

STAT 700, Midterm Exam Part II, Fall 2011

Due 5:30PM, Wednesday October 19

This assignment is a take-home midterm exam. You **may not** collaborate with *any* other person (whether in the class or not). You **may** use any reading material (class notes, books, etc.) you wish. Professor Bailey will answer questions.

Please follow the lab report directions for Homework, i.e. include commands and output you used to answer the questions.

1 Problem: 50 total points.

Problem 1. Return to enzyme kinetics. Most analyses of enzyme kinetics fit the initial velocity of the enzyme reaction as a function of the substrate concentration. In the Non-linear Regression Lab we fit a nonlinear model to data from a biochemical experiment where the initial rate or velocity of a reaction was calculated for different concentrations of the substrate are given in the data frame `Puromycin`. We will use the “treated” dataset.

It is clear from inspection of these data that velocity increases with concentration, seeming to “level off” at high concentration levels. However, the relationship may begin to “drop off” at very high levels of concentration, rather than “leveling off,” perhaps reflecting a breakdown in the process.

A standard model postulated to describe the mean relationship is the Michaelis-Menton model

$$f(x; \theta) = \frac{\theta_1 x}{x + \theta_2} \quad (1)$$

where  $x$  is concentration. A model that allows for the possibility of the “drop off”, the quadratic Michaelis-Menton model is

$$f(x; \theta) = \frac{\theta_1 x}{x + \theta_2 + \theta_3 x^2}, \quad (2)$$

where the additional term in the denominator allows for a “downturn” in mean response. Another possible model allowing for “shifting” is

$$f(x; \theta) = \theta_3 + \frac{\theta_1 x}{x + \theta_2}. \quad (3)$$

1. Fit model (1), model (2), and model (3) with nonlinear least squares. You can use the variable names in the data frame. Superimpose the fitted values with three different line types over the scatter plot of the data. Make a legend for the plot. Using your **myplotnls** from Lab, make a summary plot of the fits. For each fit, you should include model summaries. How well do the models fit the data?

2. Which model best describes the relationship between velocity and concentration of an enzymatic reaction? Parts (a)-(d) will help answer this question.

(a) Calculate the Akaike's Information Criterion: AIC model selection criterion (use the function `AIC`). What are the values of AIC for models (1)-(3). Which model is the "best" based on AIC?

(b) Calculate the Bayesian Information Criterion: BIC model selection criterion (use the function `AIC`). What are the values of BIC for models (1)-(3). Which model is the "best" based on BIC?

(c) Another model selection criterion is based on the "leave-one-out" cross-validation function. Leave out data point  $(x_i, y_i)$ , one at a time, re-fit based on the remaining  $n - 1$  data points, and then predict the  $y_i$  value at the deleted  $x_i$  value. Do this for each data point. Construct the sum of squares

$$CV = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}^{-i}(x_i))^2$$

where  $\hat{f}^{-i}(x_i)$  indicates the fit at  $x_i$ , computed by leaving out the  $i$ th data point.

What are the values of CV for models (1)-(3). Which model is the "best" based on cross validation? Include your source code or function.

(d) Note that if  $\theta_3 = 0$ , then models (2) and (3) reduce to model (1). In the situation of nested models we would be interested in the simplest model which adequately fits the data. To decide which is the simplest nested model to fit the data, you can assess the extra sum of squares due to the extra parameter involved in going from the partial model (1) to the full model (2) and (3). To help answer this question you should use formal statistical inference (as in linear regression) and test  $H_0 : \theta_3 = 0$  vs.  $H_1 : \theta_3 \neq 0$ . For nonlinear models, as we might expect, the analysis is only approximate because the calculated mean square ratio will not have an exact  $F$  distribution.) Compute the ratios in R using the `anova` function. What is the approximate  $p$ -value for your tests and your conclusions?

3. To compare the nonlinear models to a linear model that is cubic in the  $x$ -variable we consider

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + error. \quad (4)$$

Fit model (4) with least squares using the `lm` function. Hint: There is an `inhibit I` function that will prevent R from simplifying the quadratic and cubic terms. Superimpose the fitted values over the scatter plot of the data. Make a summary plot of the fit and include the linear model summary. How well does the model fit the data?

Repeat 2 (a)-(c) for model (4). How does the linear model compare to the nonlinear models in describing the relationship between velocity and concentration of an enzymatic reaction?