Investigating indicators of low birthweight

I. Identification

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 (VLBW) 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."

Their 1995 review noted that much progress had been made since the introduction of neonatal intensive care in the 1960s, before which relatively few infants born with VLBW survived. Current information provided by Stanford Children's Hospital notes that while modern hospital care in developed countries gives LBW babies better chances of survival with minimal negative effects, "nearly all low birthweight babies need specialized care in the Neonatal Intensive Care Unit."[2] This delays release and adds significantly to the cost of childbirth.

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. It states that currently in the United States, 8 percent of babies are born with LBW and that this rate is increasing.

Determining the significance of the factors identified by the hospital and in various studies, including education level and exposure to pollutants, could suggest interventions aimed at decreasing the rate of LBW, thus decreasing the risk of developmental problems as well as health-related expense.

Two studies that produced results relevant to this problem are referenced in the Benchmark section. But neither approached LBW from a national perspective using locally aggregated data. One analyzed predictive factors for a single county, using individual data, while the other looked at the effects of carbon monoxide exposure and found a correlation with LBW.

My interest in this matter stems from a desire to use machine learning to help improve people's quality of life and to work with real data in approaching a real problem. Among areas that the project suggested exploring, health care seemed to provide such an opportunity, and looking over the CHSI data, it seemed that LBW was a good candidate for investigation, since it has not been given the same level of attention as other conditions or diseases.

Problem

The problem is to determine the relative significance of potential contributing factors to or predictors of LBW, in order to be better able to design interventions aimed at decreasing it.

1Maureen Hack, Nancy K. Klein, and H. Gerry Taylor, "Long-Term Developmental Outcomes of Low Birth Weight Infants," The Future of Children 5, no. 1 (1995), doi:10.2307/1602514. Abstract at https://www.ncbi.nlm.nih.gov/pubmed/7543353

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

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 that 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 VLBW along with various demographic and health-related measures.

State_FIPS	County_FIPS	County	State	Strata_ID	LBW	Min_LBW	Max_LBW	CI_Min_LBW	CI_Max_LBW	VLBW	Min_VLBW	Max_VLBW	CI_Min_VLBW
1	1	Autauga	Alabama	29	8.1	6	8.1	7.1	9.1	1.6	0.8	1.5	1.2
1	3	Baldwin	Alabama	16	8.6	6.3	9.1	7.9	9.4	1.9	0.9	1.9	1.6
1	5	Barbour	Alabama	51	11	6.7	11.9	9.5	12.4	1.9	0.9	2.7	1.2
1	7	Bibb	Alabama	42	8.7	5.1	10.3	7.7	9.8	1.7	1	2.1	1.2

The CHSI dataset actually comprises several .csv files. The MEASURESOFBIRTHANDDEATH.csv snippet above shows the available stats for LBW and some for VLBW. The table extends to include many other measures. The national mean for LBW (for all counties reporting this measure) is 7.57 with a standard deviation of 1.85, a minimum value of 2.1 and a maximum of 15.6. The Min and Max columns give the minimum and maximum values of the indictor being reported among all the counties in the same strata. In the case of the first entry, Autauga County, the reported rate of 8.1 percent of babies being born with LBW is the highest among counties in strata 29. The CHSI strata divide the counties into 88 peer groups of 14-58 counties each, based on similarity in population and other measures as determined by an advisory committee. The CI_Min and CI_Max columns give the lower and upper limits of the (unspecified) confidence interval for each reported value, so the CI for Autauga's 8.1 ranges from 7.1 to 9.1. This pattern repeats throughout the set.

Much of the data is in the form of percentages, but the file DATAELEMENTDESCRIPTION.csv gives short descriptions of each measure and whether it is a percentage, count or indicator. DEFINEDDATAVALUE.csv provides additional information on the values used for missing data and indicators. Narrative information can be found in CHSI_data_definitions.pdf.

The CHSI data does include measures for environmental quality, but these are sparsely populated and unusable. Nonetheless, these are potential factors worth exploring. The Environmental Protection Agency's Environmental Quality Index (EQI) dataset Eqi_results_2013JULY22 gives measures for overall as well as air, water, land, built and sociodemographic environmental quality for all counties. It is important to note that in this data, higher numbers mean worse environmental quality. Supplementary information is available in the file EQI_Overview_Report. The composite Sociodemographic_EQI and Built_EQI indices would be difficult to interpret and did not make it into the final feature set.

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 dataset adds median household income as well as poverty percentages for the overall, under 17 and under five population segments. The Education dataset gives percentage measures for various education levels.

Two datasets are available in CVS format and two in Excel only. All four use a consistent identification system for the more than 3,000 US counties, and all were published within the past seven years. The names of the 29 features initially selected are mostly self-explanatory. Outside of the environmental quality indices, median household income and presence or absence of health centers (the only binary feature), they give either the percentage or number of people that fit in that category in the general population. Four refer specifically to women who gave birth — Under_18, Over_40, Unmarried and Late_Care — the last of these pertaining to women who received no prenatal care during the first trimester of their pregnancy.

Solution

A good fit obtained by using a regression model to predict LBW or VLBW from the other data would show that the selected features are adequate as a comprehensive set of indicators. This would suggest that interventions targeting factors with large coefficients could reduce LBW. Various models could be used, but the weight coefficients obtained from a linear or logistic regression model (splitting the LBW measure into low/high) would rank the features in terms of importance as well as whether they increased or decreased probability. Tree-based feature selection would indicate feature importance, but not whether the influence was positive or negative.

Benchmark

I am not able to find a benchmark model for this problem. However, a 2008 article reported a similar investigation focusing on Bibb County, Georgia. 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.[3] 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 for factors that vary little in how they affect the sub-population of women who had children and the general population as a whole. Racial composition and education might fall into this category.

Additionally, an article on the effects of exposure to carbon monoxide (CO) cited a study that determined a 1.22 odds ratio for CO exposure as a factor in LBW.[4] CO is a major component of auto exhaust, in turn a major source of air pollution. This project includes data on environmental quality, which, unlike the population percentage measures, would be expected to have a relatively consistent effect across the population of any given county, increasing its potential feature value.

Evaluation Metrics

The metrics applied depend on the method being used. For multiple linear regression, I use R^2. At first I considered using the adjusted $R^2_{adj} = 1-(1-R^2)(n-1)/(n-p-1)$, where n is the sample size and p the number of features, to counter inflation as p increases when comparing models with different numbers of features. But with $n >> p$, the term $(n-1)/(n-p-1)$ is very close to 1, so R^2 closely approximates R^2_{adj} and is sufficient for the investigation.

For logistic regression, I used both the Matthews correlation coefficient and ROC-AUC (area under the receiver operating curve) to measure the model's ability to classify data points as representing areas with a high incidence of low birthweight or not. The Matthews coefficient provides in a single number a measure that incorporates the information contained in all four cells of a binary classification confusion matrix — true positives, true negatives, false positives and false negatives. This makes it robust to class imbalance. ROC-AUC similarly provides a single-number measure of how well a model fits the data independent of the decision function value chosen to separate the two classes. Taken together, the AUC score can indicate which model performs best overall while the Matthews coefficient can show which model best fits the data for a particular decision function threshold. Odds ratios are easily derived as the exponentiated logistic regression coefficients.

---

3 Jackson H, Wei Y, Chen F. Quantitative Data Analysis of Multiple Factors Associated with Low Birth Weight in Bibb County, Georgia. *Journal of the Georgia Public Health Association.* 2008;1(1):24. https://www.gapha.org/wp-content/uploads/2015/11/Jackson-Lowbirth-Weight-2008.pdf.

4 Townsend C, Maynard R. Effects on Health of Prolonged Exposure to Low Concentrations of Carbon Monoxide. *Occupational & Environmental Medicine.* 2002;59(10):710. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1740215/pdf/v059p00708.pdf.

II. Analysis

Data preprocessing and exploration

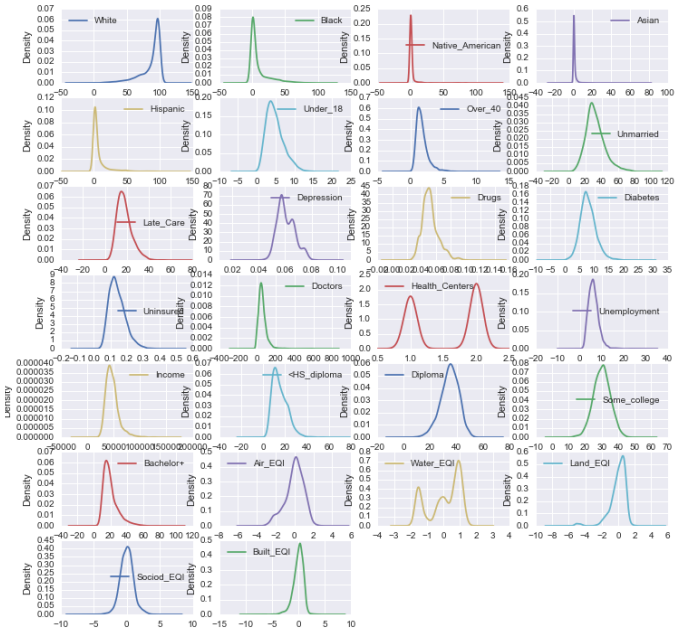
Datasets in Excel format were first converted to .csv files. Selected columns from the files were then read into Pandas dataframes, some of them renamed in the process. These features were selected manually as being those that seemingly could affect births. They comprised demographic data on racial composition (White, Black, Native American, Asian, Hispanic), education (<high school diploma, diploma, some college, bachelor's degree or higher), economics (unemployment rate, median family income), health (obesity, high blood pressure, smoking, depression, drug abuse, diabetes, uninsured), health services (doctors per capita, presence or absence of health centers), environmental quality indices (air, water, land, sociodemographic environment, built environment) and statistics on women who gave birth (under age 18, over age 40, unmarried, little or no prenatal care in the first trimester). The data also contained the statistics for LBW and VLBW births, which became the targets for the regression analyses.

All the datasets use the Federal Information Processing Standard code to identify the approximately 3,000 counties in the 50 US states (only two counties did not appear in all the sets), comprising one or two digits for the state and three for the county. But these were given separately or combined, with or without leading zeroes, so the numbers were converted to strings and later standardized. Additionally, some data was missing and the different sets, as well as different components of the CHSI set, handled this differently. So they were searched for occurrences of any of the several markers or null values, all of which were replaced with -9999.0, which could not occur in the actual data.

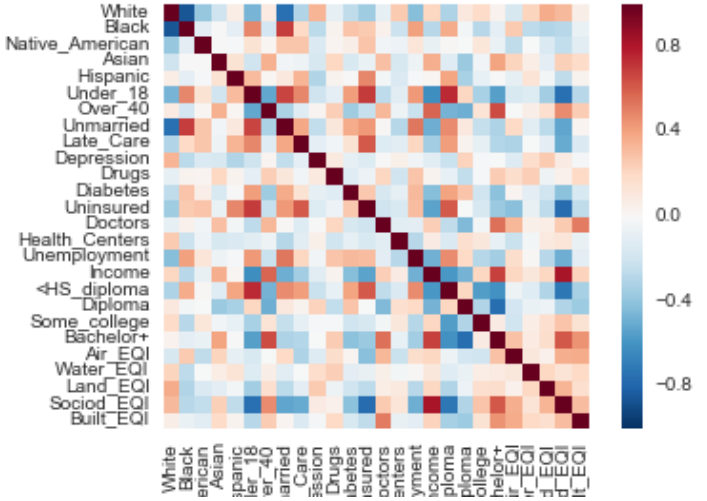
The missing data was dealt with in several ways. Three of the factors — Obesity, HBP (high blood pressure) and Smoker — were dropped because more than 25% of the values were missing for each. Missing values for other factors were interpolated, using the mean value for each factor for the strata each county belonged to. Diabetes, missing 422 values, was a candidate for being dropped, but was also an interesting feature. Further analysis showed that these missing values were from 35 states and 59 strata. Since they were so well spread out, they too were interpolated. The feature Late_Care was missing 109 values. Since this was one of the factors identified as significant in the Bibb County study, I did not want to interpolate these values if this could be avoided. Instead, the records containing these values were removed from the dataset pending determination of which factors would be screened out in the initial stage of the investigation. They would not be used if Late_Care turned out to be significant, but brought back in otherwise, since the feature could be dropped.

Another matter that needed attention was that while most of the data was in percentage form, some values (Depression, Drugs, Uninsured) were counts, so these needed to be divided by the county population. The environmental indices were in normalized form, and the Health_Centers data was binary, either 1 or 2, so these could be left as was.

As a first step in analyzing the data, distribution plots were generated for each of the 26 remaining features (after Obesity, HBP and Smoker were removed). The full-sized plot can be seen in the Jupyter notebook, but the general shapes in the figure to the right show that although several of the variables were moderately skewed, in different directions, nearly all were fairly symmetric. Since the ultimate aim was to evaluate the relative strength of the predictors rather than to predict values for the dependent variable, any data transformations that were not uniform across all the variables would have biased the results. There was no systemic pattern in the data that a single transformation could addressed, so no transformation was applied.

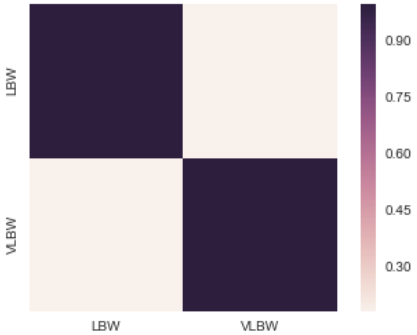


A significant problem was the potential for correlation and collinearity among the predictors. This was analyzed using a heat map and variance inflation factor (VIF) scores. The VIF scores measure how much the contribution of an independent variable could be inflated by the presence of other variables, essentially by removing each in turn and measuring how well the remaining variables can predict it. Opinions differ, but it is generally agreed that VIF scores above the 8-10 range are a cause for concern, although not necessarily a cause for outright disqualification. The predictors with the most extreme VIF scores were eliminated first: White (11,755), Diploma (2,015) and Some_college (1,410). It is easy to see that the race predictors were collinear as were the eduction predictors. In each case, almost every member of the population was in exactly one group, so together they added up to close to 100 percent. With regard to race, removing White removed the problem, as the heat map shows the other races did not correlate with one another. With regard to education, however, the heat map shows strong correlations among all four categories. Only one could remain, and <HS_diploma had the lowest VIF score and potentially the greatest value as an actionable predictor.



Depression also had a very high VIF score, 141, and was seen as likely be an unreliable measure from county to county. Unmarried also had a high score, 55, and can be seen to be strongly correlated with a number of other predictors. Sociod_EQI did not have a high VIF at 8.75, but does have strong correlations with other features and is a complex predictor that would be hard to interpret. All three were dropped. The VIF scores for the remaining 19 predictors were recalculated, and although several still had scores in the 20s, and one above 30, they were kept for further exploration.

A heat map also was generated for the potential target variables, LBW and



VLBW. The map, to the right, shows that these are not correlated. Hence it made sense to investigate the relationship of the predictors to both targets.

The next step was to explore the data for outliers, but since outliers in the targets could be important, only the predictors were investigated. The standard definition of an outlier was used — more than 1.5 times the Inter Quartile Range below the first quartile or more than 1.5 times the IQR above the third quartile. By this measure, the 2,999 records contained 2,524 outlier values. The number of records with two or more outlier values was 637, still almost a quarter of the data set. The number with three or more outlier values was 246, only 7 percent of the records, so these were dropped. The data was now ready to process with a random forest algorithm. Afterward, it would be scaled for use with linear and logistic regression algorithms.

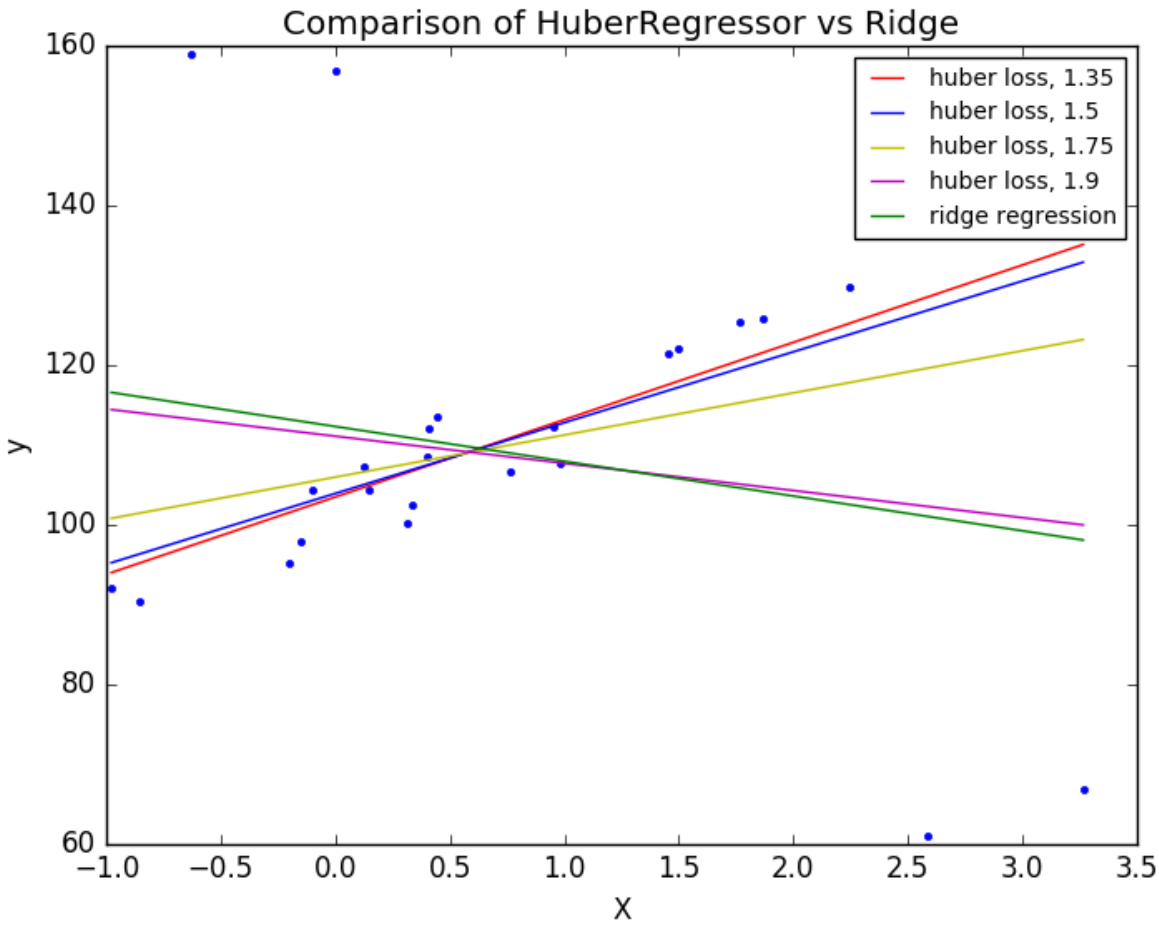
Algorithms and techniques

Concern over the effect of possible collinearity and the presence of numerous outlier values were prime considerations in selecting the algorithms to use with the data, in addition to the condition that the results provide numerical measures of the importance of the various factors relative to one another. The need for interpretability of the feature weights ruled out algorithms such as SVM, neural networks and nearest neighbors.

A random forest algorithm was used first to eliminate the weakest features, due to its ability to rank feature importance and its robustness to collinearity. This robustness arises from the random selection process the algorithm uses to select subsets of features to group together for each tree. Given enough trees, all possible permutations of the feature set should be used. Cross-validation with random shuffling would see these permutations applied to a multitude of different subsets of the data, and collinear factors would not always stay together. The features that account for the greatest information gain on average receive the highest importance scores. But the algorithm does not differentiate between factors that increase the probability of a given result and those that decrease it.

Linear regression algorithms were used with the reduced feature sets. But two different approaches to handling the problem of outlier data values were employed. One was to use robust scaling with a regular linear regressor, and the other was to use standard scaling with a Huber regressor, which handles outliers differently. Compared to standard scaling, which centers around the mean and scales to the standard deviation, scikit-learn's robust_scale centers around the median and scales to the interquartile range, so outliers will have less of an effect on a least-squares regression than they would using standard scaling. This is not necessary with the Huber regressor, which uses squared error for errors less than its epsilon parameter and absolute error for errors above that value. According to the scikit-learn documentation, "This makes sure that the loss function is not heavily influenced by the outliers while not completely ignoring their effect." Huber also scales errors using an alpha parameter. For use with the standard-scaled data I chose scikit-learn's ridge regressor. This implements an alpha parameter that controls the L2 regularization, which penalizes algorithms by summing the squares of the coefficients that they produce. According to the documentation, this makes the coefficients "more robust to collinearity."

The graphic below, from the Scikit Learn documentation, shows how outliers strongly affect a ridge regression (green line) while varying the epsilon parameter of a Huber regressor can diminish this effect and allow the regression to better fit the data without removing the outliers. As epsilon gets smaller, algorithm switches to an absolute error term closer to the target value.



In the event that the resulting linear regression models did not fit the data well or were not robust, which turned out to be the case, the investigation would be turned into a classification problem by classifying the LBW and/or VLBW targets as high or low based on some threshold value and then using a logistic regression algorithm. These regression coefficients can be converted to odds ratios by exponentiating them.

III. Methodology

Implementation

All of the algorithms were implemented using scikit-learn's GridSearchCV, both for the ability to select parameters that give the best performance and for the built-in cross-validation. As well, they were run on the full data and three subsets. One of these comprised the samples with VLBW targets, since 183 records were missing this data. The other two were the LBW and VLBW sets with racial features removed. Racial features would potentially be useful in identifying at-risk groups, but identifying significant factors irrespective of race was likely to be even more valuable. Since the code for the various runs of each algorithm was essentially the same, functions were defined and used so that they could be called and supplied the necessary inputs rather than repeating blocks of code.

GridSearchCV uses kfold cross-validation to separate the training data into k folds, or subsets, and then perform k training runs using each of the k folds in turn as the target data and the other k-1 folds as training data. For regression it uses regular kfold, while for classification it uses a stratified kfold, which keeps the proportion of each class in each fold the same as in the training set as a whole. It has an option for random shuffling, which I used for the random forest regressions. It was not necessary for the other algorithms, because I used a shuffle split on the data beforehand. The random forest algorithm

has a bootstrap option that creates new training sets the same size as the original through random selection with replacement, and related out-of-box validation on the samples that were not selected. I experimented with this, but did not like the duplication in the training set, so I changed to kfold cross-validation instead. It was simpler to add shuffling in at this step. I began with five folds, so the algorithms could train on 80 percent of the training sets at a time and validate against 20 percent, but tried three folds in an attempt to reduce overfitting by training on less data each time. While this did not reduce the discrepancy between training and holdout scores, neither did it reduce the holdout score, so I stayed with 3-fold since it is less computationally intensive.

For the random forest regression, I used GridSearchCV's parameter grid to vary the number of estimators (trees) along with the maximum depth of each tree. I held the number of features selected constant at seven — roughly one-third of the 19 features remaining after the initial selection through VIF scores and the heat map — because less than that resulted in lower scores and more than that would have the same features appearing together in combinations too often to effectively sort out the best. First I used the LBW targets for the regressions. The algorithm was strongly overfitting the data, but despite optimizing the parameters in the grid and then decreasing the maximum depth, the best performance stayed relatively constant with an R^2 of about 0.66 on the holdout data. Several trials produced an optimal number of estimators in the 350 range, while one produced a significantly higher number. Even then, however, the best scores stayed around 0.66. For these runs, the same random seed was used for consistency. Afterward, as throughout the rest of the investigation, several different random seeds were used for comparison. After determining the optimized parameters, the regression was also run on the data without the racial predictors, and these same two regressions were also run against the VLBW targets. The holdout scores for all variants were lower than for the LBW data with racial features. Importantly, the same features consistently had the lowest ranking. (The one exception was Unemployment, which had a much higher ranking in the runs with no racial features, so this was kept to simplify further handling of the data.) Based on these results, the number of features in the full feature set was reduced from 19 to 14, and in the set without racial features from 15 to 11. Interestingly, Late_Care was among the features dropped, despite lack of care in the first trimester being identified in one of the earlier cited studies as a significant factor. So I was able to merge back into the dataset the records I had removed because they had no data for this feature.

After paring down the feature sets, linear regressions were run on the data. Random forests can handle unscaled data, but linear regressions require consistent scaling across the features to perform well. As mentioned, two approaches were tried out in light of the large number of outlier values. Robust scaling was used on the data for the ridge regression and standard scaling for the Huber regression. For the ridge regression, GridSearchCV was used to optimize the alpha parameter, while for the Huber regression, optimization was done on the epsilon parameter. The parameter ranges were progressively narrowed to fine tune these parameters. The sklearn implementations of both of these regressors use R^2 scoring by default, and the results were very similar. The feature rankings produced by the Huber regressions, however, were more consistent with each other. As with the random forest regressions, the best scores were obtained using the LBW data. Moving from the LBW to the VLBW data, the optimal parameters also grew very large, showing that the algorithms were straining to fit the data. The runs for the datasets without racial features produced even worse results and were aborted. A sanity check was performed by regressing the population data, and population density data generated from it, against the LBW targets to see if either of these was significant and, hence, skewing or even dominating the outcomes. But the regressions showed essentially no correlation.

Logistic regression was then used on the data by turning the LBW targets into binary labels for either high or low LBW values. GridSearchCV was used to optimize the C parameter that controls the regularization term, as well as to select between L1 (linear) and L2 (squared) regularization.

Refinement

The process of refinement in this project primarily involved the selection and combination of features, targets and algorithms to produce a model that fit the data as well as possible while being stable and interpretable. The effect of correlations and possible collinearity among the candidate features was a large concern, so reducing the feature set was important. Seven of the initial 26 features were discarded based on analysis of the VIF scores and correlations. The random forest algorithm, refined through parameter optimization and adjusting the number of cross-validation splits, showed that six features consistently ranked lowest in predicting both the LBW and VLBW targets, and that when racial features were removed, only one of these six ranked higher. The other five were dropped, leaving 14 features, three of them racial, for further analysis. The random forest trials showed a better fit to the data using the LBW targets than the VLBW targets, as well as a better fit using the racial features than without them.

For the linear regression, two approaches were compared: a Huber regressor with standard scaling, and a ridge regressor with robust scaling. Trials of each were run with the same three random seeds. The results were similar, but the alpha regularization parameter for the ridge regression was consistently very high, the smallest value being 1,000 and the largest 5,623. The Huber regression produced epsilon parameters of 2.2 and 3.8 — as well as one high value of 11 — which are more in line with the default "standard" value of 1.35. It also produced more stable rankings, so it was used for further trials. Even so, performance was not very good, with test R^2 scores on the LBW data — which produced the best results — mostly falling below 0.70. A new feature set was constructed by multiplying all of the distinct pairs of features to generate interacting features. Rather than producing some strong new features and many weak ones, the resulting coefficients were spread along a gradient of slight variations that would be very hard to interpret. Given that the combinations would also increase the problems of correlation and collinearity, this was not a successful refinement.

The major successful refinement was to turn the regression problem into a classification problem that used a logistic regression model to classify LBW values as either high (true) or low (false). Setting the high/low threshold at the mean of the LBW values worked better than setting the value at the mean plus one-half of the standard deviation. The Matthews correlation scoring was appropriate for comparison, since it is not sensitive to class imbalance. The first trial on LBW data with racial features achieved an ROC-AUC score of .82, which is very good, but it did prove sensitive to perturbations of the training data. The data without racial features achieved a slightly worse but still good fit, and was stable.

An attempt to further refine the results of the stable logistic regression model, however, failed to bring significant improvement. Through visual analysis of the residuals from the decision function, it was determined that values less than -2 or greater than 3 were outside the band of closely clustered points. I wondered if removing the records that produced these outlying residuals and training on the remaining data could allow the algorithm to fit the data points closer to the decision boundary better, with the removed points likely to still fall on the proper sides of the boundary (similar to an SVM algorithm maximizing the distance to the decision boundary by working with the points closest to it). For this a custom implementation was employed that trained on only the data with moderate residuals, cross-validated against the held out training data combined with the removed data, and tested against combined training and removed data that was not used in the training or cross-validation. The results were mixed. Performance might have improved slightly, but not enough to justify the added computational overhead for a large number of trials, so the regular algorithm was used to produce the final results.

No major complications were encountered, although several minor issues arose. These included learning how to pass sample weights into GridSearchCV using the `fit_params` parameter, and refactoring the code to keep a raw dataset to fall back on and subset this to work on.

Summary of results

Initial features *VIF / heat map analysis*	*Random Forest regression*	*Huber regression (with linear coefficients)*		*Logistic regression (with odds ratios)*			
				Original data		*Shuffled data*	
~~White~~	Black	Black	+0.833	**Black**	9.13	**Under_18**	6.49
Black	Native_Am	Uninsured	+0.514	**Uninsured**	2.12	**Black**	5.39
Native_American	~~Asian~~	Doctors	+0.157	Diabetes	1.62	Air_EQI	1.33
Asian	Hispanic	Diabetes	+0.149	Doctors	1.23	Diabetes	1.22
Hispanic	Under_18	<HS	+0.130	Air_EQI	1.07	Doctors	1.02
Under_18	~~Over_40~~	Native_Am	+0.048	Native_Am	1.00	**Uninsured**	1.00
Over_40	~~Late_Care~~	Air_EQI	+0.043	Hispanic	1.00	Drugs	1.00
~~Unmarried~~ →	Drugs →	Unemploy	+0.024 →	<HS	1.00	<HS	0.98
Late_Care	Diabetes	Water_EQI	-0.004	Drugs	1.00	Water_EQI	0.95
~~Depression~~	Uninsured	Hispanic	-0.074	Water_EQI	1.00	Native_Am	0.95
Drugs	Doctors	Drugs	-0.090	Land_EQI	0.99	Hispanic	0.85
Diabetes	~~Health_Cen~~	Income	-0.100	**Under_18**	0.97	Land_EQI	0.75
Uninsured	Unemploy	Under_18	-0.109	Unemploy	0.80	Unemploy	0.71
Doctors	Income	Land_EQI	-0.177	Income	0.71	Income	0.71
Health_Centers	<HS	↓		↓			
Unemployment	Air_EQI						
Income	Water_EQI						
<HS_diploma	Land_EQI						
~~Diploma~~	~~Built_EQI~~	Under_18	+1.295	**Final model:**		Under_18	9.72
~~Some_College~~		Air_EQI	+0.519	Mean		Air_EQI	3.63
~~Bachelor+~~		Doctors	+0.286	ROC-AUC		Doctors	1.36
Air_EQI		Unemploy	+0.245	score: **.75**		Diabetes	1.21
Water_EQI		Diabetes	+0.167			Unemploy	1.11
Land_EQI		Income	-0.025			Drugs	0.93
Built_EQI		Drugs	-0.049			Uninsured	0.79
~~Sociodemo_EQI~~		Water_EQI	-0.173 →			Income	0.74
		Uninsured	-0.216			Water_EQI	0.74
		Land_EQI	-0.219			<HS	0.56
		<HS	-0.314			Land_EQI	0.43

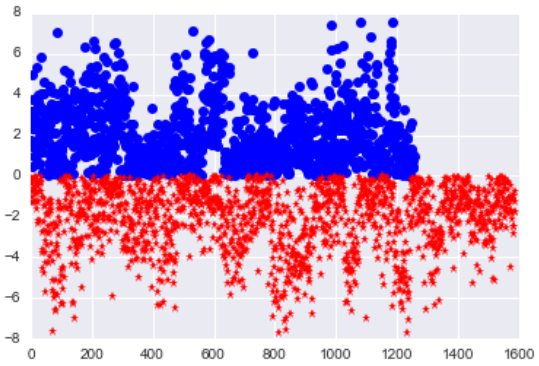
The Huber and logistic regression rankings show features that increase or decrease the probability of high LBW

IV. Results

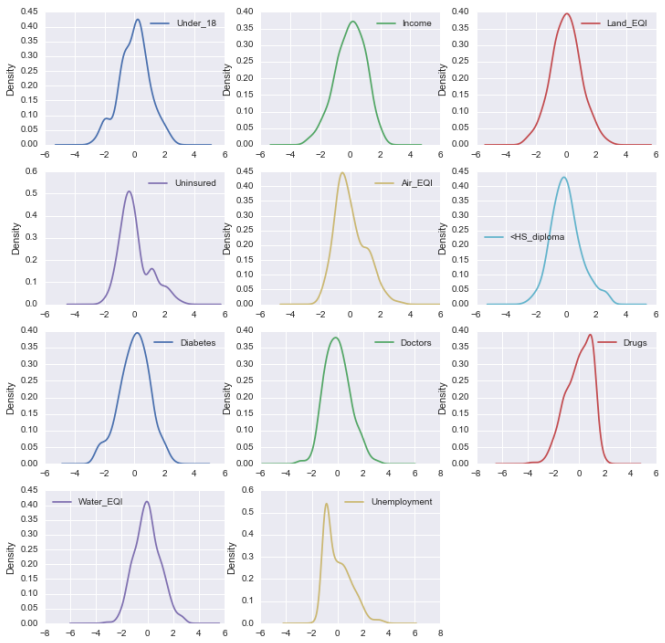
The process of arriving at the final model has been explained in different sections, so the whole process — illustrated in the table above — will be recapped. The initial dataset contained 26 features. This was reduced to 19 by eliminating those with the highest VIF scores and correlations. These were then fed into a random forest algorithm, based on which five more features were discarded. These five consistently ranked lowest across the four dataset variants (LBW and VLBW with/without racial features). Both in the random forest regression and the ensuing Huber and ridge regressions, the best results were obtained using the data with the LBW targets. Since R^2 results were not very good, however, logistic regression was tried to classify the LBW results as either low or high. Good results were obtained for the data with racial features, but shuffling the data changed the output significantly. The model was not stable.

So the racial features were removed and logistic regression run on the reduced set. The results were not as good as with the racial features, the ROC-AUC score for the first trial dropping from 0.82 to 0.77. Trials with several random seeds produced comparable results, with ROC-AUC scores staying at or

above 0.75, which is considered good if not very good. Importantly, however, trials with shuffled data and the same random seeds produced exactly the same results, both for scores and coefficients. The LBW models produced by logistic regression on the dataset with racial features removed (the lbw_nr dataset in the Jupyter notebook) are stable with respect to this perturbation of input. The plot at right shows that the residuals from the decision function of one of these runs are distributed in a visibly random pattern, and are clustered within the same range above the boundary (in this case 0) as below it, which was not the case for the LBW data with racial features. This is the kind of pattern I expected for an unbiased model.



Based on these results, the algorithm was run 300 times with randomly generated random seeds and the coefficients, as well as the ROC-AUC scores, collected. The final model comprises the mean values of these coefficients, with a mean ROC-AUC value of 0.75. As mentioned, this is generally considered a good score. An analysis of the coefficient values shows that all appear to be distributed normally, which would be expected of a good-fitting algorithm run against randomly sampled data, which is what the randomly generated seeds should produce, since they control the train-test splitting. As a final statistical assessment of the results, VIF scores were computed for the final feature set. This time, the highest score was 3.33, and only three of the other 10 features had scores greater than 10. This indicates that the features are strongly independent, which is necessary to produce a robust model.



Some of the results are completely in line with expectations. The feature having the strongest positive correlation with a high LBW rate is the percentage of births to young women under the age of 18. Maternal age was cited in the Stanford Children's Health information referenced above: "Teen mothers (especially those younger than 15 years old) have a much higher risk of having a baby with low birthweight." It should also be pointed out that the initial-features heat map showed Black and Under_18 having a moderate correlation, and that when Under_18 rose to the highest-ranked feature when the LBW data set with racial features was shuffled, Black decreased by a comparable amount. This instability made the data set with racial features unsuitable, but the one thing that consistently stands out in the regressions using racial data is that a high percentage of African-Americans in a population is a very strong predictor of high LBW, which is consistent with the benchmark Georgia study that identified the mother being African-American as the second-strongest predictor of LBW. (Note: The removed White feature had a strong inverse correlation with Black.)

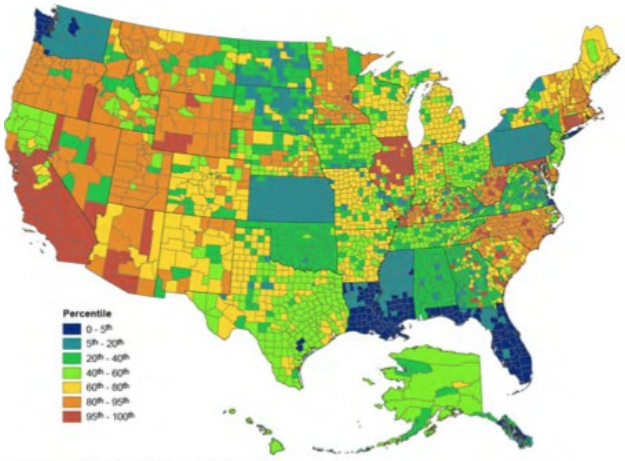
The second-strongest predictor in the final model is poor air quality (higher Air_EQI scores), which is in line with the cited benchmark study on the effects of carbon monoxide. Other results that seem reasonable are that higher incidence of diabetes (but maybe other medical conditions as well) or higher unemployment correlates with an increase in LBW rates, and higher median income with a decrease.

Other results, however, do not seem reasonable at first, such as Late_Care being one of the least significant predictors according to the random forest regression. Lack of early prenatal care was the strongest predictor of LBW in the Georgia study. The Late_Care feature, however, might be problematic in this investigation, because a closer look shows that it includes women for whom no information was available as well as women who reported not receiving care. So it would be useful only if the no-information group is consistently small compared to the no-care group. The model also has a higher percentage of people not graduating from high school as predictive of a *lower* rate of LBW, again seeming to contradict the Georgia study, a higher percentage of people being uninsured as predictive of a lower rate, and a higher number of doctors per capita as being predictive of a *higher* rate of LBW. These results seem opposite to what would be expected, as do worse water and land environmental quality leading to lower rates of LBW.

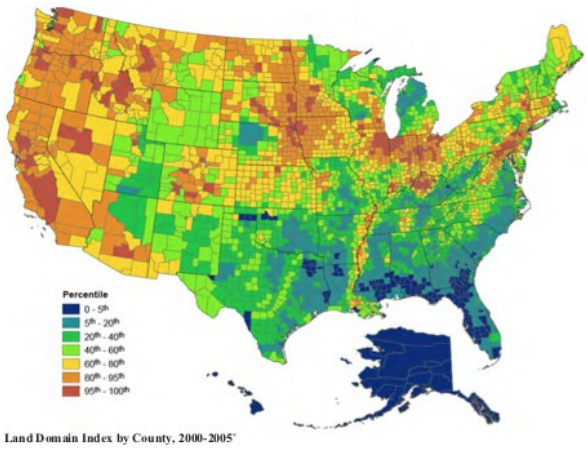
Part of this discrepancy arises from the use of aggregate data in this investigation, because a particular factor could affect a high percentage of the population, but relatively few pregnant women. Drugs are a good example. A woman who abuses drugs might have a higher chance of having an LBW baby, but relatively few pregnant women might abuse drugs in a community with an otherwise high incidence of drug abuse. So drugs could have a high impact individually but a low impact as an aggregate factor.

Other factors could confound one another. As a theoretical possibility, a relatively high number of young women in areas where high school graduation rates are low might have their first child under the age of 18. LBW might be relatively common for these babies, but that experience might make the mothers more mindful and informed, so that children they have when they are older are born at regular weight while the mother's education level remains unchanged. Or Hispanics, the racial group most strongly correlated with no insurance, could have strong family/community ties that provide effective support for pregnant women, leading to low rates of LBW. The increase in doctors per capita also could be misleading if access to them is the significant factor instead. So aggregate features in general are hard to interpret when the matter under investigation affects a subset of the population that is not chosen at random. They need to be looked at carefully.

On the other hand, factors that are tightly related to women, especially pregnant women, or that affect people more or less equally, such as environmental conditions, can reasonably be expected to serve as reliable predictors. The Under_18, Air_EQI and Income (median household income) features performed as expected. Women over the age of 40 having babies does not seem to increase LBW incidence, and that too is reasonable given that the Over_40 feature correlated positively with higher income and higher education level. But high water and land EQI scores predicting lower rates of LBW need further investigation. The EPA maps that accompany the dataset are useful. The Land Domain map (lower) shows an almost solid block of the lowest scores, which represent the best reported environmental quality, stretching from South Carolina down through Florida and across Louisiana up into Arkansas. The LBW data shows that nearly all the counties in these states have LBW rates above the mean. The Water Domain map (upper) is similar, though not as concentrated. Hence, even if the land and water EQI



. Water Domain Index by County, 2000-2005*



Land Domain Index by County, 2000-2005*

scores are accurate environmental measures, they are also de facto encodings for regions of the country, and hence include socio-demographic factors unrelated to environmental quality. This severely compromises their viability as distinct features. The Water Domain map raises an additional concern in that several states appear as solid or almost solid blocks of one percentile range, different from adjacent states, which raises the question of how accurate or useful this particular measure is. The Air Domain map, included in the EQI Overview Report, does not show the questionable block phenomenon, with the worst quality concentrated around urban centers, especially the Northeast Corridor, as expected.

As a further check on these results, additional logistic regressions were run omitting several features in turn from the final feature set. The omission of Air_EQI or Under_18 resulted in decreased performance from the earlier regressions run with the same random seeds, consistent with their ranking as the features most indicative of high LBW, although the change in the ROC-AUC score was smaller than expected for the two strongest positive features. Omission of Land_EQI also had a noticeable effect. But omitting Doctors or Water_EQI had essentially no effect on the ROC-AUC score.

Conclusion

In light of the interpretive difficulties and concerns that I have detailed, I think the only conclusive results of this investigation are that efforts to decrease pregnancy among girls aged 17 or less and to improve poor air quality would likely lower LBW rates. Pregnant women taking steps to reduce exposure to polluted air, perhaps by wearing pollution-filtering masks when traveling on city streets and using air purifiers at home, would also be suggested. Raising awareness of LBW among teenage girls and women with diabetes would also would be in line with the findings. I do not think this solution is significant enough to solve the problem, as these correlations were either already known or are obvious, although the identification of air pollution as a significant factor strengthens the earlier finding that exposure to carbon monoxide increases the risk of LBW. But other results, if corroborated, would raise significant questions for further investigation. These could include: Why would communities in which a high percentage of people do not graduate high school have relatively low LBW rates? Why would the rate of LBW rise with an increasing number of doctors per capita? In fact, putting together the unexpected results for both more doctors per capita and lack of early prenatal care suggests even more strongly that the link between health care and LBW needs further investigation. If answering these anomalies surfaced previously undetected relationships, then the investigation would have made a much more significant contribution toward solving the problem.

The graphic above illustrates the effect of the various features according to the final model. With LBW as the reference, the relative *height* of the font used for each feature indicates the strength of the effect, with red for features that increase the rate of LBW and green for features that decrease it. Lighter coloring is used for features that have either a slight or questionable effect. (Note: The feature Air_EQI has been relabeled 'Poor air quality' to avoid any ambiguity, while the Land_EQI and Water_EQI features have not been included because of their hidden socio-demographic links.)

Reflection

The project began with the selection of an area of interest from the given choices. I chose health care and looked at the available suggested datasets. The CHSI set was appealing, in that it has extensive data for all US counties. Looking at the data, I got the idea to try to predict LBW rates from other data while also looking for predictors that could support intervention strategies. I searched for other datasets with national county-level data to add to the CHSI data, and found the other three used in the investigation. Understanding and processing the data was time consuming, but feature selection was probably the greatest challenge. The initial selection was manual and aimed to gather all the features that I thought could relate to LBW. From there, however, I needed to narrow these down to useful features. The difficult task was to reduce strong correlations and collinearity among the features, since these destabilize models. A heat map and VIF scores were the initial analytical tools. Both were useful, but especially the VIF scores.

The variance inflation concept is very interesting, but the VIF scores also proved hard to work with. After removing the features with the highest scores, I started removing the remaining features one at a time to see if this would lower the other scores. I did not expect, however, that removing a feature would cause some of the other scores to *increase,* and without any apparent pattern. And removing two or more had equally unpredictable effects. So I used a random forest algorithm to rank the features in terms of their discriminatory capacity. Interestingly, at the end of the investigation I ran VIF scores on the features in the final model, and the highest was just over 3. This increased my confidence in the model and the stability it showed compared to the penultimate model.

After selecting a reduced feature set through the random forest regressions, I used linear regression on this data, trying two approaches to handling the large number of outlier data values. Robust scaling with a ridge regressor was a more standard approach, but I was intrigued by Huber regression and its use of an adaptive loss function that keeps the outliers but reduces their impact. The scores produced by the two algorithms were almost identical, but the Huber produced more stable feature rankings with greater separation among the features, due in part, I would think, by the smaller regularization terms it required. This validated its unique approach.

In the end, though, I did not think the scores were good enough, so I used thresholding to convert the investigation into a classification problem. At first it seemed the results for the features including the racial data were excellent, with ROC-AUC scores above 0.80, but these results were very unstable. The feature ranking changed significantly when the input data columns were shuffled. Dropping the racial features provided stability for the final model.

The final model fell short of my expectations, however, since the results for several features were very counter-intuitive. The reasons are detailed in the previous section, but a simple and clear example is that there is no obvious reason why an increase in doctors per capita would increase the odds of a high LBW rate. As I see it, the aggregate nature of the data has serious limitations in this investigation, since it misses the individual behaviors and influences that affect whether or not a mother gives birth to an LBW baby. I do not think I sufficiently considered this at the outset. Having said that, I think the

investigation was successful in identifying poor air quality, an environmental factor that affects people in a given area more or less equally, as a very strong predictor of high LBW rates. In terms of intervention, as mentioned, this could suggest that pregnant women decrease exposure to polluted air, but it also buttresses the importance of regulatory efforts aimed at ensuring adequate air quality standards. And the anomalies in the results could point toward as-yet-unidentified dynamics at play.

Improvement

I think that the most significant improvement to the project would lie in assessing carefully what the various potential features represent in the data, and in choosing the features that work best together. This would involve searching for and selecting data related to factors directly associated with pregnancy or birth, or that produce relatively equal affects among all the people in a given area. The accuracy of the data also needs to be considered. The EPA maps certainly raise questions about accuracy for the water quality data, and possibly the land quality data. Another source of uncertainty is exemplified by the data on lack of early prenatal care.

In terms of feature selection, it would be good to revisit the process with no racial features from the beginning, since their removal was key to reducing the VIF scores to good levels. It would be worth starting with the final features, minus the questionable water and maybe land quality data, and trying out other features one at a time, adding those that did not inflate the other scores or have high scores themselves. This could allow the linear regressions to achieve higher scores with more stable performance, although I think that classification better fits the nature of the problem.