Tamar Feldman

Biol419, Winter ‘16

Project Proposal

1. My datasets are from the World Bank and the Global Health Observatory data repository of the World Health Organization. I am using datasets on sanitation, population density, healthcare infrastructure and workforce, literacy, and neonatal mortality.
2. I am investigating a series of social, economic, and environmental indicators that potentially affect a country’s health. I will use neonatal mortality as a measure of health. My questions are:
3. Does the variability in the indicators reflect the variability in health across countries?
4. Are there countries whose health is not well explained by the indicators selected by PCA and, if so, why?
5. Can these indicators be used to build a classifier that predicts a country’s health?
6. Split the countries into categories of health (poor, fair, good, excellent) based on neonatal mortality rate
7. Identify the most recent year for which there is the most complete data in every indicator.
8. Perform descriptive analysis on each indicator. If there is an indicator with low variability, exclude it because it will not explain variability in health.
9. Perform PCA on the indicators. Determine if the underlying variability reflects the variability in health (similar to the PCA seed type example). Are there countries that have unusually good or bad health for their performance by indicator?
10. Use the indicators that explain 95% of the variance to do LDA using the classify function. Check the model with cross-validated accuracy.
11. Histogram of neonatal mortality in the chosen year