Psychology 613: Multivariate Statistics (Data Analysis III)
Problem Set 4

The following problem set is due at **5pm** on **Thursday, May 26th**, via Canvas.

The document that you turn in should consist of a write-up (including screenshots of relevant output) and selected bits of relevant code ***embedded in the write-up*** (i.e., as part of each problem and not in a separate section or document). Questions that ask you to interpret results of analyses should be phrased in terms that are substantively meaningful. For example, instead of saying something like “the predictor was significant,” you should phrase interpretations in terms of the measured constructs, e.g., “aggression was significantly positively related to alcohol use.”

For the first two questions use an intimate partner violence dataset. It contains the following variables:

CHILDS	Self-reported childish behavior when facing conflict (past month frequency)
PHYSS	Self-reported physical aggression against partner (past month frequency)
AnyPhys	Dichotomized physical aggression (0=No, 1=Yes)
SELFDET	Self-determined (intrinsic) motives to stay in the relationship
ALCFREQ	Frequency of binge drinking each month (5 categories: 1=0 times, 2=2-3 times, 3=3-4 times, 4=6-7 times, 5=8+ times)
BP	The Buss-Perry aggression scale (higher scores mean more aggression)

Logistic regression (***Complete problems 1-2. Problem 3 is optional***)

1. Run a logistic regression predicting whether there is any physical aggression in a relationship or not (*AnyPhys*) based on the Buss-Perry Aggression scale. Interpret your findings in terms of odds. [HINT: Remember to first convert the DV, *AnyPhys*, into a categorical variable using the factor() function.]

2. Re-run the logistic regression from #1 but this time as a hierarchical regression. Enter the Buss-Perry scale in the first step, and then include frequency of binge drinking in the second step. (Hint: remember that binge drinking is also a categorical variable.) Based on the output, report the point at which the probability of any physical aggression reaches 50% (i.e., the threshold) for any of the binge drinking groups (your choice which one!).

3. Write down the equations to predict the probability of any physical aggression for each of the five binge drinking groups. Plot the probability curve for each using R.

CONTINUED ON NEXT PAGE

Machine Learning (***Complete problems 4-5. Problem 6 is optional)***

Suppose you are interested in foraging wild mushrooms. Unfortunately, you don't know anything about which mushrooms are poisonous or edible. You begin recording information about the mushrooms you collect, and your friend agrees to taste them to determine whether they make him sick or not. Based on this information, train a model to be able to determine whether mushrooms no one has tasted yet are poisonous or edible.

The outcome variable is "class," with "e" for edible and "p" for poisonous. The training set is "04. trainMushroom.csv" and the test set is "04. testMushroom.csv." For more information on this dataset see: https://www.kaggle.com/uciml/mushroom-classification. Credit: Max Drascher.

4. Run a logistic regression model on the training data to predict whether each mushroom is edible or poisonous. Use the parameters from this model to predict whether the mushrooms in the test set are poisonous or edible. How accurate is this model?

5. Repeat the previous problem with a linear SVM classifier. First, **set a random seed**. Next, use **k-fold cross-validation** to determine which tuning parameters to use and **plot the results**. What value will you use? Finally, **apply this model to the test data**. How accurate is this model? If you find a poisonous mushroom that you haven't seen yet, if you apply this model, how likely are you to incorrectly think it is edible?

6. Bonus: Repeat the previous problem with a different classifier.