Data Analysis Assignment III

Dean Huang

9/11/2020

Summary

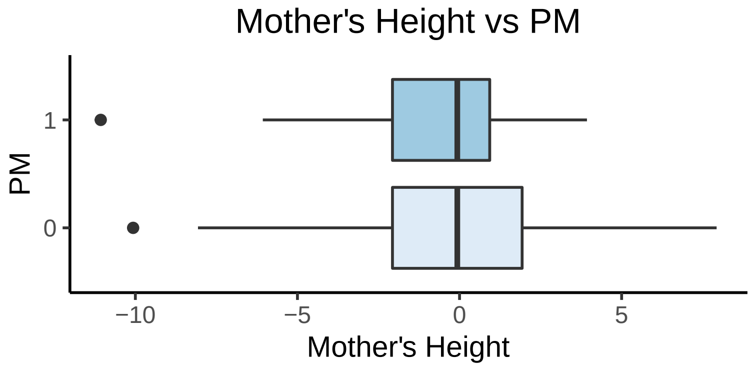
The goal of this report is to address the question "Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke?". In addition, we are trying to find out if odd ratio of pre-term birth for smokers and non-smokers differs by mother's race. Boxplots and binned plots will be used to analyze the association between binary response variable and numeric predictor variable. On the other hand, joint probability table and conditional probability table will be used to analyze the association between binary response variable and categorical predictor variable. Binned residual plot will be used to evaluate the overall fit of regression model, and check if the function of numeric predictors is well specified. Chi-squared test will be used to compare the deviance of null model and new model. Confusion matrix and ROC curve will be used to validate the performance of the model through calculation of sensitivity, specificity, accuracy, and AUC curve. VIF will be used to calculate the multicollinearity of the function. The outcome of the study shows mothers who smoke tend to have higher odds of experiencing premature birth than mother who do not smoke. However, there is not enough evidence to conclude that the association between smoking and premature birth differs by mother's race.

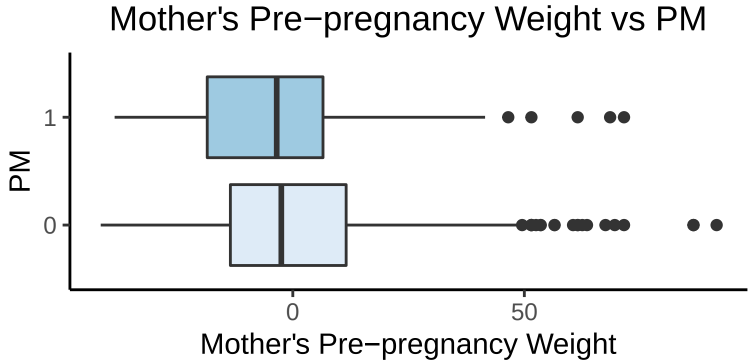
Introduction

Since the goal is to find out the association between the smoking status of mother and pre-term birth, variables that are either irrelevant or insignificant to our question will be removed. These variables include id, birth, gestation, bwt, drace, dage, ded, dht, dwt, marital, time, and number. The response variable we are interested in looking to is premature, and the predictor variables we are interested in looking to are parity, mrace, mage, med, mht, mpregwt,income, and smoke. To facilitate the EDA, data modeling, and model validation processes, we will remove data rows with one or more missing values. After completing data modeling and model validation, the estimated range for the difference of odds ratio in pre-term birth for smokers and non-smokers will be calculated from the final model. Interesting associations with odds of pre-term birth will also be highlighted. The experiment will begin with EDA with the goal of checking the association of predictor variables and response variable, and highlight the preliminary concerns for the response and predictor variables. Next, preliminary logistic model fitting will be performed (excluding interactions and transformation) to understand the significance of coefficients for each selected predictor variables. Binned residual plots will be utilized to assess the overall fit of regression model and check if the function of predictors is well specified. Next, preliminary model validation will be performed to understand the fit of model for making prediction. Confusion matrix and ROC curve will be utilized to validate the performance of the model through calculation of sensitivity, specificity, and accuracy. Through careful analysis of the results from EDA, preliminary model fitting and preliminary model validation, transformation will be performed, and interactions will be added to improve the fit of the model. Logistic model fitting and model validation processes will be performed again to justify each modification made to the model. In addition, chi-squared test will be implemented to compare the performance of new model to the original. Lastly, stepwise function will be performed to find the optimal model with the lowest AIC score. To ensure the final model fulfill the assumptions of logistic regression, model validation will be performed again. Last but not least, VIF will be used to check for any multicollinearity for the final model.

Data

Since there are no continuous predictor variables, binned plot is not suitable to use for this EDA. Boxplot will be used to determine the significance of discrete predictor variables, which include parity, mother's age, mother's height, and mother's pre-pregnancy weight. According to the boxplot of parity vs premature, the distribution and median for both 0 and 1 binary response appears the be the same; therefore, parity appears not to be a significant predictor variable for calculating premature. The boxplot of mother's age vs premature shows the distribution for both 0 and 1 binary response appears the be the same; however, the medians for both binary outcomes are different (more testing is needed to determine if mother's age is significant for predicting premature). The boxplot of mother's height vs premature shows the median for both 0 and 1 binary response appears the be the same; however, the distribution for both binary outcomes are slightly different (more testing is needed to determine if mother's height is significant for predicting premature). The boxplot of mother's pre-pregnancy weight vs premature shows the median for both 0 and 1 binary response appears the be the same; however, the distribution for both binary outcomes are slightly different (more testing is needed to determine if mother's pre-pregnancy weight is significant for predicting premature). Since mother's race, mother's education, income, and smoking status are categorical variables, conditional probability table and chi-squared test will be utilized to determine the significance of these predictor variables. The conditional probability table for mother's race shows the difference in distribution of conditional probability for some races like mix and white. In addition, the chi-squared test reaffirms the observation by having a p-value of 0.0036 (more testing is needed to determine if mother's race is significant for predicting premature). The conditional probability table for mother's education shows the difference in distribution of conditional probability for some races like mix and white. In addition, the chi-squared test reaffirms the observation by having a p-value of 0.0005 (more testing is needed to determine if mother's education is significant for predicting premature). The conditional probability table for mother's income shows no difference in distribution of conditional probability for different income levels. In addition, the chi-squared test reaffirms the observation by having a p-value of 0.9. The conditional probability table for smoke shows some difference in distribution of conditional probability for smoker vs nonsmoker. The chi-squared test has a p-value of 0.07, which is above 0.05 but below 0.1 (more testing is needed to determine if smoke is significant for predicting premature).





Model

After conducting EDA, the next step is to perform preliminary logistic model fitting and model validation for the predictor variables picked for this study (excluding transformation and interaction), which are parity, mrace, mage, med, mht, mpregwt, income, and smoke. The summary table of the preliminary model shows only race (black) and mpregwt are significant, the rest of the predictor variables have a p-value above 0.05. All points of binned raw residuals versus predicted probabilities plot are within the standard error bound and the overall plot appears to be random. For parity, the binned residual plot appears to be random with all points except one within standard error bound. There are not a lot of points in the graph so one point not within standard error bound is significant. For mage, the binned residual plot appears to be random with all points within standard error bound. For mht, the binned residual plot appears to have a trend going up and down with all points within standard error bound. For mpregwt, the binned residual plot appears to be random with all points except three (95% of points still in the bound) within the standard error bound. According to the confusion matrix and ROC curve of the preliminary model, the optimal specificity and sensitivity is (0.620,0.622), the accuracy is 0.62, and AUC is 0.667, which are not ideal. Improvement of preliminary model is required.Through analysis of results of binned residual plots, transformation for mht appears to be necessary because the plot appears to have a trend going up and going down. Both log transformation and quadratic transformation are performed on mht, and the binned residual plot for quadratic transformation produced better result. After performing quadratic transformation, the shape of binned residual plot for mht appears to be more random. However, the confusion matrix and ROC curve are the same with no improvement on AUC, and optimal sensitivity and specificity. Also, the square of mht variable has a p-value of 0.8, which is way above 0.05. There is a possibility that binned residual plot of mht might be exhibiting pattern because there is simply not sufficient data for each bin.

The next step will be to investigate potential interactions in logistic regression. Since the main purpose of this study is to find the association of mother's smoking status and pre-term birth, interaction of smoke and each numeric variable will be investigated. Binned plots for mpregwt vs premature by smoke and mage vs premature by smoke shows a difference in distribution between mother who smoke and mother who does not smoke. To answer the question whether ratio of pre-term birth for smokers and non-smokers differs by mother's race, the interaction of smoking status and mother's race is also investigated. According to Pearson's Chi-squared test results, the p-value for smoke vs premature for white ethnicity is lower than 0.05. Therefore, mrace*smoke appears to be a potential limitation.

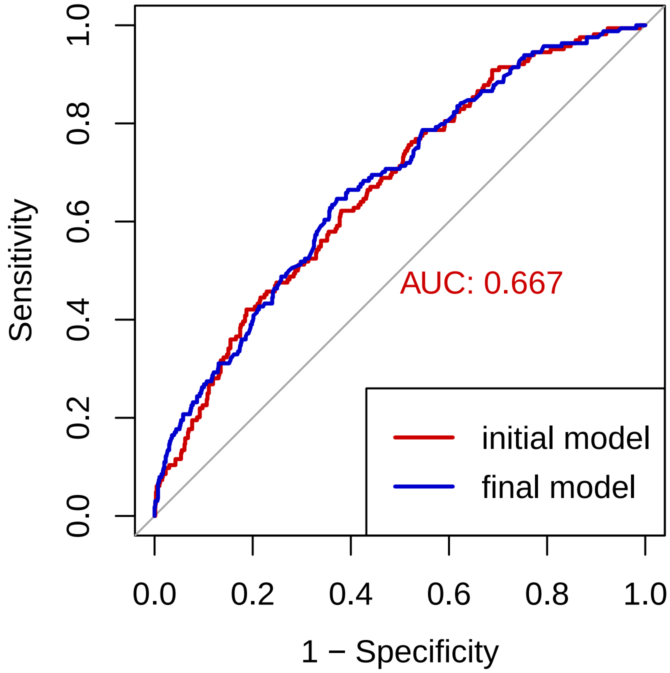
For this report, stepwise method will be implemented to find the lowest AIC because BIC generally places a heavier penalty on modes with more than 8 variables. The single predictor variables we will include for our full models are parity, mrace, mage, med, mht, mpregwt,income, and smoke. Besides the single predictor variables, we will include seven interactions including 1) smoke and parity 2) smoke and mage 3) smoke and race 4) smoke and mht 5) smoke and income 6) smoke and med 7) smoke and pregwt. After performing stepwise selection, the final model ended up having four predictor variables: med, mrace, mpregwt, and smoke. These four predictor variables match the findings from EDA. However, all interactions are dropped from the model. Through the results of EDA and potential interaction investigation, three potential interactions are identified: 1) mpregwt vs premature by smoke 2) mage vs premature by smoke 3) smoke vs premature by mrace. F-test is conducted to decide whether to drop the interactions, and the result of f test shows including these interactions have a high p-value compare to excluding these interactions. The binned residual plots look random, and 95% of the points are within the standard error bound. The next step is to check the multicollinearity of the final model, and all vif value except education are approximately 1.1, which is acceptable. The reason for the high multicollinearity for education might be because of the insufficient of data for each education group. Also, according to the cooks's distance and leverage score, there are many points with high leverage score but no influential points. According to the confusion matrix and ROC curve of the final model, the optimal specificity and sensitivity is (0.698,0.524), accuracy is 0.61, and AUC is 0.667, which are still not ideal and around the same as the preliminary model.

$$log(\frac{\pi_i}{1-\pi_i}) = \beta x_i; Bernoulli(\pi_i).$$

This is the equation of our final model.pi/(1-pi) is the odds of premature for observation i, and x_i is the vector containing the corresponding values for mother's pre-pregnancy weight in pounds, mother's race, and smoke.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9237	0.9568	-0.97	0.3344
racem6	0.1874	0.6292	0.30	0.7658
racem7	1.0552	0.3058	3.45	0.0006
racem8	0.8273	0.4947	1.67	0.0944
racem9	-13.5150	413.9505	-0.03	0.9740
mpregwtc	-0.0127	0.0048	-2.62	0.0087
sm1	0.3971	0.2279	1.74	0.0814
edu1	-0.5592	0.9639	-0.58	0.5618
edu2	-0.9064	0.9602	-0.94	0.3452
edu3	-0.7412	1.0144	-0.73	0.4650
edu4	-1.5744	0.9739	-1.62	0.1060
edu5	-1.0632	0.9769	-1.09	0.2764
edu7	1.8392	1.5062	1.22	0.2220
racem6:sm1	-0.0325	1.1125	-0.03	0.9767
racem7:sm1	-0.5652	0.4241	-1.33	0.1826
racem8:sm1	0.3170	0.8451	0.38	0.7076
racem9:sm1	14.4624	413.9524	0.03	0.9721

The table below shows the summary of model including the interaction of smoke and race.



Conclusion

The intercept of the final model shows mother who has an ethnicity of white, does not smoke, with an education level of less than 8th grade, and with an average pre-pregnancy weight has an odds of having premature birth of **0.41**. As the pre-pregnancy weight of the mother increases by one pound, the odds of having premature birth will decrease by approximately **1%** with all other variables constant. Compare to mother who **does not smoke**, the odds of having premature birth for mother who **does smoke** will increase by approximately **33%** with all other variables constant. It is interesting that only the p-values of the coefficients of Asian group and African American group are lower than 0.05. Compare to mother with an **ethnicity of white**, the odds of having premature birth for mother with an **ethnicity of black** will increase by approximately **187%** with all other variables constant. Compare to mother with an **ethnicity of white**, the odds of having premature birth for mother with an **ethnicity of Asian** will increase by approximately **130%** with all other variables constant. With 95% confidence, the range of the increase of odds of having premature birth for smokers compare to non-smokers with other variables constant is **between -7% and 92%**. Therefore, mothers who smoke tend to have higher odds of premature birth than mothers who do not smoke. Since the p-value for the F-test for including interaction of mrace and smoke is bigger than 0.05, there is not enough evidence to conclude that the association between smoking and premature differs by mother's race. One potential limitation we have for this study is there are not enough data for some population groups like mothers with an education less than 8th grade to accurately calculate the association of the corresponding predictor variables to odds of premature birth. Also, since there are not sufficient data for some bins in the binned residual plot, the average residual for those bins might be inaccurate.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8767	0.9403	-0.93	0.3512
racem6	0.1549	0.5164	0.30	0.7643
racem7	0.7707	0.2227	3.46	0.0005
racem8	0.9060	0.4077	2.22	0.0263
racem9	-0.7528	1.0515	-0.72	0.4740
mpregwtc	-0.0121	0.0048	-2.51	0.0121
sm1	0.2889	0.1843	1.57	0.1171
edu1	-0.5410	0.9490	-0.57	0.5686
edu2	-0.8876	0.9407	-0.94	0.3454
edu3	-0.7067	0.9931	-0.71	0.4767
edu4	-1.5479	0.9558	-1.62	0.1054
edu5	-1.0646	0.9585	-1.11	0.2667
edu7	1.8257	1.4841	1.23	0.2186

The table above shows the summary of final model.

	2.5 %	97.5 %
(Intercept)	-2.95	0.97
racem6	-0.95	1.11
racem7	0.33	1.20
racem8	0.08	1.69
racem9	-3.67	0.90
mpregwtc	-0.02	-0.00
sm1	-0.07	0.65
edu1	-2.41	1.54
edu2	-2.74	1.18
edu3	-2.66	1.44
edu4	-3.42	0.54
edu5	-2.95	1.03
edu7	-0.92	5.26