Assignment 3

Pranav Manjunath

9/12/2020

SUMMARY

The goal of this analysis is to understand the relationship between premature delivery of a child (pre term birth) and the mother's smoking habits along with other attributes. To understand these relationships, I performed EDA on the data and then fitted a logistic regression model (formula = premature ~ med + mrace + mpregwtc + smoke +smoke:mrace) to answer the questions of interest. Interestingly, the odds ratio of pre birth occurring if the mother smokes is 1.48 times more than if the mother does not smoke. Variables such as mother's race (mrace), pre- pregnancy weight (mpregwt), smoking habits (smoke), and mother's education (med) are statistical significant in predicting pre term birth.

INTRODUCTION

The Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA addressed the issue of pregnancy and smoking. The researchers interviewed mothers early in their pregnancy to collect information on socioeconomic and demographic characteristics, including an indicator of whether the mother smoked during pregnancy. This data is now used to help analyze and answer the following questions.

- Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the odds ratio of pre-term birth for smokers and non-smokers?
- Is there any evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race? If so, characterize those differences.
- Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

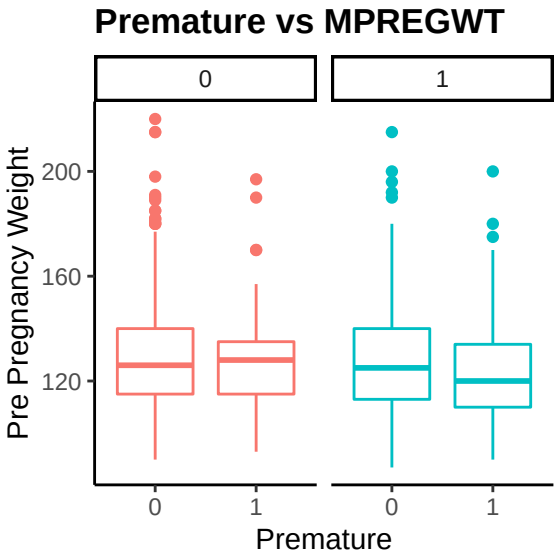
To help answer these questions, EDA plots and a logistic regression model are built incorporating the attributes that are statistically significant with the response variable (premature).

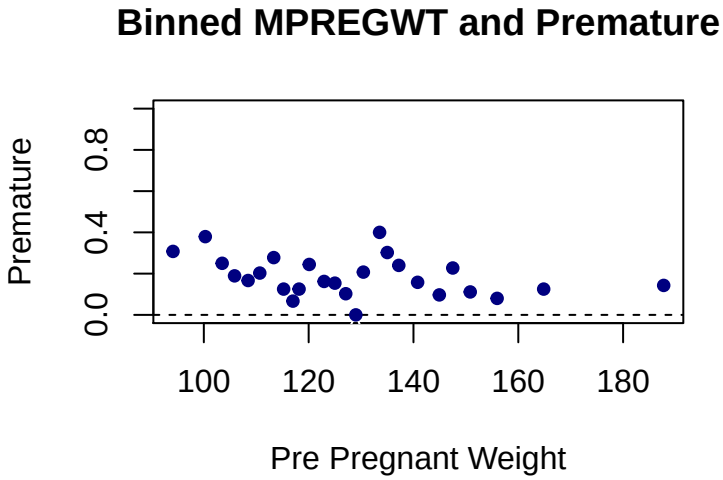
DATA

The data I have considered for the analysis consists of 869 observations and 12 variables. The following columns are: id, date, gestation, bwt.oz, parity, mrace, mage, med, mht, mpregwt, inc, and smoke. As Premature/ Pre Term Birth is indicated by gestation period < 270 days, I created a new variable as Premature, keeping this as the response variable. There are 705 non premature observations and 164 premature observations (18.87% of data consists of premature observations). The columns id, date, and gestation will not be used in this analysis. I did not perform any log transformation to any of the variables. The data transformations performed was i)Converting the mrace variable values 0,1,2,3,4,5 to 0 (White race) ii)Converting categories 6 and 7 of the med variable into one category, Trade School. The variables Smoke, med, and race variables have been converted into factor variables.

Exploratory Data Analysis

For this report, I have included two plots that I felt showed interesting relationships. The box plot below indicates the interaction between mpregwt (mother's pre pregnancy weight) and smoke with respect to premature. This interaction seemed to be important as the trend was not constant among smoking habits. The binned plot shows a slight non linear pattern could be because of less data.





	Smoke0	Smoke1
0	0.8348	0.7841
1	0.1652	0.2159

The table above represents the relationship between smoking and pre term birth. The index (0,1) in the table indicate Pre Term Birth=0 and Pre Term Birth=1 respectively. The table shows that the probability of pre term births increases if the mother smokes. To test the significance of smoking habits on pre term birth, a Chisq test was used. The results displayed a p value of 0.0694 (significant at 0.1 level).

Observations from EDA:

- *Smoke vs Premature:* - As noticed in the probability table and Chisq test above, there is an increased probability of pre birth for mothers who smoke. The Chisq test indicates that the variable smoke is significant with respect to the 0.1 significance level.
- *Parity vs Premature:* - By observing the box plot (Parity is Continuous variable), I observed that there is not a significant relationship between parity and premature (medians and boxes are aligned). When observing the binned plot, there seemed to be no apparent pattern.
- *Mother's Race vs Premature:* - While observing the probability table, I noticed that the probability of premature varies amongst different races. The Chisq test provided a p value of 0.0037 indicating that it is a significant variable.
- *Mother's Age vs Premature:* - The box plot representing the relationship between Mother's age and premature seems to have minute differences. The median age of the non premature mother (pre mature=0) seems to be slightly higher than the median age of premature mother. When observing the binned plot, no apparent patterns were detected.

- *Mother's Education vs Premature:* - As there are 7 discrete values for education, there was not sufficient data in each discrete value to clearly understand relationship between education and Premature. When observing the probability table, the chances of Premature occurring varied depending on the education level of the mother. The Chisq test indicated that med is a significant variable, having a p value 0.000547.
- *Mother's Height vs Premature:* - The box plot representing the relationship between the mother's height and premature occurrences illustrates no significant patterns, the median mother's height for both premature and not premature birth seems to be similar. The binned plot also indicates no observant pattern.
- *Mother's Income vs Premature:* - The box plot representing the relationship between the mother's income and premature occurrences illustrates no significant patterns, the box size and median mother's height for both premature and not premature birth seems to be similar. The binned plot also indicates no observant pattern.
- *Mother's Pre Pregnancy Weight vs Premature:* - On observing the box plot, there seems to be a slight shift in the boxes between mother's pre pregnancy weight and the pre birth occurrence. It is noticed that the median pre pregnancy weight for mother with no pre birth is slightly higher than the median pre pregnancy weight for mothers with pre birth occurrences. There also seems to be a slight non linear pattern in the binned plot (hard to confirm due to less data points).

MODEL

Upon performing EDA, I noticed that variables such as smoke (Significant at 0.1 significance level), mrace, mpregwt and med seem to have some relationship with the response variable. Looking at interactions, I observed various interactions between smoke and parity, weight, height, age, and med. For the analysis, I have centered the continuous variables. I used forward AIC for model creation. The Null Model used included the response (premature) and the variable smoke, mrace, and the interaction smoke:mrace. The Full Model included all the variables in the dataset (except date, id, and gestation) along with all the interactions with smoke. The full model has an AIC score of 833.83.

After performing forward AIC, the model outputted included med, mrace, mpregwtc, smoke, and the interaction between smoke and med & smoke and mrace as the predictor variables. While building this model and performing multicollinearity test, I noticed that smoke:med interaction seemed to inflate all the VIF scores for the variables, indicating high multi-collinearity. One reason for this is that there is not sufficient data in each education value and hence I have removed the interaction between med:smoke in the final model.

The final regression model used is:

$\hat{premature} = \hat{\beta}_0 + \hat{\beta}_1$ * smoke + $\hat{\beta}_2$ * mpregwtc+$\hat{\beta}_{3:6}$ * mrace +$\hat{\beta}_{7:12}$ * mrace:smoke +$\hat{\beta}_{13:18}$ * med

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.9237	0.9568	-0.9654	0.3344
med1	-0.5592	0.9639	-0.5801	0.5618
med2	-0.9064	0.9602	-0.944	0.3452
med3	-0.7412	1.014	-0.7307	0.465
med4	-1.574	0.9739	-1.617	0.106
med5	-1.063	0.9769	-1.088	0.2764
med7	1.839	1.506	1.221	0.222
mrace6	0.1874	0.6292	0.2978	0.7658
mrace7	1.055	0.3058	3.451	0.0005595
mrace8	0.8273	0.4947	1.672	0.09444
mrace9	-13.51	414	-0.03265	0.974
mpregwtc	-0.01268	0.004833	-2.625	0.008678

	Estimate	Std. Error	z value	Pr(>|z|)
smoke1	0.3971	0.2279	1.742	0.08143
mrace6:smoke1	-0.03248	1.113	-0.02919	0.9767
mrace7:smoke1	-0.5652	0.4241	-1.333	0.1826
mrace8:smoke1	0.317	0.8451	0.3751	0.7076
mrace9:smoke1	14.46	414	0.03494	0.9721

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	841.8 on 868 degrees of freedom
Residual deviance:	790.7 on 852 degrees of freedom

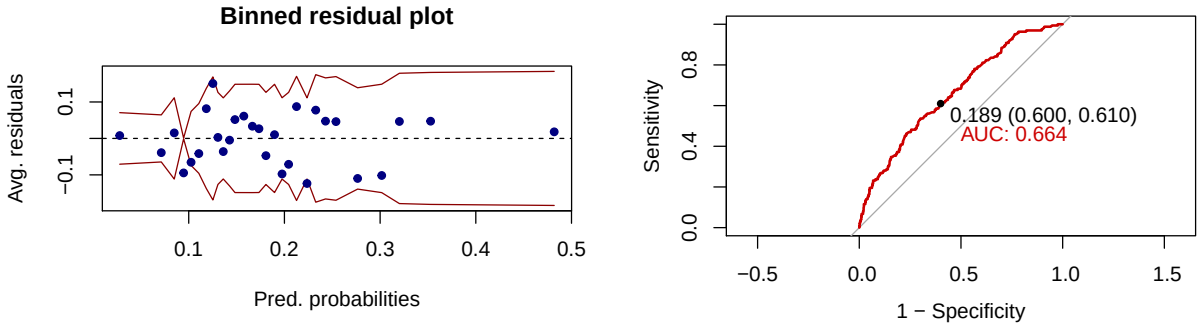
The baseline values taken in the intercept is smoke=0 (non smoking mothers), med=0 (education level below 8th grade) and mrace=0 (race=White). As the model has been centered, the intercept can be interpreted as, the odds of a non smoking mother of white race, pre-pregnancy weight of 128.48 pounds having pre birth is decreased by 60.29%.

Keeping the other variables constant and observing only the statistically significant variables,

- The odds ratio of premature birth occurring if the mother smokes is 1.48 times more than if the mother does not smoke.
- A unit increase in the mother's pre-pregnancy weight tend to decrease the odds of pre term birth by 1.26%
- The odds ratio of pre term birth occurring if the mother belongs to Black race is 2.87 times more than if the mother belongs to a White race.
- The odds ratio of pre term birth occurring if the mother belongs to Asian race is 2.29 times more than if the mother belongs to a White race.

In the model, mrace, smoke, mpregwt, and med are statistically significant (tested by observing p value of model and F Tests). I have kept the interaction between mrace:smoke, though it is not statistically significant as it is used to answer a question of interest.

The RSME Value of the model is 2.76. The Null deviance is 841.83 and the Residual deviance is 790.73.



The above left plot describes the average residuals and predicted probabilities of the model. As noticed, only 2 point is beyond/outside the 95% confidence level and remaining points does not seem to follow a clear pattern (high randomness). We can conclude that this model follows necessary requirements.

According to the ROC (top right), the Area under the Curve (AUC) is 0.664 (66.4%). The threshold used to identify the confusion matrix and ROC is the mean(data$premature) = 0.189. The accuracy is 0.60184 (60.18%). The Sensitivity and Specificity values are 0.6010 and 0.60 respectively.

The final model above has an AIC score of 824.73. There are leverage points present in the data(leverage score below 0.5). However, there are influence points (none of the points have a cook's distance above 0.5) present. Hence, no outliers and influence points have been removed from the data.

The VIF values of majority of the variable is range between 1-5 (moderately correlated). The VIF values for education are relatively high (above 10) however as it is a factor variable, I have not removed this variable from the dataset.

The answers to the questions from the logistic regression model:

1. Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the odds ratio of pre-term birth for smokers and non-smokers?

- **Yes, mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke**. This question can be answered by the EDA and justified by the regression model. The variable smoke does seem to be statistically significant (at 0.1 significance level). It is observed that keeping the other variables constant, the odds ratio of pre term birth occurring when the mother smokes is 1.48 times more than if the mother does not smoke. Using the 95% confidence intervals, premature birth occurrences from mothers who smoke tend to be a minimum of 0.95 times and a maximum of 2.34 times more than mothers who do not smoke.

2. Is there any evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race? If so, characterize those differences.

- To determine this, I ran an ANOVA test to determine the the significance of the interaction between smoke and race. **According to the test, the interaction between race and smoke with respect to occurrence of pre term birth is statistically insignificant (p value of 0.2689) Hence, there is no evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race**

3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

- By noticing the regression model output, **along with mother's race (mrace) , mother's education (med), and mother's pre pregnancy weight (mpregwt) seem to have a strong statistic association with birth weight**. This is noticed by the small p value of each of the variables in the regression model.

CONCLUSION

The study uses EDA and a logistic regression model to understand the relationship between the premature birth and the mother's smoking habits, along with identifying other predictors and interactions that are statistically significant to the response variable premature. The analysis concluded that interestingly, smoke is statistically significant in the model and the odds ratio of pre birth occurring if the mother smokes is 1.48 times more than if the mother does not smoke. Variables such as mother's race (mrace), pre- pregnancy weight (mpregwt), smoking habits (smoke), and mother's education (med) have statistical significant associations with pre birth.

Limitations: The dataset is not evenly distributed between the two classes of response variable pre birth, only 18% of the dataset represented pre birth=1. Hence it is statically difficult to get an accurate model. While running Chisq tests, a warning that the Chisq test would be inaccurate was displayed, indicating the less data. To improve the model, we need to introduce more features, variables, and data points.