Homework 6

Analyze a grouped data set of your choice. That is, obtain a data set where the observations fall into two well–defined subgroups, and where it is of interest to try to understand, model and/or predict group membership, initially using at least three predictors (as always, you should try to find what you think is the best description of the data you can obtain, subject to the limits of your data and the software, involving however many predictors that may be). The groups should come naturally from the context of some question of interest. The groups can be based on the discretization of some underlying continuous measure if you wish (that is, they can be defined as corresponding to some numerical variable being less than or greater than a specified value), but **only** if that discretization is justifiable from the nature of the problem, and not just done arbitrarily for convenience (that is, there's a good reason to think of modeling this measure in grouped form, rather than in its inherent continuous form). If you do that, you **must** give an explicit justification **from an outside source** for why using the response variable in a categorical form makes sense (and provides additional insight compared to using it in its original numerical form) in the context of the problem. Please note that it is **extremely** unlikely that splitting on the basis of being above or below the mean or median of a numerical response is an appropriate thing to do (that is, having a daily stock return as the response and modeling whether the return is positive or negative is **not** appropriate, since the magnitude of the return matters, not merely the direction); similarly, "beating the industry average" is not appropriate, since it is actual performance that typically matters, not just being above or below average. Your data should have at least 25 "events" in total (that is, if the data are entered as having a 0/1 response variable, there should be at least 25 rows).

Perform a complete and full analysis of the data using the logistic regression model. Discuss what you find. Even if you believe that there are violations of assumptions in your final model(s), you should still discuss the implications of those models, while also (of course) noting the potential limitations of those implications.

I strongly urge you to only use numerical or 0/1 variables as predictors in your model. Logistic regression generalizes to categorical predictors, but we might not cover that situation fully in class. You should understand that if you choose to use categorical predictors in your model you are basically on your own in making sure that your analysis is complete;

I will **not** be able to provide guidance related to that material before it is covered in class. For that reason I also suggest that you do not use only 0/1 variables as your predictors, since if you only use 0/1 variables you are effectively fitting a model with only categorical predictors. Good examples of the kind of data I recommend that you use is the dataset used in the bankruptcy analysis in class.

A reminder: get your data from an original data source. As was stated in the syllabus, you should **not** take your data from a textbook, a journal article that includes (logistic) regression analysis of the data, an online digest of data sets that have been put together for teaching or expository purposes, or a data analysis competition (see Homeworks 2 and 3 for lists of the kinds of sites this refers to). This also includes digests of data sets from textbooks and articles specifically gathered together to be examples of logistic regression (that is, **don't** try to find data by doing a Google search of "logistic regression data" or use data from a web page with the title "Logistic regression data sets").

Late assignments for this homework will **not** be accepted (so don't wait until the last minute to gather your data or start analyzing it; remember, computer malfunctions or unavailability are **not** valid excuses for lateness). I would be very happy to receive the homework earlier than the due date. A hardcopy of the homework can be turned in at the last class session on December 19 or left in my mailbox on the 8th floor of KMC up to the due date and time given below. I will also accept it electronically via e-mail. Submitting the homework early is not only permitted, but encouraged, but submission in any form (hardcopy or electronic) **must occur by the date and time given below**. **No** late submissions will be accepted. If you do turn it in electronically, **be sure that the version of the homework you attach is the correct one**; once the deadline below has passed the assignment I am going to grade is the one you have submitted, even if you later discover that it was not the correct version.

Whether you submit the homework in hardcopy or electronically, please be sure to include a cover page for your homework that has your name on it. It needn't have anything else on it, but it should definitely not contain any of the text of your homework.

Due date and time : December 19 by 4:00 PM