Data Analysis Assignment 3

Aiman Jawaid Haider

September 15, 2020

QUESTION 1

SUMMARY

The report attempts to analyze the harmful impact of smoking on the chances of pre-term birth among newborns with the help of a logistic regression model. This is done using the "smoking.csv" dataset, and by developing a model after considering the AIC and BIC step-wise and forward-built models, taking into account the accuracy and reasonableness. The model thus formed suggests that the chances of pre-term birth are significantly associated with smoking, the pregnant mother's race, weight, education and her smoking habit given her education. It, thus, finds that the chances of pre-term birth are associated with demographics, physiology and smoking. Moreover, it finds that smoking during pregnancy may be a characteristic associated with higher pre-term births, but is not very significant and that demographic and physiological factors have stronger associations .

INTRODUCTION

A pertinent question in the domain of Maternal and Neonatal Health has been that of the effect of smoking during pregnancy. This report strives to analyze the impact of smoking on the chances of pre-term birth, taking into account the presence of some other socio-economic and demographic factors.

In order to analyze the impact of smoking on the chances of pre-term birth, we need to look at questions such as, do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not and what is the likely range for the odds ratio of pre-term birth for smokers and non-smokers. Besides, we also need to understand if the odds ratio

of pre-term birth for smokers and non-smokers differs by mother's race or some such other variable. And also, if there are other significant associations with the odds of pre-term birth.

The report uses a logistic regression model on the odds of a pre-term birth[1] for the analysis. It begins with an EDA of the data, tries building a model by exploring models built with the help of AIC and BIC criteria using forward and stepwise model building and chooses the most suitable model on the basis of accuracy and plausibility to answer the above questions.

DATA

The data used is from the "smoking.csv" file, which is already cleaned and organized as mentioned in the codebook[2]. This data is explored for trends first. From the data, it can be observed that the cohort having premature babies is a little younger (*mage*) than those with normal term births. However, there is no such noticeable trend with respect to the parities of the two.[3] But, there are slight differences with respect to mother's height and pregnancy weight among the two cohorts. With respect to mother's race (and also income, individually), while the conditional probabilities for premature births are lower than for the normal-term ones, there does seem to be some difference in the trends across the categories i.e. different races (or income brackets). With Mother's education too, there does seem to be a difference in the trends across the levels, this time showing some major differences. It is also observed that smokers have higher probabilities of premature birth than non-smokers (However, the Chi-square for this does not turn out to be significant).[4] The report then

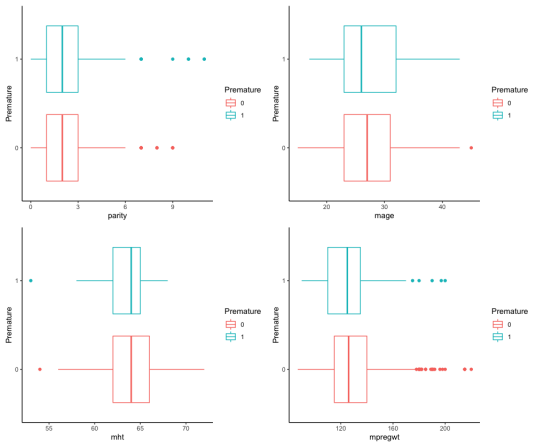

Figure 1: Some EDA Plots

looks at some possible interactions between variables, particularly those concerning smoking (the main interest). It is observed that the trend between pre-term births and mother's

[1] As it is a binary variable on the occurrence of an event.

[2] See Appendix for important variables used for the report.

[3] Mothers with pre-term births and those with normal terms.

[4] Chi-square tests for *mrace* and *med* seems to be significant, but is not significant for *smoke* and *inc*. Less number of observations may impact.

height, age*, preg.weight*, parity, education, income and race* differs by smoking[5]. Besides, it is also observed that the trends between pre-term births and mother's parity, age*, pregnancy weight* and education differs by mother's race.[6]The trends between pre-term births and preg. weight also seem to differ by parity. And so does the relation between pre-term-birth and mother's education differ by parity and age. From the EDA, it can, thus be understood that the above-mentioned variables should be given a consideration.

MODEL

On the basis of the above, the logistic regression model is developed after considering the stepwise and forward[7] built models using the AIC and BIC criteria and then selected. The Null Model used is premature ~ smoke+ smoke*mrace [8] and the Full Model used is: *premature ~ smoke+mrace+smoke*mrace+med+mht+mpregwt+inc+parity+smoke*mrace+smoke*med +smoke*mht+smoke*mpregwt+smoke*parity+mrace*mht+mrace*mage+mrace*med+mrace*inc+mpregwt*parity+med*mage+med*parity*[9]. Feeding these models, the AIC Forward Model found is: *premature ~ smoke + mrace + mpregwt + med + smoke:mrace + smoke:med*, the AIC Stepwise Model is: *premature ~ smoke + mrace + mpregwt + med + smoke:mrace + smoke:med*,while with BIC Forward is: *premature ~ smoke + mrace + smoke:mrace* and the BIC Stepwise Model is:*premature ~ mht*.

From the above, the AIC Forward, AIC Stepwise and the BIC Forward Models are considered as they contain the variable of interest-smoke. On comparing these models with each other and the NullModel using the ANOVA Chi-square test, it is found that the AIC Stepwise Model is highly statistically significant (Refer to Appendix for all values). Also, a few other models by adding variables like *mage* and *parity*[10] and by dropping a few others(*smoke:med, med*) are compared through the ANOVA Chi-square test. From these, it turns out that the AIC Stepwise Model is the most succinct and statistically significant with the least AIC value (820.5). Although, it has a slightly higher deviance (782.5) than the AIC Forward Model, the latter is not found to be significantly different from it (p-value:0.2659). Besides, the only difference between the two (*term-smoke:mrace*) is not found to have significant p-values in the Forward model (itself), is not supported by the EDA and even jeopardises the coefficients by burgeoning the VIFs. Also, after some consideration, the reference level of education is kept at level 2 as it is the most preponderant; and hence, acts as a "normal" (common) baseline. It helps in the interpretation of the model and keeps the VIFs in check, too. This model is then assessed using the binned plots for residuals versus predicted probabilities and it is found that 95% of the points are within the limits and are random, indicating that there are no issues. The residual versus *mpregwt* (the continuous

[5]For categorical variables, a particular value of category is fixed and the conditional probabilities seen.

[6]* show minor diff but are retained to be explored further. Similar interactions are not considered twice; based on corelated variables.

[7]To consider the smoke*mrace predictor.

[8]*smoke*mrace* is an important requisite in the analysis and retained.

[9]Based on the observations in EDA.

[10]It is often suggested that the no. of pregnancies and lower age may cause pre-term births. Ref:here

variable) plot supports this as well. The VIFs for all the variables are also in check ensuring no concerns of multicolinearity. Further, it is found that the ROC gives a threshold of 0.182, sensitivity of 0.628, specificity of 0.646 and accuracy of 0.6340 with AUC being 0.674.[11]. Given the limitations of data and the question at hand (inference), the confusion matrix seems to be fairing well too (See Appendix). Intuitively too, the predictors make sense as pregnancy weight and race have been associated with pre-term births. For the mother's education, it turns out that it has a strong association with age [12](which is correlated to parity) and *inc*, thus acting as a proxy for them besides being an indicator of awareness. The interaction variable, thus, also becomes comprehensible e.g. an xyz educated woman may choose to smoke but keep that in check. The final model, thus, selected is:

$$Pr(premature_i|[smoke_i, mrace_i, mpregwt_i, med_i]) \sim Bernoulli(\pi_i);$$
$$log(\pi_i) = \beta_0 + \beta_1 smoke_i + \sum_{k=6}^{9} \beta_2 1[mrace_i = k] + \beta_3 mpregwt_i + \sum_{k=1}^{6} \beta_4 1[med_i = k] + \sum_{k=1}^{6} \beta_{5k} 1[smoke_i = 1, med_i = k]$$

Output of Linear Regression						
Variable	Coefficient Estimate	Std. Error	z val	Pr(>|z|)	Conf Int(2.5%)	Conf Int(97.5%)
Intercept	-2.328e-01	6.653e-01	-0.350	0.72641	-1.5251	1.0859
smoke1	5.088e-01	2.940e-01	1.731	0.08352 .	-0.0641	1.0924
mraceMexican	-3.785e-02	5.531e-01	-0.068	0.94544	-1.2388	0.9721
mraceBlack	7.331e-01	2.255e-01	3.250	0.00115 **	0.2866	1.1725
mraceAsian	8.798e-01	4.118e-01	2.136	0.03265 *	0.04289	1.6706
mraceMix	-8.017e-01	1.057e+00	-0.759	0.44812	-3.7202	0.86702
med5	-4.366e-02	3.610e-01	-0.121	0.90375	-0.76973	0.65349
med0	-1.398e+01	8.343e+02	-0.017	0.98663	NA	94.7836
med4	-7.330e-01	3.891e-01	-1.884	0.05961 .	-1.5320	0.00628
med1	7.888e-01	3.855e-01	2.046	0.04071 *	0.01906	1.5382
med3	1.131e+00	5.365e-01	2.109	0.03496 *	0.03878	2.1693
med6	2.526e+00	1.253e+00	2.016	0.04382 *	0.1266	5.6276
mpregwt	-1.273e-02	4.928e-03	-2.582	0.00981 **	-0.0226	-0.0033
smoke1:med5	-2.726e-01	5.286e-01	-0.516	0.60606	-1.3234	0.7579
smoke1:med0	3.057e+01	1.323e+03	0.023	0.98156	-106.8049	NA
smoke1:med4	1.729e-01	5.314e-01	0.325	0.74494	-0.8697	1.2245
smoke1:med1	-7.215e-01	5.015e-01	-1.439	0.15022	-1.7045	0.2666
smoke1:med3	-1.765e+00	7.898e-01	-2.235	0.02545 *	-3.3877	-0.2469
smoke1:med6	1.436e+01	1.455e+03	0.010	0.99213	-268.7379	NA

From the above, the log odds of having a premature baby seems to be strongly associated with *smoke, mrace,mpregwt, med,smoke:med.* Based on statistical significance of at least 0.1 and with other variables held constant, it can be observed that for every unit (pound) in-

[11]Although not the most accurate, these strike a great balance with near closeness to the most accurate full model (accuracy=0.71) with slight bent towards specificity

[12]glm(med-mage,family = "binomial",data=smoking) is highly significant.Corrplot in Appendix.

crease in *mpregwt* the odds of pre-term birth decreases by 1.2649%. Similarly, if the mother is Black or Asian then the odds increase by 108.15% and 141.04% (compared to the baseline of White mothers) resp. And, if she has studied high school without graduating or graduated high school and gone to trade school or only gone to trade school or gone to college then the odds increase by 120%, 209.98%,12.5 times and decrease by 51.96% respectively (compared to the baseline of being only a high school graduate). Further, if the mother smokes and has gone to high school and trade school, then the odds are likely to additionally reduce by 82.88% (compared to a non-smoker high school graduate, over the main effects of smoking and education).

From the above, it can also be seen that smoking can increase the odds of pre-term birth by 66.33% and is statistically significant at 0.1 and hence it does seem to suggest that mothers who smoke may tend to have higher chances of pre-term birth than mothers who don't. The likely range for the odds ratio of pre-term birth for smokers vis-a-vis non-smokers ranges from 0.93 to 2.96 (0.96 to 2.96 times more likely; CI). From these, it can observed that smoking is associated with the odds of pre-term birth, but not as significantly.

From the output, it can also be seen that the odds ratio of pre-term birth for smokers and non-smokers does not differ by mother's race. This can be understood from the results of ANOVA Chisquare test- The model containing the interaction term is not (statistically) significantly different from the final one(p-value:0.2659), with the interaction term being the only difference. And even in the model that contains it, the p-values are greater than 0.2. These indicate that there is no such interaction effect. However, *mrace* is certainly strongly associated with the odds as seen above. Other interesting associations with the odds ratio of pre-term birth are the mother's pregnancy weight, her education level and her smoking habit given her education. Of these, weight and race have higher significance. Thus, it indicates that the physiological and demographic characteristics are more strongly associated with the odds of pre-term birth than smoking.

CONCLUSION

The findings suggest that demographic and physiological characteristics have note-worthy associations with the odds of a pre-term birth. However, it also points out that the association of the odds with smoking is significant, yet not striking.

The report, though, suffers from limitations in terms of the limited data available especially for subcategories of variables (like higher income, xyz race, etc.). Also, it misses out on a number of variables like medical conditions that could have potential impacts on the model formation and hence, the analysis. Besides, Regression does not help one gauge the cause and effect among variables. It merely suggests such possibilities.

Nonetheless, the findings, on the whole, suggest that smoking does have an association with the odds of a pre-term birth, but not a very strong one.

APPENDIX

TABLES

IMPORTANT VARIABLES IN THE DATASET:

Variable	Description	Treated as(Type)
parity	total number of previous pregnancies, including fetal deaths and still births. (99=unknown)	Discrete
mrace	mother's race;0-5=white;6=mexican;7=black; 8=asian ;9=mix;99=unknown	Categorical
mage	mother's age in years at termination of pregnancy	Discrete
med	mother's education;0 = less than 8th grade;1 = 8th to 12th grade. did not graduate high school;2 = high school graduate, no other schooling;3 = high school graduate + trade school;4 = high school graduate + some college;5 = college graduate;6,7 = trade school but unclear if graduated from high school;9 = unknown	Categorical
mht	mother's height in inches	Discrete (as numbers are rounded)
mpregwt	mother's pregnancy weight in pounds	Discrete (as numbers are rounded)
inc	family yearly income in 2500 increments. 0 = under 2500, 1 = 2500-4999, . . . , 9 = 15000+. 98=unknown, 99=not asked	Categorical
smoke	Does mother smoke? 1=yes,0=no	Categorical
Premature	1 = baby born before gestational age of 270, and 0 = otherwise.	Categorical

COMPARISON OF MODELS USING ANOVA CHI-SQUARE TEST:

Model A	Model B	Significance
Null	BIC Forward	0
Null	AIC Forward	0.0001114 ***
Null	AIC Stepwise	5.137e-05 ***
AIC Forwarlind	AIC Stepwise	0.2659
Model with med	Model without med(Based on a basic model with smoke, mrace and mpregwt)	0.002328**
Model with smoke:med	Model without(Based on a basic model with smoke, mrace,mpregwt and med)	0.03125 *
Model with mage	Model without(Based on a basic model with smoke, mrace and mpregwt)	0.3615
Model with parity	Model without(Based on a basic model with smoke, mrace and mpregwt)	0.3164
Final Model	Full Model	0.5083

AIC STEPWISE MODEL OUTPUT

	Premature		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.79	0.22 – 2.92	0.726
smoke [1]	1.66	0.93 – 2.96	0.084
mrace [Mexican]	0.96	0.33 – 2.85	0.945
mrace [Black]	2.08	1.34 – 3.24	**0.001**
mrace [Asian]	2.41	1.08 – 5.40	**0.033**
mrace [Mix]	0.45	0.06 – 3.56	0.448
med [5]	0.96	0.47 – 1.94	0.904
med [0]	0.00	0.00 – Inf	0.987
med [4]	0.48	0.22 – 1.03	0.060
med [1]	2.20	1.03 – 4.69	**0.041**
med [3]	3.10	1.08 – 8.87	**0.035**
med [6]	12.50	1.07 – 145.72	**0.044**
mpregwt	0.99	0.98 – 1.00	**0.010**
smoke [1] * med [5]	0.76	0.27 – 2.15	0.606
smoke [1] * med [0]	18978171893615.39	0.00 – Inf	0.982
smoke [1] * med [4]	1.19	0.42 – 3.37	0.745
smoke [1] * med [1]	0.49	0.18 – 1.30	0.150
smoke [1] * med [3]	0.17	0.04 – 0.81	**0.025**
smoke [1] * med [6]	1715293.69	0.00 – Inf	0.992
Observations	869		
R^2 Tjur	0.073		

VIFs:

```
   smoke1 mraceMexican   mraceBlack   mraceAsian     mraceMix         med5         med0         med4         med1
 2.659104     1.067746     1.168320     1.144550     1.013632     2.330302     1.660591     2.578544     2.973320
     med3         med6     mpregwt smoke1:med5 smoke1:med0 smoke1:med4 smoke1:med1 smoke1:med3 smoke1:med6
 2.036885     1.040194     1.137974     2.265731     1.660591     2.742211     3.332581     2.035387     1.000001
```

Confusion Matrix:

447	60
258	104

The False Positives are lower. The False negatives are a little high, but are limited by the nature of the problem-Inference. It, however, strikes an overall balance with slightly more specificity, which is welcome.

AIC FORWARD MODEL OUTPUT

	Premature		
Predictors	*Odds Ratios*	*CI*	*p*
(Intercept)	0.80	0.22 – 2.95	0.738
smoke [1]	1.82	0.99 – 3.36	0.055
mrace [Mexican]	1.10	0.32 – 3.79	0.884
mrace [Black]	2.78	1.49 – 5.20	**0.001**
mrace [Asian]	2.14	0.80 – 5.77	0.131
mrace [Mix]	0.00	0.00 – Inf	0.983
med [5]	1.02	0.50 – 2.08	0.965
med [0]	0.00	0.00 – Inf	0.991
med [4]	0.47	0.21 – 1.01	0.053
med [1]	2.06	0.96 – 4.45	0.065
med [3]	3.05	1.04 – 8.90	**0.041**
med [6]	12.71	1.08 – 149.35	**0.043**
mpregwt	0.99	0.98 – 1.00	**0.007**
smoke [1] * mrace [Mexican]	0.64	0.05 – 8.32	0.732
smoke [1] * mrace [Black]	0.57	0.24 – 1.35	0.201
smoke [1] * mrace [Asian]	1.51	0.28 – 8.22	0.632
smoke [1] * mrace [Mix]	5020931.93	0.00 – Inf	0.982
smoke [1] * med [5]	0.69	0.24 – 2.01	0.501
smoke [1] * med [0]	195794777255777.38	0.00 – Inf	0.988
smoke [1] * med [4]	1.25	0.43 – 3.60	0.678
smoke [1] * med [1]	0.55	0.20 – 1.51	0.245
smoke [1] * med [3]	0.17	0.04 – 0.83	**0.028**
smoke [1] * med [6]	4438722.21	0.00 – Inf	0.995
Observations	869		
R^2 Tjur	0.077		

Some Additional Plots

Some EDA Plots

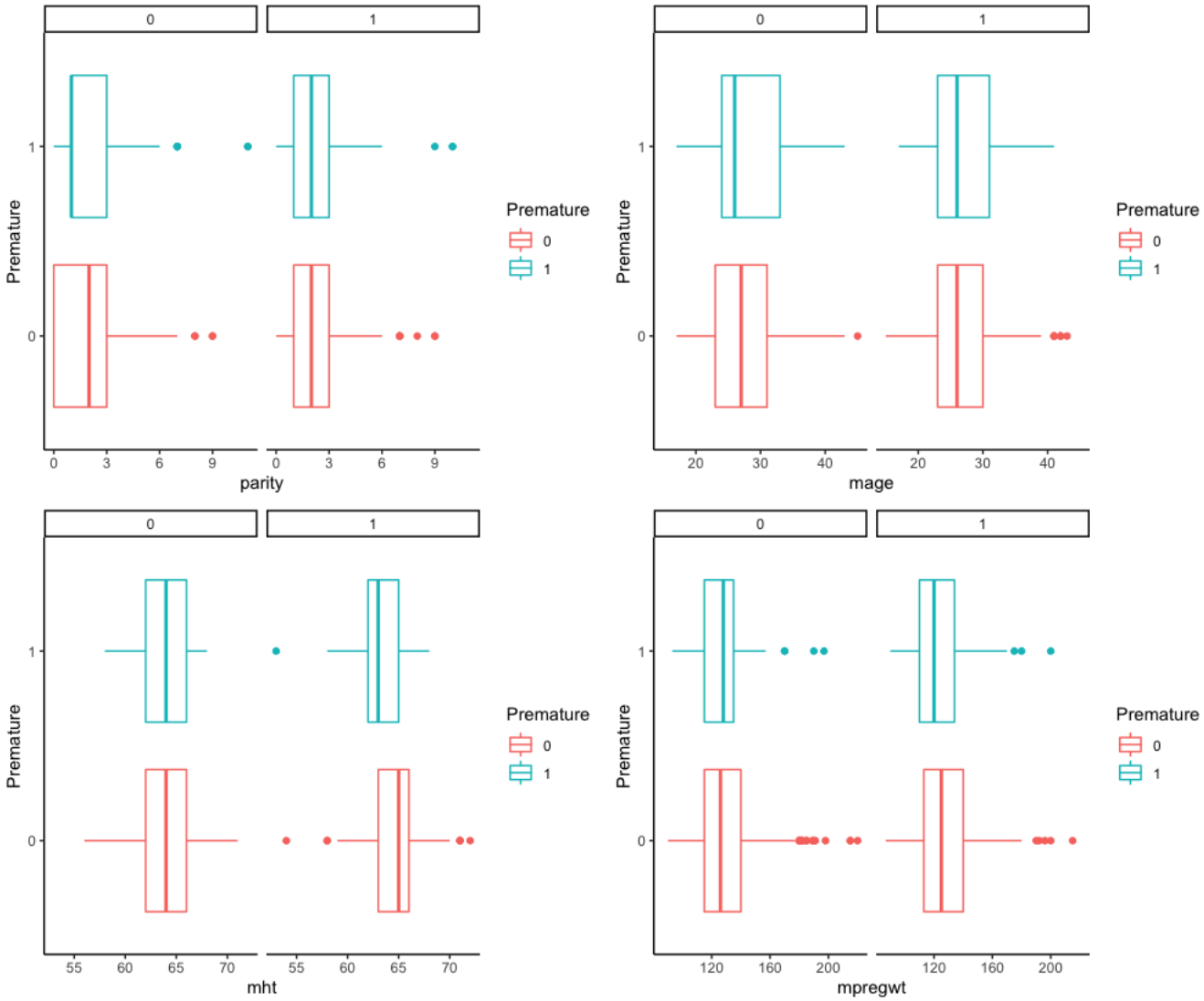

Figure 2: EDA for Interactions

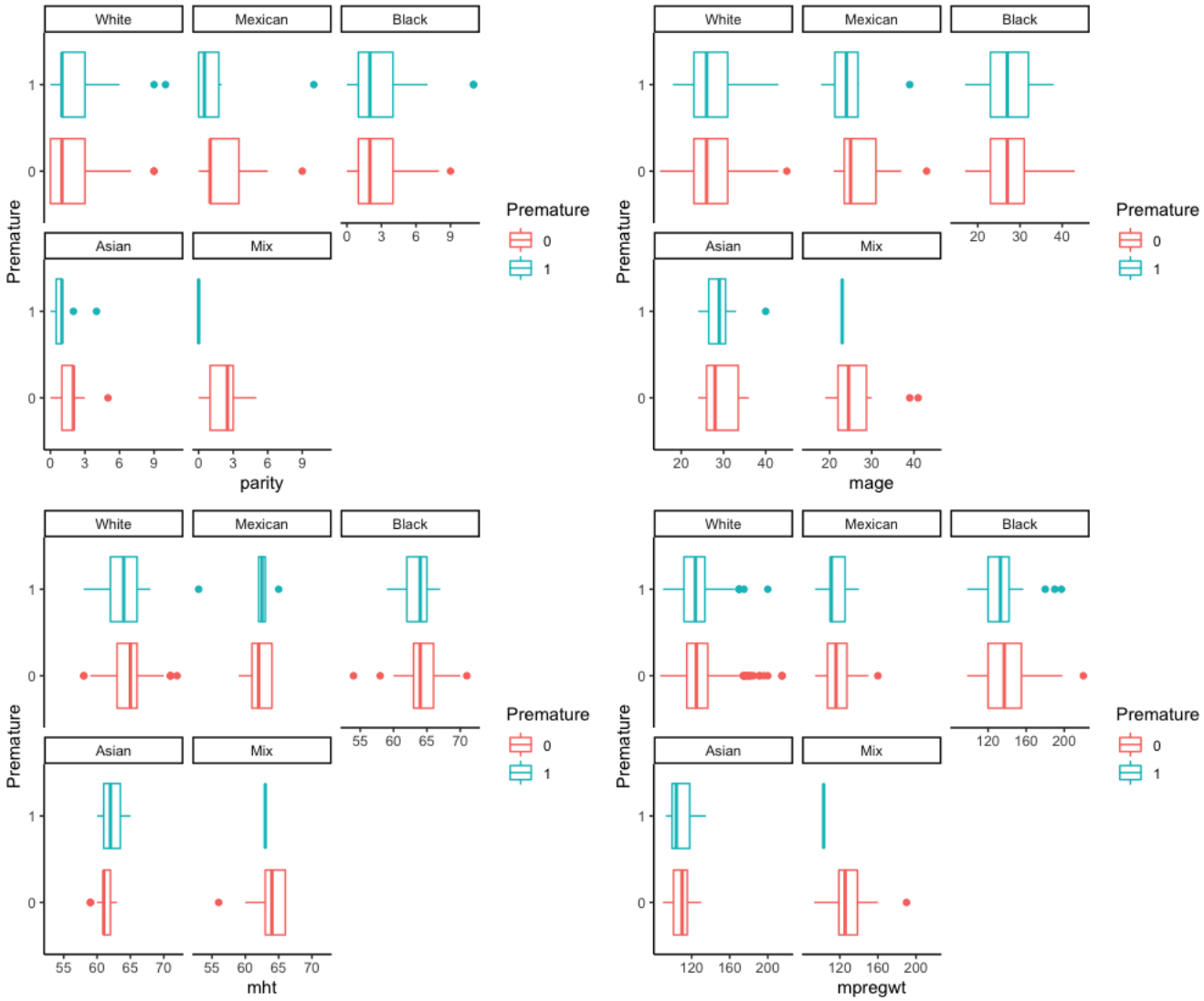

Figure 3: EDA for Interactions

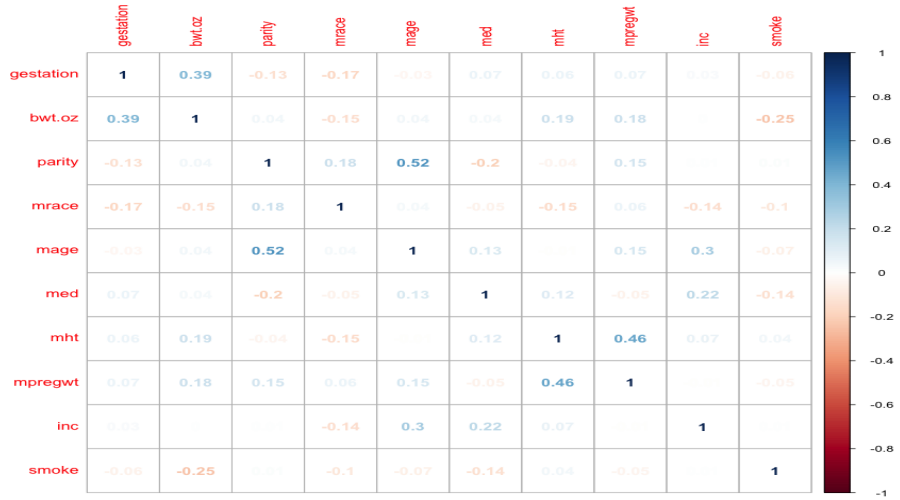

Figure 4: Correlation Plot

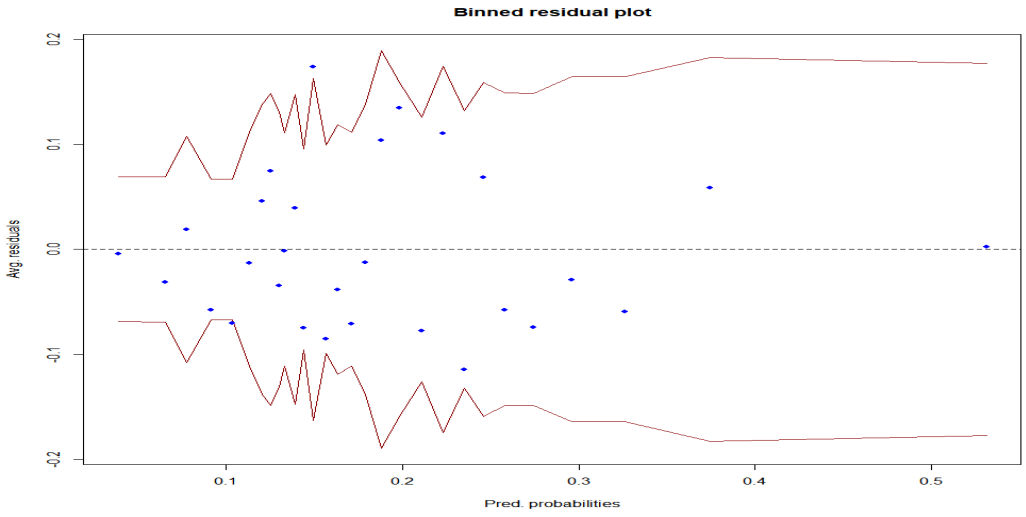

Figure 5: Model Assessment

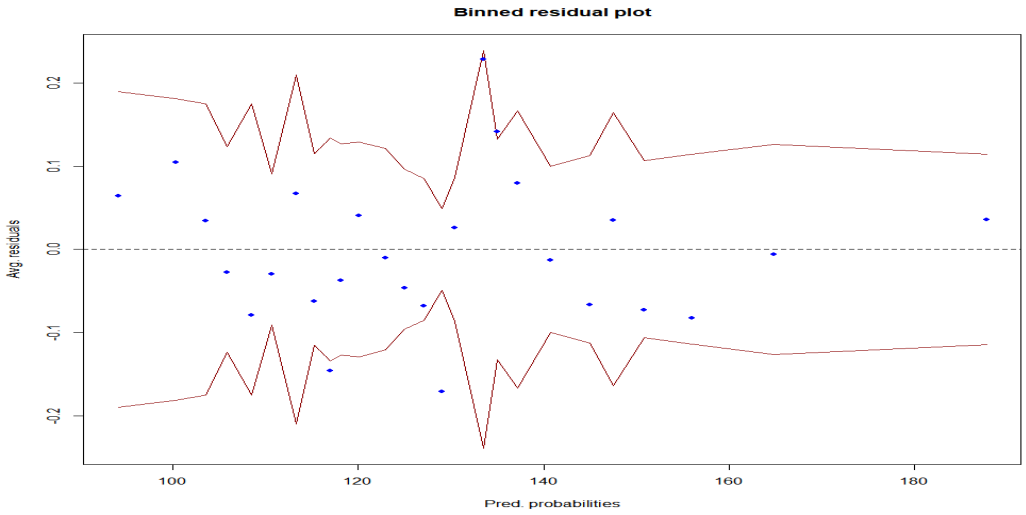

Figure 6: Model Assessment with Predictor

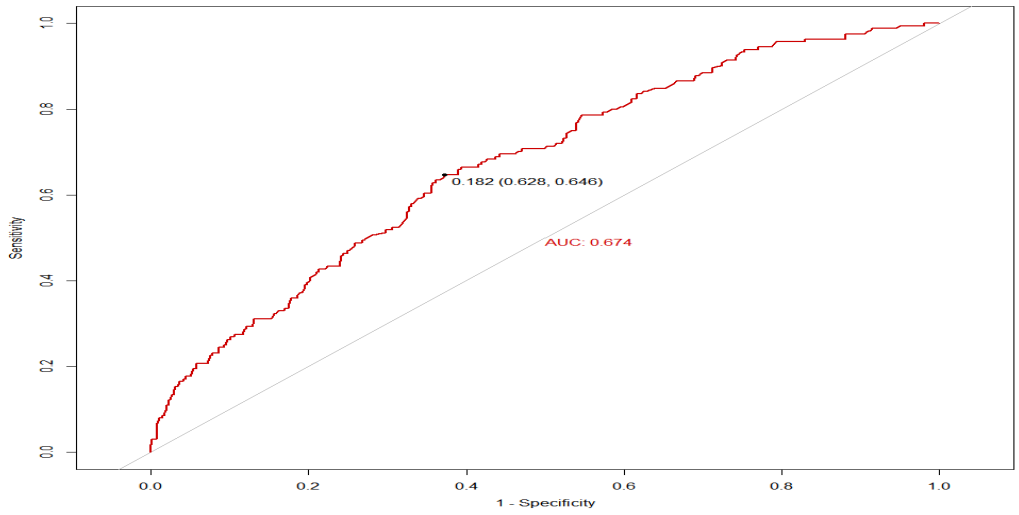

Figure 7: ROC Plot