DATA 102 FINAL PROJECT, FALL 2022
A STUDY OF THE 2018 US BIRTHS DATA *

Wenhao Pan, Ziyue Deng, Anni Kang, Huanzhong Jia
University of California, Berkeley
{wenhao1102, ziyue_deng, annikang, huanzhong0625}@berkeley.edu

1 Introduction

A baby's birth weight is correlated with mortality risk and potential future developmental problems. Rapaport (2017) found that babies with birth weights of less than 5 pounds are more likely to experience health complications and even a lower intelligence quotient as children. Thus, it makes sense for healthcare workers and parents to want to predict a baby's birth weight based on current information. Intuitively speaking, a baby's birth could be mainly affected by a lot of factors about itself and its parents, such as the health conditions of the parents, the sex of the baby, the mother's pregnancy records, etc.

In this project, we use multiple applied statistics and data science techniques taught in Data C102 of Fall 2022 (Course Staff Team, 2022) to answer two research questions about what might cause a baby's birth weight to change and how to predict a baby's birth weight. Specifically, we use causal inference to answer the second question and use GLMs and nonparametric methods to answer the first question. The full code and data are available from `https://github.com/WenhaoP/Data102-Fa22-Proj.git`.

2 Data Overview

We use a publicly available Kaggle dataset that contains information about all of the childbirth in the United States in the year of 2018. It can be downloaded from `https://www.kaggle.com/datasets/des137/us-births-2018`. The dataset was created by using the raw data stored at the website of the National Center for Health Statistics, and they can be downloaded from `https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`.

The dataset contains information about 3.8 million childbirths in the US in 2018, and each row contains information about a single baby. There are 55 columns, and we can group them into the following categories:

- Delivery situation ex) place of birth, number of people around, birth time
- The baby's health information ex) period of gestation, birth weight
- Parents information ex) marital status, education, race
- Parents health records ex) smoking history, age
- Mother's pregnancy records ex) number of prenatal visits, prior births

These 55 columns have already contained abundant information about childbirth, so we do not think extra information is necessarily needed for our study. The column `DBWT` represents babies' birth weight which is our response or outcome variable, and all other variables will be used as potential explanatory variables.

A User Guide that contains detailed explanations of the dataset and each column in the dataset can be found at `https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/UserGuide2018-508.pdf`. According to the User Guide, the raw data is a credible census as it is collected from the US birth registration system and the US Standard Certificates of Birth. Thus, we believe that there should not be any

*We thank the course staff team for their feedback on our project. All errors are our own.

groups that were systematically excluded from the data; all participants should be aware of the collection of this data, and there should not be any significant concerns such as selection bias, measurement error, or convenience sampling in the data.

3 Research Questions

3.1 Causal Inference

Our first research question is: *How does a mother's smoking behavior cause a change in her baby's birth weight?* The real-world decisions related to this question include deciding the effect of smoking on pregnant women and also their babies. If smoking is proven to cause a decrease in a baby's birth weight, certain protection and regulations can be made to restrain pregnant women from smoking in order to protect newborns. Causal inference techniques, including randomized experiments, outcome regression, and inverse propensity weighting, are a good fit for the question because we are trying to study the causal effect of smoking on babies' weights.

3.2 Prediction

Our second research question is: *How to predict a baby's birth weight from its mother's age and the number of prenatal visits?* By answering this research question, we can help doctors to predict a baby's birth weight before the baby is born, which might help the doctor to decide what is the best delivery method for the mother: c-section might be better if the baby has a very high birth weight. We are trying to make a prediction, so GLM (parametric) or decision trees (non-parametric) are a good fit.

4 Exploratory Data Analysis

4.1 Data Preprocessing

We first remove all observations with any missing value so that the number of observations drops from 3.8 million to 2.8 million. This may introduce bias to our analysis as the missing values may follow a pattern, but the size of the remaining data should still be sufficient for us to conduct a meaningful analysis.

Next, we conduct feature engineering by recoding some features to create new features that are potentially more helpful for later analysis based on domain knowledge. For example, we add the feature `PREG_LEN` to directly calculate the pregnancy length instead of having to look for the difference between the last normal menses month `DLMP_MM` and the birth month `DOB_MM`. We binarize the number of cigarettes smoked before pregnancy (`CIG`) as 0 or 1 to simplify the later analysis. We recode the feature `PRECARE` and rated it according to the number of months of prenatal care. Simplifying this feature could make our analysis and calculation easier. We had also binarized `PRIORDEAD`, `PRIORLIVE`, and `PRIORTERM` features, so any prior deaths, livings, or other terminations of babies were either categorized as 1 (True) and 0 (False).

Then, we drop columns whose over 99% entries have the same values because they probably can not provide important insights for our analysis. Finally, for the computational time concern, we randomly sample 10000 observations to create a subsampled dataset and use it for all later analyses.

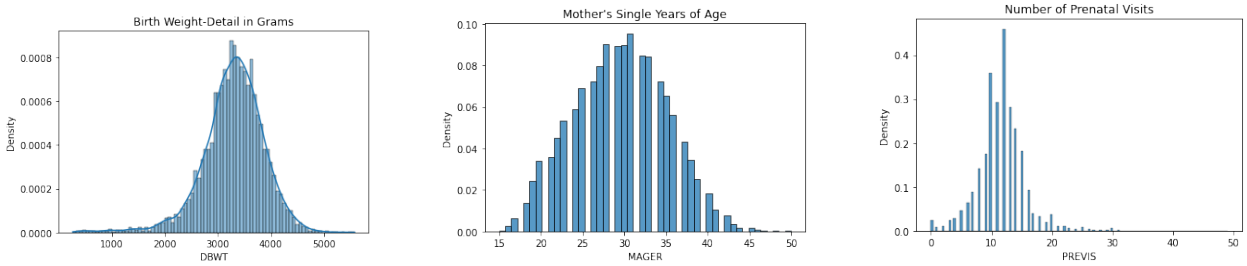
4.2 Data Visualization

We create data visualizations for a few selected quantitative and categorical variables. They serve to help us better understand our data and better study our research questions.

4.2.1 Quantitative Variables

From the histogram in Figure (1a), we observe that the birth weight is roughly normally distributed with a mean of 3292 and a standard deviation of 578. We can see that it is slightly skewed to the left, with a minimum weight of 269; this is an interesting trend that we can follow up in order to answer our first research question (3.1). Our second research question (3.2) is attempting to predict birth weight, so we can use this histogram to help us perform model checking.

From the histogram in Figure (1b), we see that the mother's age is roughly normally distributed as well with a mean of 19.5 and a standard deviation of 5.6. We can see that it is slightly skewed to the right, with a maximum age of 50; this is an interesting trend that we can follow up on for our second research question ("Predicting a baby's birth weight from



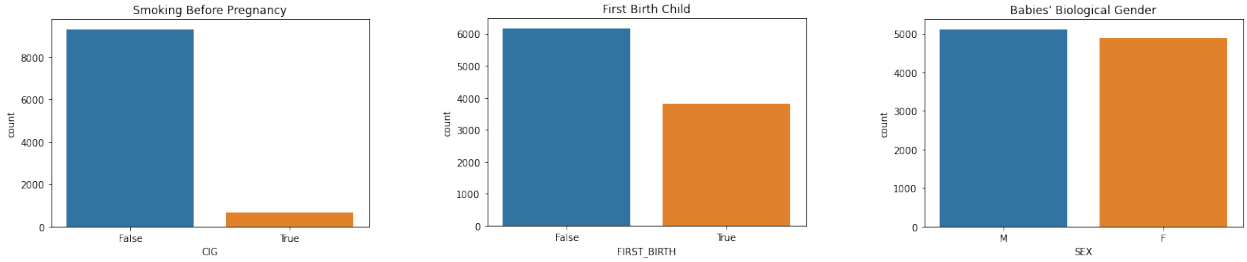
(a) Birth Weight (DBWT) (b) Mother's single years of age (MAGER) (c) Number of Prenatal Visits (PREVIS)

Figure 1: Data visualization of quantitative variables

Mother's Single Years of Age and Number of Prenatal Visits, comparing GLMs to nonparametric methods."). We can use this to help us answer if an older mother is likely to have babies with lighter birth weights.

From the histogram in Figure (1c), we observe that the number of prenatal visits is roughly normally distributed with a mean of 11.65 and a standard deviation of 3.96. We can see that there is an outlier in the data: one family has 49 prenatal visits. This is an interesting observation for our research question. We can use this to help us answer if more prenatal visits are likely to increase a baby's birth weight.

4.2.2 Categorical Variables



(a) Smoking before pregnancy (CIG) (b) First birth (BIRST_BIRTH) (c) Sex of Infant (SEX)

Figure 2: Data visualization of quantitative variables

In the data preprocessing, the number of cigarettes before pregnancy (CIG-0) is binarized into True (having greater than or equal to 1 cigarette) and False (not having any cigarettes). From the bar chart in Figure (2a), we can see that 93.18% of mothers don't smoke before pregnancy, while 6.82% do smoke. The significant difference between the numbers of smokers and non-smokers is something we need to take into account when analyzing the effect of smoking on baby weights. This data will help us address our first research question (3.1).

In the data preprocessing, the time interval since the last other pregnancy (ILLB_R) is binarized into True (the baby is the first child) and False (not being the first child). From the bar chart in Figure (2b), we can see that there are more families giving birth to their non-first child than their first child: 61.77% of the family has already had one or more than one baby, while 38.23% gives birth to their first child. This helps us investigate the effect of being the firstborn on birth weights as it could be a potentially useful feature to predict baby weights.

From the bar chart in Figure (2c), we can see that there isn't a big difference between the number of female babies and the number of male babies, with 51.19% of them as males and 48.81% of them being females. The biological sex of a baby is a useful feature that may influence a baby's birth weight.

5 Analysis of Question 1 - Causal Inference

We consider smoking before a mother's pregnancy as the treatment Z, so CIG_True == 1 means a mother smokes before pregnancy, and vice versa. For the convenience of analysis, we have binarized the number of cigarettes a mother smokes before pregnancy. Any mother who smokes before her pregnancy, no matter how many cigarettes, is classified as 1, while the rest are classified as 0. While this binarization helps simplify our analysis, it may raise questions about

the Stable Unit Treatment Value Assumption (SUTVA). We suggest further investigation into this problem in future studies.

We consider birth weight as the outcome Y. There can be multiple confounding variables in our study, which can influence both the treatment and the outcome. In this study, we consider `PRECARE` and `BMI` as potential confounding variables. For example, `PRECARE` measures the level of prenatal care a mother receives. We believe that mothers who receive higher levels of prenatal care from family and friends are more likely to have better mental conditions and less stress. Therefore, their need for cigarettes will be less than mothers who receive a lower level of prenatal care. Besides, we assume that mothers who receive higher levels of prenatal care are more likely to have more nutritious meals as well, which in return can provide better nutrition to babies, leading to their birth weights being higher. Therefore, `PRECARE` affects both the mother's tendency to smoke and the babies' birth weight and is considered one of the confounders.

We do not find any convincing colliders. Colliders are variables that are affected by both the treatment and the outcome. Since all variables are measured before the delivery, they can not be affected by the outcome, and thus, can not be colliders.

5.1 Randomized Experiment

To analyze the causal relationship between smoking and birth weight, we start with a randomized experiment analysis as the initial try. Randomization ensures that with a sufficiently large sample, all potential confounding variables have the same average value between different groups. Since these variables do not differ by group assignment, they cannot correlate with the treatment and thus cannot confound our study.

We use the Fisher Randomized Test (i.e., Permutation Test) with the simple difference in means as the test statistics

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i)Y_i \tag{1}$$

and the null hypothesis

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i = 1, \ldots, n \tag{2}$$

which says that the treatment and control outcomes come from the same distribution. It is also known as *sharp/strong null hypothesis*. The alternate hypothesis is that the treatment and control come from different distributions.

The observed test statistics is -115. The result is around -115, showing that smoking causes a lower birth weight. Since we have 10000 units with 682 treated units, going through all possible permutations – there are $\binom{10000}{682}$ different permutations – will be too time-consuming. Thus, we use Monte Carlo to approximate the true p-value, and the approximated value is 0. Based on the approximated p-value and the distribution of the test statistics in Figure (3), it is clear that we should reject the null hypothesis. Thus, we claim that a mother's smoking behavior causes a change in her baby's birth weight.

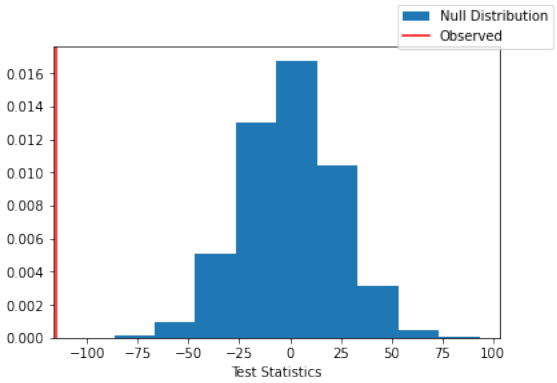


Figure 3: The distribution of test statistics.

However, the permutation test cannot tell either qualitative (*i.e.*, whether smoking causes lower or higher birth weight) or quantitative (*i.e.*, how much birth weight increase or decrease can be caused by smoking) causal effect. Moreover,

the data is a census and thereby an observational study. We next show our application of several observational study techniques for quantitatively estimating the causal effect.

5.2 Observational Study

We use outcome regression and inverse propensity weighting (IPW). In observational studies, confounders are included as control variables. This way, the effect of confounders on the outcome can be shown in the result for us to analyze.

5.2.1 Outcome Regression

In outcome regression, for the estimated coefficient of treatment from OLS to be an unbiased estimate of the average treatment effect, we need to make two assumptions. First, we need to assume unconfoundedness given both confounders `BMI` and `PRECARE`. Second, we need to assume there is a linear model that correctly describes the interaction between the variables. Our linear regression model is

$$DBWT = \tau \cdot Z + a \cdot BMI + b \cdot PRECARE + c \tag{3}$$

where τ is the causal effect. The information of the fitted model is in Figure (4).

OLS Regression Results

Dep. Variable:	DBWT	R-squared:	0.008
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	25.89
Date:	Sat, 10 Dec 2022	Prob (F-statistic):	1.12e-16
Time:	15:03:40	Log-Likelihood:	-77750.
No. Observations:	10000	AIC:	1.555e+05
Df Residuals:	9996	BIC:	1.555e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3142.3565	27.896	112.644	0.000	3087.674	3197.039
CIG_True	-119.8484	22.890	-5.236	0.000	-164.717	-74.979
BMI	6.2407	0.869	7.178	0.000	4.536	7.945
PRECARE	-9.0407	11.608	-0.779	0.436	-31.795	13.713

Omnibus:	1600.050	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4452.710
Skew:	-0.862	Prob(JB):	0.00
Kurtosis:	5.778	Cond. No.	138.

Figure 4: Information of the fitted outcome regression model.

We can see that the coefficient of cigarettes on birth weights is around -119, meaning that smoking an extra cigarette will decrease the baby's weight by 119 grams.

5.2.2 Inverse Propensity Weighting

We also use the IPW estimator to estimate the causal effect. The propensity score $e(X_i)$ calculates the probability that a unit was treated and conditioned on a particular set of confounders (*i.e.,* $e(X_i) = P(Z = 1|X_i = x)$). The IPW estimator is calculated as

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i:Z_i=1} \frac{Y_i}{e(X_i)} - \frac{1}{n} \sum_{i:Z_i=0} \frac{Y_i}{1 - e(X_i)} \tag{4}$$

where n is the number of observations. Using a logistic regression model to fit the propensity scores, we have the IPW estimator being -121. This means that smoking by a mother will cause her baby's birth weight to decrease by 121 grams.

One thing to note is that usually, we need to trim data to include points with propensity scores between 0.1 and 0.9. Anomalies happen if some observations are rare in the treatment group, which may cause IPW to be enormous. However,

since the majority of propensity scores for both the treatment and the control are less than 0.1, excluding these data points will make our data more biased. Therefore, we decide to include them still.

5.3 Discussion

There are several limitations of our methods. First of all, we binarize cigarette consumption into true or false, which inhibits our ability to discover a more accurate causal relationship between mothers' smoking behavior and their babies' birth weights. We make the Stable Unit Treatment Value Assumption (SUTVA). For example, we assume every mother smokes the same type of cigarette, but this cannot be true. Different cigarettes may contain various amounts of nicotine and affect mothers' health and babies' birth weights accordingly. We also assume that units don't affect each other. But in reality, many variables can be related. For example, a woman's age may affect her BMI.

In addition, we believe the frequency of smoking, the average amount of nicotine in a cigarette, and the number of cigarettes during pregnancy are all useful data that aren't included. Having them can help us narrow down the casual question and lead us to a more convincing result.

Overall, we are quite confident that there is a causal relationship between a mother's cigarette consumption and her baby's birth weight since three different techniques (randomized experiment, outcome regression, and inverse propensity score weighting) show that smoking has a negative effect on the baby's birth weight.

6 Analysis of Question 2 - Prediction

We are trying to use various models including GLMs and Nonparametric methods to predict babies' birth weights. Using the EDA results and for the purpose of simplicity, we proposed two features that will be helpful for constructing the model, the Mother's Single Years of Age `MAGER` and the Number of Prenatal Visits `PREVIS`.

`MAGER`: Using our common knowledge, we think that younger mothers are likely to have healthier babies and thus higher birth weights.

`PREVIS`: Using our common knowledge, we think that more prenatal visits mean that the family pays more attention to the pregnancy and thus is likely to have babies with higher birth weights.

6.1 Nonparametric methods

6.1.1 Decision Tree

Why Decision Tree For nonparametric methods like decision trees, we do not need to consider selecting the features. Since the impurity and impurity reduction will do the heavy lifting. The first few depths will select the best features and threshold that decrease the impurity the most, which helps us to see which features are the best for the prediction.

Assumptions Since it is not a probabilistic model, there's no assumption about the data. However, by using decision trees, we assume that the relationship is nonlinear and complex.

Details about selecting model and data Since y is continuous, we will be using regression trees from CART, the DecisionTreeRegressor. We need to realize some mechanics of the model in order to make some assumptions about our data: The model takes in X and y as input, and it tries to iterate through all the possible features in X and iterate through all the values(threshold) that particular feature can be. And calculate the impurity reduction of such a split, where

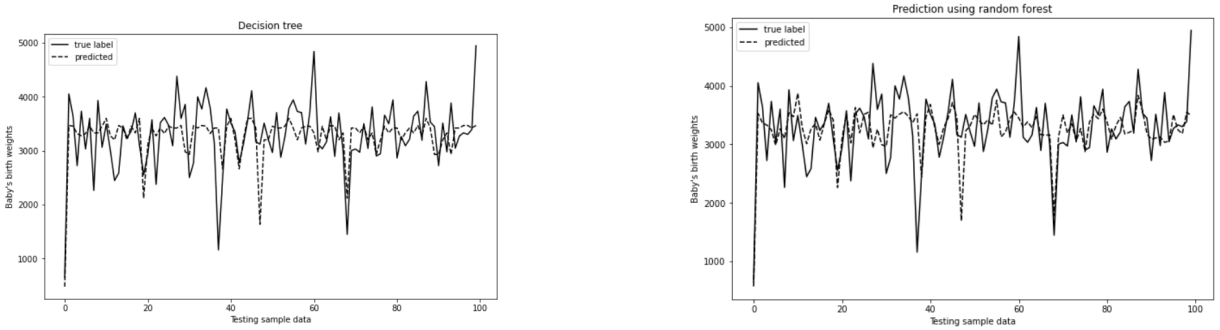
$$Impurity = \frac{1}{N_n(t)} \sum_{i:X_i \in R_i} (y_i - \mu_n(t))^2 \tag{5}$$

and Impurity reduction:

$$\Delta_I(t) = Impurity(t) - \frac{N_n(t^{left})}{N_n(t)} Impurity(t^{left}) - \frac{N_n(t^{right})}{N_n(t)} Impurity(t^{right}) \tag{6}$$

Impurity calculates the weighted sum of differences between the mean and each sample's label squared in a tree node, which basically tells how pure a particular node is. The Impurity reduction calculates the difference between the Impurity before splitting and the weighted sum of the Impurity of nodes after splitting. Since there's mean involved in the model, we need to make all the data continuous, we will achieve this by performing one-hot encoding.

How will you evaluate each model's performance? We will calculate the training and testing score and plot the prediction and observed results in Figure (5a).



(a) Decision tree

(b) Random forest

Figure 5: Results of nonparameteric methods

Decision Tree result As the plot in Figure (5a) indicates, although the score of the model is not the best, it actually performs decently. The overlap between the true label and the predicted label is fair considering they are continuous.

6.1.2 Random forest

Why Random forest Random Forest is one of the Ensemble learning algorithms. It takes n decision tree as its weak learner and learns the dataset, and then at the prediction phase, it asks all the weak learners to vote, and then "summarizes" the opinions and gives a prediction based on that. Since a lot of opinions are involved and the wisdom of the crowd is very strong, the random forest usually performs better than the decision tree does. More specifically, a decision tree is likely to overfit itself with weird splitting. However, for a random forest, the opinions of 200+ decision trees will kill the "outlier" and reduce the variance.

Assumptions Like a decision tree, Random forest is not a probabilistic model, so there is no real assumption since the splitting model doesn't need any requirement for the dataset. However, note that random forest has bootstrapping and bagging. Bootstrapping is a process that randomly samples from the data set with replacement. So the sample should be representative of the entire dataset.

How will you evaluate each model's performance? We will do the same as the decision tree, calculate the training and testing score and plot the prediction and observed results.

Random Forest Result See the plot in Figure (5b).

6.2 GLMs

Our choice of model is Linear Regression.

From EDA, we see that `DBWT`, `MAGER`, and `PREVIS` plots are all roughly normal distributed, and we are predicting real-valued outputs, so the best choice of model here is Linear Regression. Specifically, the inverse link function is Identity, and the likelihood is Gaussian.

This means that we will model birth weight as $W_i \sim N(\beta_0 + \beta_1 M_i + \beta_2 P_i, \sigma^2 I_n)$ where M_i is `MAGER` and P_i is `PREVIS`.

6.2.1 Assumptions

For our model, we are making the following assumptions:

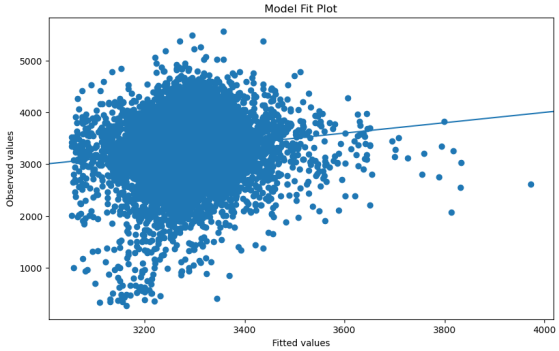
- There is a linear relationship between the response variable (`DBWT`) and explanatory variables (`MAGER` and `PREVIS`)
- Constant Variance

- Birth weights are assumed to be normally distributed (The histogram from EDA shows that this is valid).
- The birth weights are independently distributed.

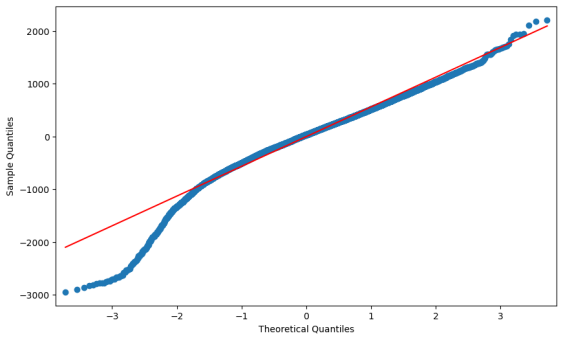
6.2.2 Frequentist Regression

Result Based on the information in Figure (6c), the frequentist model predicted that $W_i = 3014.8158 + 2.0604M_i + 18.5685P_i$, where W_i is the baby's birth weight, M_i is the mother's age, and P_i is the number of prenatal visits. This means that older mothers have babies with higher birth weights, and mothers who go to more prenatal visits also have babies with higher birth weights.

Model Checking First, we plot the model fit plot in Figure (6a). The prediction it provides is fine but there's still a lot of room for improvement, so we might want to try a different set of features to see if it gives better predictions (For example, using `PREG_LEN` and `M_Ht_In`, which are the best features from the decision tree).



(a) True values versus predicted values



(b) QQ plot

Generalized Linear Model Regression Results

Dep. Variable:	DBWT	No. Observations:	10000
Model:	GLM	Df Residuals:	9997
Model Family:	Gaussian	Df Model:	2
Link Function:	identity	Scale:	3.2879e+05
Method:	IRLS	Log-Likelihood:	-77704.
Date:	Mon, 05 Dec 2022	Deviance:	3.2869e+09
Time:	21:30:21	Pearson chi2:	3.29e+09
No. Iterations:	3	Pseudo R-squ. (CS):	0.01708
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	3014.8158	34.136	88.318	0.000	2947.911	3081.721
MAGER	2.0604	1.031	1.998	0.046	0.039	4.081
PREVIS	18.5685	1.454	12.775	0.000	15.720	21.417

(c) Statistical properties

Figure 6: Results of frequentist GLM

Then, we check the Q-Q Plot in Figure (6b) and find it roughly linear. It means that the assumption that the residuals are normally distributed is satisfied.

Uncertainty Quantification The parameter we are estimating is a fixed quantity but our estimate is random because it depends on our data and the data is random. According to the information in Figure (6c), the 95% confidence interval for the intercept is $[2947.911, 3081.721]$, for the coefficient of `MAGER` is $[0.039, 4.081]$, for the coefficient of `PREVIS` is $[15.720, 21.417]$. The standard error for the intercept is 34.136, for the coefficient of `MAGER` is 1.031, for the coefficient of `PREVIS` is 1.454.

6.2.3 Bayesian Regression

Choice of Priors For standard deviance, we set the prior to be exponential(0.01), because this must be non-negative so exponential distribution is a good fit; according to the histogram we plotted in EDA, the SD of the birth weights is

about 6, so the parameter 0.01 makes sense. For intercept and coefficients, we reference the result from the frequentist model and set relatively larger variances so that the posterior will be more dependent on the data instead of the priors.

Result The posterior prediction given by the bayesian model is $W_i = 3015.256 + 2.086M_i + 18.471P_i$ based on the output in Figure (7). This is very similar to the one given by the frequentist model. This again means that older mothers have babies with higher birth weights, and mothers who go to more prenatal visits also have babies with higher birth weights.

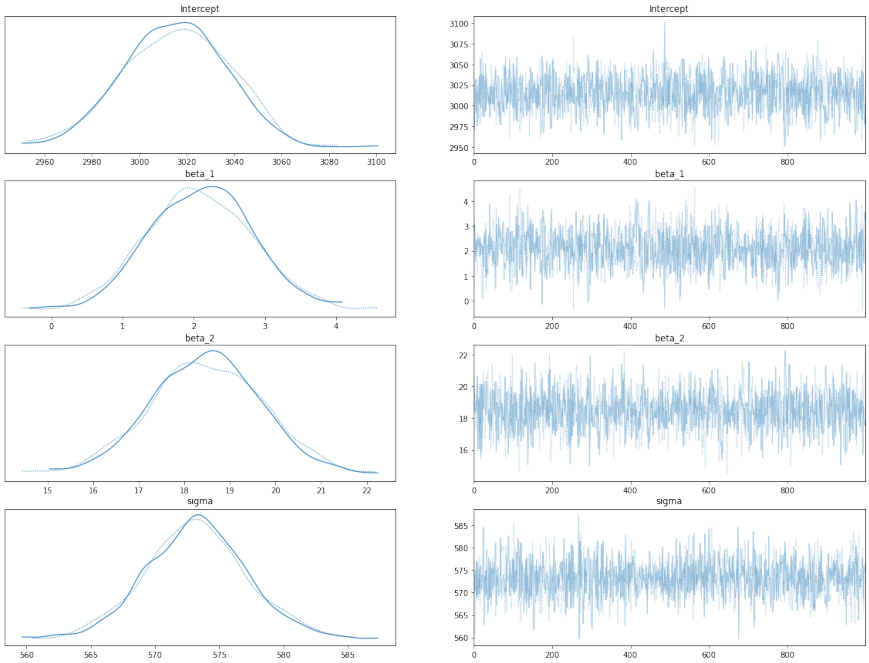


Figure 7: Results of Bayesian GLM

Model Checking We plot the histogram of the PPC samples in Figure (8a), and we observe that the PPC samples have a similar distribution as the data in Figure (8b), so the model is a reasonable fit for the data.

Uncertainty Quantification We see that the highest density interval (another term for credible interval)(8c)for the intercept is [2974, 3053]. The HDI for the coefficient of `MAGER` is $[0.605, 3.41]$. The HDI for the coefficeint of `PREVIS` is $[16.1, 20.8]$.

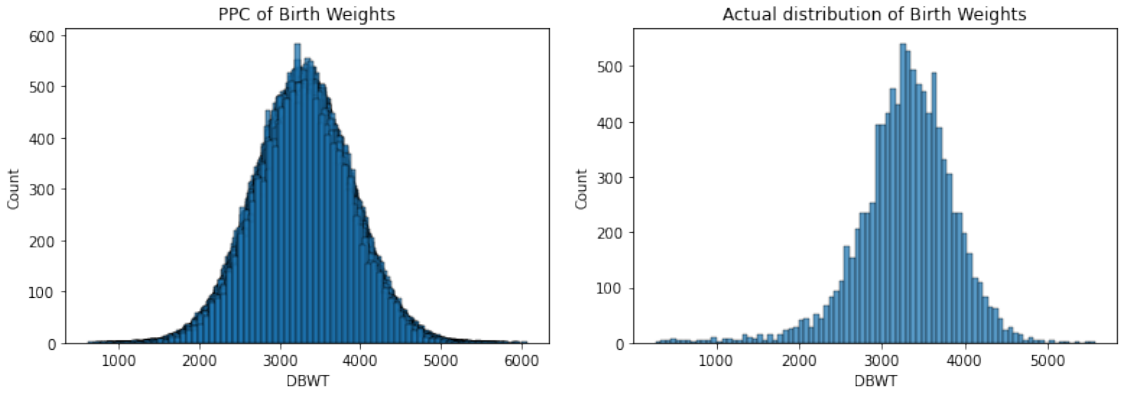
6.3 Discussion

We plot the predictions from both our GLM and nonparametric method in Figure (9). As can be observed in the plot, the random forest model does a better job of predicting the baby's weight. We are fairly confident to apply the random forest model to the future dataset since it follows the true labels plot decently.

For the non-parametric model, the decision tree model did not fit the data very well since it has both a training and validation score of around 0.25. On the other hand, the random forest has done a much better job, with a 0.9 training score and 0.3 testing score. The high training score is because we set the decision tree max_depths very high, between 10-20. However, this act does not make our testing score go lower, because the 200 learners reduce the variance, so the testing score is still better than a depth 4 decision tree. We can see from this fact that random forest does have a lower variance, and if we can adjust the hyperparameter well, the bias will also be lower.

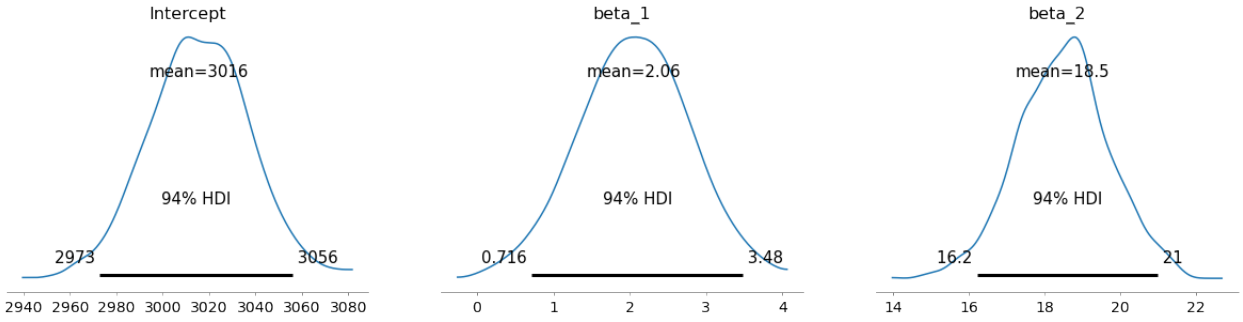
For the GLMs, we can see that the frequentist model did not fit the data very well as we can see from the model fit plot. On the other hand, the Bayesian model did a better job as we see that the PPC plot is quite similar to the original data.

Comparing the frequentist and bayesian model implementations, we see that we do not need to provide any prior for the frequentist model while we need priors for the bayesian model. This means that the frequentist model entirely depends



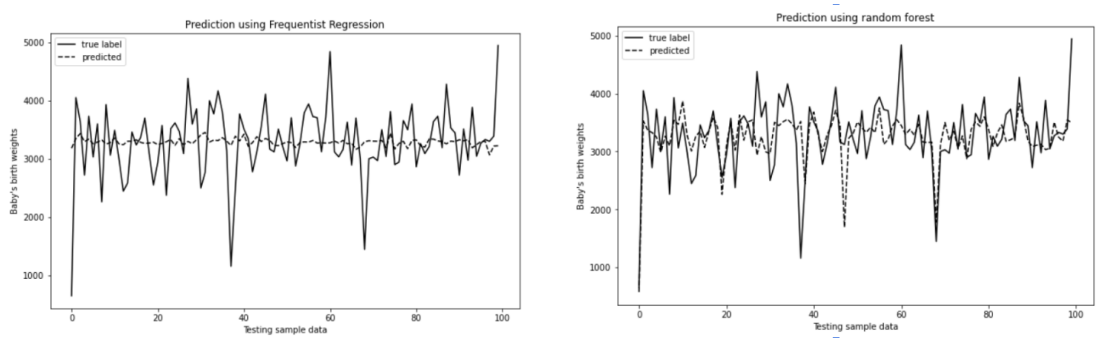
(a) PPC of Birth Weights

(b) Actual distribution of Birth Weights



(c) Highest Density Intervals

Figure 8: Bayesian Model Checking



(a) Prediction using Frequentist GLM

(b) Prediction using Random Forest

Figure 9: Comparing GLM and nonparametric Models

on the data we provided, while the bayesian model also depends on the prior. We have referenced the frequentist model when deciding on the priors for the bayesian model, so both of our models resulted in similar predictions.

From the nonparametric models, we learned that features like the number of pregnancy months (`PREG_LEN`) and the Mother's height (`M_HT_In`) help the most. This makes sense because the longer baby stays in the mother's body, the more likely it can get more nutrition from nutrient transfer and thus heavier, and the taller mother may have a higher chance of having a taller baby, and it will make the weights change.

From the GLM models, we learned that older mothers have babies with higher birth weights, and mothers who go to more prenatal visits also have babies with higher birth weights.

Both decision trees and GLMs have some limitations. The two nonparametric models have a testing score of around 0.3, which does not indicate that they fit the model very well. However, there are limitations to the models. Since these

are both just splitting algorithms, they do not apply fancy combinations and functions like RELU like neural networks, so it may not catch some complex relationships. For the GLMs, we have to first satisfy the assumptions, which do not hold in most scenarios.

To improve our models, we may need features that will have higher impurity reduction. For example, if feature A is 0, then the baby's weight cannot exceed 2500, and vice versa. Since we only used 2 features for the GLM, we can try different sets of features and see which gives the best prediction.

7 Conclusion

In causal inference, we discover that there is a negative relationship between a mother's cigarette consumption and her baby's birth weight. In prediction with GLMs and non-parametric methods, we discover that older mothers, mothers who go to more prenatal visits, longer pregnancy months, and taller mothers have babies with higher birth weights. The result is pretty generalizable since it is derived from data that contains Information about all of the childbirth in the United States in the year 2018. However, we realize that the pandemic has an influence on the worldwide population and that there may be more variables that affect the treatment and the outcome nowadays. What's more, since the data is about the US population, it may not generalize well to populations in other nations.

Based on the result from the causal inference, smoking by a mother has a negative effect on her baby's birth weight. Therefore, to prevent adverse impacts on babies' health conditions, we suggest that families of pregnant females, stores that sell cigarettes, and pregnant women themselves pay close attention. Families should provide more care to pregnant women and look after them so that they can stay away from cigarettes. Stores should not sell cigarettes when asked to be purchased by a pregnant woman. Moreover, pregnant females should be responsible for themselves and their babies and avoid smoking during pregnancy. Based on the result from the GLM models, mothers who go to fewer prenatal visits are likely to have underweight babies, so we would suggest pregnant mothers go to prenatal visits more frequently in order to have healthier babies.

We did not merge different data sources together, but if we can find reliable sources that entail information such as the frequency of smoking or the average amount of nicotine contained in a cigarette, our analysis will be more well-rounded and convincing. For the causal inference question, future research can build on the missing information; moreover, instead of analyzing the effect of smoking before pregnancy, we can study the effect of smoking during pregnancy and how it affects babies' birth weights; we can also study whether that effect will change in different nations or after the pandemic. For the prediction question, future research can use a larger dataset with more features and see if other features give better predictions.

References

[Dataset] Course Staff Team (2022). Data 102 fall 2022

[Dataset] Rapaport, L. (2017). Birth weight may impact intelligence throughout life