Logistic Regression: Effect of Smoking on Odds of Pre-Term Birth

Raza Lamb

9/21/2021

Maternal Smoking and Preterm Birth

Summary

The goal of this analysis is to examine whether maternal smoking during pregnancy affects whether or not a baby is born premature (before 270 days gestation time). To answer this question, I utilized logistic regression. The final model included smoking status, race, the mother's pre-pregnancy weight, education, and an interaction between smoking and race. Interestingly, the final model indicates that the smoking status of mothers during pregnancy does not effect the probability of having a premature birth. Similarly, the interaction effects between smoking and mother's race are also insignificant.

Introduction

It is now well known that smoking during pregnancy increases the risk of numerous health problems for the child. This can include preterm birth, low birth weight, birth defects, and sudden infant death syndrome (SIDS). However, this relationship was not fully investigated until the 1960s. Since this discovery, massive public health actions have been undertaken in the United States to curb smoking during pregnancy. For this exercise, I investigated a dataset of male births (with extensive data on the mothers) from an observational study in the 1960s. Using this data, I will attempt to confirm what is now known: "Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke?" And, if this is the case, how large is the effect? Additionally, I will determine whether the odds ratio of pre-term birth for smokers and non-smokers is different by mother's race.

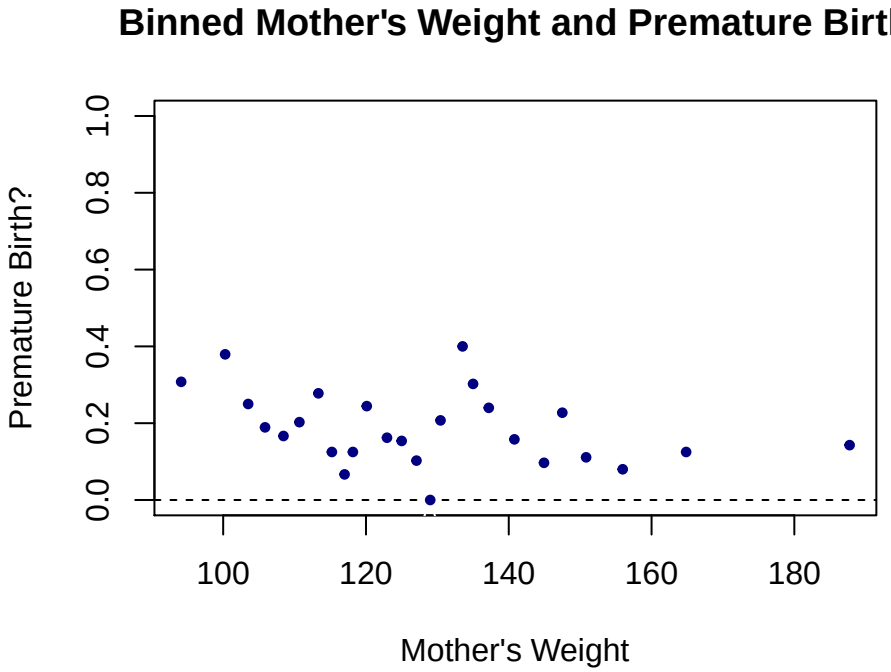
Data

The statistical analysis and modeling that follows is based on the modified data contained in the "smoking.csv" file. This dataset contains observations on 869 male births, and contains 12 variables. I created a binary variable for pre-term birth based on the variable *gestation*, with values less than 270 coded as "1" and all other variables coded as "0." Other variables in the data include: unique ID, date, birth weight, previous number of pregnancies, mother's race, mother's age, mother's education, mother's height, mother's pre-pregnancy weight, income, and smoking status during pregnancy. Date and ID were excluded from the analysis due to relevance. Additionally, I also did not include birth weight—because birth weight is also a potential outcome variable, it would not make scientific sense to include it in this model.

Next, I performed an exploratory analysis on the dataset in order to determine a model and identify potential predictors. During this analysis, I noticed several features of the data that are worth mentioning. First of all, there are 4 numerical variables: previous number of pregnancies, mother's height, mother's age, and mother's pre-pregnancy weight. The categorical variables include mother's education, income, mother's race, and smoking status during pregnancy. One concern I had regarding the categorical variables is that some have many unique levels, which may make interpretation harder, due to small sample sizes within each category. For example, the education variables has 7 different levels, while income has 9. In addition, some categorical variables, such as race, are highly concentrated on one level, with White mothers representing approximately

72% of all observations. When examining the response variable, *premature*, I noticed that the observations are not distributed equally. In fact, 164 births were premature, and 705 births were not premature.

After examining box plots and binned plots for continuous potential predictors and tables for categorical variables, I did notice a few variables that seemed to have an effect on the probability of a mother having a premature birth. Specifically, I noted that predictors I found convincing were mother's race, mother's education, mother's weight, and smoking status. However, when running a chi-squared test on the table for smoking status and premature births, I found that the p-value was greater than 0.05, however it was relatively close. Below is a binned plot for one of the continuous potential predictors, mother's weight, and probability of pre-term birth. There is a decreasing trend visible in the data, meaning that as a mother increases in weight, the probability of having a pre-term birth may decrease.



After graphing initial interactions, I also investigated potential interaction terms. As mentioned earlier, some variables, such as income and education, are very difficult to interpret interactions for, due the large number of categories. And, for other variables, such as race, it can be difficult to visualize changes when certain levels have very few observations (i.e. there are only 25 Mexican and 34 Asian mothers). However, with this in mind, I was able to identify several potential interactions to investigate: height and race, previous number of pregnancies and education, weight and smoking, and height and smoking. In addition to these interactions, I also added smoking and race to the list of interactions, due to its direct relation to the analytical questions for this assignment.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Model

In order to determine my final model, I started with a null model and a full model. In the null model, I included smoking status and the interaction between race and smoking. I wanted to ensure that during model selection, both of these were kept in the model, due to their importance to the main questions of interest. In the full model, I included all potential predictors, and the 5 interactions mentioned in the initial exploratory analysis.

Model Selection

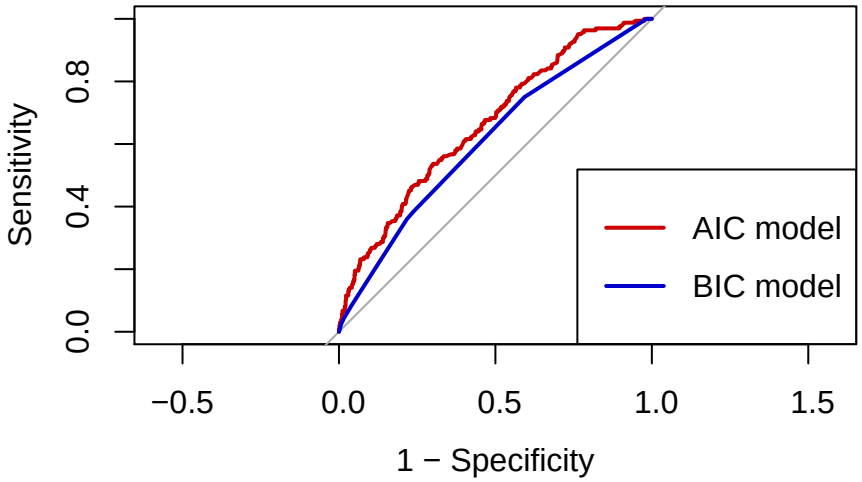
Before fitting the model, the four numeric variables were centered to their means, in order to improve interpretation of the intercept. Then, the full model and null model were fit to the data. Both the null model and the full model follow the general formula below, where y_i is the binary variable encoding whether or not the birth was premature, x_i is a vector of predictors, and β is a vector of coefficients.

$$y_i \mid x_i \sim Bernoulli(\pi_i) \;\; \log(\frac{\pi_i}{1-\pi_1}) = x_i\beta$$

After fitting the null and full model, I performed stepwise selection with AIC and BIC in order to attempt to select a final model. When using BIC, the model selected was simply the null model, including only smoking status, mother's race, and the interaction between those two terms. When using AIC, the final model included smoking status, mother's race, mother's education, mother's pre-pregnancy weight and the interaction between smoking and mother's race. The equation for this model is included below.

$$premature_i \mid x_i \sim Bernoulli(\pi_i) \;\; \log(\frac{\pi_i}{1-\pi_1}) = \beta_0 + \beta_1 smoke_i + \sum_{j=2}^{5} \beta_{2j} 1[mrace_i = j]$$
$$+ \sum_{j=2}^{7} \beta_{3j} 1[med_i = j] + \beta_4 mpregwtc_i + \sum_{j=2}^{5} \beta_5 smoke_i : [mrace_i = j]$$

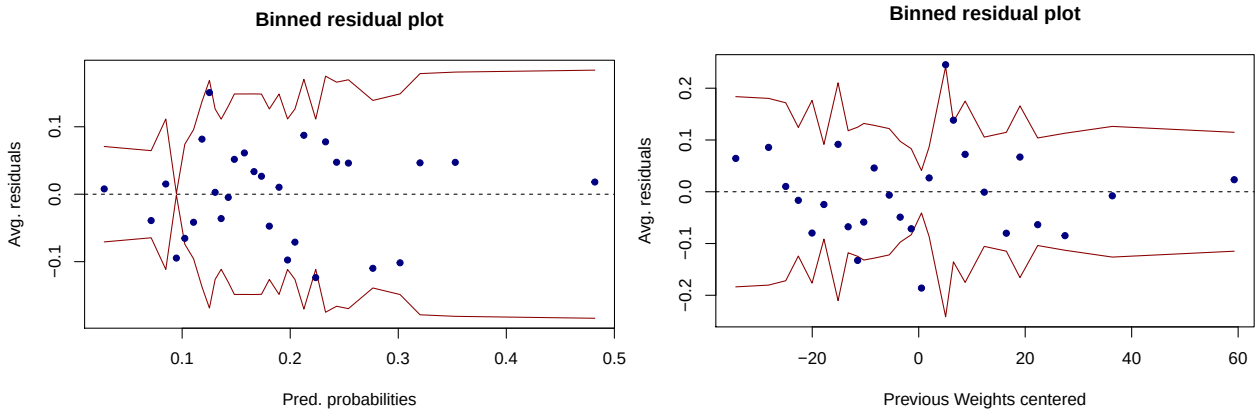
In order to select between these two models, I constructed confusion matrices for both, using the mean value of the binary premature variable as the cutoff. The accuracy of the model selected using AIC was higher, at 0.60, vs. 0.47 for the model selected using BIC. In order to graphically compare the two models, I plotted the ROC curves for both models on the same plot window. The AUC for the BIC-selected model is 0.61 and the AUC for the AIC-selected model is 0.66. Continually, the residual deviance is 817.42 for the BIC-selected model and 790.73 for the AIC-selected model. As is visible in the plot, the red line, corresponding with the AIC model has higher sensitivity and specificity for the entire path. This, along with a higher AUC and a lower residual deviance all indicate that it is likely a better model, thus I selected this model and moved forward to model assessment.



Model Assessment

To assess my model, I analyzed binned residual plots (included below). First I plotted the residuals binned by probability. In this plot, the assumptions for logistic regression are met. There is no visible trend in the residuals, and only 2 points fall outside the 95% confidence band, which is approximately 6.9%, given 29 bins. Subsequently, I plotted average binned residuals vs. continuous predictors. For this model the only

continuous predictor is the mother's pre-pregnancy weight (centered). Again, in this plot, there is no trend visible. However, there are 3 points that fall outside the 95% confidence band, which is 10%. This is relatively high, but still within a reasonable level. The next issue I addressed in model assessment is the potential for multicollinearity. Using the vif() function in R, I was able to determine that there are no issues in this model with multicollinearity.



Model Interpretation

After selecting and assessing my final model, next I moved onto interpretation. The results from this model are included in table 1 below.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8301	0.2033	-9.00	0.0000
smokeSmoking	0.3971	0.2279	1.74	0.0814
mraceAsian	0.8273	0.4947	1.67	0.0944
mraceBlack	1.0552	0.3058	3.45	0.0006
mraceMexican	0.1874	0.6292	0.30	0.7658
mraceMixed	-13.5150	413.9505	-0.03	0.9740
med8th-12th Grade	0.3472	0.2535	1.37	0.1708
medCollege	-0.1567	0.2677	-0.59	0.5582
medHigh School plus Trade School	0.1652	0.3825	0.43	0.6658
medLess than 8th Grade	0.9064	0.9602	0.94	0.3452
medSome College	-0.6680	0.2687	-2.49	0.0129
medTrade School	2.7456	1.1762	2.33	0.0196
mpregwtc	-0.0127	0.0048	-2.62	0.0087
smokeSmoking:mraceAsian	0.3170	0.8451	0.38	0.7076
smokeSmoking:mraceBlack	-0.5652	0.4241	-1.33	0.1826
smokeSmoking:mraceMexican	-0.0325	1.1125	-0.03	0.9767
smokeSmoking:mraceMixed	14.4624	413.9524	0.03	0.9721

Table 1: Final Regression Model

First of all, we can interpret the intercept in this model as the log odds of a premature birth for a White mother of average pre-pregnancy weight with a college education who does not smoke. Converting this to the odds scale, we can see that the odds of having a premature baby given the previously stated baseline is approximately 0.16, or about 1:6. For all of the remaining significant coefficients, I will interpret them on the log scale, that is $\exp \beta_i - 1$ represents the percentage change in the odds of a pre-term birth. First, a black mother has 187% higher odds than a white mother, all else constant, of a pre-term birth. Continually, mother with some college, compared to a mother with only high school education, has 49% lower odds of a pre-term birth, and a mother with trade school education has 1357% higher odds of a pre-term birth. Finally, a one pound increase in a mother's pre-pregnancy weight equates to a 1.3% decrease in the odds of a pre-term birth,

all else constant. However, one note on the coefficient for trade school: there are only four mothers who went to trade school, which means that this coefficient, while statistically significant, is not likely scientifically significant.

Chi-Squared Test

Because I included the interaction between smoking status and race in the null model, when using stepwise selection it was included in the model regardless of the significance. None of the interaction coefficients in the final model are significant, which is indicitive that race does not change the impact of smoking on the odds of premature birth. However, to confirm this, I also fit a model with all the same terms as the final model, but without the interaction term, and conducted a chi-squared test. The results of the test are included in the table below. As is visible, the p-value is greater that 0.05, so we fail to reject the null hypothesis, and conclude that race does not change the effect of smoking on the odds of premature birth.

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	856	795.91			
2	852	790.73	4	5.18	0.2689

Table 2: Chi-Squared Test

Conclusion

Using the final model output from table 1, I can now answer the questions of interest. First of all, I concluded that at the p = 0.05 level, we fail to reject the null hypothesis that smoking does not effect birth weight. However, the p-value is relatively close to 0.05 at 0.08—to fully confirm that smoking during pregnancy does not effect the odds of premature birth, more data could be utilized.To demonstrate a potential range of the odds ratio of premature birth for smokers to non-smokers, we can calculate the 95% confidence interval for the coefficient for slope and exponentiate it. This gives us $(0.95 - 2.34)$, which is the final model's best guess at a possible range for the odds ratio. As is visible, it includes values greater than one (smoking increases odds of premature birth) and less than one (smoking decreases the odds of premature birth). The remaining question of interest is: does the effect of smoking on the odds of pre-term birth differ by mother's race? I can definitively answer here that the data we have does not suggest such an interaction, as initially seen by examination of coefficients Table 1, and then confirmed via chi-squared test. In terms of additional interesting associations, we can see that black mothers, on average, are more likely to have pre-term births. There is also a significant effect for mother's pre-pregnancy weight, with a higher weight leading to a lower chance of pre-term birth. Finally, there are two education levels that are significant to high school, but from a scientific standpoint, I would like to see further data confirming these.

While the final model presented here can answer our initial questsions, there are important limitations of the dataset to consider when interpreting the results. First of all, the dataset only contains male babies (births). This is an incomplete dataset when comparing the population (i.e. all births), and it is therefore difficult and even potentially inappropriate to make inferences about the population. Secondly, while one of our main questions of interest refers to race and smoking, we only really have sufficient numbers of Black and White mothers for analysis. In order to feel more confident about the interaction between smoking and race, larger number of Mexcian, Asian, and Mixed mothers would be necessary. Finally, one last limitation is that this model could be potentially improved significantly with inclusion of complete paternal data, which is not at our disposal with this dataset.