Problem Set 2

MaCCS 201 - Fall 2025

2025-10-14

Tentative Due Date: October 21

Please submit markdown file named [last_name]_[first_name]_ps2.Rmd or a pdf with all code and answers.

##Part A: Which songs become popular?

You are working for Spotify. You are feeling pretty good about yourself about this gig. Your senior VP walks in and says "Hey fancy stats person, can you help us figure out which factors affect the popularity of a song? The predictive analytics shop is doing a pretty good job at predicting, but I want to truly understand what factors drive popularity, instead of some black box ML algorithm output." She Whatsapps you a dataset of 17835 songs she found on Kaggle. The dataset is called `song_data_2025.csv`. It has a number of characteristics of each song in it.

1. Create a nice looking summary statistics table in R, which lists the mean, standard deviation, min and and max for each of the variables.
2. We will focus on the variables `song_popularity, song_duration_ms, acousticness, danceability, energy, tempo`. Make a nice correlation matrix figure. I used `ggpairs`, but knock yourself out. What do you see? Any interesting correlations?
3. Using the `lm` package, regress song_popularity on the other variables from the previous question. Produce some decent looking regression output. Compare the slope estimates from this regression to the correlations between the outcome and the right hand side variable from question 2. Any sign changes? Interpret the coefficient on the variable `danceability` correctly (in actual words!) What does your F-Test tell you for this regression you ran?
4. Plot the residuals from this regression against the predicted values. Are you worried about heteroskedasticity?
5. Formally test for heteroskedasticity using the White test from `skedastic` package (turns out `whitestrap` is absolute garbage). What do you conclude?
6. Using the `felm` package show regression output using the white robust standard errors. Did the coefficients change? Did the standard errors on the coefficients change? Did any of your t-statistics change? Did your F-Statistic Change?
7. Now for the fun part. Packages are great. But let's do the White Test by hand, so we can understand what is happening. The first step is to generate your outcome variable for your White regression. Create a variable called e2, which is the squared residuals. Then create new variables, which are each the square of song_duration_ms, acousticness, danceability, energy, tempo. Then create interactions between these variables. These interactions are all possible pairwise products of these. e.g. acousticness x danceability and energy x danceability. Then run a regression of the squared residuals on the right hand side variables, their squares and the interactions you generated. Can you replicate the test statistic from step 5?

##Part B:Fully Optional and no extra credit for it.

Sometimes students will argue that once you adjust for heteroskedasticity, your standard errors will always be bigger. This is not true. Can you come up with a simulation in R, which results in the standard errors being smaller after adjusting for heteroskedasticity? You can google around a bit. Fun little Monte Carlo exercise!