Noah Randolph
21 October 2017

CORRELATION OF HOSPITAL AVERAGE MEASURE SCORES WITH PATIENT SURVEY RESPONSES

Spearman's rank correlation coefficient = 0.491

Based on Spearman's rank correlation coefficient, the hospital measure scores are positively correlated with the patient survey responses. However, the correlation is not strong.

Spearman's rank correlation coefficient was selected for correlation because the hospitals rankings are on ordinal scales. The coefficient was determined by the following operations in Spark-SQL:

1. The hospital ranking by survey responses were ranked, first based on the patient_experience_of_care_domain_score, which is the sum of the hcahps_base_score (max score = 80) and the hcahps_consistency_score (max score = 20), for a total possible score of 100. To reduce the amount of tied rankings, the survey responses were further ranked by the overall_rating_of_hospital_dimension_score, the communication_with_doctors_dimension_score, and the communication_with_nurses_dimension_score. These were viewed as the most important survey rankings, since the doctors and nurses are directly responsible for the patients' well being, while the overall rating captures other general categories broadly, enabling the cleanliness, quietness, and discharge information scores to be left out. The pain management score was intentionally left out, due to the U.S. epidemic of opiate abuse, which may be influenced by the medical system and, therefore, is a point of contention.

2. Next, the hospitals were numbered based on the ranking in the best_hospitals.sql script.

3. With each hospital converted to a rank, Spearman's rank correlation coefficient can be computed similarly to the Pearson correlation coefficient, except the covariance and standard deviations are based on ranks, rather than raw data (https://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient). Although there were several ties in the rankings (prohibiting use of the popular and quick version of Spearman's rank correlation coefficient) there was clear enough of a ranking amongst the more than 2000 pairs of data points to establish the presence of a correlation.

4. Per the long version of the formula for Spearman's rank correlation coefficient, the standard deviations for both the hospital measures ranking and the patient surveys ranking were computed. First the mean ranks were computed, then cross-joined to another table to complete the formula for standard deviation.

5. Similarly, the covariance between the two rankings was determined by inner-joining the two rankings and cross-joining the means to compute the covariance formula.

6. Finally, Spearman's rank correlation coefficient was determined by cross-joining the covariance with the two standard deviations (covariance/(standard deviation A * standard deviation B)).