IDS 702 - Data Analysis Assignment 3

Aarushi Verma

21/09/2021

MATERNAL SMOKING AND BIRTH WEIGHTS

Summary This report covers the analysis of data derived from the Child Health and Development Studies, a comprehensive study of children born between the years 1960 to 1967 at the Kaiser Foundation Hospital.Our analysis, considers a subset of the original study which covered 15000 families. We removed the observations with missing data (such as information related to the fathers) and our dataset now contains 869 observations. Our objective is to assess the association between the gestational age of a child with the smoking habits of a mother. We want to evaluate whether mother's who smoke have a higher chance of pre-term birth.

Introduction In this report we perform exploratory data analysis and data processing to determine the relationship of different variables with gestational age. The objective of this study is based on Surgeon General's claim that mother's who smoke have increased rates of premature delivery (before 270 days). Through this analysis we will check and determine if there is an association between smoking and the pre-term birth. The variable gestation is recorded as 'Premature' which a is a binary variable. Here if gestational age $<$ 270 days we have assigned the value 1 and considered it premature, and 0 otherwise. We will also evaluate the other variables such as Mother's age,race, education, income, height and weight etc. Since our response variable 'Premature' is binary we will use Logistic Regression

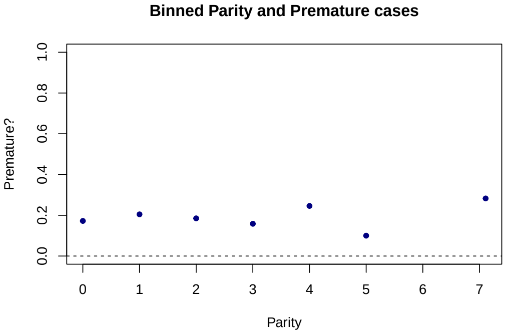
Data

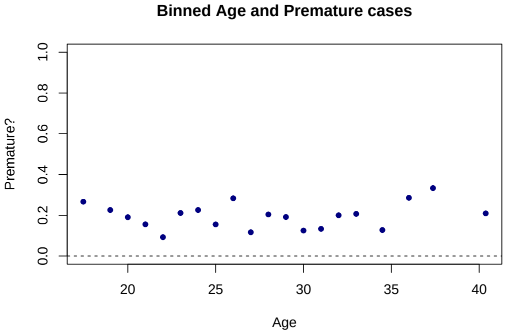
Data Pre-processing Before analysis, we performed a few transformations on our data to prepare it for our analysis. The steps taken are as follows:
1. We converted the variables, education, race, income and smoke from integer to factor variables.
2. We collapsed levels for some variables and renamed them based on the labels provided in the data dictionary.
* Race: Levels 0-5 were collapsed to 5 and renamed to 'white'
* Education: Levels 6-7 were collapsed to 7 and renamed to 'trade school'
3. We also dropped the columns - ID, Date and bwt.(Children's birthweight) from our data set. ID and Date do not add any value to our study in this context. Birth weight is similar to the Gestation period and both can be response variables. In this case we are focusing on Gestation Age.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
gestation	869	278.507	15.699	148	272	286	338
parity	869	1.953	1.882	0	1	3	11
mage	869	27.295	5.708	15	23	31	45
mht	869	64.069	2.534	53	62	66	72
mpregwt	869	128.479	20.778	87	113	140	220
premature_num	869	0.189	0.392	0	0	0	1

EDA We perform EDA to understand the underlying relationships between the independent and response variables. Firstly we plot our response variable Premature with other continuous variables. Based on the binned plots for the continuous variables we look at whether there is any trend followed between these variables.





There seems to be no obvious trend between Premature cases and Binned parity. The probability is constant initially, seems to rise a little and then fall. No transformation is needed. For Premature and Height, the predicted probability again does not seem to follow any specific trend, it declines initially but then remains constant only to decline again followed by a rise. Similarly, for age there is no trend in the probability ad age increases.Finally for weight, we can see that the probabilities are completely random. There is no discernible trend in the predicted probabilities. There is no need to transform any of the variables. Next, we plot contingency tables for the discrete variables against our response variables

Table 2: Contingency Table (Premature vs. Smoke)

	0	1
0	0.835	0.784
1	0.165	0.216

Table 3: Contingency Table (Premature vs. Race)

	white	mexican	black	asian	mix
0	0.839	0.760	0.734	0.676	0.933
1	0.161	0.240	0.266	0.324	0.067

Our main variables of interest are smoke and race. Looking at the contingency table for smoke we can observe that 21% of premature births occur for mother's who smoke whereas 16% occur for mother's who don't smoke. We can see that 16.13% of white mothers have pre term births. We must also note that our data set contains most data points about white mothers itself. We perform a chi-square test to see whether the variable smoke and race are dependent on each other. The chi-square test indicated a dependence between the variables with a pvalue of 0.002. This means we can include it as an interaction variable in our model. We also run Chi-square tests between other variables such as education, smoke and race to check if they are dependent on each other. We can observe that both have very low p-values indicating that the *med* has interaction with smoke ($p = 0.0003$) and mrace ($p<0.0001$).

Based on our EDA we then move on to Model Building and incorporate the effects we think affect our response variable

Model

Model Building We build the first model with the main effect of every variable and linear predictors. To improve our interpretation we also mean center the continuous predictors. Here our response variable is **premature** and the predictors are **smoke, parity, mother's race, mother's age, mother's height, mother's pregnancy weight, mother's education and income**.

Based on model 1 we see that only mrace = Black is significant. None of the Income levels are significant. On performing chi-square test between Income and Premature we see that there is no dependence between the variables. We can remove income as a predictor of pre-term birth. While the other variables are not significant based on our model summary, we will proceed to retain them based on our chi-square test results which indicated that the variables may be dependent on each other. Based on the results of Model 1 we drop income from our model and incorporate the interaction effect between smoke and race and build the next model.

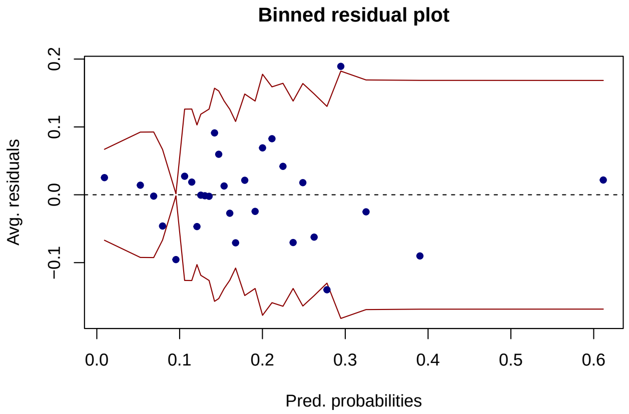
Based on model 2 we can notice very slight changes. Centered pregnancy weight of mother's is also a significant predictor now. However to proceed we will use step-wise selection and based on the AIC we will arrive on significant variables for our models. Next, we construct Model 3 using stepwise selection. A null model and a full model are determined, where the null model only considers our Model_2. The full model consists of all variables (except income) and all the possible interactions between other variables and smoke and race.

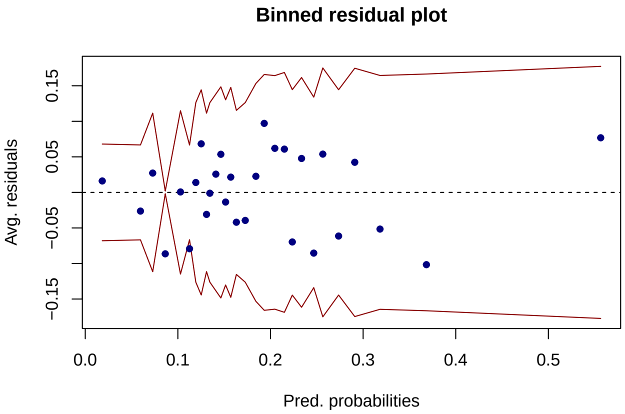
Based on AIC our model is premature_fac ~ parity_c + mrace + med + preg_c + smoke + med:smoke + parity_c:mrace. Our main interaction of interest between smoke and race has been excluded from the model however since we think it is scientifically significant and pertinent to this experiment we will retain it.

On examining the significant predictors, we can see that similar to our previous model, only mrace = black and pregnancy weight are significant. To improve our model, we can further drop parity since it has a very high p-value. Our model then becomes :

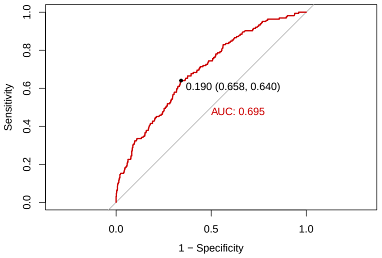
Model Assessment **Checking Residuals for Model_3 and Model_4**

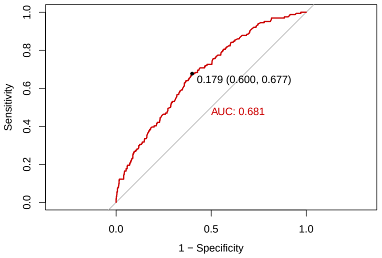
First, we check binned residual plots to check how the relationship between the predictors and the response is being captured. We look for randomness of data points and ensure that almost all points fall inside of the 95% lines. Both in Model 3 and Model 4, we see that data points are randomly scattered, with more than 95% of points within the $ +-2SE $ red lines.





Model Validation: Model_3 and Model_4 On creating a confusion matrix using the mean probability of the data as a threshold, Model 3 gives us accuracy of 64.67% and Model 4 gives us accuracy of 63.52%. The Sensitivity and Specificity of Model 3 is 64.82.9% and 64.02% respectively, and that of Model 4 is 63.68% and 62.80% respectively. Model 3's AUC is 0.69 and Model 4's is 0.68.





For the new model let us do an ANOVA test i.e. change in deviance test with the model with only the main effects of the individual variables. the results do not show any significant improvement based on the p-values. Based on the low pvalue (0.027) we can see that Model 3 performs better than Model_4.

Final Model: Based on these observations we have considered the following predictors for our final model: centered Parity, Mother's Race (mrace), Mother's Education (med), Centered Weight (preg_c), smoke, interaction between Smoke and education, parity and race and smoke and race. Including the summary of Model_3 in the appendix due to space constraints.

Model Interpretation Our Final model (Model_3) has an accuracy of 63.68%. The AUC is 0.69 and sensitivity and specificity are 64.9% and 64% respectively. The significant variables are (based on p-values) centered weight and Mother's race = Black. The model estimate for mrace= black is 0.977 on the log scale which converts to 2.64 on the odds scale. This can be interpreted as that the odds for for a premature birth increased by 2.64 for a mother whose race is black when compared to one that is white. The model estimate for centered weight is -0.0013 on the log scale. On exponentiation we get 0.99 on the odds scale. This indicates that the odds for premature birth fall by 0.99 for a 1 pound increase in the mother's weight. The interaction terms are not significant. This implies that the odds ratio of pre-term birth for smoking mothers and non-smoking mothers does not vary by mother's race.

Conclusions

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? Based on our final model, the coefficient of the smoke variable is 42.17 . On exponentiating this we get a very large number (2.06 e+18) which indicates that the odds for pre-term birth given the mother is a smoker is very high as opposed to a mother who does not smoke. The confidence interval for the odds ratio of smoke (level 1) is [1.59e+03 and 1.36e+03]. However we must also keep in mind that this variable is not significant and p value is very high.
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. Our final model includes the interaction effect between smoke and race, however based on the summary the interaction effect is not significant. The pvalue is greater than 0.05. On the basis of this we can conclude that the odds ratio of pre term birth for smokers and non-smokers does not differ by mother's race
3. Are there other interesting associations with the odds of pre-term birth that are worth mentioning? Our final model points towards mean centered pregnancy weight of the mother and mother's race being black as significant predictors of pre term birth. We have also include an interaction term between parity and race however the same is not significant. Additionally when checking the dependence of

variables on each other during EDA we noted that the variable mother's education is dependent on smoke as well as race. However based on the step wise model this interaction was not significant for prediciting pre-term birth.

Potential Limitations

1. The original study includes observations from 15000 families however we have done our analysis only 869 families. This is a very small subset of the population and may not be fully representative of the population.
2. The data for our response variable *premature* is imbalanced. It is skewed towards non premature births with 83% of the children being born after a gestation period of >270 days.
3. The data for the independent variable *mrace* is also imbalanced. Most data points are associated with the Mother's race = white and there are fewer points for mother's belonging to the other race. This might impact the interaction effects related to smoke and might be biased for white mothers.
4. Due to the missing values present in the complete data set pertaining to information about the fathers, we are disregarding the role the paternal factors could play in the premature birth of a child. A more thorough study could be performed with updated details of fathers and incorporating those effects in our model as well.

Appendix

Table 4: Final Model Results

	Dependent variable:
	premature_fac
parity_c	0.026 (0.064)
mracemexican	−0.860 (1.143)
mraceblack	0.971*** (0.334)
mraceasian	−0.011 (0.705)
mracemix	−31.203 (1,186.036)
medonly HS	16.793 (2,247.563)
medHSgrad	16.056 (2,247.563)
medHSgrad+TradeSc	17.240 (2,247.563)
medHSgrad+college	15.274 (2,247.563)
medCollgrad	16.008 (2,247.563)
medTradeSc	18.580 (2,247.563)
preg_c	−0.013*** (0.005)
smoke1	42.175 (3,143.599)
medonly HS:smoke1	−42.189 (3,143.599)
medHSgrad:smoke1	−41.574 (3,143.599)
medHSgrad+TradeSc:smoke1	−43.483 (3,143.599)
medHSgrad+college:smoke1	−41.416 (3,143.599)
medCollgrad:smoke1	−41.864 (3,143.599)
medTradeSc:smoke1	−25.297 (5,053.076)
parity_c:mracemexican	−0.992 (0.664)
parity_c:mraceblack	−0.001 (0.103)
parity_c:mraceasian	−0.835* (0.452)
parity_c:mracemix	−9.524 (405.839)
mracemexican:smoke1	0.571 (1.679)
mraceblack:smoke1	−0.530 (0.447)
mraceasian:smoke1	1.323 (1.027)
mracemix:smoke1	28.856 (1,260.956)
Constant	−17.965 (2,247.563)
Observations	869
Log Likelihood	−382.103
Akaike Inf. Crit.	820.205
Note:	*p<0.1; **p<0.05; ***p<0.01