Assignment_3

Anshupriya Srivastava (netID: as996)

9/25/2019

Introduction

The data for this exercise has been obtained from an Observational Study conducted by Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. The original dataset contains details of about 15,000 families but we have taken into account 869 male single births where the baby lived for at least 28 days. The mothers were interviewed quite early in their pregnancy to collect information on socioeconomic and demographic characteristics, with an indicator of whether the mother smoked during pregnancy. Our aim is to analyze this data and identify any associations between smoking and birth weight. By the doing so, we aim to 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?

Data

The data for this exercise was made available Sakai - filename "smoking.csv" has been used to answer the questions mentioned above. The list of variables present in the dataframe:

*id: (numeric): id number
*date: birth date where 1096 = January1, 1961
*gestation: length of gestation in days
*bwt.oz: birth weight in ounces
*parity: total number of previous pregnancies, including fetal deaths and still births
*mrace:mother's race or ethnicity
*mage: mother's age in years at termination of pregnancy
*med: mother's education
*mht: mother's height in inches
*mpregwt: mother's pre-pregnancy weight in pounds
*inc: family yearly income in 2500 increments
*smoke: does mother smoke?

After evaluating the structure of the data, we can see that apart from mother's race (mrace), all other variables are numeric. The following **transformations** has been made:

*Race category 0 - 5 has been collapsed into one category for race = white.
*A new column Premature has been added where 1 indicates the birth of a premature baby and 0 indicates that the baby is not premature. This has been calculated using gestation where if gestation < 270, Premture = 1 else Premature = 0.

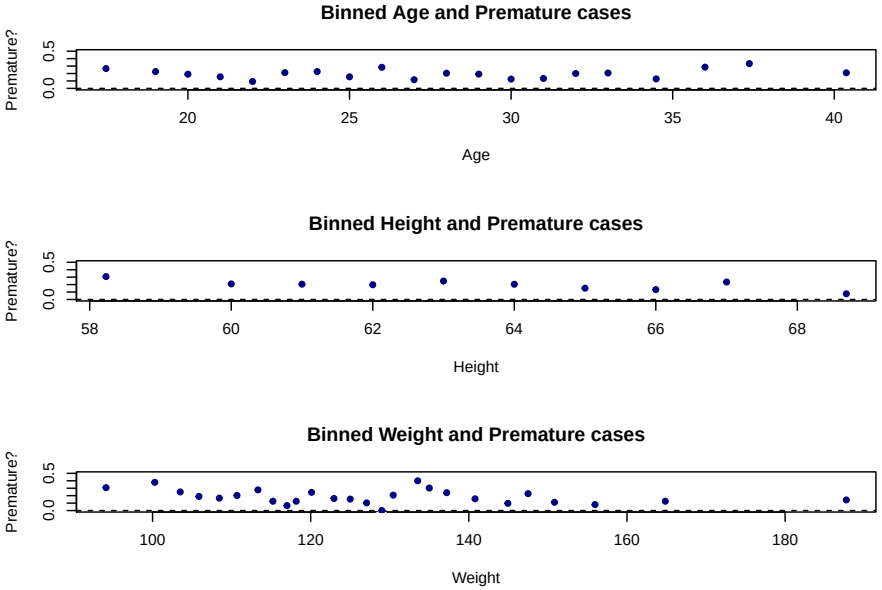
This assignment aims to understand the association between smoking and premature birth. The first analysis is performed on these two variables show that the conditional probability that the mother will have a premature child given she smokes is 0.165 or 16.5%.

Table 1: Relationship of continuous variables with Premature Birth

Variable	p.value	Is.Associated
mrace	0.003561	Yes
smoke	0.0694	No
parity	0.01125	Yes
med	0.0005476	Yes
inc	0.9087	No

The next set of analysis is between Premature and Categorical/Discrete variables. Using Chi-square test we can determine if they are associated. In this case since the discrete variables are treated as categorical variables. The p-values help in determining association. If the p-value is less than 0.5 we can say that there is a strong evidence against the null hypothesis that suggests that no relationship exists in the categorical variables in the population; they are independent. Table1 shows the associations of race, smoke, parity, education and income with Premature births. Even though the association between smoke and premature is False we proceed with using it in our model because that is the requirement of this exercise.

The graphs below show the associations between Premature and Continuous variables. Using a binned plot we can determine if they show any interesting relationship. Height, Weight and Age are continous variables.



The graphs with Age and Height follow an almost linear relationship. It seems like the data points in the third graph comparing mother's pre-pregnancy weight and the probability of premature birth is showing a non-linaer trend and is concentrated towards one side, but that can be due to lesser number of data points.

The boxplots in Appendix B graphs for Premature birth relationship with age, weight and height based on whether smoke = 1 or 0 to explore interaction. Visually the relationship with height is showing a slight difference. Apart from that there is no other relationship of interest.

Model Selection

For model selection the continuous variables are being centered. This helps in better interpretation of the intercept. Since none of these variables have a meaningful 0 value we use a centered value.

1. **Model_all**: The initial model is created using all the variables and then a model is chosen by AIC in a stepwise Alogrithm.
2. **Model_AIC**: The results of the stepwise algorithm show that variables race, weight, education and

Table 2: Model Comparison

Model.Name	Accuracy	Sensitivity	Specificity	AUC
model_all	60.07%	57.93%	60.57%	63.8%
model_AIC	60.3%	59.76%	60.43%	63.1%
model_Interaction	59.15%	59.76%	59.01%	63.8%

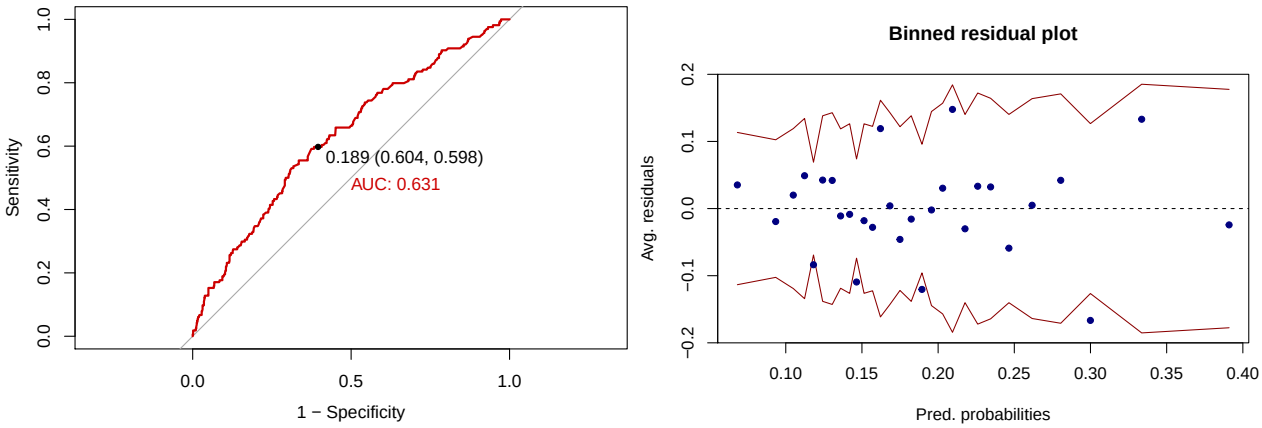
smoke are significant. The next model is created using the results of AIC. These variables are - *mrace*, *med*, *mpregwt_c* and *smoke*.

3. **Model_Interaction**: The third and final model is used by *interacting smoke and race* along with *mrace*, *med*, *mpregwt_c* and *smoke*.

Looking at **Table2** we can see that the second model [formula = premature ~ mrace + med + mpregwt_c + smoke] has a higher accuracy and specificity (True Negative). The AUC and sensitivity (True Positive) values are comparable. Thus choosen model is model_AIC. [Please refer to Appendix B for the other models]

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```



The ROC Plot shows that the AUC value is 63.8%. It tells that model is capable of distinguishing between classes 63.8% of the times.

The binned residual plot final model has no identifiable pattern. There is one data point that lies on the confidence interval boundary. There are three data points that lie outside the confidence interval.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.4497753	0.2445024	-5.9294938	0.0000000
mracemexican	0.4030628	0.4933486	0.8169939	0.4139319
mraceblack	0.7114898	0.2132887	3.3358060	0.0008505
mraceasian	0.9438757	0.4023839	2.3457092	0.0189909
mracemix	-0.8932369	1.0453244	-0.8545068	0.3928242
med	-0.1343014	0.0650226	-2.0654588	0.0388796
mpregwt_c	-0.0111149	0.0047249	-2.3524048	0.0186525
smoke	0.3091987	0.1810804	1.7075219	0.0877251

Model Comparison using ANOVA.

Using chi.square test in ANOVA a high p-value of 0.3028 shows that adding the an interaction does not affect the model. So, we proceed with model_AIC.

Conclusions

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?

In the chosen model [model_AIC] the coefficient is 0.309. Thus the log odds that if Smoke = 1 increases the

This implies that the chance of having a premature baby increases by a factor of 1.36 or 36% if smoke = 1.

At 95% confidence level - (0.96, 1.95)

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.

In the model model_Interaction we have used the Interaction between smoke and Race. Since the p-value is high we cannot determine any significant relationship between them. If we look at the regression output Appendix A we can see that race = Black and race = Asian are the two significant groups. Overall there seems to be no significance of adding the interaction term. Also, in comparison to the model_AIC the chisquare result was 0.3028. Thus the interaction terms are not significant.

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

In the final model model_AIC, we are using **med**, **mpregwt** and **mrace** apart from **smoke**.

1. For every level increase in education the odd's ratio decreases by 12.6% keeping all the other variables constant. At 95% confidence level - (0.77, 0.99)
2. For every unit increase in mpregwt the odd's ratio decreases by 1.1% keeping all the other variables constant. At 95% confidence level - (0.98, 0.99)

For race white is the baseline level. Also, the following results are exluding race = mexican and race = mix since they are not signicant.

3. For every level change in mrace (white -> black) the odd's ratio increases by 103% keeping all the other variables constant. At 95% confidence level - (1.33, 3.08)
4. For every level change in mrace (white -> asian) the odd's ratio increases by 156% keeping all the other variables constant. At 95% confidence level - (1.13, 5.56)

Limitations

1. The dataset does not contain enough data points for some categories. For example our binned plot for weight shows concentration in one direction. This skewness causes discrepencies.
2. We do not have enough data points for Premature = 1. Only 18% of the data represents Premature = 1.
3. Race also does not contain enough datapoints for all races apart from the baseline group i.e. American. Thus the result of the interaction between Smoke and Race is not reliable.
4. The chisquare test was giving a warning that the results might not be accurate due to lesser number of data points.

Appendix A

Regression Output

model_all

term	estimate	std.error	statistic	p.value
(Intercept)	-1.4725647	0.3241747	-4.5425038	0.0000056
parity	-0.0024367	0.0581759	-0.0418858	0.9665897
mracemexican	0.3339064	0.5024619	0.6645407	0.5063443
mraceblack	0.6973142	0.2242664	3.1093116	0.0018752

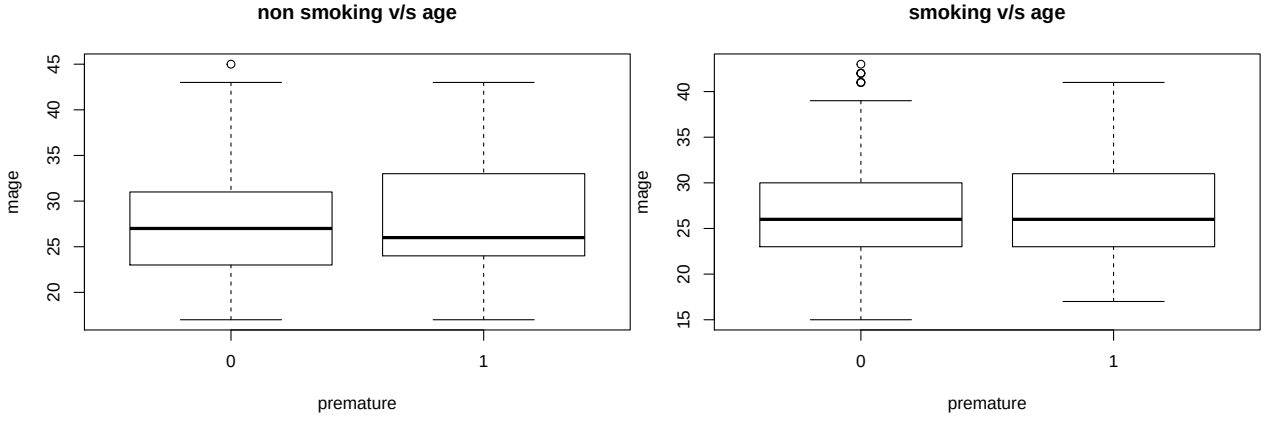
term	estimate	std.error	statistic	p.value
mraceasian	0.8363604	0.4100381	2.0397137	0.0413789
mracemix	-0.9171866	1.0488436	-0.8744742	0.3818601
mage_c	0.0162837	0.0198199	0.8215792	0.4113165
med	-0.1351143	0.0698689	-1.9338260	0.0531345
mht_c	-0.0377231	0.0412395	-0.9147315	0.3603326
mpregwt_c	-0.0098473	0.0053788	-1.8307753	0.0671341
smoke	0.3237795	0.1819187	1.7798038	0.0751081
inc	0.0074416	0.0421524	0.1765409	0.8598690

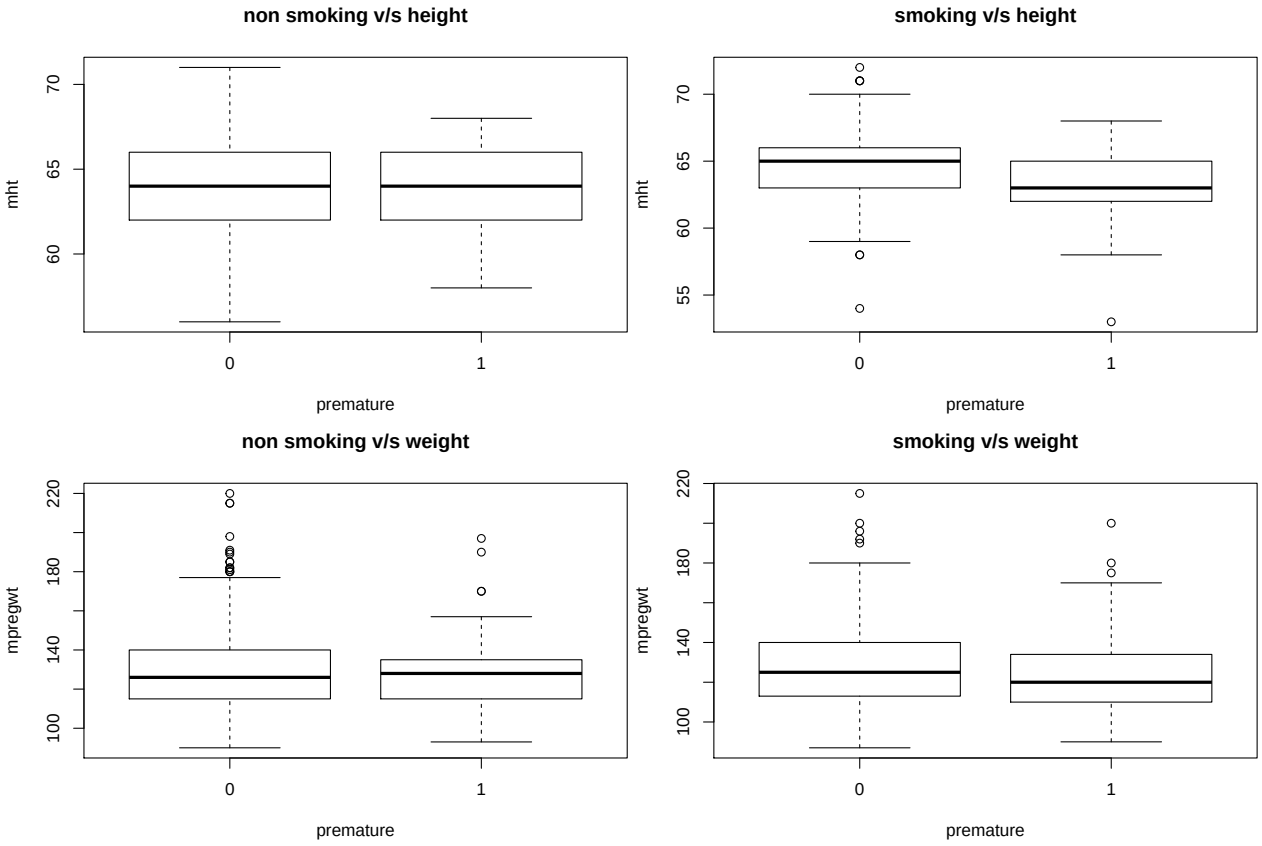
model_Interaction

term	estimate	std.error	statistic	p.value
(Intercept)	-1.5104366	0.2583407	-5.8466849	0.0000000
mracemexican	0.4492760	0.5926977	0.7580189	0.4484397
mraceblack	0.9613186	0.2951844	3.2566713	0.0011273
mraceasian	0.8752855	0.4889209	1.7902394	0.0734154
mracemix	-13.6958902	417.7754167	-0.0327829	0.9738477
med	-0.1324332	0.0654268	-2.0241425	0.0429555
mpregwt_c	-0.0115803	0.0047123	-2.4574664	0.0139921
smoke	0.4039585	0.2233468	1.8086604	0.0705038
mracemexican:smoke	-0.0777030	1.0647841	-0.0729753	0.9418258
mraceblack:smoke	-0.5041850	0.4175270	-1.2075508	0.2272201
mraceasian:smoke	0.2843847	0.8396047	0.3387125	0.7348263
mracemix:smoke	14.5168746	417.7772958	0.0347479	0.9722808

Appendix B

EDA Plots to check for interactions of variables with smoking

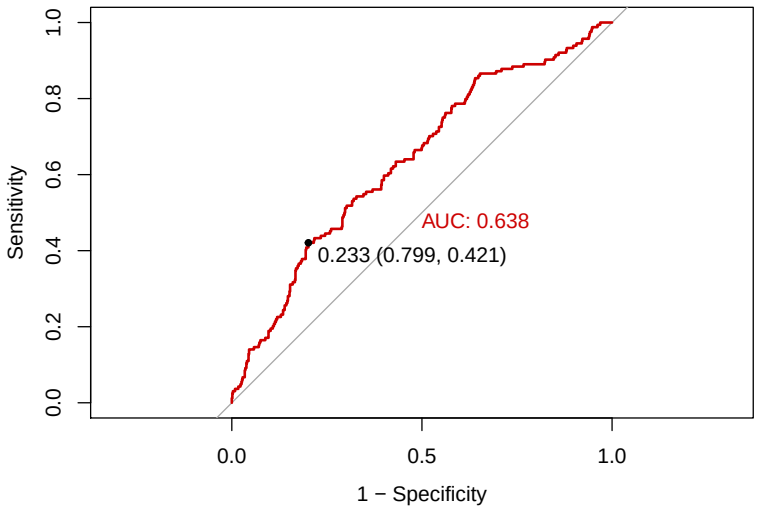




ROC

Model_all

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Model_Interaction

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```