Investigating indicators of low birthweight

Background

Low birthweight (LBW) is a mostly preventable condition with harmful consequences. Birthweight is considered low if it is less than 2,500 grams, and very low if it is less than 1,500 grams. Although only a small minority of LBW children have developmental problems as severe as mental disabilities or cerebral palsy, "as a group they generally have higher rates of subnormal growth, illnesses, and neurodevelopmental problems."[1] The same authors observe that these problems increase with decreasing birthweight, and long-term follow-up studies have indicated that "the adverse consequences of being born low birth weight were still apparent in adolescence."

Modern hospital care in developed countries gives LBW babies better chances of survival with minimal negative effects, but Stanford Children's Hospital says, "Nearly all low birthweight babies need specialized care in the Neonatal Intensive Care Unit."[2] This, of course, could add significant expense.

LBW is primarily associated with premature birth, according to the hospital, which lists the mother's race, age, health and socioeconomic status as contributing factors, as well as the level of prenatal care. Other sources also include education level. All of these factors could be addressed in order to decrease the rate of LBW, thus decreasing the risk of developmental problems as well as associated expense

Problem

The problem is to determine the relative importance of potential contributing factors to LBW, which would be useful in targeting efforts to decrease it.

Datasets

To obtain a result with the greatest relevance for the US as a whole, data covering the country is needed. This can be found scattered throughout various datasets compiled by government agencies, which give statistics for each US county. The Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer dataset published by the Centers for Disease Control, provides statistics on premature, LBW and very low birthweight along with various demographics and health-related measures.

It does include measures for environmental quality, but these are very sparsely populated and hence not usable. But I think it is very interesting to explore whether this might be a factor. The Environmental Protection Agency's Environmental Quality Index (EQI) dataset titled Eqi_results_2013JULY22 gives measures for overall as well as air, water, land, built and sociodemographic environmental quality for all counties.

Two other datasets, both published by the US Department of Agriculture, further supplement the CHSI information, which gives the number of people living below the poverty line as the only economic measure. The Poverty Estimates database adds median household income as well as poverty percentages for the overall, under 17 and under 5 population segments. The Education database gives percentage measures for various education levels.

[1] Hack M, Klein N, Taylor H. Long-term developmental outcomes of low birth weight infants. *The Future of children.* 1995;5(1):176–96. https://www.ncbi.nlm.nih.gov/pubmed/7543353. Accessed November 7, 2016.

[2] Low birthweight - Stanford children's health. http://www.stanfordchildrens.org/en/topic/default?id=low-birthweight-90-P02382.

The first two datasets are in CVS format and the latter two in Excel format, which I will need to convert. All four identify the more than 3,000 US counties by their unique FIPS number, and all were published withing the past 2-3 years.

Solution

A good fit obtained by using a regressor model to predict LBW, very low birthweight or premature births from the other data would show that the selected features are adequate as a comprehensive set of indicators. This would suggest that addressing these factors could reduce LBW. Various models could be used, but the weights obtained from a linear model or the coefficients of a logistic regression model (splitting the LBW measure into low/high) would add a measure of relative importance among the features. Tree-based feature selection or PCA would also be indicative of this.

Benchmark

I am not able to find a benchmark model for this problem. However, a 2008 article in the Journal of the Georgia Public Health Association reported a similar investigation focusing on Bibb County, Georgia, which is in all the datasets for this project. That investigation used a logistic regression analysis that ordered the most significant factors, along with their odds ratios, as: insufficient prenatal care, 3.6; mother of African-American race, 2.0; maternal education not beyond high school, 1.7; and use of tobacco during pregnancy, 1.7. These factors are all available in the datasets, but as percentages of the total population, whereas the Bibb County study used information on individual women. While these measures do not necessarily correlate, I would expect similar results at least on average, which could be observed using peer counties if the single county result diverges significantly from the study.

Evaluation Metrics

The metrics applied will depend on the method being used. For multiple linear regression, I will use the adjusted R^2: $R^2_{adj} = 1 (1 - R^2)(n - 1)/(n - p - 1)$, where n is the sample size and p the number of features. Without the adjustment, regular R^2 values are not comparable as they suffer from inflation as p increases. The goal will be to maximize R^2_{adj} so long as plots or other analyses show the model fitting the data well.

For logistic regression, the `F1` score will be used to measure the model's ability to classify the data points as representing low birthweight or not, using the Healthy People 2010 target as the divider between classes.

Project design

Preprocessing

- Combine the datasets, performing an initial manual feature selection in the process. Most of the CHSI dataset can be discarded, except for data on demographics, LBW and premature birth, and smoking. Data on education, economics and environmental quality will be added from the other three sets, keying on the FIPS number.
- Check data for completeness, and delete counties with missing data, which seemingly are few if any. If more data than expected is missing, consider alternate sources.
- Convert data into comparable formats, e.g. raw numbers to percentages.

Exploration

- Obtain statistical measures for the data to identify and remove outliers.
- Scale the data.

- Use a scatterplot to determine whether the data for each feature is normally distributed, and if not transform it.
- Use the scatterplot or a heat map to identify highly correlated features and whether any of the three potentially interesting targets — low birth rate, very low birth rate and premature births — are highly correlated.

Partitioning

- Randomly partition the data 80/20 into training and test sets, so none of the model-building procedures has access to the test data.

Feature identification

- Use Scikit Learn's ExtraTreesRegressor and PCA to identify relevant features and the amount of variance in the data that can be accounted for. This will not be used for feature selection but for comparison with the features selected by the models.

Modeling

I plan to explore two approaches to modeling the data: a multiple regression to predict a continuous target variable (potentially all three of the possible targets in turn), and logistic regression to predict the target classes of high or low rates of the target variable, after suitable transformation of the target data. In both cases, Scikit Learn implements variants that employ stability selection.

For the multiple linear regressor I will use Scikit Learn's RandomizedLasso regressor. Using GridSearchCV I will search for optimal settings for the alpha and scaling parameters.

For the logistic regression, my plan is to use Scikit Learns RandomizedLogisticRegression classifier, varying the C and scaling parameters with GridSearchCV. If I can learn enough R to implement this in R, however, I will try that, since the additional statistical data that the R package implementation provides will enable a more detailed comparison with the benchmark study of Bibb County in Georgia.

Conclusion

The features suggested for action to reduce LBW will be those ranked the highest by the most accurate model.