2.296-3.476). A plot of the best fit model looks like it reasonably predicts the data, and a plot of the residuals doesn't exhibit any obvious patterns that would cause us to reject this model out of hand.



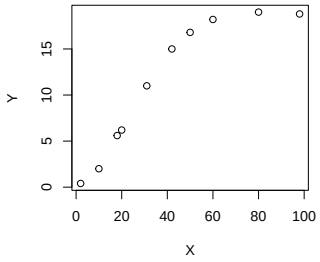
So far, everything looks good.

2 Fitting a Michaelis-Menten Model to Hemoglobin Binding Data

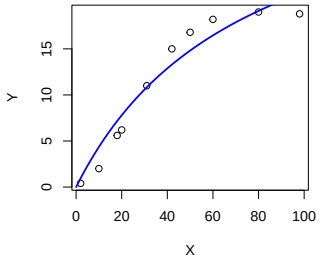
We will now repeat this exercise for similar data taken for hemoglobin. The experimental observations are:

P_{O_2} (torr)	2	10	18	20	31	42	50	60	80	98
$[O_2]$ (mL/dL)	0.4	2.0	5.6	6.2	11.0	15.0	16.8	18.2	19.0	18.8

These data, plotted below, are available in the supplementary file `hemoglobin.txt`.



The plot of the raw data already suggests a sigmoidal shape that may not be consistent with our model. However, this could just be noise in the model, so we proceed objectively with a similar fit as before.



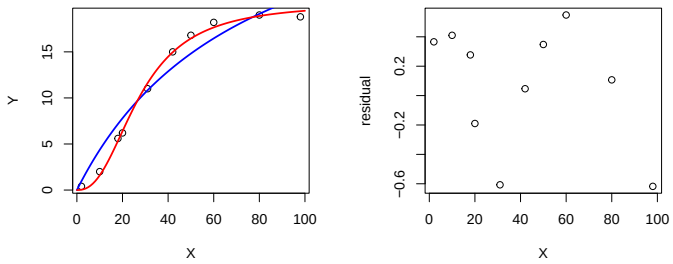
Here we see that the model curve does not fit the data too well. While we could proceed with our quality control plots, for current purposes we'll stop here and reconsider the model. It turns out that if you take into account that hemoglobin is a multimeric protein, and assume that affinity for binding at different sites is not independent, you get a more elaborate form of the Michaelis-Menten relationship, called the Hill model:

$$Y = \frac{B_{max} \cdot X^h}{K_d^h + X^h} \tag{3}$$

The exponent, h, is called the Hill exponent, and is an indication of the degree of cooperativity the system exhibits. If $h > 1$, the system is said to exhibit positive cooperativity; if $h < 1$, the system exhibits negative cooperativity.

We also see that when $h = 1$, the model reduces to the Michaelis-Menten model. In other words, the Michaelis-Menten model is a special case of the Hill model. This relationship between the models is important, and has a specific term: the models are said to be 'nested'.

If we ask Prism to fit the hemoglobin data to a Hill model (the saturation model named "Specific binding with Hill slope" choice), we get a best fit curve (in red), and a residual plot as shown below.



When performing non-linear fits like this, you should be aware that Prism uses an iterative, trial-and-error procedure to determine the model parameters that minimize the SSQ of the residuals. If you explore the tabs on the setup, you'll note that there is a default of a maximum number of 1,000 iterations, and that you can adjust the strictness of the convergence criteria (see the Diagnostics tab). Additionally, and more importantly, these iterative numerical procedures require a starting guess of the parameters. If you visit the initial values tab, you'll see what Prism chose. If you run into convergence problems, you may need to take manual control. Getting such optimization to converge is sometimes more art than science, and occasionally it is not possible. Some tips to aid in convergence are:

1. Plot the curve predicted by the model at the initial guess, and adjust the parameters "by hand" to get a decent starting guess.
2. Try fitting the model with one or more of the parameters fixed. Then use the optimized values for the remaining parameters as starting points for a full optimization.

Prism has built in rules for initial guesses of the standard equations, so you may not run into these problems too often if you stick to the forms of equation that Prism has baked in. If you introduce your own algebraic equation, you'll need to address this starting point issue as well.

Also note that the statistical models that are used to estimate confidence intervals are designed to work with real data that contain some noise. If the data that you fit to were

generated from a function and all of the residuals were zero, Prism won't compute CIs. This may seem counter-intuitive, as you would expect most optimizations to perform well when the error is zero, but the CIs need variance in the data relative to the model, and can't proceed if there is none. This is not likely to happen in the real world, but be aware that if you use generated data, you'll need to add some degree of noise.

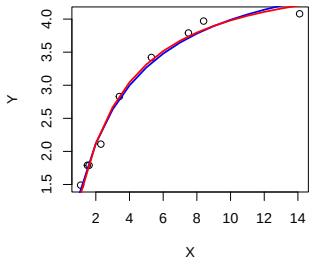
The best fit value of B_{max} for the Hill model is 20.300 (95% CI: 19.145-21.788), that of K_d is 27.529 (95% CI: 25.383-30.265), and that of h is 2.435 (95% CI: 2.046-2.879). We should note here that the physiologically accepted value of the Hill constant for hemoglobin is between 2.5 and 3.0.

3 A Return to Myoglobin

Given the success of our Hill model, and given that regular Michaelis-Menten kinetics are a special case of the Hill model, one might wonder why we don't just always use the Hill model. After all, if the system does not demonstrate cooperativity, the regression will tell us by reporting a Hill exponent of unity.

Let's try this approach with our myoglobin data by asking Prism to fit the myoglobin data to the Hill model.

Our initial guess is informed by our previous run:



Here we see that the two parameter Michaelis-Menten model fits the data almost exactly as well as the three-parameter Hill model. This can be quantified by looking at the sum of the squares of the residuals.

$$SSQ_{MM} = 0.108$$
$$SSQ_{Hill} = 0.094$$

Inspecting the CIs for the parameters, and comparing them to the results from the Michaelis-Menten mode, is informative as well:

parameter	Michaelis-Menten	Hill
B_{max}	5.117 (95% CI: 4.747-5.535)	4.777 (95% CI: 4.197-6.157)
K_d	2.828 (95% CI: 2.296-3.476)	2.439 (95% CI: 1.886-4.543)
h	-	1.140 (95% CI: 0.790-1.532)

The first thing that you should notice is that the CI for the Hill exponent is quite wide for this model; we could have reasonably significant positive or negative cooperativity. Comparing the CIs for the other parameters with those from the two-parameter model shows that the uncertainty in the three-parameter model is significantly larger. This alone is a reason for rejecting the three parameter model if we can; it will reduce the uncertainty in the parameter CIs. However, a more compelling argument is that of maximum parsimony, or Occam's razor. Given a choice between two models, if we don't have good evidence to support the more complex model (such as the cooperative Hill model), we should prefer the simpler one.

Another way of looking at the problem is to keep in mind here that we only have nine data points for myoglobin. A two parameter model therefore has seven degrees of freedom, while a three parameter model has six. This is not an insignificant change, and there is a real possibility that the Hill model represents an over-fit of the limited available data.

While choosing between models is often a judgment call that should integrate all available scientific information, there are tools that help us in the decision. We will consider two, the F-test and an interesting, non-statistical approach called AIC.

4 The F-test for Model Comparison

Using the F-test to compare two models follows the classical framework for statistical testing. You state a null hypothesis, assume it is true, and then compute a p-value that gives the probability of observing your data (or something more extreme) under that assumption. If the probability is low enough, you reject the null hypothesis.

We know that the more complex model will always fit better because we have more parameters. If we start with the more complex model and remove a parameter, we expect that the SSQ will go up. In fact, we can quantify this expected change in SSQ: if the simpler model is the correct one, then we expect that the relative change in the SSQ should be about equal to the relative change in the degrees of freedom. In other words, we expect this ratio to be more or less equal to one if the complex model is just fitting to noise a little better:

$$F = \frac{\left(\frac{\mathrm{SSQ}_{simple} - \mathrm{SSQ}_{complex}}{\mathrm{SSQ}_{complex}}\right)}{\left(\frac{\mathrm{DF}_{simple} - \mathrm{DF}_{complex}}{\mathrm{DF}_{complex}}\right)} \tag{4}$$

However, if the complex model really does represent the system better, we expect F to be much larger.

The p-value computed by the F-test answers the question: assuming that the simpler model is the correct one, what is the probability that we see a change in SSQ at least large as the one we observed when we simplify the complex model. If this p-value is low, then we reject the null hypothesis and accept the more complex model.

From a purely statistical point of view, if the p-value is above our pre-determined cutoff, we could not make any conclusion. However, since either model is considered a viable candidate for explaining our data, we apply the principle of maximum parsimony, and accept the less complex model.

Performing an F-test is Prism is straightforward. Begin by choosing a non-linear model to fit to – you can start with either the more complex or the simpler model. Then, on the compare tab, choose the option "For each data set, which of two equations (models) fits best?" Make sure you choose the "Extra sum-of-squares F test" as the comparison method, and then select the alternative model (equation).

If you compare the Michaelis-Menten and Hill models using the provided myoglobin data, you should get an F-value of 0.906. The F-test yields a p-value of 0.378. This means that the observed F-value is not surprising, and consistent with an incremental improvement in SSQ due to better fitting of random variation; the principle of maximum parsimony then dictates that we accept the simpler model (given this data).

On the other hand, if you compare these models using the hemoglobin data, you should get an F-value of 104.778. The F-test yields a p-value of 1.83e-05. The interpretation is that this is a very surprising improvement in SSQ if it is due to just a better fit to random variation. Thus we conclude that the added complexity of the Hill model is worth it, and adopt that model.

It is very, very important to know that the F-test is only applicable for nested models, and only when you are fitting them to the exact same data. You can't compare unrelated models with it, and you can't compare a transformed and non-transformed model with it (the data are not the same).

5 Using AIC to Compare Models

The derivation for Akaike's Information Criteria (AIC) is well beyond the scope of this course. It involves information theory, maximum likelihood theory, and entropy. We can get a rough feel for what the method is doing by looking at the resultant formula.

$$\mathrm{AIC} = n \cdot \ln\left(\frac{\mathrm{SSQ}}{n}\right) + 2\,(P+1) \tag{5}$$

Here, n is the number of observations, and P is the number of model parameters in the regression. We observe that the larger the SSQ, the higher the AIC will be. Also, the AIC increases as we add parameters to the model. Therefore, we can conclude that lower AICs are better. We can imagine that if we add a model parameter (increment P), the SSQ will go down. If the parameter was worth adding, the increase in the second term will be more than offset by a decrease in the first term.

By itself, the AIC is meaningless. The astute observer will realize that the SSQ has units of measure, and therefore there is an implicit standardization. We can therefore make the numerical value of the AIC whatever we like by altering the units of SSQ (or the standard value).

This ostensible shortcoming is overcome, however, when we look at the difference between the AICs of two models:

$$\Delta\mathrm{AIC} = n \cdot \ln\left(\frac{\mathrm{SSQ}_B}{\mathrm{SSQ}_A}\right) + 2\,(P_B - P_A) \tag{6}$$

The problem of the units of SSQ goes away. In practice, however, we can compute our AICs using consistent units, and select the model with the lower value.

A correction to AIC is necessary when n is not much greater than P. The corrected AIC equation is:

$$\mathrm{AIC}_C = \mathrm{AIC} + 2\frac{(P+1)\,(P+2)}{n-P} \tag{7}$$

We can use the AIC to compare any two models fitted to the same dataset. The models do not need to be nested; this makes the use of AICs a very powerful technique for comparing unrelated models.

You are much more likely to see F-tests in the literature. Because these tests are based on the classical statistical framework, many people feel more comfortable with them. How-

ever, comparison of AICs can be more powerful, especially when dealing with non-nested models.

It turns out that the difference in AIC (or AIC_C) is related to the probability that one model is correct relative to another. A difference of about 6 corresponds to a 95% chance that the lower scoring model is correct. Therefore, if a more complex model has a lower score than a simpler model, but the difference is less than 6, you may still want to stick with the simpler model, because the evidence in favor of the complex one is not overwhelming. Given two non-nested models (perhaps with the same number of parameters), you might simply choose the one with the lower AIC score, but appreciate that the difference between the models is not 'significant'.

6 Confidence Intervals with the F-Test [Optional; not in Prism]

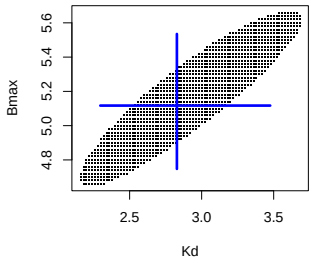
An interesting application of the idea behind the F-test is that it can be used as an alternative means of estimating the uncertainty in model parameters. The basic idea is to use what we learned about F-tests to compare a model with zero parameters to our best fit model.

Consider the Michaelis-Menten myoglobin model. The SSQ for this model is 0.108325. We had $n = 9$ data points and $P = 2$ parameters, so we had $\texttt{DF} = 7$ degrees of freedom. Take this as model A.

Now consider a hypothetical model with no floating parameters (because we've arbitrarily chosen values for B_{max} and K_d); we would have $n = 9$, $P = 0$, and $\texttt{DF} = 9$. Take this as model B.

We could compare these models in the usual way – sort of. In Prism, compare the model to itself, and set constraints on model B! This would tell us if the combination of parameters for B_{max} and K_d that we chose is plausible, given the data. We could (although not easily in Prism) do this repeatedly for many different combinations of B_{max} and K_d. If we record which combinations are plausible and which are not, we can get a map of the plausible combinations.

For the myoglobin data here, such a map for the two parameter Michaelis-Menten model looks like this (we've overlaid the 95% CIs of the B_{max} and K_d parameters as reported by Prism):



If you think about it, this plot makes sense. The combination of parameters B_{max}/K_d is the initial slope of the Michaelis-Menten curve. Based on the data we have (look back at the original plot), we have a pretty good idea of what that should be. However, determining the maximum (plateau) of the curve is quite difficult from our data (this is notoriously difficult experimentally; you need to go to very high values of X). To get K_d, which is the half-maximal concentration, we need a decent idea of B_{max}. So while there is significant uncertainty in both K_d and B_{max}, we do expect that they have a relationship (i.e., they are not independent).

7 Homework

Prism has a more complete model for binding data analysis that includes terms for background and non-specific binding (the "One site – total" model). Compare this model to the "One site – Specific binding" model for the myoglobin data used here. Which model do you think best represents the data? Explain the method you used and your reasoning for doing so.

