Capstone Project

Roy Hu

Machine Learning Engineer Nanodegree

August 30th, 2016

Casual Effect of Age on Birth Weight Outcomes

Definition

Project Overview

Machine learning methods and algorithms are revolutionizing analytics across a variety of industries: from education, to healthcare, and to finance. Yet despite their rising popularity, I believe there is a tremendous amount of interesting and important work still yet to be done. An area that I am particularly interested in is the intersection of machine learning and econometrics. At times, the goals of machine learning prediction models can be fundamentally at odds with those of a lot of social science work on causal inference: a foundation of supervised ML model is predicting an outcome and selecting a model based on an optimized fit on a test sample. However, in econometrics, "just prediction" is *not* the same thing as causal inference. For instance, in empirical economics we often see confounding situations in which higher home prices are strongly positively associated with higher crime rates, but mainly because cities generally have both higher real estate prices and higher crime rates. Yet obviously, all else equal, we should expect higher home prices to be negatively correlated with higher crime rates.

Thus, a perfectly valid goal when doing empirical work in economics is to estimate the most fundamental *causal* effect of some variable on an outcome, e.g., find the causal effect of crime on housing price (presumably negative), as opposed to estimating simple correlations (usually positive) between the two—since, as everyone knows, correlation does not imply causation.

But just because textbook machine learning prediction methods are not immediately available for causal inference does not mean that they are not useful. The goal of my capstone project will be to adapt these ML methods and algorithms for causal inference. **In particular, I attempt to use machine learning methods to estimate the causal effect of age on the birth weight of the infant—a key indicator of infant health.** The birth weight of an infant is a key indicator for short term neonatal health, and has even been linked with later-life conditions, including diabetes, obesity, tobacco smoking, and intelligence. But as with the housing price example, it can be difficult to estimate the causal effect because since women who waited longer to have children also likely have more established careers, come from better economic circumstances, and have access to better prenatal care. So when building models to predict birth outcomes, it might appear on the surface that age is positively correlated with birth weights.

I use data from the CDC's National Survey for Family Growth, a particularly rich and publicly-available data set that gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. It is my hope that combining the machine learning algorithms with econometrics can uncover new insights into the causal effect of later pregnancies on the birth weight outcomes for infants.

Problem Statement

The classic framework for the causal inference problem is "potential outcomes" model developed by Neyman [1] (1923) and extended by Rubin [2] (1974) to non-experimental settings. Suppose we have a random sample $i = 1, 2, 3, ..., N$ from some population. For each unit i in the sample, let $T \in \{0,1\}$ indicate whether the unit received the treatment of interest (T= 1) or the control treatment (T = 0). We are interested on the effect of the treatment T on an outcome Y. Let $Y_i(1)$ be the potential outcome for individual i under treatment, and $Y_i(0)$ the potential outcome for the same individual if he or she had not undergone the treatment.

Of course, the difficulty here, usually called the "Fundamental Problem of Causal Inference" by Holland [3] (1986), is that we can never actually observe *both* potential outcomes for the *same* individual at the *same* point in time; that is, we can only observe either $Y_i(1)$ or $Y_i(0)$, but never both. For example, if we wanted to estimate the effect of attending college on future earnings, we could only ever observe the wages ("the outcome") for if an individual had gone to college ("the treatment"), or if he or she had not, but not for both the observed observation and its counterfactual simultaneously. Under idealized scientific experimental conditions, randomization of the treatment and control groups will take care of the problem. However, in social sciences doing such is usually either impossible or at least highly unethical: one cannot simply randomly assign someone to attend college or not.

To address the Fundamental Problem, the experimental design literature has developed various methods for balancing the control and treatment groups to mimic an ideal experiment ex-post, the most popular method being the propensity score matching. Propensity score matching splits causal inference into a **two-stage problem**. The **first stage** is to estimate the propensity score as the conditional probability of receiving the treatment, e.g., the probability of going to college, given the set of covariates, or $p(x) = \Pr(T=1 \mid X=x)$. Assuming "unconfoundedness," or the assumption that there are no additional confounding variables which affect both the treatment and outcome, then the causal inference problem $Y_i(T) = (T \mid X)$ can now be simply rewritten and estimated as $Y_i(T) = (T \mid p(x))$. That is, instead of conditioning upon a potentially huge vector of covariates, we only need to worry about a single scalar score that takes all that information into account. While there are always arguments for additional confounding variables, this assumption is more plausible if the covariates in the data set are considered to be "rich." As we will see later, I believe the CDC's data set is extremely rich, and it is actually quite hard to come up with any additional confounding variables. Moreover, due to the data set's huge feature space, it becomes extremely useful to summarize the information as a single propensity score. In the **second stage**, we use the propensity score to construct an "inverse propensity treatment weight" (IPTW), and we simply estimate a weighted regression of the outcome Y (birth weights) on the treatment T (age) using the IPTW. Weighing by the IPTW creates a pseudo-population in which there is no longer any association between those covariates and treatment (and therefore no confounding). The goal of weighting, therefore, is to get a contribution to the average outcome value that each individual makes. Thus, the coefficients from the second stage will give you the true causal effect of the treatment on the outcome.

Machine learning algorithms can be easily adapted to causal inference here, since the first stage of estimating the propensity score probability is essentially a classic supervised learning prediction problem. In classic causal inference, the treatment is a binary variable (either “received treatment” or “did not receive treatment”), and so the propensity score is estimated using a logistic regression. Subsequent research has shown that, however, non-parametric, machine learning techniques like a generalized boosted model often outperform the basic logistic regression, and can capture non-linear relationships. In my project, I hope to extend the research even further by estimating a Generalized Propensity Score (Hirano and Imbens 2004) using machine learning techniques, because my treatment variable, mother's age of birth, is a continuous variable, rather than a simple binary treatment. **It is my expectation to find that age should have a negative relationship on the birth weight of an infant.**

Metrics

Because this is a regression problem, I choose mean squared errors because minimizing the quadratic loss makes sense because some errors are better than others: errors that are closer to the "true" parameter should not be penalized as much as errors that are further away. Moreover, the MSE will asymptotically give us the minimum variance and unbiased estimators.

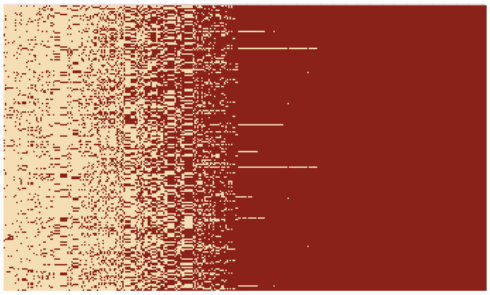
Analysis

Data Exploration

The data comes from the CDC's National Survey of Family Growth (NSFG) 2011-2013, which contains two data sets: one containing pregnancy, birth, and general demographic information for different mothers, and the other containing her detailed interview responses to the survey. My plan is to join together the **Demographic data** and the **Survey Response** together after some pre-processing.

The data has an enormous feature space. The **Demographic data set** has 9543 observations with 278 covariates, and the **Survey Response** has 3096 covariates. Both data sets suffer from having many features which have a severe missing data problem. If it were missing at random, then imputation techniques like KNN or median might make sense, but I strongly suspect this not to be the case. Below is a missing map of the Survey Response data before dropping the *NA* columns (the Demographics one looks similar). Because the main purpose of this project is to estimate causal effects from the demographic data set, I am comfortable with trying to extract some signal from the survey data set, and will use PCA for dimensionality reduction to create orthogonal features that can provide some additional background information for each mother in the demographic data set.

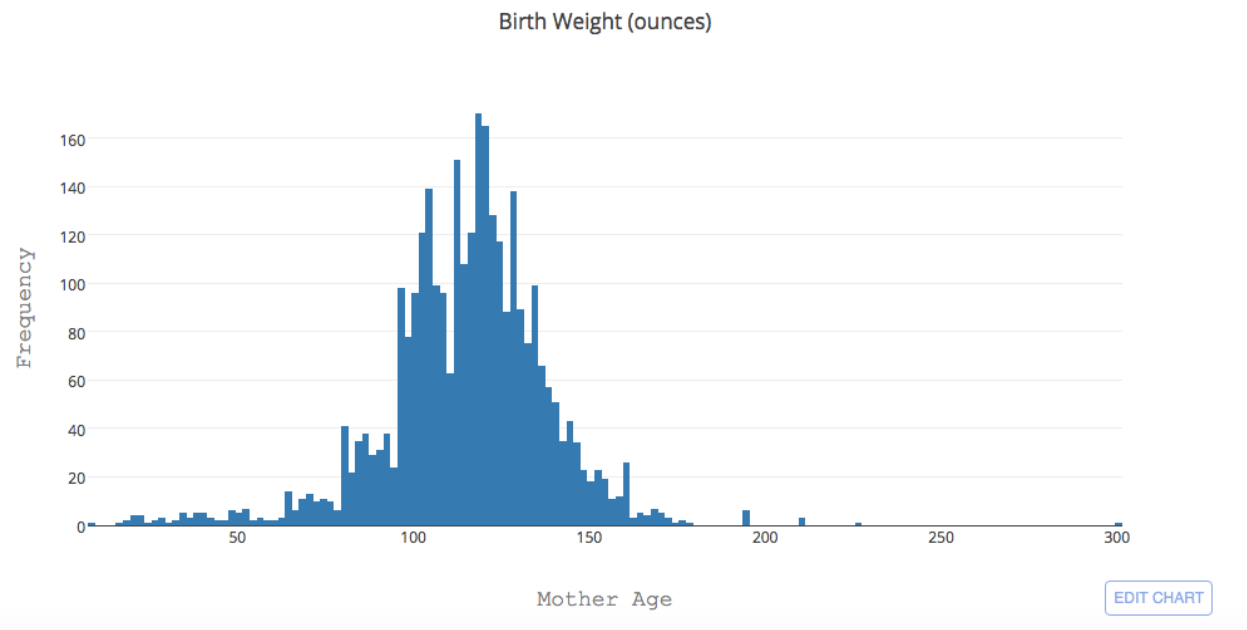
Missingness Map



Below are some summary statistics for the **Demographics** data set. Of particular relevance this problem is the mother's age and birth outcome (in ounces). The mean for the mother's age is 26.85 years, with a standard deviation of 5.5 years. The youngest mother who gave birth was just 12.75 years old, and the oldest was 43.5 years old.

	CASEID	PREGORDR	WKSGEST	BABYDOB_M	BABYDOB_Y	HPAGELB	CMFSTPRG	AGEPREG
count	3135.000000	3135.000000	3135.000000	3135.000000	3135.000000	3135.000000	3135.000000	3135.000000
mean	55190.815311	2.608293	38.431260	6.627432	2016.288676	30.123126	1233.283254	26.851537
std	2976.896115	1.469210	2.614996	4.998112	285.392377	9.578921	323.901959	5.538927
min	50002.000000	1.000000	17.000000	1.000000	1983.000000	14.000000	991.000000	12.750000
25%	52649.500000	1.000000	38.000000	3.000000	2003.000000	24.000000	1171.000000	22.660000
50%	55163.000000	2.000000	39.000000	7.000000	2007.000000	29.000000	1233.000000	26.410000
75%	57718.500000	3.000000	40.000000	9.000000	2010.000000	34.000000	1287.000000	30.750000
max	60414.000000	9.000000	46.000000	99.000000	9999.000000	99.000000	9999.000000	43.500000

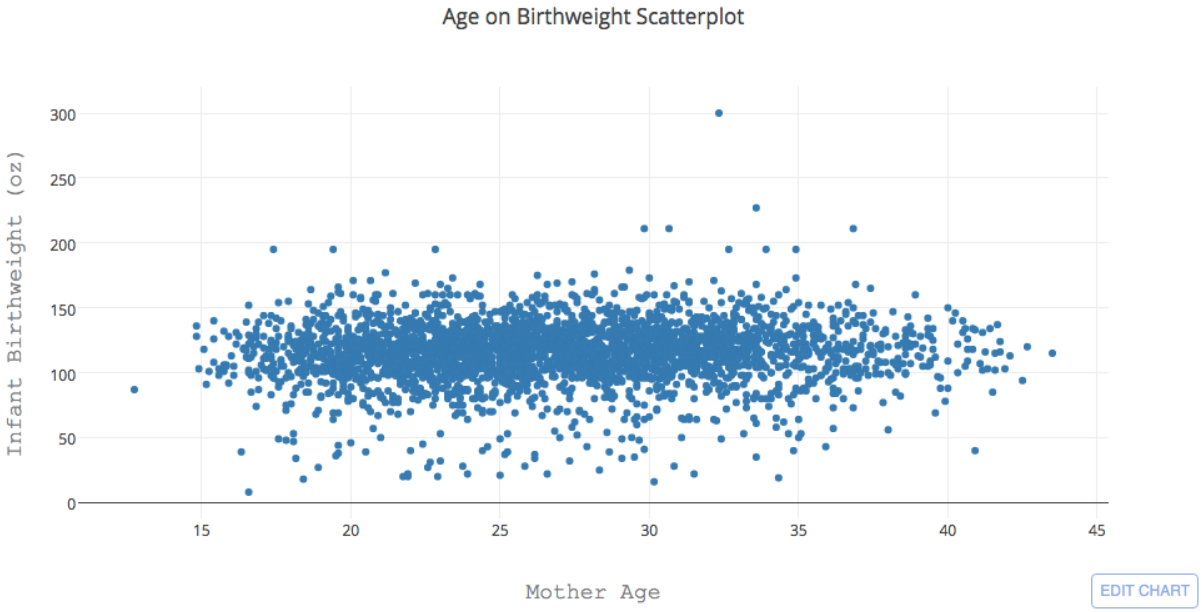
The average weight for infants was 115 ounces (about 7 pounds and 3 ounces), with a standard deviation of 23 ounces, which looks to be approximately normally distributed:



I was surprised to see so many low birth weight observations at the tail end of the spectrum. The lowest birth weight infant was only 8 ounces, although that could certainly be a data entry error. 3 pounds and 4 ounces (52 ounces) considered is the cutoff for "very low birth weight" infants.

Exploratory Visualization

Looking into the data a bit deeper, we can plot the outcome (birth weights in ounces) against the treatment (age in years) to get a sense of the apparent relationship:



Interestingly enough, visually there seems to be a slightly positive effect of age on birth outcomes. A simple OLS estimate of the coefficient for age is 0.265, statistically significant at the 1% level, meaning that for each additional mother's year, we should expect the infant to be born 0.265 ounces heavier. Of course, both theory and intuition would say that all else being equal, age should have a negative effect on birth weights. In the econometric literature, researchers will often add in age squared or a wage squared term, to take into account any changing effects. As described previously, there are many confounding factors, such as socio-economic statuses that may increase as the mother gets older. The goal of this project will be to find the true causal effect of birth weights on age.

Algorithms and Techniques

As described above, the **first stage** is a straight-forward prediction problem for the propensity score. To estimate the propensity score, I will implement a simple ensemble model with three different regressors all with equal voting share. Even with cross-validation, algorithms like decision trees and random forests are known to over-fit, so my ensemble will serve as a check on overfitting. To refine the models further, I will also implement grid search to identify the best parameters for each algorithm. The algorithm that I chose to implement are below:

- *Decision Tree* – A decision tree iterates through different decision nodes, and splits based on some criteria of maximum "information gain." Information gain could be something like a Gini coefficient, which indicates the degree of "purity" of the leaf nodes. A decision tree regressor can be tuned by optimizing its maximum "depth," or the longest distance from root to leaf.

- *Random Forest* – A random forest is an ensemble of decision trees. As with a decision tree, the random forest wants to split on maximum information gain. However, one

advantage of a random forest versus a decision tree is that each tree doesn't generalize very well; moreover, each tree doesn't see the exact same labeled training points. This helps with generalizing the learning. The parameter to be tuned here is the number of trees in the forest.

- *Gradient Boosting* – As with all boosting models, Gradient Boosting is about sequentially combining many "weak" learners into a "strong" one. And just like gradient descent in math, Gradient Boosting works by gradually minimizing some given loss function. There are a whole host of parameters to tune, but for this project I will stick with just the number of estimators.

Benchmark

As will become apparently shortly, it's important that the errors for the ensemble be normally distributed. Once I have the ensemble predictions, I will be able to create a residual plot versus the actuals that demonstrates this.

For this project, eventually I will need to run a weighted regression of the outcome (birth weights) on the treatment (age) to find the true causal effect. The quick OLS from above showed a statistically-significant positive relationship between birth weight and mother's age, which does not seem to match our intuition. Thus, my benchmark thresholds for finding a "successful" solution will be to find 1) the **correct sign** of the coefficients after each observation has been properly weighed, and 2) **statistical significant critical value at 95%**, or a t-statistic of 1.665.

Methodology

Data Preprocessing

To reiterate, the CDC posted two data sets: the Demographics data set and a Survey Response set. The following is a summary of the pre-processing:

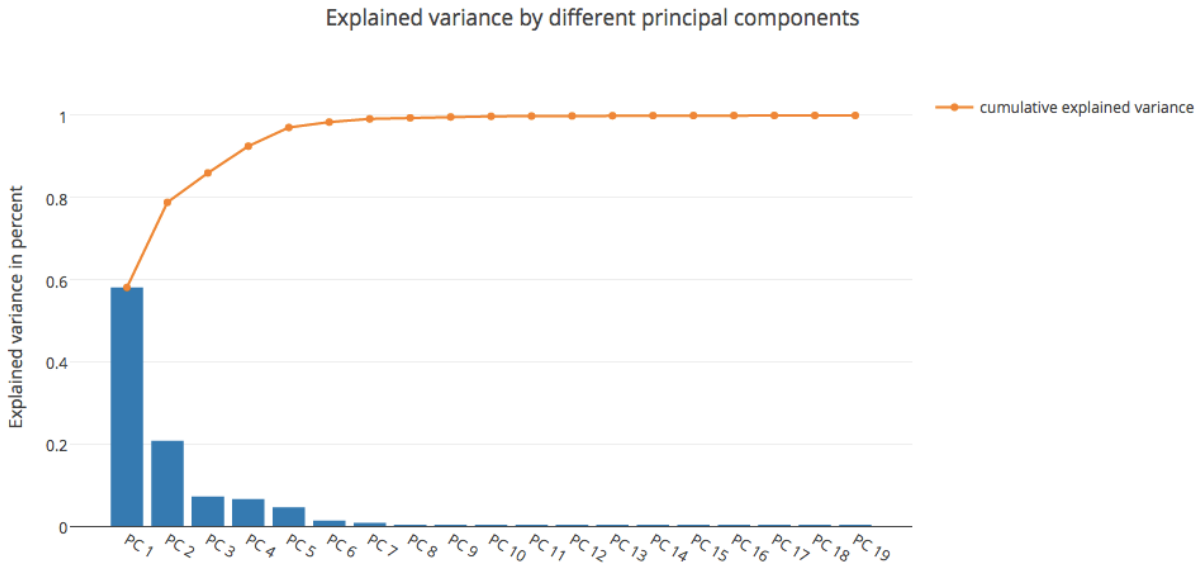
- **Demographic Data Set**
  - I only kept the observations for which the mother wanted to have the infant, and successfully completed the pregnancy. Thus, I removed observations for which the pregnancy did *not* end in a successful C-section or vaginal delivery, i.e., excluded observations that ended in an abortion, miscarriage, stillbirth, or *NA* value.
  - I removed columns which had nearly 0 variance because they are not informative. This dropped the feature space from 278 to 121 covariates.
  - I removed columns which had more than 10% missing values. This dropped the feature space from 121 covariates to 88 covariates.
  - Using the variable descriptions from the CDC, I removed features that were redundant to avoid issues of multicollinearity; I also removed a few columns that were not relevant to my analysis. Examples of redundant variables are:

addition metrics (e.g., different units) for age, birth weight, dates, education, pregnancy order. This was admittedly a judgement call, but I am fairly confident in my methodology in preserving enough covariates (40) that the data set could be still considered "rich" enough for the weakly unconfoundedness assumption to hold.
    - Implemented a simple mean imputation to fill in missing values.
    - Finally, following a lot of economic literature, I created a variable for the square of the mother's age. This is because the effect of age on birth weights may not stay constant.

- **Survey Response Data Set**
    - I removed columns which had nearly 0 variance because they are not informative. This dropped the feature space from 3096 covariates to 1065 covariates.
    - I removed columns which had more than 10% missing values. This dropped the feature space from 1065 covariates to 388 covariates.
    - Implemented a simple mean imputation to fill in missing values.

After the initial pre-processing, the final PregDemographics data set has 3,135 observations with 40 covariates; the PregSurveyResponse has 3,135 observations with 388 covariates.

Leaving the **Demographics** data set for a moment, the next step in pre-processing will be to take the **Survey and Response** data set, which has a fairly sizeable 388 features, and apply PCA to reduce the dimensionality and create orthogonal PCAs. Below is a plot of the percentage of variance explained for each principal component along with a cumulative explained variance graph:



As visualized above, the first principal component explains 58.09% of the variation, decreasing with each subsequent PC. I'm using the "elbow rule" to decide how many PCs to use; to me, it looks like after the 5th PC it doesn't seem worth it to include additional components. 5 principal components explain a total of 97.05% of the variation.

Finally, I combined the two data sets together, and converted the factor and categorical columns into dummies. The total features (including the dummies for all the factors) is now only 128 features.

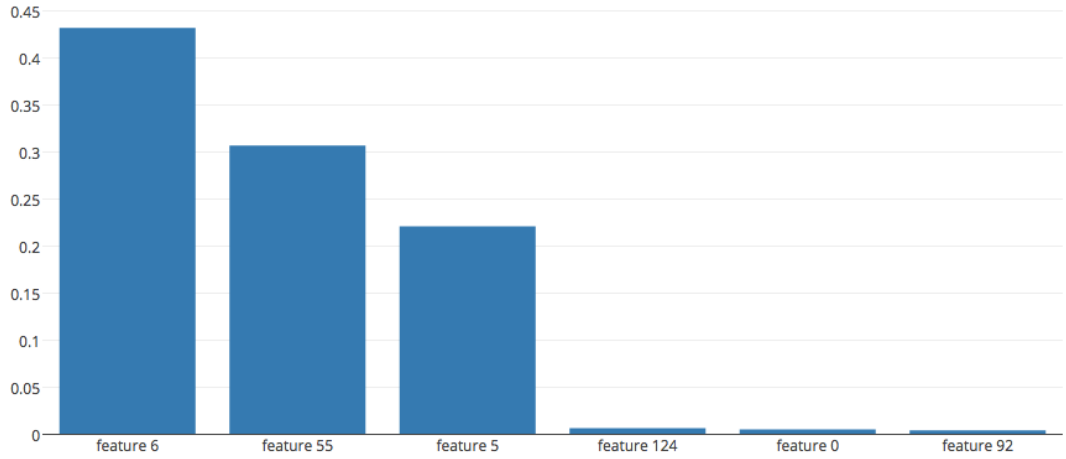
Implementation and Refinement

Now that I have everything in one big data frame, I randomly separate the data frame into features and targets in case that there was some ordering bias. As described previously, this is a **two-stage problem**. The **first stage** will be to estimate the propensity *p(x)* of treatment (Mother's age) given the vector of covariates (excluding the outcome, birth weight).

I use Grid Search to iterate through the different depths for the decision tree and found that the optimal level was 11. I did the same for the random forest (optimal number of tree was 150) and for gradient boosting (optimal rounds was 450). Below is a table of the means square errors before and after refinement with grid search to demonstrate that tuning the parameters does make a big difference:

ML Algorithm	*Before Grid Search (MSE)*	*After Grid Search (MSE)*
Decision Tree	4.339	0.426
Random Forest	0.132	0.120
Gradient Boosting	1.342	0.148

One algorithm I wanted to take a closer look into is the random forest. Like the decision tree, the random forest chooses to split on the node which gives it the maximum information gain, and the sklearn version has a method that can output a rank-order of the most important features:

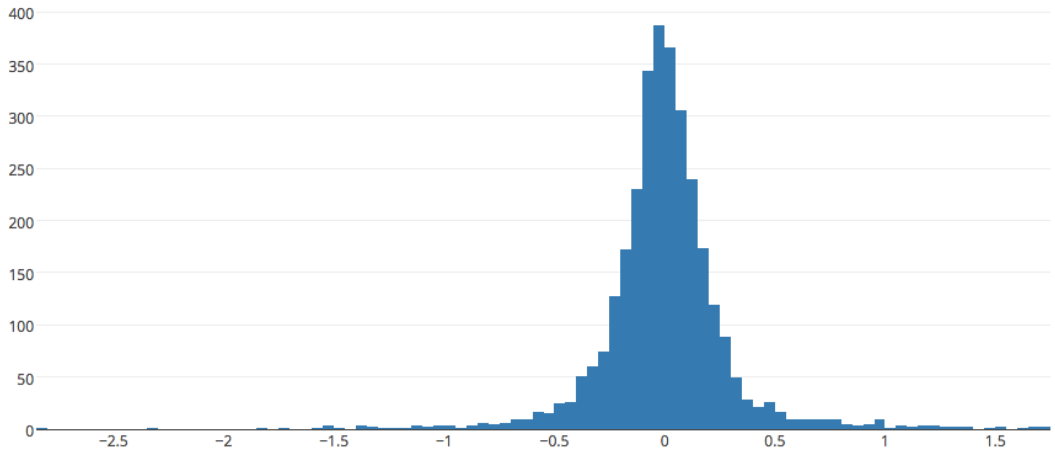


According to the algorithm, the most important features for predicting the mother's age is the Feature 6 (Baby DOB).

Using the predictions for each algorithm, I take the average to create an ensemble prediction, giving each one the same voting share. Following the methodology of Zhu, Coffman, and Ghosh (2014) [4], ideally we would like the errors of the residuals to be normally distributed.

$$T = m(\mathbf{X}) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2).$$

Where T is the treatment (age), and $m(\mathbf{X})$ is the mean function of the treatment given the covariates, i.e., the ensemble predictions. A plot of my residuals is below:

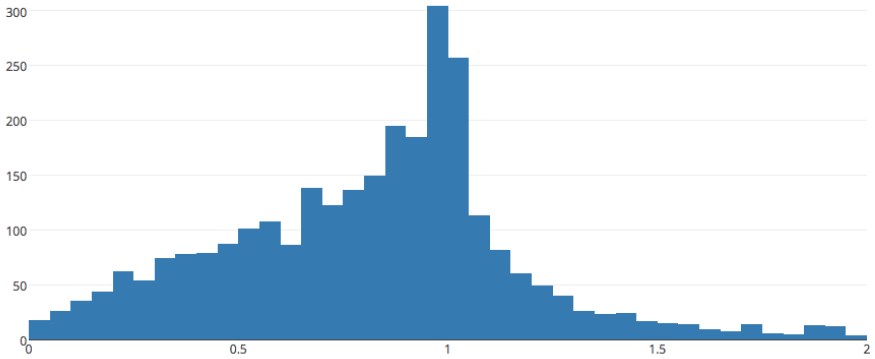


The residuals for my ensemble predictor are not exactly normally distributed, but they are approximately normally distributed, with no obvious skewness, and still useful for the analysis.

Finally, moving on to the **second stage** of the analysis, now that we have the propensity of the treatment (age), we can calculate the Inverse Propensity Treatment Weight following Zhu (2014) and Thoemmes (2016) [5]:

$$\varphi\Big(\mathrm{E}(Z)\Big)/\varphi\Big(\mathrm{E}(Z|\mathbf{X})\Big)$$

The equation uses the normal probability density function. The numerator for the IPTW simply means the pdf of the unconditional expectation of the treatment *Z* (which we had been previously calling *T*), or basically a regression of the treatment (age) against the intercept only without predictors. The denominator is the pdf of the conditional expectation of the treatment (age) given the set of covariates, i.e., what we had been estimating previously in this project. Looking at the weights, there are some extreme values, which I traced back to poor predictions in the denominator. Zhu (2014) recommends dropping these observations. Extreme weights are harmful to the analysis because they increase the variance of the causal estimates.



The weights, above, though not exactly normal, are probably good enough for this analysis. Now that we have the IPTWs for each observation, we can run a weighted linear regression to mimic a pseudo-random experiment in our data. Hirano and Imbens (2004) implement a similar model specification:

WLS Regression Results

Dep. Variable:	birth	R-squared:	0.006
Model:	WLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	8.461
Date:	Thu, 01 Sep 2016	Prob (F-statistic):	0.000217
Time:	14:19:26	Log-Likelihood:	-14572.
No. Observations:	2865	AIC:	2.915e+04
Df Residuals:	2862	BIC:	2.917e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
agepreg	1.7309	1.541	1.123	0.261	-1.290	4.752
intercept	85.9265	20.461	4.199	0.000	45.806	126.047
agesq	-0.0203	0.029	-0.708	0.479	-0.076	0.036

Omnibus:	3504.704	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	682884.173
Skew:	6.354	Prob(JB):	0.00
Kurtosis:	77.559	Cond. No.	2.28e+04

Unlike the scatterplot above, which suggested that age had a statistically significant positive effect on birth weights due to a host of endogeneity issues, the weighted model here is much more ambivalent about the true causal effect of age on birth outcomes. The model suggests that for each additional mother's year, we can expect the infant to be 1.7309 ounces heavier; however, the effect is not constant and over time deceases per the negative age squared term. More importantly, however, the standard errors are large enough such that at the 95% confidence interval captures 0. I wish the results provided a more unequivocal answer, but at the very least we can't say that age has a positive relationship with infant birth weights. I did try to adjust the covariance to something that may help with heteroscedasticity, e.g., HC3, but it didn't do much for the model.

Results

Model Evaluation and Validation

Regression models can sometimes be misleading because they may be a function of the data and thus not represent the true hypothesis. In Bayesian terms, the results we see are usually the probability of the data given the hypothesis, when a lot of the times we really want to know the reverse case. As a way to address this, we sometimes test the sensitivity of the model. Just as an experiment, I disturbed the data by randomly removing a few hundred observations to see if it changes anything:

WLS Regression Results

Dep. Variable:	birth	R-squared:	0.005
Model:	WLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	5.346
Date:	Thu, 01 Sep 2016	Prob (F-statistic):	0.00483
Time:	14:38:25	Log-Likelihood:	-10339.
No. Observations:	2000	AIC:	2.068e+04
Df Residuals:	1997	BIC:	2.070e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
agepreg	1.6660	1.994	0.836	0.403	-2.244	5.576
intercept	87.0757	26.507	3.285	0.001	35.091	139.060
agesq	-0.0186	0.037	-0.502	0.616	-0.091	0.054

Omnibus:	2501.617	Durbin-Watson:	2.028
Prob(Omnibus):	0.000	Jarque-Bera (JB):	446859.837
Skew:	6.589	Prob(JB):	0.00
Kurtosis:	75.032	Cond. No.	2.28e+04

Again we find some pretty similar results. The coefficients of each term is the same are fairly stable, but the standard errors are large again. With regards to my original benchmark, I found some mixed success. I did find the correct signs for my regression coefficients; however, due to the standard errors my critical t-statistic does not meet the benchmark of statistical significance at the 95% level.

Justification

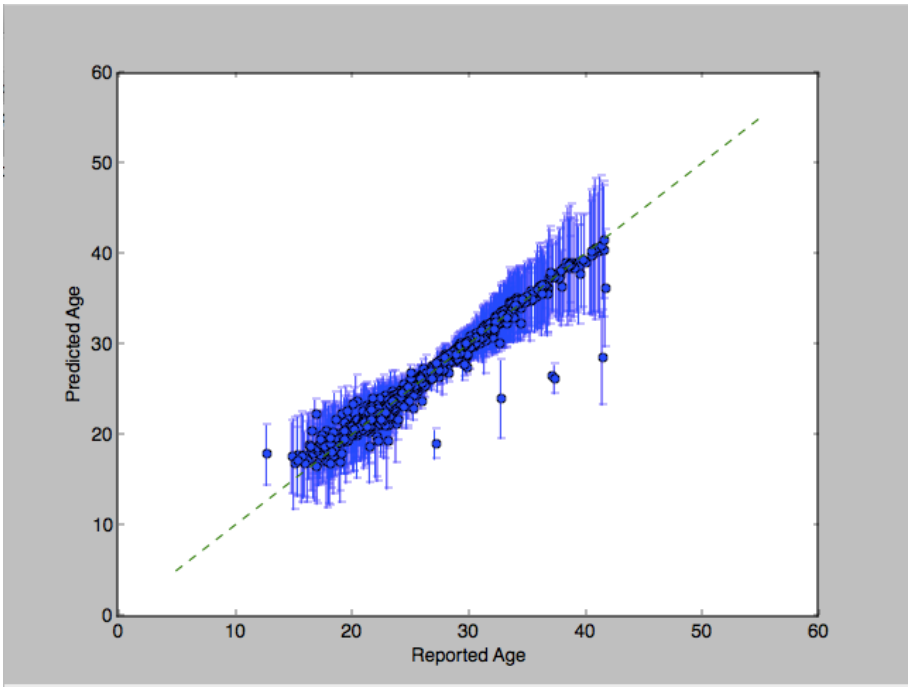
I believe that the regression results are closer with our intuition than the original scatter plot. Of course, ideally we would have liked to see that age had a statistically significant negative effect on birth outcomes. For me, I think it comes down to the standard errors. As demonstrated below, the standard errors for the ensemble mean function was not exactly homoscedastic. And so in a linear regression, in the presence of heteroscedasticity the estimators may not totally be efficient.

Conclusion

Free Form Visualization

An important aspect of my project that I wanted to highlight was some of the difficulties of using traditional statistics with machine learning. An interesting line of work that I came across when researching machine learning and econometric is some research [6] being done by Stefan Wager, a Ph.D candidate at Stanford University, which investigates the classical asymptotic properties of the random forest algorithm. I believe that bridging machine learning and traditional statistics might be one of the next advances in machine learning—things like central limits for machine learning. Wager wrote a paper about bootstrapping the standard errors for the random forest algorithm using the infinitesimal jackknife approach. While the exact math is a little above my

head, I was able to implement his approach using the sciforest package in Python, which is Wager's original R package translated by Ariel Rokem. Below is the output of my standard errors:



Based on the graph, it does seem like algorithm is doing a reasonably good job at predicting the age of the mother, especially around the central tendencies. It does seem, however, that the predictions degrade around the extreme values of age. It is interesting that the standard errors are not constant; the algorithm is more confident at some values than others. A lot the distributions in my project I've had to live with just "good enough" or "normal enough," though clearly based on this visualization there is still work to be done. Unfortunately, a lot of statistical and econometric analysis still requires some normality and homoscedastic assumptions. Having heteroscedastic errors is less than ideal because although it won't change the bias of the prediction, it does affect the efficiency of the estimators.

Reflection and Improvements

I personally found this project to be pretty rewarding because I was able to combine two of my interests, machine learning and econometrics, into one analytics project. In the pre-processing step, I was able to implement both an unsupervised machine learning technique (PCA) to massively reduce the dimensionality of the Survey Response data set into useable components, capturing over 97% of the variance in just a few principal components. Then in the first stage, I was also able to implement and refine a supervised machine learning techniques (decision tree, random forests, gradient boosting) to make fairly accurate ensemble predictions for the mother's age.

That said, there is definitely room for extension. One challenge that was just inherent in the data set was the missing value problem. I did feel like I potentially may have dropped columns that

were informative; there likely was a reason why someone might not be able fill in a response in the Survey Data set, or provide an answer to the interview question in the Demographics data set. Unfortunately, despite some significant time spent looking into different imputation techniques, I'm not sure there would have been a way to impute without assuming "missingness at random."

Another interesting extension could be more of the kinds of work that Wager is doing with bootstrapping asymptotic for machine learning algorithms. For example, I would have liked to have standard errors for gradient boosting, but there does not appear to be any published research for something like that at the moment. In the end, I strongly believe machine learning is growing field that will benefit greatly from more research back into its statistical foundations—while at the same time pushing both fields forward.

Works Cited

[1] Neyman, Jerzy. *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes.* Master's Thesis (1923)

[2] Rubin, Donald (1974) "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66 (5), pp. 688–701.

[3] Holland, Paul W. *Statistics and Causal Inference.* Journal of the American Statistical Association, Vol. 81, No. 396 (Dec., 1986), 945-960.

[4] Wager, Stefan. "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife."

[5] Zhu, Coffman, and Ghosh. "A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments." (2014).

[6] Thoemmes and Ong. "A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models" (2016).