# Homework 6: How the *Hyracotherium* Got Its Mass

36-402, Advanced Data Analysis

Due at 11:59 pm on Monday, 25 February 2013

INSTRUCTIONS: Submit a single PDF including all your written responses and all your figures; include your Andrew ID in the file name. Put all R code in a separate plain text file, also named with your Andrew ID. When a question asks you to write code, indicate where the code is in your R file, and briefly describe, in your own words, how your code works. As always, raw computer output, missing or inadequate explanations, spurious precision, and Word files are all unacceptable.

AGENDA: Using nonparametric smoothing to check parametric models; more practice with simple simulations and function-writing.

We continue to work with the fossil data set from homework 4. As mentioned there, some paleontologists have suggested that the right curve relating change in log mass to ancestral log mass should be piece-wise linear and homoskedastic: a downward-sloping line for small ancestral log masses, flat for larger ancestral masses, and constant conditional variance:

$$\begin{aligned} Y &= \begin{cases} a + bx + \epsilon & \text{if } x \leq d \\ c + \epsilon & \text{if } x \geq d \end{cases} \\ \mathbf{E}[\epsilon|x] &= 0 \\ \mathrm{Var}[\epsilon|x] &= \sigma^2 \end{aligned}$$

In the last problem set, you fit that model; in this one, you will see whether the data support non-linear corrections.

You will first need to load the data set from homework 3, and add the column of change in log mass to the data frame. (See solutions that problem set, if necessary.)

The `mgcv` package is recommended for the additive model in Problem 5. Earlier problems call for spline smoothing, and can be done with either the `smooth.spline` function or with the `gam` function. If you want to use a different smoother, ask.

1. (10) *Plotting the Parametric Model*

   (a) (5) Make a scatter-plot showing the change in log mass as a function of the log ancestral mass.

   (b) (5) Add the estimated piecewise linear model from homework 4. You may refer to the solutions for code and parameter estimates, but must explain, in your own words, any code you borrow from there.

2. (25) *Residual inspections*

   (a) (5) Calculate the residuals of the estimated piecewise linear model and plot them against the log ancestral mass. Describe any patterns to the plot in words; you should address whether the model systematically over- or under- predicts in certain ranges of ancestral mass, but there may be other important features.

   (b) (5) The column `first_appear_Mya` lists how many millions of years ago each species first appeared. Plot the residuals against this variable; describe any patterns.

   (c) (5) Plot the squared residuals against the log ancestral mass. Add a smoothing spline. Explain whether the scatter-plot and the spline show evidence of heteroskedasticity.

   (d) (5) Plot the squared residuals against date of first appearance and add a smoothing spline. Explain whether the scatter-plot and the spline show evidence of heteroskedasticity.

   (e) (5) Plot the histogram of the residuals (not the squared residuals). Are they Gaussian? Should they be, under the model?

3. (10) *A nonparametric alternative*

   (a) (7) Fit a spline regression of the change in log mass against log ancestral mass. Plot this spline on the same graph as the data and the estimated piece-wise linear model. Compare, in words, the shape of the spline to that of the parametric model.

   (b) (3) Find the in-sample root-mean-square error of both the parametric model and the smoothing spline. Which fits better?

4. (20) *Testing parametric forms*

   (a) (3) Write a function to fit the smoothing spline to a data set. Check that it works by making sure it gives the right answer on the original data.

   (b) (2) Write a function to calculate the MSE of a fitted smoothing spline. Check that it works by making sure it gives the right answer on the original data.

(c) (5) Write a function to take in a data set and return the difference in MSEs between the parametric model and the smoothing spline. Check that it works by making sure it gives the right answer on the original data.

(d) (5) Write a function to simulate from the estimated piecewise-linear model by resampling the residuals. You can borrow from the solutions to homework 4, but must explain, in your own words, how that code works. How can you check that the simulation works?

(e) (5) Combine your functions to draw 1000 samples from the distribution of this test statistic, under the null hypothesis that the parametric model is right. What is the $p$-value of this test of the null hypothesis?

5. (25) *Additional Variables* The piecewise linear model implicitly assumes that the relationship between ancestral mass and change in mass is the same at all times. An alternative is that this relationship has itself evolved.

(a) (5) Estimate an additive model which regresses the change in log mass against the log ancestral mass and the date of first appearance. Plot the two partial response functions, and describe, in words, the shape of the curves. Compare the shape of the partial response function for log ancestral mass to the spline curve from Problem 3a.

(b) (4) Does the estimated additive model support or undermine the idea that the relationship between ancestral mass and descendant mass is invariant over time? Explain.

(c) (1) What is the in-sample root-mean-square error of the additive model?

(d) (10) Explain what you would have to change from your code in Problem 4 to test the piecewise-linear model against the additive model, and what pieces of code could stay the same.

(e) (5) Write the new code called for by Problem 5d and run the test. What is the $p$-value?

6. (10) Is the piecewise-linear, homoskedastic parametric model an acceptable representation of the data? Justify your answer by referring to your work above.