IDS 702: Data Analysis Assignment 3

Peining Yang

9/15/2021

Summary

In this assignment, we used logistic regression to model the odds that a baby is born premature by the mother's smoking behavior among other demographic information of the mother. The goal of this study is to identify the variables associated with whether or not a baby will be born premature. Results showed that the mother's smoking behavior, race, pre-pregnancy weight and education level are influential in the odds of a baby being premature. Although we've identified several factors, there are some concerns to statistical significance of the variables and the predictive abilities of the final model, therefore, further analysis could be needed.

Introduction

Previous research has shown that pregnant mothers who smoke cigarettes can lead to many health issues in their babies. In this assignment, we will examine whether a mother's smoking behavior is correlated with pre-term births, which are infants born with a gestational age of less than 270 days. If we do detect a relationship, we will look for the odds ratio of pre-term births for smokers and non-smokers. On top of that, we will also investigate whether the odds ratio is influenced by the mother's race. Lastly, we will also explore other variables associated with the mother and determine whether there are any significant associations to pre-term births.

Exploratory Data Analysis

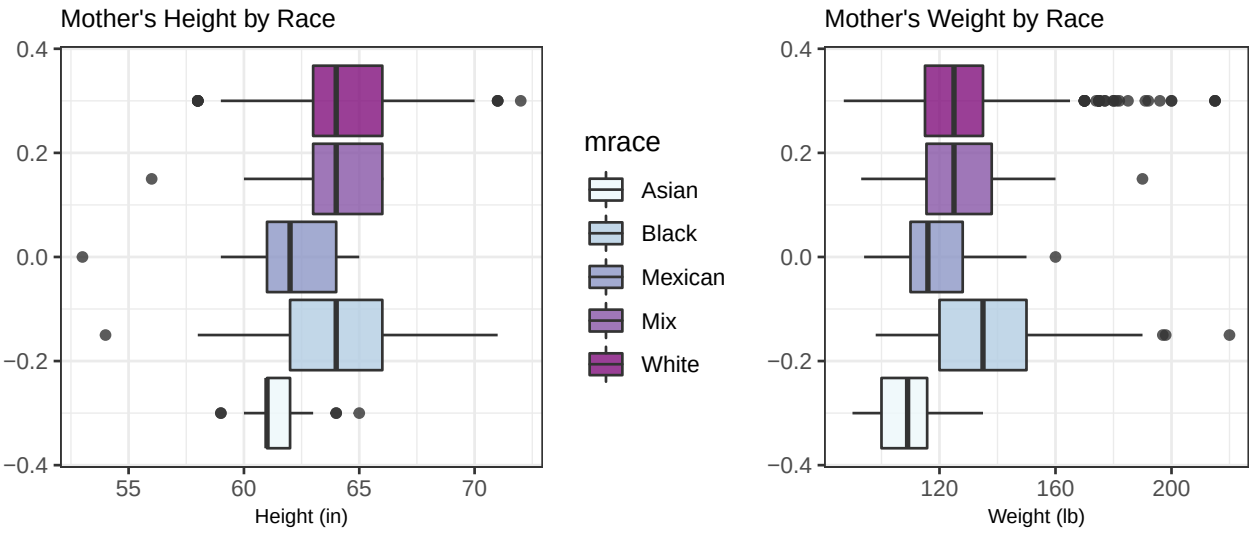
After data wrangling, the final dataset we will use contains 869 observations and 11 variables. The response variable is *Premature*, which is an binary variable with "1" representing a baby that is premature and "0" for one that is not premature. The predictor variables are *parity*, which is total number of previous pregnancies, *mrace*, which is the mother's race, *mage*, which is the mother's age in years, *med*, which is the mother's level of education, *mht*, which is mother's height in inches, *mpregwt*, which is mother's pre-pregnancy weight in pounds, *inc*, which is the family's yearly income and *smoke*, which is whether the mother smokes or not. We centered the *mpregwt*, *mht* and *mage* variables for the sake of model interpretation.

From the table below, we can see that 18.99% of the data have observations with premature babies, which is arguably a low proportion of the data. There are slightly more premature births when the mother is a smoker.

In addition, we also plotted several predictor variables against each other in order to detect patterns. It is important to note that from the boxplots below that both the mother's height and weight are slighly influenced by the mother's race. There are also many outliers in the plots, especially in the mother's weight for white women. We will take note of this as we proceed into the analysis.

Table 1: Exploratory Data Analysis of Premature Births and Mother's Smoking Behavior

Premature	Smoke	Count
Premature	Non-Smoker	77
Premature	Smoker	87
Non-Premature	Non-Smoker	389
Non-Premature	Smoker	316



Model

For our initial model, we included all predictor variables with an additional interaction term of *smoke:mrace* since we are interested in whether the mother's race has an impact on smoking behavior and premature births. As seen from the EDA, we will also explore the interaction effects between *mpregwt:mrace* and *mht:mrace*.

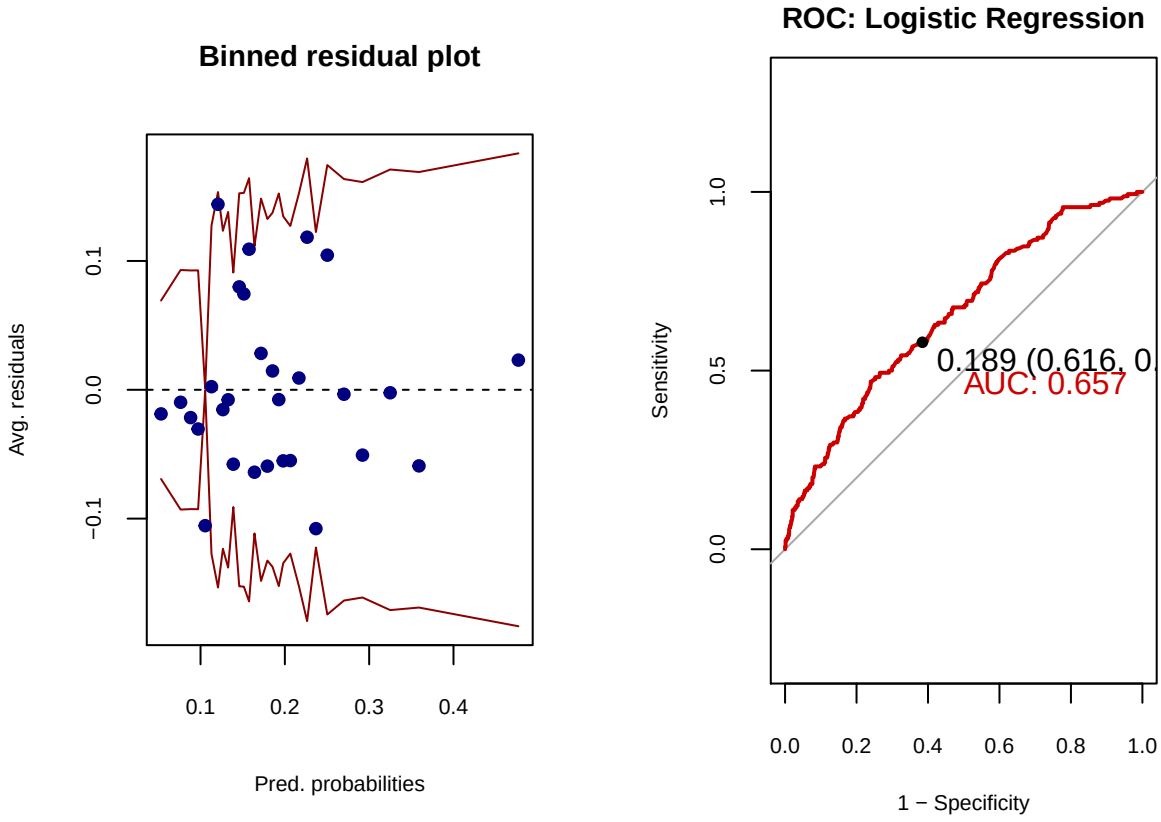
The initial model yielded a Residual Deviance of 774.59 with only *mrace = White* as a statistically significant variable. The binned residual plots show that the residuals are mostly randomly scattered around zero with 3 points lying outside of the 95% confidence interval. The residuals are split equally positive and negative for lower predictive probabilities and there aren't many points in the higher predictive probabilities range.

We then performed model selection with a backward, forward and stepwise algorithm using the Akaike's Information Criterion (AIC). Stepwise AIC produced the model with the smallest number of variables and no interaction terms, which includes *smoke*, *mrace*, *med* and *mpregwt_c*. In addition to these variables, forward AIC also selected *mrace:smoke* and backwards AIC also selected *mrace:smoke* and *mrace:mht_c*. We first eliminated the backwards AIC model as it contains too many variables to be efficient.

The only difference between the stepwise and forwrad models are the interaction term between race of mother and their smoking behavior. However, all the interaction levels yielded a p-value of greater than 0.05, indicating that these levels are not statistically significant. To test whether we should include the interaction terms, we performed a chi-squared test that yielded a p-value of 0.387. This is above the 5% significance level, therefore, we will exclude the interaction terms and select the stepwise AIC model as our final model.

The figures below shows the binned residual plots for the predictive probabilities and the Receiver Operating Characteristic (ROC) curve of the final model. The residuals are randomly scattered around 0 and showed

improvement compared to the initial model as there are now less points outside of the 95% confidence interval. This shows that the logistic regression assumptions are met. Since there is only one continuous variable of *mpregwt_c* in the final model, we do not have to worry about issues of multicollinearity.



Using the mean of *Premature* for the dataset as the threshold, we generated a confusion matrix. The final model achieved an accuracy of 0.609, sensitivity of 0.579 and specificity of 0.617. This means that the model predicted 60.9% of the data correctly. Given that a baby is premature, the model has a 57.9% probability of predicting it as premature. Given that a baby is not premature, the model has a 61.7% probability of predicting it as not premature. As we can see from the second figure, the model achieved an Area Under the Curve (AUC) score of 0.657.

Results

The final model is given by the following equation:

$$Premature_i|x_i \sim Bernoulli(\pi_i) log(\frac{\pi_i}{1-\pi_i}) = x_i\beta$$

where $Premature_i$ is the binary response variable indicating whether the baby is premature or not and x_i includes all predictor variables of *smoke*, *mrace*, *med* and *mpregwt (centered)* as the main effect.

The model coefficients are shown in the table below.

Table 2: Smoking vs. Premature Logistic Regression Model Output

	Odds Ratio	Std. Error	t-value	p-value	95% CI
(Intercept)	0.029	1.015	0.029	0.977	(-2.16, 2.01)
mraceBlack	-0.135	0.45	-0.301	0.764	(-1, 0.77)
mraceMexican	-0.751	0.635	-1.183	0.237	(-2.05, 0.47)
mraceMix	-1.659	1.118	-1.484	0.138	(-4.64, 0.19)
mraceWhite	-0.906	0.408	-2.222	0.026	(-1.69, -0.08)
med1	-0.541	0.949	-0.57	0.569	(-2.41, 1.54)
med2	-0.888	0.941	-0.944	0.345	(-2.74, 1.18)
med3	-0.707	0.993	-0.712	0.477	(-2.66, 1.44)
med4	-1.548	0.956	-1.619	0.105	(-3.42, 0.54)
med5	-1.065	0.959	-1.111	0.267	(-2.95, 1.03)
med7	1.826	1.484	1.23	0.219	(-0.92, 5.26)
mpregwt_c	-0.012	0.005	-2.508	0.012	(-0.02, 0)
smoke1	0.289	0.184	1.567	0.117	(-0.07, 0.65)

Note:

Residual Deviance: 795.91

Discussion

Although the model selection process included the variables in the above table, only the variables of the mother's pre-pregnancy weight and when the mother is white are statistically significant. Keeping all else constant, when the mother is a smoker, we increase the odds for the baby being born premature by a multiplicative effect of 1.34. This means that the odds for mothers who smoke is 33.5% higher than for mother's who do not smoke. This also answers our main research question as mothers who do smoke tend to have a higher chance of pre-term birth than mothers who do not smoke. However, we would like to note that in our final model, *smoke* is not a statistically significant variable as the p-value is above the 0.05 threshold. Looking at the confidence interval, we are 95% confident that keeping all else constant, the range of the odds that a baby is premature will increase by a multiplicative effect of between 0.93 to 1.92. Since the interaction terms between *smoke* and *mrace* were selectd out and deemed statistically insignificant through the chi-squared test, we can conclude that the odds ratio of pre-term birth for smokers and non-smokers do not differ by the mother's race.

There are also other statistically significant variables in the final model. Keeping all else constant, with every unit increase in the mother's pre-pregnancy weight, we increase the odds of for a baby being born premature by a multiplicative effect of 0.988. If the mother also smokes, with every unit increase in weight, we increase the odds by a multiplicative effect of 1.319. Within the race variable, the only statistically significant level is when the mother is white. When the mother is white and is a smoker, the odds that their baby is premature increases by a multiplicative effect of 0.540.

Limitations

We acknowledge that there are still limitations to the analysis that requires further investigation. Since many of the predictor variables are of categorical nature, this inhibited us from exploring its interaction effects because certain interactions just don't have enough observations in the dataset. For example, there is no data for a Mexican mother whose income is above $15,000. There is also no data for Asian mothers whose education level is less than 8th grade or trade school. In addition, our final model only yielded an accuracy of 60.9% and many of the variables are statistically insignificant. I believe this issue could only be solved by having a larger dataset or potentially including the father's information on the analysis.

Conclusion

Looking at the big picture, the purpose of this study is to identify the significant factors of the mother that causes premature births in order to ultimately seek solutions to eliminate this problem. So far, we have examined smoking and its effect on birth weight (previous assignment) and premature births. However, there are many other potentially more severe and long-term damages to the baby by smoking during pregnancies. Therefore, results of our analysis will hopefully raise awareness on this issue. It will also help identify other influential factors such as race, income, education levels, etc., and can help devise effective intervention programs that will be best suited for different communities.