Case Study 1

Ekim Buyuk; Debra Jiang; Katie Tsang; Steven Yang; Bihan Zhuang

9/4/2018

Data Import and Cleaning

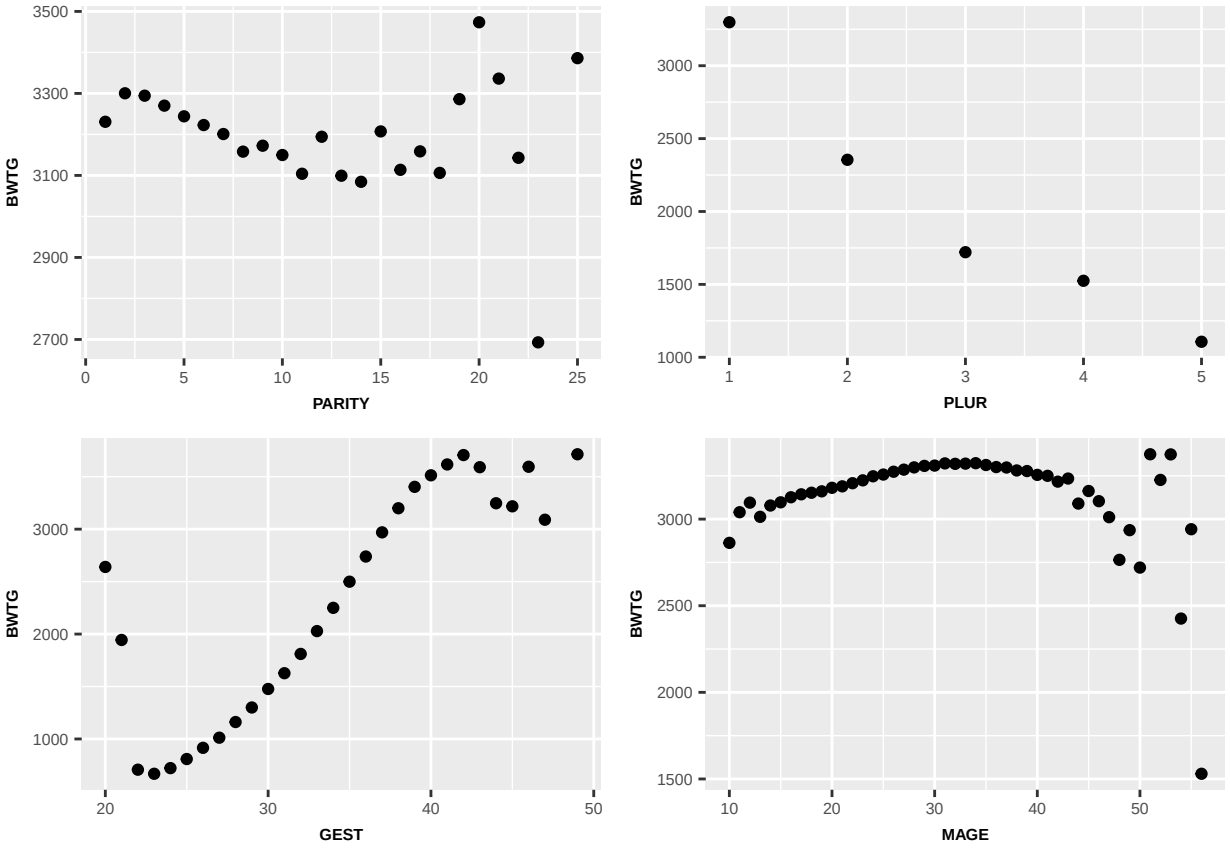
At the start of our data cleaning, we read in the .csv file with birth weight data from North Carolinian mothers in 2011-2016. In the data, missing values were expressed as "9", "99", or "999", so we replaced these occurrences with NAs. We also accounted for outliers by turning illogical values to NAs. For example, we found it implausible for gestation period to take less than 20 months or more than 50 months, and a birth weight of less than 500 grams would be near impossible. Afterwards, we removed the NA values to create a clean data set. We also made sex into a binary variable called "male," in which "male=1" means the baby was male and "male=0" means the baby was female. This was to make modeling, analysis, and interpretation easier.

Factors Associated with Birth Weight

To begin determining which factors are associated with birth weight, we first investigated each variable and made a decision on whether we wanted to treat it as continous or categorical. We then did some preliminary exploratory data analysis using scatter plots, box plots, correlation matrices and histograms. We present our findings on the relationship between birth weight and each of these variables below.

Continous Variables: Age, gestation, plurality, parity

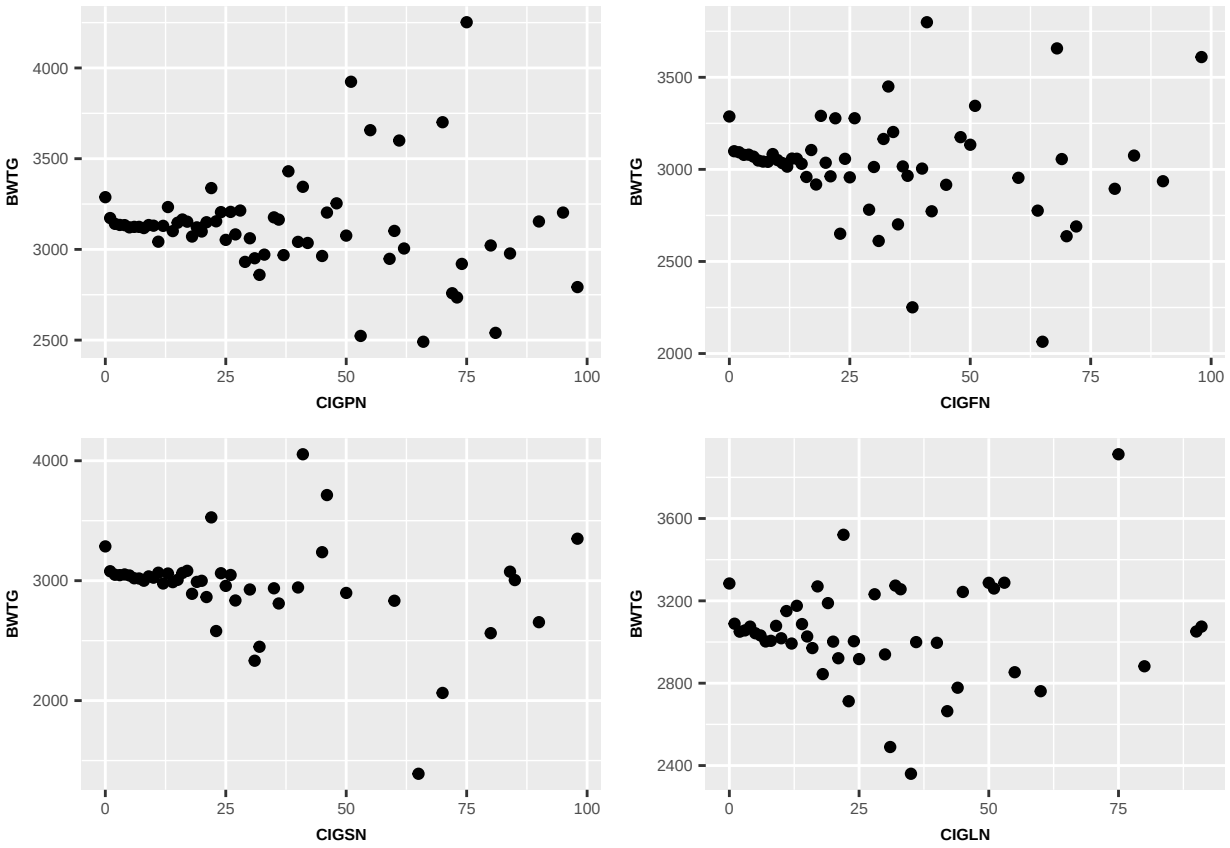
In order to explore the possible relationships between numerical variables and birth weight, we paired and plotted the values in a matrix. We noticed that the birth weight vs. parity and birth weight vs. plurality plots fanned left, suggesting non-constant variance. We could not notice any clear correlation between birth weight and a mother's age. The strongest correlation we saw was in birth weight vs. gestation, which seems to have a slight curve in its positive trend. This would suggest that we need to use a logistic regression for this variable, or transform the gestation variable by squaring it before performing linear regression.



Investigating Categorical Variables: Counties, Race, Hispanic Origin and Cigarettes Smoked

Birthweight vs Average Number of Cigarettes Smoked Per Day

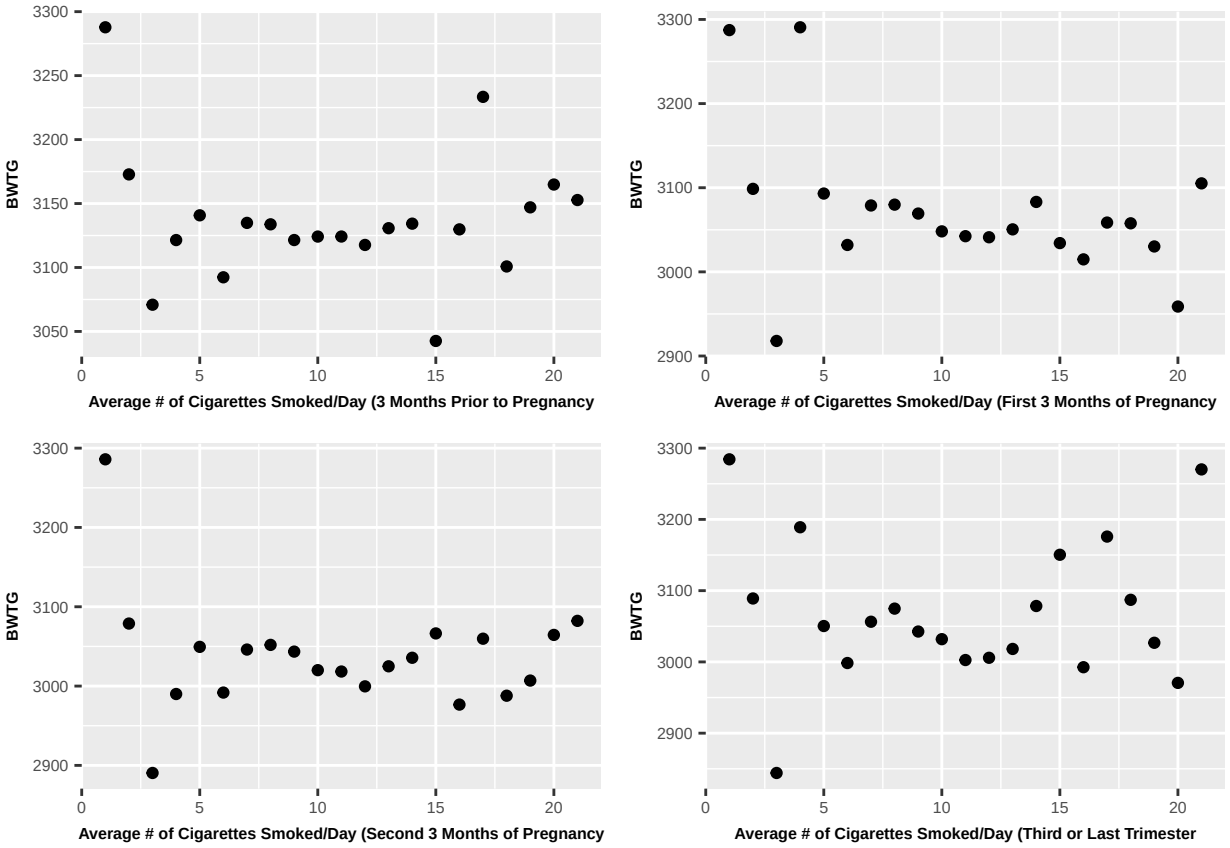
Due to the sheer number of points in our data set, it was difficult to determine trends precisely, especially where points overlapped. Therefore, we took the average birth weight for each value of each numerical variable. We plotted the average birth weight against these numerical values and found some interesting trends. Parity seems to have a slightly quadratic trend, with higher birth weights at low and at high parity, with a dip in the middle. Greater plurality seems to negatively correlate with birthweight, also possibly with a quadratic trend, with the greatest difference in birthweight between having one child and having twins.For gestation, there seems to be a quadratic correlation with birthweight. The graph suggests that we may need to perform a transformation with an order greater than 2. Finally, for mother's age, there seems to be a negative quadratic relationship with birth weight. I.e. very young and older mothers have lower birth weights than mothers in the middle. In conclusion, these graphs of averages suggest that we will need to test out polynomial transformations for all these variables.



Birthweight vs Average Number of Cigarettes (Grouped in Bins)

We noticed a strange phenomenon when we looked at the frequency of reported values for average number of cigarettes smoked per day. It appeared that the number reported was higher for multiples of 10, and specifically very high for the number 20. At first, we were convinced there was some reporting error, but investigating further, we realized that most cigarette packs have either 20 or 25 cigarettes in them. Thus, we realized that responders were most likely responding with the equivalent of one-half, one, one-and-a-half, or two packs a day of smoking in their heads. With this in mind, it would be more logical to look at cigarettes by pack, rather than account for each individual cigarette.

Moreover, when viewing the average birthweight for across each number of cigarettes, we noticed a fan right pattern for all smoking data. This is likely because very few people will smoke an average of, for example, 60 or 80 cigarettes per day, which leads to higher variability. This fanning started especially around 20 cigarettes, which makes sense considering that a typical pack contains 20 cigarettes. So, we binned 20+ cigarettes and re-plotted. We noticed that for all 4 cases of cigarette data, the birth weight for zero cigarettes was much higher, while the data for 1+ cigarettes fell mostly around the same range.



Therefore, in order to work with this variable, we found it illogical to treat cigarettes as continuous. Instead, we will consider two possibilities: 1) We will distinguish between non-smokers and smokers, making smoking a binary variable. 2) We will distinguish between non-smokers, light smokers, and heavy smokers, making this a categorical variable and creating the bins as follows: 0 cigarettes smoked / 1-20 cigarettes smoked (<1 pack) / 21-40 cigarettes smoked (More than 1 pack but less than 2 packs) and so on.

Gender of Child vs Birthweight

Taking a look at the relationship of the gender of the child v. the birthweight, it does appear that male babies have slightly higher weights than female babies. The median male baby is appxomately 3,374 grams v. 3,250 grams for a female baby. Furthermore, conducting a simple difference in means test, we do find a statistically significant difference between the mean of male babies (~3,322 grams) and female babies (~3,206 grams). Thus, we do think that we should include this variable in our linear regression model, and it does make sense to treat as a categorical variable (female, male).

Race of mother vs birthweight

It appears that the average birthweight is consistent throughout all the categories of mother's race, except that African mothers on average give birth to lighter babies than mothers of other races. Because of these observations we decided to do an ANOVA test and result suggests that at least one mean is significantly different from the others, thus there could be a relationship between birthweight and the race of mother. It will be interesting to figure out which race deviates most from others in this aspect. Currently we decided to keep race of mother as categorical predictor of birthweight. Alternatively we may also group by ethnicity to reduce the number of categories in this predictor.

Hispanic origin of mother vs birthweight

Similarly, average birthweight appears to be consistent across different categories of the mother's hispanic origin. An ANOVA test again suggests that there is at least one mean that is significantly different from the others, which means there could be some relationship with between birthweight and the type of hispanic origin of the mother. Again, it will be interesting to figure out which category deviates most from others. We decided to use hispanic origin of mother as a categorical variable in our model. In the mean time, we also acknowledge the number of non-hispanic mothers is much higher than that of hispanic origin. Depending on the model fit, we are open to an alternative – dichotomize this predictor into hispanic and non-hispanic origins.

Year of Birth

Finally, we took a look at the distributions and means of birthweights across different years. Looking at the means, we found no reason to believe that the year of birth had an impact on the birth weight of the mother, and any significance found through an anova test, could probably be attributed to the large sample size. Thus, we chose not to include this variable as a predictor moving forward.

Linear Regression Model

By the end of our exploratory data analysis, we determined that we wanted to work with all of the variables except for year of birth, and we had transformed a few of the numerical variables to categorical variables. Below we present our process for model selection and validataion.

1st model: BWTG_C ~ male + CORES + CIGPN_bucket + CIGFN_bucket + CIGSN_bucket + CIGLN_bucket + GEST_C_center + PLUR_new + MAGE_center + MRACER + PARITY_new

Our first simple linear model yielded a model with an R^2 around .4993 with and R^2 adjusted of .4992. Here, we kept the cigarettes smoked in separate buckets.

We also tried a simple model with cigarettes smoked as a binary variable, which yielded a similar R^2.

Trying More Interesting Models

Squaring gestation model: BWTG_C ~ male + CORES + CIGPN_dichotic + CIGFN_dichotic + CIGSN_dichotic + CIGLN_dichotic + GEST_C_center + GEST_C_center_sq + PLUR_new + MAGE_center + MRACER + PARITY_new

In this model,we achieved an R^2 around 0.4994. Based on this model, we also saw that the variable CIGPN_dichotic was not significant.

Let's Try Some Interaction Variables

We added interaction variables between age and cigarette smoking, age and plurality, age and parity and cigarette smoking and race. Using backward selection based on the P-Value criteria and systematically removing variables who have the highest p-value in each iteration, we end up with the model below, which has an R^2 of 0.4999 and and R^2 adjusted of 0.4998. We used this type of model selection to arrive at a model with high predictability.

Model with Interaction Variables: BWTG_C ~ male + as.factor(CORES) + CIGFN_dichotic + CIGSN_dichotic + GEST_C_center + GEST_C_center_sq + PLUR_new + MAGE_center

+ MRACER + PARITY_new + PARITY_new*MAGE_center + PLUR_new*MAGE_center + CIGFN_dichotic*MAGE_center + CIGLN_dichotic + CIGLN_dichotic*MAGE_center + CIGFN_dichotic*MRACER + CIGSN_dichotic*MRACER

Interpretations of the Model: