STAT 700, Midterm Exam Part II, Fall 2016

Due 4:00PM, Wednesday October 19

This assignment is an individual 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.

2 Problems: 25 total points.

Problem 1. Physical fitness measurements on 31 men in a fitness program was collected. In addition to age (years) and weight (kg), we are interested in oxygen update rate (ml per kg body weight per minute) (Y), time to run 1.5 miles (minutes) (X_1), heart rate while resting (X_2), heart rate while running (X_3), and maximum heart rate while running (X_4).

The data are available off the class web page:

`http://www.rohan.sdsu.edu/~babailey/stat700/fitness.dat`

We will consider the model,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}' = (\beta_0\ \beta_1\ \beta_2\ \beta_3\ \beta_4)'$. Assume that the ε_i are independent $N(0, \sigma^2)$ random variables.

We want to test the null hypothesis that $\beta_2 = \beta_4 = 0$. Suppose that prior information suggests that the intercept β_0 for this group of men of this age and weight should be 90. We will test the composite null hypothesis that $\beta_2 = \beta_4 = 0$ and $\beta_0 = 90$. From class, the general linear hypothesis is

$$\mathrm{H}_0 : \boldsymbol{K'\beta} - \boldsymbol{m} = \boldsymbol{0}$$
$$\mathrm{H}_1 : \boldsymbol{K'\beta} - \boldsymbol{m} \neq \boldsymbol{0}.$$

(a) Write down in matrix notation indicating all the individual elements of the composite **null** hypothesis.

(b) In class we derived the sum of squares for this hypothesis and it is

$$Q = (\boldsymbol{K'\hat{\beta}} - \boldsymbol{m})' \left[\boldsymbol{K'(X'X)^{-1}K}\right]^{-1} (\boldsymbol{K'\hat{\beta}} - \boldsymbol{m}).$$

Let $\hat{\boldsymbol{\beta}}$ be the least-squares estimate. Let the rank of $\boldsymbol{K}$ be denoted by $rank(\boldsymbol{K})$ and the residual mean square be s^2. Then, the F-test for the null hypothesis is

$$F = \frac{Q/rank(\boldsymbol{K})}{s^2}$$

with numerator degrees of freedom equal to the $rank(\boldsymbol{K})$ and denominator degrees of freedom equal to the degrees of freedom in s^2. Use R matrix operations to compute the above F. What is your conclusion at the $\alpha = 0.05$ level?

Problem 2. 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 Nonlinear 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 "untreated" dataset (not the "treated" dataset).

It is clear from inspection of these data that velocity increases with concentration, seeming to "level off" at high concentration levels. A a standard model postulated to describe the mean relationship is the Michaelis-Menton model

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

where x is concentration. Another possible model that allows for "shifting" is

$$f(x;\theta) = \theta_3 + \frac{\theta_1 x}{x + \theta_2}. \tag{2}$$

(a) Fit model (1) and model (2) with nonlinear least squares. Superimpose the fitted values with a smooth line over the scatter plot of the data. Make a legend for the plot. Using the **myplotnls** from Lab 3, make a 2×2 summary plot of the fits. You should be able to call your plot function with ONLY the fitted object and use no objects defined outside the function. You should include model summaries. How well do the models fit the data? Explain.

(b) Note that if $\theta_3 = 0$, then model (2) reduces to model (1). We can bootstrap the data to construct a percentile interval for θ_3. Use the bootstrap function from the Nonparametric Bootstrap Lab. You will have to make your own function for the theta argument.

(i) Using 500 bootstrap replicates, bootstrap pairs to obtain the bootstrap estimates of θ_3. Make a histogram of the 500 bootstrap estimated coefficients. Use the R `set.seed` function to set the seed and use `set.seed(1)` before calling the bootstrap function. Construct a 95% percentile interval for θ_3. What is your conclusion about null the hypothesis that θ_3=0?

(ii) Compare the length your intervals in (i) with the 95% CI constructed with normality assumptions using your fit from `nls` and no bootstrapping.

(c) Which model (1) or (2) best describes the kinetics? To answer this question we will consider two different model selection criteria.

(i) Calculate the Akaike's Information Criterion: For the AIC model selection criterion, make sure to use the function AIC **not** extractAIC. What are the values of AIC for models (1)-(2)? Which model is the "best" based on AIC?

(ii) Calculate the Bayesian Information Criterion: For the BIC model selection criterion, make sure to use the function AIC **not** extractAIC. What are the values of BIC for models (1)-(2)? Which model is the "best" based on BIC?