LINEAR REGRESSION MODELS W4315

Take-Home Midterm QUESTIONS

November 23, 2011

Instructor: Frank Wood

You are terribly afraid of pumpkins. Somebody who is not your friend has assembled a ballistic pumpkin shooting device and has been using it to discourage you from moving into a particular neighborhood by launching pumpkins at the only apartment building into which you can move. The pumpkin launching device has a number of peculiar characteristics, not least of which is the fact that it can launch pumpkins using two or more initial configurations. These configurations are characterized by differing initial horizontal velocities alone.

The world, for the purposes of this problem, can be assumed to be frictionless, airless, and so forth. The pumpkins may also be assumed to be points. The pumpkins have mass in that they obey gravity; however, the apartment building and its windows are strong enough that you are in no danger. The pumpkins splat harmlessly against the wall. This, however, is sufficient to terrorize you anyway.

You decide that you are willing to move into any apartment that has a less than a 5% chance of getting hit by a pumpkin. Each apartment in the apartment building is ten length-units high. There are ten apartments arranged vertically on top of one another in the building (it's a high-rise apartment building). All apartments are vacant and you can afford to move into any of them. The apartment building wall facing the pumpkin onslaught is at position $x_T = 0$. The pumpkins are being fired from position $x_0 = -1000$.

Because you are truly terrified of pumpkins, you have spent the time and money to measure pumpkin flight trajectory information for pumpkins that have been fired towards the building in the past. You used a cheap and noisy vertical position measurement system to take 100 measurements from points in between the launch site and the apartment building. Each measurement consists of a single $\{x, y\}$ pair and which kind of horizontal velocity v_x setting was using on launch. Unfortunately, before you could analyze this data, your enemy infected your computer with a virus which added corrupting features to your observations. So, instead of having a design matrix for a regression problem with two non-transformed features, one real-valued, the other categorical, the virus-corrupted design matrix was filled with features that have nothing to do with typical pumpkin flight at all in addition to the measurements you took.

Your general task is to use linear regression techniques to determine which apartments you are willing to move into.

You will accomplish this task by going through the list of questions provided, answering each one to the best of your ability. Your answers should be provided on no more than two concisely and clearly typed or written pages. Supporting material must also be handed in. This material must show evidence that you and you alone got the answers you provide. You must print the matlab code used to obtain your answers and hand it in as supporting material too.

The data for this midterm is available for download from

`http://www.stat.columbia.edu/∼fwood/w4315/Midterm`

There you will find a webpage containing a list of uni's. Clicking on your uni will cause your browser to download a datafile unique to you that you will use for this exam. You are to run your analysis on the dataset in your individual datafile. Your answers will differ from your classmates' answers, so communicating about answers is likely to cause you significant confusion. Copied answers will be readily identifiable.

The datafile is organized in the following manner. You should find it has 100 rows and 101 columns. The first column are the noisy observed pumpkin heights. The remaining columns are features, two of which are legit, the others of which are noise from the virus.

As many of you may not have even a rudimentary physics background, it may be helpful to note that particles following a ballistic trajectory obey basic rules of motion, namely

$$\begin{aligned} y(t) &= y_0 + v_y t + g t^2 \\ x(t) &= x_0 + v_x^{(\ell)} t \end{aligned}$$

where g is "gravity", v_y is the initial y ("vertical") velocity, $v_x^{(\ell)}$ is the initial x ("horizontal") velocity, t is time, y_0 is the initial vertical position of the pumpkin (placement of the pumpkin cannon) and x_0 is the initial horizontal position of the pumpkin cannon. You may assume (and actually could infer) that $y_0 = 0$ and $x_0 = -1000$. The superscript ℓ is a hint that suggests where differing initial velocities might enter this system of equations (it is not an exponentiation operation, rather, it is a selector).

This system of equations should significantly inform the class of regression models one might choose to test between.

To be clear: your submitted answers should fit on two or fewer typed or neatly written pages. We reserve the right to not read beyond the first two pages of answers. Be concise and precise. This is a statistics examination. We are looking for evidence of your ability to perform a multivariate regression analysis. Supporting evidence that you did your own work must be provided.

1. (5 points) What is your uni? What are the numbers in the first three columns of the first row of your data matrix.

2. (20 points) Propose a regression function for the described problem. Label your variables. This regression function should include transformations of variables that are supported by the application domain and categorical variables to encode varying different initial x velocities.

3. (10 points) Write a paragraph describing the variable selection strategy you used.

4. (30 points) Labeling the apartments 1 on the bottom and 10 on the top, into which apartments would you be willing to move? Why?

5. (10 points) How many different configurations of the pumpkin cannon are there? Why?

6. (15 points) Which columns of the data matrix were included in your final regression function?

7. (10 points) Give your final, learned regression function, clearly noting which variable transformations, which variables, and which categorical variables are included. Give 95% confidence intervals for all learned parameters.

8. (0 points) Extra credit: can you derive confidence bounds for g? If so, what are they? If not, why not?