MATERNAL SMOKING AND PRE-TERM BIRTH ANALYSIS

SUMMARY:

In this analysis, we want to explore what variables will affect pre-term birth using a logistic regression. We found smoking doesn't have great effect on pre-term birth and smoking also doesn't have noticeable effect on pre-term birth by different race. Mother's race and mother's pregnancy weight have bigger effect on pre-term birth.

INTRODUCTION:

In this report, I'm going to work on the same modified data set "smoking.csv" to explore the relationship between maternal smoking and pre-term birth(gestational age less than 270 days). The data set is collected by Child Health and Development Studies back to 1960s with 869 observations and 13 variables. The purpose of the analysis is to decide whether mothers who smoke would have a higher chance of pre-term birth compared to mothers who don't smoke; examine if there is evidence showing the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race and explore other interesting association with the odds of pre-term birth.

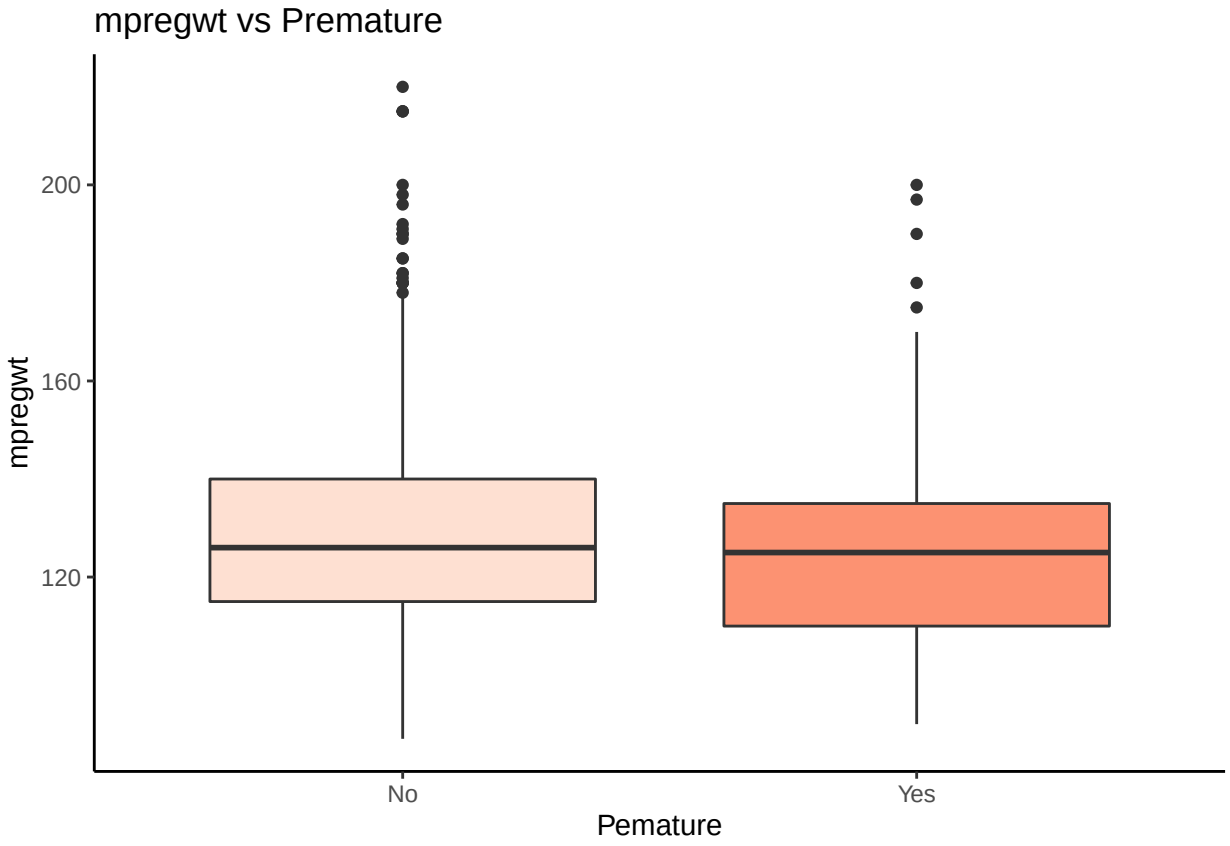
DATA:

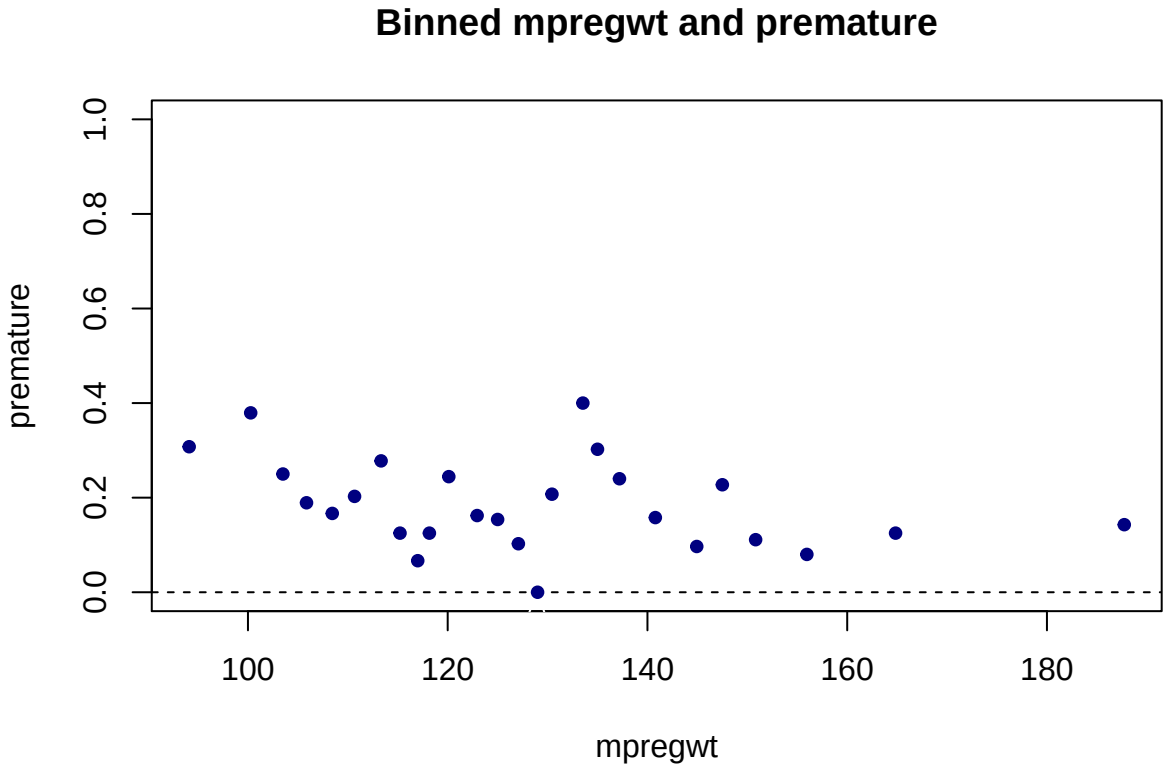
smoking.csv is out data set, which is a pre-processed data set with missing value and father's feature removed from the original data set. There are 869 observations and 14 columns in the data.**id**, **date**, **bwt.oz** and **gestation** are irrelevant to our analysis. **parity** is a continuous variable ranging from 0 to 11. **mrace** is a unbalanced categorical variable with **mrace=0** greater than the sum of other races. **mage** is a continuous variable with mean equals to 27.29. **med** is a categorical variable describing mothers' education levels.**mht** is a continuous variable with mean equals to 64.07. **mpregwt** is a continuous variable with mean equals to 128.5. **Inc** is categorical variable and **smoke** is a binary variable. Finally **premature** is our response variable for the analysis but it's extremely unbalanced with non-premature=705 and premature=164 so our model might be affected by the fact.

premature	smoke	parity	mrace	mage	mpregwt	med	mht	inc
No :705	0:466	Min. : 0.000	0:626	Min. :15.00	Min. : 87.0	0: 5	Min. :53.00	1 :153
Yes:164	1:403	Median : 2.000	6: 25	Median :26.00	Median :125.0	1:130	Median :64.00	2 :146
		Mean : 1.953	7:169	Mean :27.29	Mean :128.5	2:321	Mean :64.07	3 :136
		Max. :11.000	8: 34	Max. :45.00	Max. :220.0	3: 47	Max. :72.00	7 :111
			9: 15			4:203		4 :105
						5:159		5 : 98
						7: 4		(Other):120

After plotting boxplots of **mage vs premature**, **mht vs premature**, **parity vs premature**, there are no obvious difference in the distribution between premature and non-premature cases. They all have roughly the same range and medium value. However, the boxplot of **mprgwt vs premature** shows that non-premature group has a higher medium value and the distribution of non-premature cases is higher than the premature cases. So I think that mprgwt might have relationship with the response variable. The binned plot of mpregwt vs premature also suggests that as a mother's pregnancy weight increases the probability of having premature baby will decrease. However, the trend is not linear so I might need log transformation on it. Furthermore, I found premature cases differ by mother's race. The probability of having premature babies of Asian and Black mothers are 8% higher than White mother. And the probability of have premature babies decreases as mother's education level increases

Since we are interesting in whether smoking mothers are more likely to give pre-term birth, I used a chi-square test on **smoke** column and **premature** column. The p-value of the test is $0.07 > 0.05$, which suggests smoke is not significant to pre-term birth. I'll also check smoke in my model later.





MODEL:

I used a Logistic Regression model to fit the data using formula:

$$y_i|x_i \ \sim Bernoulli(\pi_i)$$

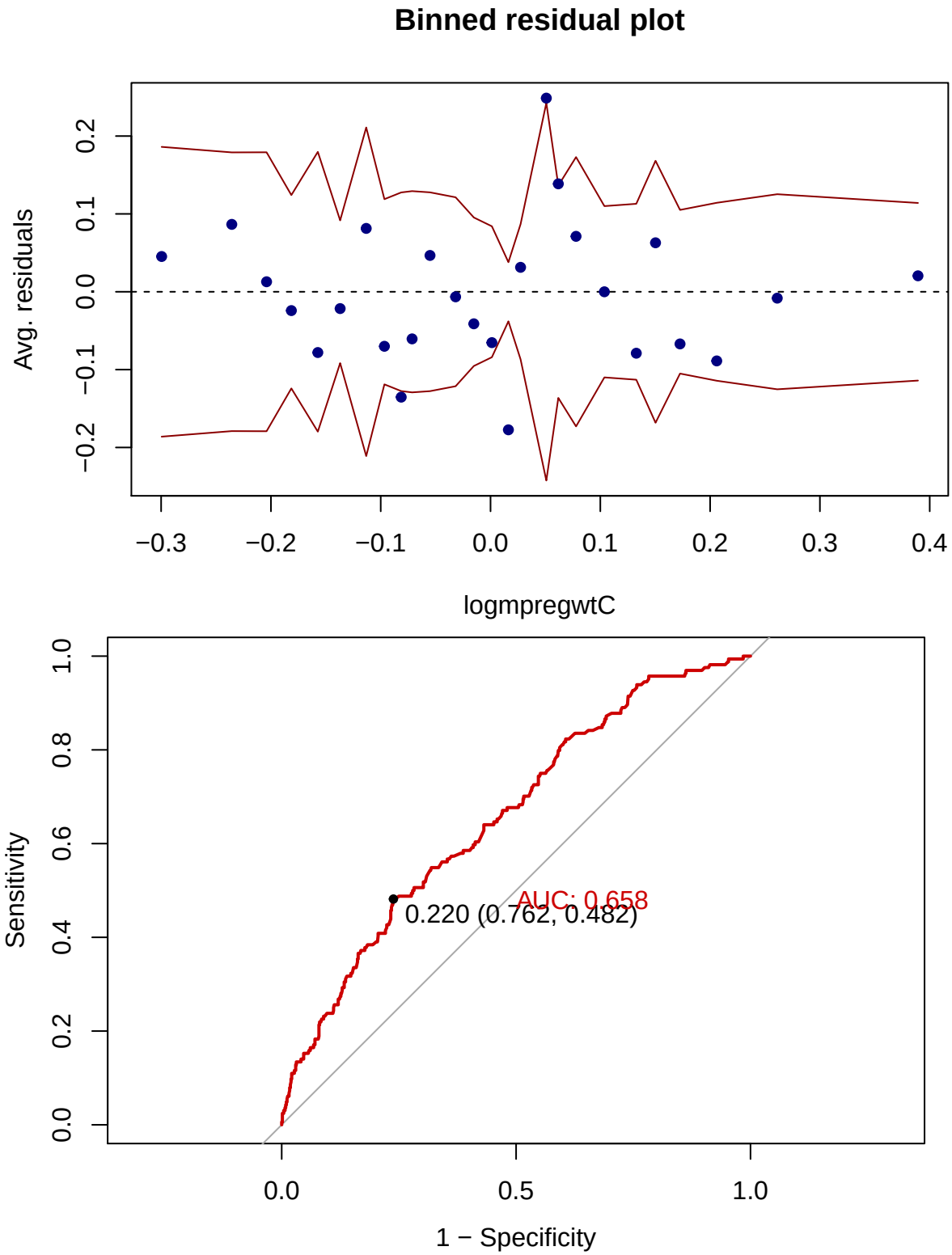
$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4}$$

where y_i is binary response variable *premature*, x_{i1} is a binary predictor *smoke*, x_{i2} is a categorical variable *mrace*, x_{i3} is a categorical variable *med* and x_{i4} is a continuous variable *logpregwtC*, which is log of centered mpregwt.

Starting with the null model equals to $glm(premature \sim smoke + mrace + smoke * mrace)$, I utilized step-wise selection within the scope $premature \sim parity + mhtC + logmpregwtC + smoke + mrace + med + inc + smoke * mrace$ with AIC as the selection criteria. The final model has AIC equals to 821.59. From the binned residual plot versus predicted probabilities, the points are randomly distributed and 3% of points out of 95% bands. From the binned residual plot versus logpregwtC, 11% points are outside the bands. There are some points have leverage scores higher than the threshold(2*13/869==0.03) but none of them are influential points so we don't have outliers in our analysis. For multicollinearity, the model doesn't have high multicollinearity because the VIFs < 2 for all predictors(smoke:1.07,mrace:1.35,med:1.21,logpregwt:1,14).

I used threshold=0.5 to compute the confusion matrix and the model has accuracy=0.81, sensitivity=0.99 and specificity=0.02. The extremely low specificity makes sense since the response variable is unbalanced and I set the threshold to be 0.5. From ROC curve, the best threshold to predict y-label is 0.22 and the model has AUC=0.658, Sensitivity=0.762 and 1-Specificity=0.482, which seems to a valid result. Lastly, the residue-deviance for the final model is 796(<842 null deviance) with change-in-deviance equals to 46.

To understand if the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race, I ran the anova test between final model and final model plus the interaction term $mrace * smoke$. The p-value is 0.27 which means the interaction term is not significant. Therefore, the odds ratio of pre-term birth for smokers and non-smokers doesn't differ by mother's race.



Result: logpregwt, mrace are the only significant predictors in the model so I'm going to interpret the two terms. For every 1% increase in a mother's pregnancy weight and everything unchanged, the odds of pre-term birth is multiplied by 0.2. For Black mothers and everything else unchanged, the odds of pre-term birth is multiplied by 2.16. For Asian mothers and everything else unchanged, the odds of pre-term birth is multiplied by 2.41. Other information regarding the non-significant predictors is included in the table below. The 95%

confidence interval for the odd ratios of logmpregwtC is between -2.89 and -0.42. The 95% confidence interval for the odd ratios of African mothers is between 0.33 and 1.20. And the 95% confidence interval for Asian mothers is between 0.05 and 1.67. And because the p-value for smoke is equal to $0.12 > 0.05$, we know that smoke is not significant to pre-term birth. And, the 95% range for the odds ratio of pre-term birth for smokers and non-smokers is between -0.07 and 0.65

	Estimate	Std. Error	Pr(>|z|)	2.5 %	97.5 %
(Intercept)	-0.86	0.94	0.36	-2.93	0.98
smoke1	0.29	0.18	0.12	-0.07	0.65
mrace6	0.14	0.52	0.78	-0.96	1.10
mrace7	0.77	0.22	0	0.33	1.20
mrace8	0.88	0.41	0.03	0.05	1.67
mrace9	-0.76	1.1	0.47	-3.67	0.89
med1	-0.56	0.95	0.56	-2.42	1.53
med2	-0.9	0.94	0.34	-2.75	1.17
med3	-0.72	0.99	0.47	-2.67	1.43
med4	-1.6	0.96	0.1	-3.44	0.53
med5	-1.1	0.96	0.26	-2.95	1.02
med7	1.8	1.5	0.22	-0.92	5.27
logmpregwtC	-1.6	0.63	0.01	-2.89	-0.42
Residual deviance	796				

Conclusion:

The analysis suggest that there is no association between smoke and pre-term birth because smoke is not a significant term in our final model and the chi-squared test between **smoke** and **premature**from the EDA also suggests the same conclusion since the p-value(0.07) is greater than 0.05. The odds ratio of pre-term birth for smokers and non-smokers is between -0.07 and 0.65. Besides that, the analysis also suggests that there is no difference in odds ratio of pre-term birth for smokers and non-smokers among mothers' race because we performed F-test between two models that are differ by the interaction term **mrace*smoke**. The p-val from the result is 0.27, which means the interaction term is not significant in our analysis. Therefore, there is no evidence that the odds ratio of pre-term birth for smokers and non-smokers differs by mother's race.

It's interesting to notice that pre-term birth probability differs by mother race. To be specific, White mothers have less likelihood to have premature babies than Asian, African and Mexican mothers. The White mothers are 7~8% less likely to have pre-term birth.

The limitation with the analysis is that the response variable is highly unbalanced. The number of non-premature cases are 6 times the premature cases. Another limitation is the metric. The AUC from the ROC curve is only 0.658 and it can be improved to 0.67 by including more predictors. However, that will introduce more uncertainties to our analysis.