Assignment 3

Summary

Drug use behavior during pregnancy has been found to be significantly influential to the health of a new born. In this report, we look at one particular drug use (smoking) and whether the newborn is born premature (i.e., gestation age < 270 days). In order to investigte the relationship, we adopt a logistic-regression model, using whether or not the mother smokes, mother's ethnicity, mother's pre-pregnancy weights (in natural logarithm scale), as well as the interaction between race and smoke as predictors for predicting whether or not newborn will have gestation age less than 270 days.The analysis finds that while smoke and race are not statistically signifincat predictors, mother's pre-gregnancy weight in fact has considerable predictability. The model has an accuracy performance of 0.7813, AUC of 0.664.

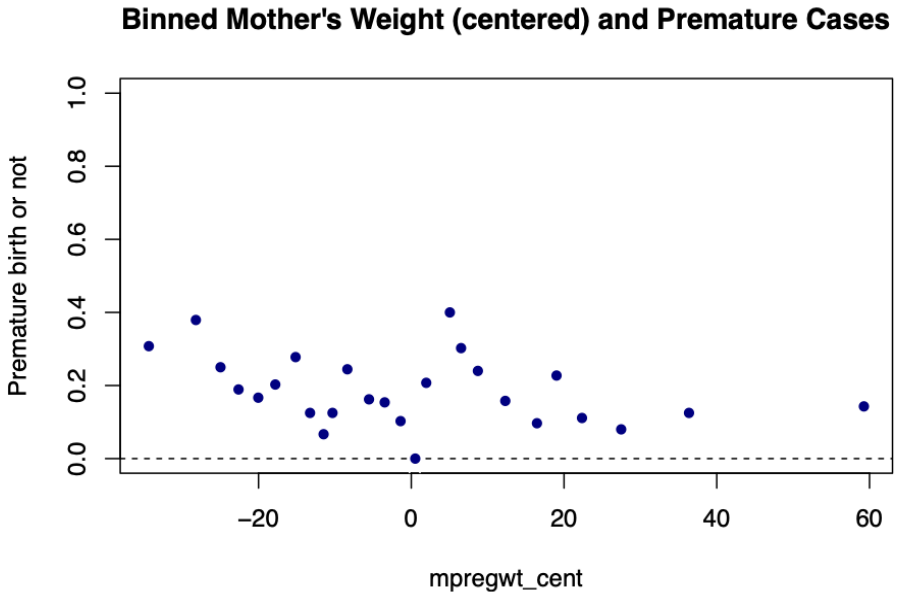
Introduction

In this analysis, we are interested in several questions: do mothers who smoke tend to deliver newborn before expected term than mothers who do not smoke? What are some other factors that play a role in influencing whether or not the baby will be born premature, or by interacting with other factors such as mother's race? If the odds ratio of a mother who smokes giving premature birth is different than those mother who do not smoke,what would be the range of differences? By investigating these questions, we hope to gain some insights about how cigarette usage during pregnancy can affect the new born, as well as other factors that can potentially exert an influence on the baby.

Data

In this dataset, variables such as mother's education and race are ordinal numerical variables. To improve the interpretability, we have factorized these variables into different levels, and collapsed race into White, Mexican, Black, Asian, and Mix. Income, as is coded into numerical levels, is factorized, Besides, we also mean-centered continuou variables such as mother's pre-pregnancy weight, age, and height.

Our EDA shows that, first of all, the distribution of premature birth is uneven, with a small amount of data in `premature = 1`, and most of the data in the category where the newborn is not delivered pre-term. Besides, information of different ethnicities are also skewed, where White has more than 600 observations, and Mix has only 15. THis is very likely to result in inflated coefficient in model. Then plotting premature against centered mother's pre-pregnancy weight using binned plot, we find that the distribution of the points follows a 'wave-y' pattern with several peaks and lows in probability across the centered-age interval.



Other binned plots such as the one with premature against centered age, premature against parity, and premature against centered height show no salient pattern or relationship between the predictor and the response variable.

For categorical variables such as race, mother's education, income, and smoke, we have used contingency table to investigate their relationship with premature. The table depicting the conditional probability of having premature newborn between mothers who smoke and those who do not smoke shows that the probability does not vary significantly between the two groups. The probability of baby being pre-term between different education groups does not vary excpet for level 7 which has a probability of 0.75. For babies whose mothers are in different income levels, the probability of being premature does not vary significantly. When looking at different ethnicity groups, we find that Asian has the highest pre-mature probability which is 0.32, whereas Mix has the lowest at 0.07.

Next, we look at the potential interaction effects that certain predictors may have. First, when looking at the relationship between smoke and premature by race, we find that the relationship between premature and smoke is significant only among people who are White, which is confirmed by chi-squard test with a p-value $< .05$. Then to explore the relationship between mother's age and premature by mother's education level, we use boxplot to illustrate and find that the pattern is similar across different levels of educatiion excpet in level 0 and level 3. For mothers whose education is in level 0, the age of those who give birth to expected-term babies tend to be higher, whereas for mothers whose education is in level 3, the age of those who give birth to pre-term babies tend to be higher. If we look at the relationship between parity and premature between smoker and non-smoker, we can find that for mothers who do smoke, delivering pre-term or at normal-term does not vary in parity. For those who do not smoke, on the other hand, mothers who give birth to normal-term babies tend to have a higher median parity.

Model

First of all, we adopt a full logistic regression model that includes all variables excluding interaction terms.

$$log(\pi_i/1-\pi_i) = \beta_0+\beta_{1:2}x_{smoke}+\beta_{3:7}x_{mrace}+\beta_{8:14}x_{med}+\beta_{15}x_{mage}+\beta_{16}x_{parity}+\beta_{17}x_{mpregwt}+\beta_{18:27}x_{inc}+\beta_{28}x_{mht}$$

First of all, by running binned residual plots, we find that the overall fitted model and residual share a relationship that seems to be non-linear. Specifically, when examining the binned residual against `mpregwt_cent`, we find a similar non-linear pattern. Given that the number of normal-term greatly is much greater than the number of premature in the dataset, we'd set the threshold at.3 to avoid inflating the number of predicted negatives. This model have an accuracy of 0.77 and ROC value of 0.6659. In order to alleviate the non-linear pattern, we decide to transform mpregwt into logarithmic scale. Fitting the new variable again, we find that the general non-linear pattern is alleviated, although not entirely eliminated, and we decide to keep this variable in the model.

By far, only two of the coefficients show stastical significance: `mpregwt_log`and `mrace_factorwhite`. We can use stepwise model selection method to help with selecting variables that are more meaningful for the model as a whole. After stepwise testing using AIC, we decide to predict premature with variables including mother's race, smoke, mother's education, and mother's pre-pregnancy weight. After fitting this model, we use binned residual plots to examine our model validity, and it turns out that the distribution of residuals are close to being random. Confusion matrix shows that this model has 0.78 accuracy with an ROC value of 0.658. Doing an anova test with the null model give s p-value $< .05$, showing that our model is significantly better than the null model.

After exploring the main effects, we should then look at interaction terms that are potentially meaningful for model improvement. Three interactions terms are included in the model: race and smoke, education and age, parity and smoke. However, after fitting the three interaction terms and comparing the model with our model with only main effects using anova test, we find that p-value $> .05$, meaning that none of these interaction terms enhanced the model. Given that the interaction between race and smoke is at our research interest, only this interaction term will be included in the final model.

Thus, we have arrived at our final moel:

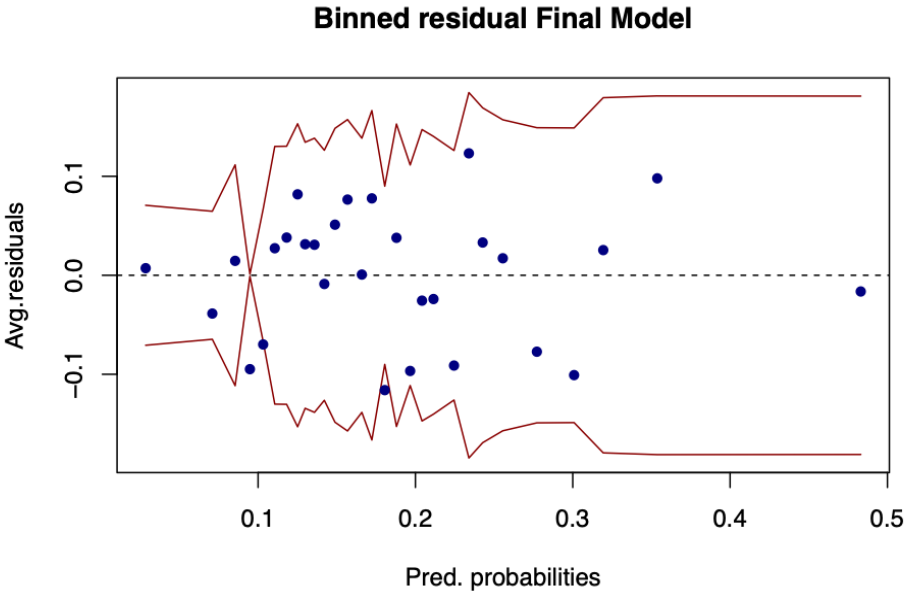
$$log(\pi_i/1-\pi_i) = \beta_0 + \beta_{1:2}x_{smoke} + \beta_{3:7}x_{mrace} + \beta_{8:14}x_{med} + \beta_{16}x_{mage} + \beta_{17}x_{mpregwt} + \beta_{18}x_{smoke} : x_{mrace}$$

term	estimate	std.error	statistic	p.value
(Intercept)	8.1435636	3.1448193	2.5895172	0.0096111
smoke	0.7127635	0.8156578	0.8738512	0.3821993
mrace_factorblack	0.2515925	0.5457956	0.4609647	0.6448240
mrace_factormexican	-0.6252034	0.7643979	-0.8179031	0.4134125
mrace_factormix	-14.3203317	413.3752161	-0.0346425	0.9723648
mrace_factorwhite	-0.8005952	0.4965021	-1.6124708	0.1068595
med_factor1	-0.5741761	0.9638448	-0.5957143	0.5513661
med_factor2	-0.9187803	0.9601530	-0.9569103	0.3386125
med_factor3	-0.7504738	1.0143218	-0.7398774	0.4593744
med_factor4	-1.5875093	0.9739523	-1.6299662	0.1031086
med_factor5	-1.0706230	0.9768165	-1.0960329	0.2730644
med_factor7	1.8446264	1.5061224	1.2247520	0.2206687
mpregwt_log	-1.7036746	0.6296516	-2.7057418	0.0068152
smoke:mrace_factorblack	-0.8795992	0.8921148	-0.9859709	0.3241474
smoke:mrace_factormexican	-0.3528315	1.3636877	-0.2587333	0.7958410
smoke:mrace_factormix	14.1210611	413.3776509	0.0341602	0.9727494
smoke:mrace_factorwhite	-0.3191232	0.8460064	-0.3772113	0.7060166

This model predicts that the odds of an Asian mother who does not smoke, received education less than 8 grades with weight before pregnancy equal to log of the average weight, giving birth to a premature baby is 126500%. In this model, only the logarithmic scaled mother's pre-pregnancy weight has a statistically significant coefficient, which predicts that for every unit incease in the logarithmic increase in weight, the odds of the child being premature is predicted to increase by 6.70%.

In order to assess the model, we can use binned residual plots and confusion matrix, ROC curve, as well as vif.

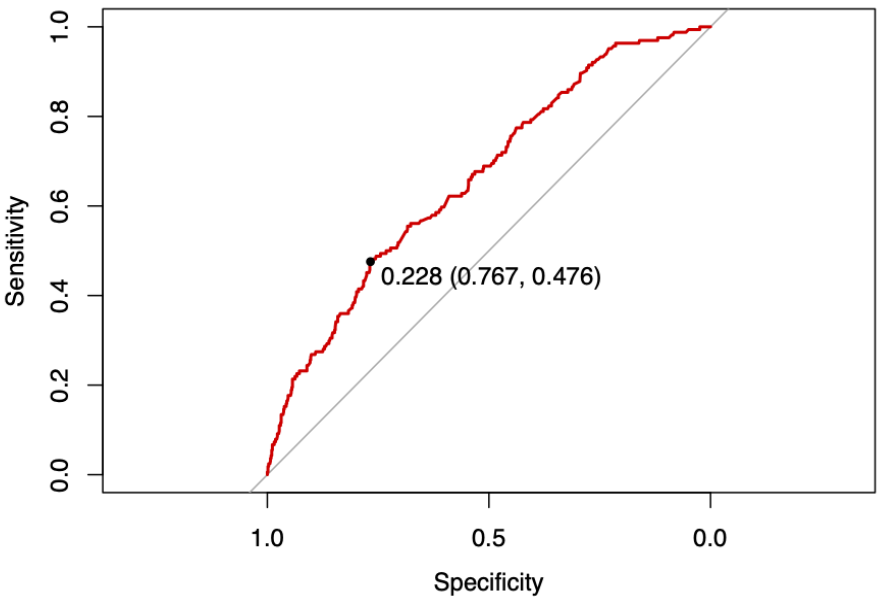
Firstly, the binned residual plot shows no pattern between the residuals and the predicted value. When plotting the residuals against `mpregwt_log`, still no relationship is observed. There are about two observations are outside the 95% interval, but given their number and distance are not significant, we'd assume that they do not exert too great an influence on the model.



We then compare this model with the null model using anova with Chi-squared test, finding that our model is significantly better than the null model with p-value $< .05$.

Then after confusion matrix, we observe that our model has accuracy of 0.7813, which is an enhanced performanced compared to the initial model with only main effects of all variables. Specifically, the final model has a sensitivity level of 0.25 and specificity level 0.904, signaling that our model is more prone to predict negative, which may be caused by the uneven number of premature and normal term in the dataset.

To make sure our variables are not redundant and do not share linear relationship, we use vif measure and find that none of the varaibles have vif value larger than 10, meaning that we can be confident that our model is not influenced by multicolinearity.



The ROC curve with the best balanced point shows that our model can reach sensitivity of 0.476 and 1-specificity of 0.767, with AUC equals 0.664.

Given these results, we conclude that mothers who smoke do not tend to have higher chances of pre-term birth than mothers who do not smoke, given that smoke is not a statistically significant predictor in this model. A 95% confidence interval for the odds ratio of pre-term birth for smokers and non-smokers would be from 0.06 to 6 * 10 ^6. Given that the interaction between smoke and race is not statistically significant in our model, we do not have sufficient evidence to support the statement. In effect, given that our model is trained from the dataset where data of premature births are disproportionately less than those of normal births, and that observations of mother who are White is dozen times more than those of minority mothers, which can adversely affect the predict power of our model. However, given this high specificity value, this model can be useful for predicting newborns who will not be born premature.

Based on our final model, it is interesting and surprising to find that the relationship between premature birth and mother's pre-pregnancy weight in log scale is statistically significant. It offers some information about how mother's pre-pregnancy health can affect the baby, and possibly shed light on obesity and it's influence the newborn. It might even help debunk some myths that mothers should eat more and gain weight in order to prepare for a healthy baby.

Potential Limitation

This analysis helps answer one of the important questions in pregnancy healthcare: will the mother's smoking behavior influence the new born? Based on our model, we do not have definitive answer. In this context, we have found that mothers who have higher pre-pregnancy weight have higher odds of giving premature birth. Additionally, smoking and it's effect does not vary across different race. However, this study is still limited: our model has a a specificty that is much higher than sensitivity, meaning that it is prone to predict negative. Thus, if given a new dataset where the distribution of premature and normal births are more even, this model may not be helpful at all.