Homework 4

NAME: Your Name
DUE DATE: April 19th, 11:59 pm

Problem 1 (ISL 7.1) [5 points]

It was mentioned in ISL chapter 7 that a cubic regression spline with one knot at ξ can be obtained using a basis of the form x, x^2, x^3, $(x - \xi)_+^3$, where $(x - \xi)_+^3 = (x - \xi)^3$ if $x > \xi$ and equals 0 otherwise. We will now show that a function of the form

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$

is indeed a cubic regression spline.

(a) Find a cubic polynomial
$$f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$$
such that $f(x) = f_1(x)$ for all $x \leq \xi$. Express a_1, b_1, c_1, and d_1 in terms of β_0, β_1, β_2, β_3, and β_4.

(b) Find a cubic polynomial
$$f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$$
such that $f(x) = f_2(x)$ for all $x > \xi$. Express a_2, b_2, c_2, d_2 in terms of β_0, β_1, β_2, β_3, and β_4.

We have now established that f(x) is a piecewise polynomial.

(c) Show that $f_1(\xi) = f_2(\xi)$. That is, $f(x)$ is continuous at ξ.

(d) Show that $f_1'(\xi) = f_2'(\xi)$. That is, $f'(x)$ is continuous at ξ.

(e) Show that $f_1''(\xi) = f_2''(\xi)$. That is, $f''(x)$ is continuous at ξ.

Therefore, $f(x)$ is indeed a cubic spline.

Problem 2 (ISL 7.10) [10 points]

This question relates to the **College** data set.

(a) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

(b) Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

(c) Evaluate the model obtained on the test set, and explain the results obtained.

(d) For which variables, if any, is there evidence of a non-linear relationship with the response?

Problem 3 (Based on GH 12.1) [3 points]

Using data of your own with more than two levels (possibly imaginary or based on previous data sets seen in the class), write a multilevel model in the five ways discussed in Section 12.5 and give an interpretation of the model.

Problem 4 (GH 12.9) [7 points]

In this exercise, you will explore the relationship between the number of observations and number of groups on the performance of a multilevel model.

(a) Take a simple random sample of one-fifth of the radon data seen in class. Fit the varying-intercept model with floor as an individual-level predictor and log uranium as a county-level predictor, and compare your inferences to what was obtained by fitting the model to the entire dataset. (Compare inferences for the individual-level and group-level standard deviations, the slopes for floor and log uranium, the average intercept, and the county-level intercepts.)

(b) Repeat step (a) a few times, with a different random sample each time, and summarize how the estimates vary.

(c) Repeat step (a), but this time taking a cluster sample: a random sample of one-fifth of the counties, but then all the houses within each sampled country.

Problem 5 [10 points]

In this exercise, you will fit a multilevel model in which within patient, pain is a function of time and temperature and between patients these relationships may depend on age, race, income, treatment, sex, occupation, working status and use of NSAIDs. Use the lme or lmer function to fit a multilevel model in which

$$Y_i = a_{j[i]} + b_{j[i]} X_i + \epsilon_i$$
$$a_j = g_0 + g_1 z_j + \eta_j^1$$
$$b_j = h_0 + h_1 z_j + \eta_j^2$$

Here, y_i is the pain of observation i at time x_i for individual $j[i]$. Further, z_i is patient level factors (age, sex, race, income, occupation, working status, use of NSAIDs).

Build a multivariable regression model treating a_i and b_i as random effects. Make sure to give a thorough writeup of your results with appropriate graphs and tables. Code should be put at the end of the assignment.