Bootstrap: Simulating The Cure

Heather Harvey

MISERICORDIA UNIVERSITY

INTRODUCTION

- In STA 486 we had projects where we used what we learned on real life data
- The analysis being done is bootstrap simulation
- Bootstrapping is a statistical procedure where several random samples from one original sample are tested
- This creates a simulation to get results like confidence intervals, standard errors, and hypothesis testing
- It can help make predictions on values that are not included in the data

BACKGROUND

- The dataset was found on Kaggle.com
- It is the entire discography of the band The Cure
- The person who created it, Xavier, created the data because he really likes the band
- He found code that takes everything that a band has on Spotify and creates a file with it
- He used this on The Cure
- The data contains record of speech level, beat, pitch, key, energy, danceability, liveness, loudness, track name, etc.

Previous Analysis

- In a previous class, the data was checked with a general linear model to see if the liveness, danceability, and energy had any effect on a song's popularity
- The results showed that these variables did have an effect on the popularity
- I was curious to see if this test would give different results than the previous linear model

Bootstrap Analysis

- For this specific project, the same data will be run through a bootstrap simulation
- The bootstrap in non-parametric
- It will check for the confidence interval of the parameters
- It will be a 95% confidence interval
- The over all sample size is 223: the amount of songs the data contained
- It will run 1000 times through the simulation
- A matrix also needed to be created so that when the values are outputted during the simulation, they will be entered into the matrix
- After all the simulations are run, the mean of all those means are taken to give a more accurate result

95% Confidence Interval

	Intercept	Danceability	Energy	Liveness
5%	53.232	-27.016	-18.432	-19.829
95%	66.624	-11.773	-6.908	-13.903

Results

- The intercept value means that when there is no danceability, energy, or liveness a songs popularity is rated between 53.23 and 66.62
- Danceability, when energy and liveness are held constantly, are expected to decrease between 27.02 and 11.77 points
- Energy, when danceability and liveness are held constant, is expected to decrease between 18.43 and 6.91 points
- Liveness, when danceability and energy are held constant, is expected to decrease by 18.83 and 13.9 points

Conclusions

- The answers were similar to the first analysis done with the linear model
- This time, they were more accurate and more precise than before
- This is due to the 1000 iterations done on the data, instead of just one
- Overall, bootstrap tend to be more precise and give better results than just a linear regression model