STAT151A Project Proposal

Rachel Chen, Wenhao Pan, Richard Shuai

November 9, 2021

Contents

1 Introduction

Baby's mass is correlated with mortality risk and potential future developmental problems. For example, researchers in Denmark 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 weight based on current information. Intuitively speaking, a baby's mass could be predicted by a lot of factors such as the health of the parents, the sex of the baby, the mother's pregnancy records, etc. In this project, we aim to answer the following two questions regarding the baby's weight.

1. What suggestions can we give a pregnant woman, such as smoking less, so that her baby's birth weight is in the healthy range?
2. Given the information about an expecting family, what is our best prediction of their baby's weight? The first question is more related to causal inference, while the second is more related to prediction.

Realistically, a baby's mass should be mostly related to genetics rather than these external factors. Unfortunately, this dataset does not contain any genetic information. Thus, it might not be easy to find satisfying answers to our research questions.

2 Data Description

This dataset was taken from the National Center for Health Statistics, and contains information about all 3.8 million childbirths in the US in 2018. There are 55 columns, so we grouped 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

The user guide on the website contains the detailed explanations of each column. We will use the baby birth weight column (`DBWT`) as the response variable, and all other variables will be used as explanatory variables.

For computational reasons, we will randomly sample without replacement for two sets of 100,000 childbirth records from the original 3.8 million childbirth records. One set, the "prediction dataset", will be used only for prediction-related tasks. We will use an 80/20 train-test split on the prediction dataset, in which we fit a model on 80% of the data and reserve the remaining 20% of the data for testing model generalization. The other set, the "inference dataset", will be only used for inference related tasks. This way, we can considerably relieve the issue of "Post-selection Inference". We will randomly sample another set of 5,000 data points from the overall dataset to plot for easier visualization in exploratory data analysis.

3 Regression Analysis Plan

3.1 Exploratory Data Analysis

First, to ensure the quality of our dataset, we will clean the data by checking for missing values in the dataset. For a given feature, we can deal with missing values by simply deleting data points with missing values from the dataset or by substituting with the mean. With a cleaned dataset, we will perform an exploratory data analysis to visualize the distribution of our variables.

We will plot the univariate distributions of numerical and categorical variables with histograms and bar plots respectively. For the response variable, baby birth weight, we will verify whether the distribution of baby weights follows the normality assumptions required for fitting a linear model. If the distribution does not fit, we will experiment with transformations of the baby weights, such as Box-Cox transformations with different hyperparameters.

We will also create a pairwise correlation matrix with the explanatory variables and the response variable, which allows us to understand the relative importance of explanatory variables for predicting birth weight. With this correlation matrix, we will also be able to detect potential collinearity between the explanatory variables. Finally, we will explore interactions between the explanatory variables in their relationship with birth weight. To do this, we will generate scatter plots or box plots of birth weight against an explanatory variable, color coded by another explanatory variable. If we see a nonlinear relation between the response variable and a numerical explanatory variable from their scatterplot, we will need a transformation to fix the nonlinearity. We will use the information gained from the EDA to gain an intuition for an initial model as a starting point for the model selection process.

3.2 Model Selection

The model selection process for the inference task is different and independent from that for the prediction task, so we describe them separately below.

3.2.1 Inference Model

To assess the significance of the explanatory variables, we will use the incremental F-test through `anova()` or `Anova()` function. Because we are interested in the causal effect of all the explanatory variables, we will not exclude any of them as long as it is statistically significant. After excluding non-significant explanatory variables, we will compute VIF to analyze possible multicollinearity in the remaining variables and make corresponding notes. Note that all the analysis work here is conducted on the inference dataset.

3.2.2 Prediction Model

An overly-complicated model with too many regressors is not desired for prediction, because it tends to overfit on the training data and thereby generalize poorly on an entirely new dataset. Moreover, an overly-complicated model requires us to collect more information than a simpler model does, so it might not be less economical than a simpler model in an actual setting. Thus, we want to simplify our model while maintaining its predictability as much as possible.

As a baseline approach, we will begin by constructing a model for predicting birth weight using ordinary least squares regression. The model will be based on handpicked features using information about relevant explanatory variables and interactions found during the exploratory data analysis. We expect this naive approach to be overly simplistic and to perform poorly, since the handpicked explanatory variables and interaction terms may not be the most informative for predicting birth weight.

To simplify our model, we will use more principled approaches. Because the dataset contains 54 explanatory variables, we would likely see overfitting when using a full model with all the interaction terms between the explanatory variables. To select regressors for a linear model, we will experiment with different methods such as LASSO, forward/backward selection, and leave-one-out cross-validation, using criteria such as AIC and BIC to evaluate the models. Then, we will use the ensemble method to decide the final model, which only includes the main effects. For example, if most methods agree to include variable X, we will include it in the final model, and vice versa. Finally, based on our EDA, we will add interaction terms to this model while obeying the Principle of Marginality. We will use a Type II Anova (`Anova()`) to determine whether these interaction terms significantly reduce the residual sum of squares. Note that all the analysis work here is conducted on the prediction dataset.

3.3 Model Diagnostics

Although the model selection process will return a statistically significant model, we still lack a comprehensive understanding of the actual performance of our model on another dataset. Because we will fit the model on the sampled training data, training on a different subset of the dataset will produce a different model. Thus, the structure of the data majorly determines the performance of the fitted model. For example, unusual and influential data points considerably affect the fitted coefficients of our model. If the empirical distribution of the response variable is skewed instead of approximately Gaussian, then it will lower the quality of conducting inference on the fitted coefficients. Thus, we need to utilize a series of model diagnostics techniques to analyze the structure of the data and thereby assess the actual performance of our model.

First, we will verify our modeling assumptions. There are four major assumptions: linearity, independent noise, constant variance, and normality. We will skip the independent noise assumption because we are not dealing with geospatial and time series data, so we can safely assume that noises are independent of each other. To verify linearity, we will plot the studentized residuals versus different explanatory variables and look for any pattern. Similarly, to verify constant variance, we will plot the residuals versus the fitted values to look for any pattern. Finally, to verify normality, we will use a quantile-comparison plot of the studentized residuals and observe the shape of the quantile line. Depending on the outcome, we will either make appropriate corrections on the data or note the potential issue. Nonetheless, we shall not expect to make too many corrections if we transformed the data well in the EDA process before.

Next, we will detect the unusual and influential data points. There are three types of them: outlier, high-leverage point, and influential point. These data points tend to determine our fitted model more than others, so we need to identify them and make suitable corrections or notes. To detect outliers, we will compute the studentized residuals and conduct testing on them with Bonferroni correction. To detect high-leverage points, we will compute the hat matrix and use its diagonal entries as the leverage measurements of the data points. To detect influential points, we will compute Cook's distance of each data point and use it to measure the influence of each data point. Similar to verifying the modeling assumptions, we will not discard any point merely because its leverage or influence measurement is higher than a pre-specified cutoff or its studentized residual has a significant p-value. Instead, we will make notes on these points because they may suggest some hidden but relevant features.

We will apply the model diagnostic techniques to both inference and prediction datasets.

3.4 Model Interpretation & Prediction

After assessing our model, we will proceed to interpret our inference model to answer our first research question. Specifically, to infer the causal effect of an explanatory variable, we will highly rely on the value of its fitted coefficients. The methods to interpret the coefficients of quantitative and qualitative variables will be different. We may consider using standardized coefficients since some quantitative variables do not share the same scale or unit. Some explanatory variables may have a relatively more significant causal effect than others. However, if they are immutable, such as the mother's race, we will exclude them from answering the first question. We will not compute or refer to the bootstrap coefficients since the size of our dataset is sufficiently large.

Finally, to answer the second research question, we will use our prediction model fitted on the training data to predict the test data. To assess the generalizability and predictability of our model, we will compare the training and test RMSEs. A model of which the test RMSE is not much higher than its training RMSE will be considered not overfitting the training data. If we find out that our model is overfitting the training data, we will use the ridge regression with k-fold cross validation to select the hyperparameter and improve the performance of our model on the unseen data. To assess the variability of the prediction, we will compute the prediction intervals of 5 test data points.

4 Evidence of successfully loading the data

```
# Template source: https://github.com/alexpghayes/rmarkdown_homework_template
knitr::opts_chunk$set(
  echo = FALSE, # don't show code
  warning = FALSE, # don't show warnings
  message = FALSE, # don't show messages (less serious warnings)
  cache = FALSE, # set to TRUE to save results from last compilation
  fig.align = "center" # center figures
)

library(dplyr)
library(tidyverse)
library(ggplot2)

set.seed(0) # make random results reproducible

# Randomly subsample CSV to 100000 records
df <- read_csv("data/US_births(2018).csv", show_col_types = FALSE)
df.subsampled <- sample_n(df, 100000, replace = FALSE)
write.csv(df.subsampled, "data/US_births(2018)_subsampled.csv", row.names = FALSE)
df.subsampled <- read_csv("data/US_births(2018)_subsampled.csv", show_col_types = FALSE)
head(df.subsampled)
```

```
## # A tibble: 6 x 55
##   ATTEND BFACIL   BMI CIG_0  DBWT DLMP_MM DLMP_YY  DMAR DOB_MM DOB_TT DOB_WK
##    <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl>  <dbl>  <dbl>
## 1      2      1  20.5     4  3400      10    2017     2      6   1417      6
## 2      1      1  19.2     3  2880       5    2017    NA      2    427      3
## 3      1      1  31.6     0  1616       5    2017     2      1   2329      1
## 4      3      1  19.1     0  3220       2    2018     1     11   1717      2
## 5      1      1  27.8     0  3751       3    2018     2     11   2141      7
## 6      1      1  24.9     0  3345       2    2018     1     11    540      4
## # ... with 44 more variables: DOB_YY <dbl>, DWgt_R <dbl>, FAGECOMB <dbl>,
## #   FEDUC <dbl>, FHISPX <dbl>, FRACE15 <dbl>, FRACE31 <dbl>, FRACE6 <dbl>,
## #   ILLB_R <dbl>, ILOP_R <dbl>, ILP_R <dbl>, IMP_SEX <lgl>, IP_GON <chr>,
## #   LD_INDL <chr>, MAGER <dbl>, MAGE_IMPFLG <lgl>, MAR_IMP <lgl>,
## #   MBSTATE_REC <dbl>, MEDUC <dbl>, MHISPX <dbl>, MM_AICU <chr>, MRACE15 <dbl>,
## #   MRACE31 <dbl>, MRACEIMP <dbl>, MRAVE6 <dbl>, MTRAN <chr>, M_Ht_In <dbl>,
## #   NO_INFEC <dbl>, NO_MMORB <dbl>, NO_RISKS <dbl>, PAY <dbl>, ...
```