Data Analysis Assignment 3

Surabhi Trivedi

20th Sept 2021

Maternal Smoking and Gestation Period

Summary

This report investigates if mothers who smoke tend to give birth to premature babies (gestation age < 270 days) compared to mothers who do not smoke (gestation age >= 270 days). To infer this I have used logistic regression. The data used contains multiple variables about the mother such as height, level of education, mother's ethnicity etc. The final model reports the variable *smoke* (binary variable indicating whether the mother smokes or not) as a statistically insignificant predictor. Additionally, the interaction effects between *smoke* and *mrace* are also found to be statistically insignificant.

Introduction

In this report I am primarily interested in establishing if mothers who smoke give birth to babies with a period of gestation < 270 days. The secondary goal of this report is to identify the odds ratio of pre-term births for mothers who smoke vs mothers who do not. Additionally, I am also interested in reporting if the effect of smoking, on a baby's gestation period, disproportionately increases or decreases based on the mother's race/ ethnicity. Furthermore, I also report findings that I found interesting but were not directly related to the mother's smoking status.

Data

The dataset of interest contains data for 869 male single births where the baby lived up to atleast 28 days. The response variable *premature* is a factor variable with levels 1 and 0, (1 - indicates pre-term birth). There are 8 predictor variables in the dataset of which 4 are categorical and 4 are continuous/ discreet. This dataset contains no missing value.

The original dataset encoded all variables as either num or int. I converted 4 of these variables to factor variables in order to make the analysis more meaningful and for the ease of interpretation. These variables were *mrace, med, inc, smoke.* I would also like to mention that for mrace, levels 0-5 were collapsed to one level: 0 (as 0-5 were white mothers). Similarly, for med, levels 6-7 were collapsed to level: 6 (as 6-7 were mothers with trade school education but unclear if they graduated from high school).

Initial EDA of the data revealed that the data collected is imbalanced with respect to race. Of the 869 mothers present in the dataset, 626 were white and 169 were black (as can be seen in the plot below). The other ethnicities are underrepresented in the dataset – this may make it harder for us to measure and interpret interaction between the variables *smoke* and *mrace.* All the other predictors are well balanced in their distribution. More importantly however, the response variable *premature* is very imbalanced. Out of the 869 data points, we have data for only 164 pre-term births compared to 705 normal births. This may effect the results of our analysis.

During exploratory data analysis, I saw that on making comparative boxplots for *premature* and *mpregwt*, there was a significant difference in the median of the mother's pre-pregnancy weight for mothers who had pre-term babies compared to mothers who didn't, prompting me to pay particular attention to this predictor during model building as it might be statistically significant. On the other hand, the chi-squared test for independence between *smoke* and *premature* was insignificant.

I interestingly also found that the probability of having a preterm baby when the mother is white is considerably lower (~16%) when compared to black mothers (~27%) as can also be seen in table 1 below. The chi-squared test was also significant for this relationship. I want to refrain from commenting on other races as we don't have enough data for those races (asian, mixed, mexican).

Table 1: Conditional Probabilities for Premature Babies vs Mother's Race

Pre-term Birth	White	Mexican	Black	Asian	Mixed
0	0.84	0.76	0.73	0.68	0.93
1	0.16	0.24	0.27	0.32	0.07

Model

For model fitting, I started with a very basic model (base model) with all the predictors shown below. This was my base model.

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

$$log(\pi_i/1-\pi_i) = \beta_0+\beta_1 mage_{i1}+\beta_2 parity_{i2}+\beta_3 mrace_{i3}+\beta_4 med_{i4}+\beta_5 mht_{i5}+\beta_6 mpregwt_{i6}+\beta_7 inc_{i7}+\beta_8 smoke_{i8}$$

I then assessed the model by plotting binned residuals vs predicted probabilities which wasn't exactly random and hence in violation of our model requirements. The plots for average residuals of *mage* and *mpregwt* with binned residuals looked good. However, the plots for average residuals of *mht* and *parity* vs binned residuals had multiple increasing/ decreasing trends and weren't random - this was due to lack of data in certain bins. For example, we don't have a lot of data for mother's with heights between 53-60 cm and 68-72 cm. I considered making *mht* and *parity* categorical but decided against it as were interested in understanding how a 1 unit increase in these predictors would effect our response variable. I proceeded to calculated the AUC for this model which came out to be 66.7%.

Next, I checked for multicollinearity between variables and the VIF for *med* were > 20. Based on this insight, I proceeded to check for variable interactions between *med* and the other variables. I found that the interaction between *med* and *smoke* was significant by conducting an F-test shown below. After including the interaction between *med* and *smoke* in my model, the VIF for all predictors were well within range. Next I checked for interaction between *mrace* and *smoke* and it was not significant. But I would like to point out that the dataset is not representative of all races and this inference may change if we could collect more data for asian, mexican, and mixed race mothers. I also checked for interactions between other variables and the interaction between *mht* and *smoke* was significant but at 0.1 significance level (F-test) which is why I did not include it in my model selection below. No other interactions were found to be significant.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
838	776.859	NA	NA	NA
844	790.212	-6	-13.353	0.038

For model selection, I used stepwise logistic model regression with BIC initially and then proceeded to use AIC for reasons mentioned below. Following are my null and full models. I included both *smoke* and *mrace* in my null model as those are the predictors we're interested in inferring. For my full model I included all other predictors and the interaction between *med* and *smoke* as no other interaction was found to be significant.

Null model:

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

$$log(\pi_i/1-\pi_i) = \beta_0 + \beta_1 mrace_{i1} + \beta_2 smoke_{i2}$$

Full model:

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

$$log(\pi_i/1-\pi_i) = \beta_0+\beta_1 mage_{i1}+\beta_2 parity_{i2}+\beta_3 mrace_{i3}+(\beta_4 med_{i4}*\beta_5 smoke_{i5})+\beta_6 mpregwt_{i6}+\beta_7 inc_{i7}+\beta_8 mht_{i8}$$

Stepwise regression with BIC returned an only intercept model so I proceeded with AIC. Below are the results from the step-wise regression with AIC, which I also chose as my final model for the following reasons:

1.The VIF for all predictors in my model were well within range

2.The predictor *mrace* for black mothers was significant at 0.001 significance level which aligned with my EDA

3.The predictor *mpregwt* was significant at 0.001 significance level which aligned with my EDA

4.The AUC for this model was **67.4%** compared to the AUC of my base model (66.7%)

5.The residual deviance for this model **(782.05)** was lower in comparison to my base model **(790.21)**

6.The binned residual vs predicted probabilities plot for my model is reasonably random as shown below. The plots for the avg residuals of the predictors vs the binned residuals are also random

7.The variables *parity* and *mht* were left out and I am happy with that because as I also mentioned earlier they did not have enough data for a lot of the values. Their plots for avg residuals vs binned values were also not random

Final Model:

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

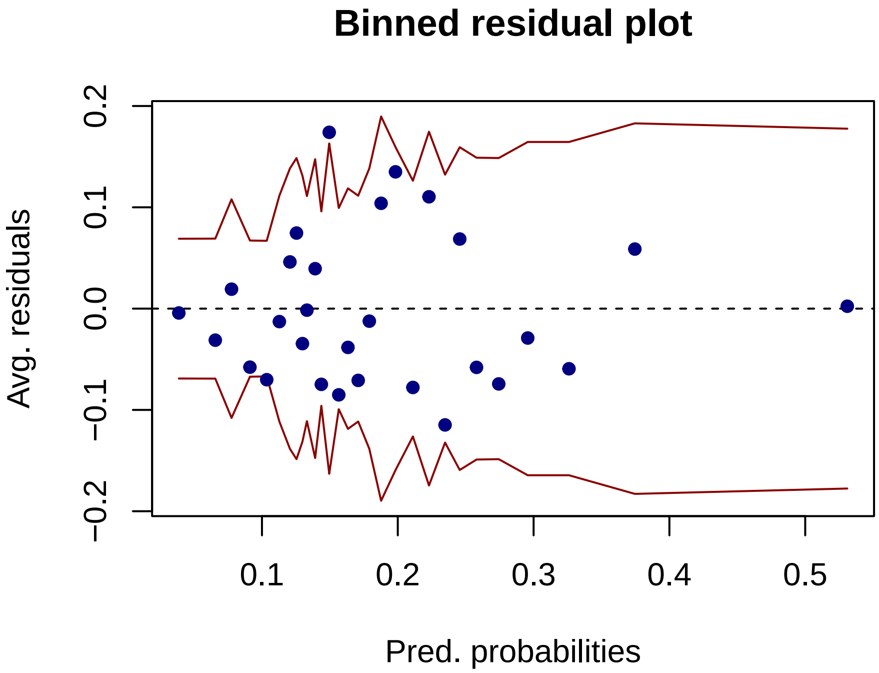
$$log(\pi_i/1-\pi_i) = \beta_0 + \beta_1 mrace_{i1} + (\beta_2 med_{i2} * \beta_3 smoke_{i3}) + \beta_4 mpregwt_{i4}$$

Observations	869
Dependent variable	premature
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(18)$	59.77
Pseudo-R² (Cragg-Uhler)	0.11
Pseudo-R² (McFadden)	0.07
AIC	820.05
BIC	910.63

	Est.	S.E.	z val.	p
(Intercept)	-14.21	834.30	-0.02	0.99
smoke1	31.08	1322.77	0.02	0.98
mrace6	-0.04	0.55	-0.07	0.95
mrace7	0.73	0.23	3.25	0.00
mrace8	0.88	0.41	2.14	0.03
mrace9	-0.80	1.06	-0.76	0.45
med1	14.77	834.30	0.02	0.99
med2	13.98	834.30	0.02	0.99
med3	15.11	834.30	0.02	0.99
med4	13.25	834.30	0.02	0.99
med5	13.94	834.30	0.02	0.99
med6	16.51	834.30	0.02	0.98
mpregwt	-0.01	0.00	-2.58	0.01
smoke1:med1	-31.30	1322.77	-0.02	0.98
smoke1:med2	-30.57	1322.77	-0.02	0.98
smoke1:med3	-32.34	1322.77	-0.02	0.98
smoke1:med4	-30.40	1322.77	-0.02	0.98
smoke1:med5	-30.85	1322.77	-0.02	0.98
smoke1:med6	-16.22	1966.70	-0.01	0.99

Standard errors: MLE



Model Interpretation: From my model I can infer that there is no statistically significant relationship between the smoking status of a mother, *smoke*, and pre-term birth, *premature*. I can also infer that the odds ratio for a black mother having a premature baby is almost twice that of white mothers **(2.081)**. Additionally, the odds of a mother having a premature baby for a 1 pound increase in the mother's pre-pregnancy weight decreases by **0.995%**.

Conclusion

In conclusion, I would like to say that there is no statistical proof that the odds ratio of giving birth to a premature baby for mother's who smoke is any higher than for mothers who do not smoke. I also report that the interaction between the mother's smoking status, *smoke* and the mother's race, *mrace* is not statistically significant.

The final model and the analyses reported above do suffer from some drawbacks however:

1.The variable *mrace* is not well represented which put's in question the significance of this variable in our model

2.The predictors *parity* and *mht* do not have enough data points for certain values

3.The predictor *med* and *inc* have a lot of 0's for a lot of the same values which makes it hard to check for interaction between these two variables. I wanted to check for this interaction because *med* has a high VIF and intuitively it makes sense for *med* and *inc* to be correlated

4.The response variable is not well balanced in its distribution: 164 1s vs 705 0s

5.The most major drawback of this model is that it does not take into account the health statistics of the father at all

---

R Code Appendix

```
###########################################################################
###########################################################################
################## Maternal Smoking and Gestational Age ###################
###########################################################################
###########################################################################

###### Clear environment and load libraries
rm(list = ls())
library(arm)
library(pROC)
library(e1071)
library(caret)
library(ggplot2)
require(gridExtra)
library(dplyr)
library(rms) #for VIF
library(MASS)

###### Load the data
smoking <- read.csv("smoking.csv")
smoking <- smoking[c(-2,-4)]

#making the response variable
smoking$premature[smoking$gestation < 270] <- 1
smoking$premature[smoking$gestation >= 270] <- 0

#Collapsing race and education categories for easier analysis
smoking$med[smoking$med == 7] <- 6
```

```r
smoking$mrace[smoking$mrace == 1 | smoking$mrace == 2 | smoking$mrace == 3 | smoking$mrace == 4 | smokin

#converting vars from num to factor
smoking[,'mrace']<-factor(smoking[,'mrace'])
smoking[,'med']<-factor(smoking[,'med'])
smoking[,'inc']<-factor(smoking[,'inc'])
smoking[,'smoke']<-factor(smoking[,'smoke'])
smoking[,'premature']<-factor(smoking[,'premature'])
smoking <- smoking[c(-2)]

#looking at the data
summary(smoking)
describe(smoking)
table(smoking$premature)
#the data may not as balanced in the response variable as we would want it to be

###### Exploratory data analysis

#EDA for continuous variables
# parity vs premature
ggplot(smoking,aes(x=premature, y=parity, fill=premature)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Parity vs Premature",
       x="Had premature baby?",y="Parity") +
  theme_classic() + theme(legend.position="none")
#no difference in median

# mage vs premature
ggplot(smoking,aes(x=premature, y=mage, fill=premature)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="mage vs Premature",
       x="Had premature baby?",y="Mother's age") +
  theme_classic() + theme(legend.position="none")
#can see some difference in median - might be a variable of interest

# mht vs premature
ggplot(smoking,aes(x=premature, y=mht, fill=premature)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="mht vs Premature",
       x="Had premature baby?",y="Mother's height") +
  theme_classic() + theme(legend.position="none")
#no difference in median

# mpregwt vs premature
ggplot(smoking,aes(x=premature, y=mpregwt, fill=premature)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="mpregwt vs Premature",
       x="Had premature baby?",y="Mother's pre-pregnancy weight") +
  theme_classic() + theme(legend.position="none")
```

```r
#can see some difference in median - var might be of interest

#EDA for categorical variables
# premature vs mrace
apply(table(smoking[,c("premature","mrace")])/sum(table(smoking[,c("premature","mrace")])),
      2,function(x) x/sum(x))
#the chances of having a premature baby when you're white are considerably lower than when you're black
# - don't want to comment on other races as the data is not sufficient for other races
#chi-square test
chisq.test(table(smoking[,c("premature","mrace")]))
#it's significant

# premature vs med
apply(table(smoking[,c("premature","med")])/sum(table(smoking[,c("premature","med")])),
      2,function(x) x/sum(x))
#can see difference in probabilities - is lower for people with higher levels of education
#but also confusing because for example mothers who are college grads have a 16% probability of having
#as compared to 23% for high school grads + trade school (level 3)
#chi-square test
chisq.test(table(smoking[,c("premature","med")]))
#it's significant

# premature vs inc
apply(table(smoking[,c("premature","inc")])/sum(table(smoking[,c("premature","inc")])),
      2,function(x) x/sum(x))
#the probabilities are almost the same ~18-19% except for level 1 (26%) income mothers and level 8 (6%)
#chi-square test
chisq.test(table(smoking[,c("premature","inc")]))
#it's not significant

# premature vs smoke
apply(table(smoking[,c("premature","smoke")])/sum(table(smoking[,c("premature","smoke")])),
      2,function(x) x/sum(x))
#can see a significant difference in probabilities :D
#chi-square test
chisq.test(table(smoking[,c("premature","smoke")]))
#it's not significant


#Let's check binned probabilities of continuous variables with premature
###Checking if we need any transformations
#premature vs parity
smoking$prem_inter <- as.numeric(smoking$premature)
smoking$prem_inter[smoking$premature == '1'] <- 1
smoking$prem_inter[smoking$premature == '0'] <- 0
par(mfrow=c(1,1))
binnedplot(y=smoking$prem_inter,smoking$parity,xlab="Parity",ylim=c(0,1),col.pts="navy",
           ylab ="Had premature baby?",main="Binned Parity and Premature cases",
           col.int="white")
#not enough data

#premature vs mage
par(mfrow=c(1,1))
```

```r
binnedplot(y=smoking$prem_inter,smoking$mage,xlab="mage",ylim=c(0,1),col.pts="navy",
           ylab ="Had premature baby?",main="Binned Mother's age and Premature cases",
           col.int="white")
#doesn't look like there's a pattern

#premature vs mht
par(mfrow=c(1,1))
binnedplot(y=smoking$prem_inter,smoking$mht,xlab="mht",ylim=c(0,1),col.pts="navy",
           ylab ="Had premature baby?",main="Binned Mother's height and Premature cases",
           col.int="white")
#not enough data

#premature vs mpregwt
par(mfrow=c(1,1))
binnedplot(y=smoking$prem_inter,smoking$mpregwt,xlab="mpregwt",ylim=c(0,1),col.pts="navy",
           ylab ="Had premature baby?",main="Binned Mother's pre-pregnancy weight and Premature cases",
           col.int="white")
#decreasing increasing trend - we'll look at binned residuals later

###### Model fitting
model1 <- glm(premature ~ parity + mrace + mage + med + mht + mpregwt + inc + smoke, data = smoking, fa
summary(model1)

#model assessment

#save the raw residuals
rawresid1 <- residuals(model1,"resp")

##binned residual plots for continuous variables
binnedplot(x=fitted(model1),y=rawresid1,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#no points outside the 95% standard error lines so that's good. But the model is not very random so let
#with the continuous variables

#mage
binnedplot(x=smoking$mage,y=rawresid1,xlab="Mother's age",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good

#mht
binnedplot(x=smoking$mht,y=rawresid1,xlab="Mother's height",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#might be a trend there - will need a transformation maybe
#categories - 60-63, 63-66
#let's look at average residuals by mht using the tapply command
plot(0:17,tapply(rawresid1, smoking$mht, mean),col='blue4',pch=10)
table(smoking[,c("premature","mht")])

#mpregwt
binnedplot(x=smoking$mpregwt,y=rawresid1,xlab="Mother's pre-pregnancy weight",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good
```

```r
#parity
binnedplot(x=smoking$parity,y=rawresid1,xlab="Parity",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#not enough data
table(smoking[,c("premature","parity")])
#maybe convert into categories - 0, 1-3,4

##tables for factor variables
#smoke
tapply(rawresid1, smoking$smoke, mean)

#mrace
tapply(rawresid1, smoking$mrace, mean)

###### Model validation
#let's do the confusion matrix with .5 threshold
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model1) >= 0.5, "1","0")),
                            as.factor(smoking$premature),positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[c("Sensitivity","Specificity")] #True positive rate and True negative rate
#we're not able to identify premature birth in mothers but we are identifying non premature births well
#for premature births?

#first, let's repeat with the marginal percentage in the data
mean(smoking$prem_inter)
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model1) >= mean(smoking$prem_inter), "1","0")),
                            as.factor(smoking$premature),positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[c("Sensitivity","Specificity")]
#this looks more balanced

#look at ROC curve
roc(smoking$premature,fitted(model1),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
#AUC = 66.7%

#check multicollinearity
vif(model1)
#med might be correlated with one of the predictors - removing med is making smoke significant at the 0
#interactions between med

#### Model transformations
#let's convert mht to two categories - DID NOT WORK
smoking$mht_new <- rep(0,nrow(smoking))
smoking$mht_new[smoking$mht > 63] <- 1
table(smoking$mht,smoking$mht_new)

model2 <- glm(premature ~ parity + mrace + mage + med + mht_new + mpregwt + inc + smoke, data = smoking
summary(model2)

rawresid2 <- residuals(model2,"resp")
```

```r
##binned residual plots for continuous variables
binnedplot(x=fitted(model2),y=rawresid2,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#feels worse than model1 - it has few outliers and isn't random

#roc
roc(smoking$premature,fitted(model2),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
#AUC = 66.8% - very very small improvement

#let's convert mht to 5 categories
smoking$mht_new <- rep(0,nrow(smoking))
smoking$mht_new[smoking$mht <= 61] <- 1
smoking$mht_new[smoking$mht > 61 & smoking$mht <= 63] <- 2
smoking$mht_new[smoking$mht > 63 & smoking$mht <= 65] <- 3
smoking$mht_new[smoking$mht > 65 & smoking$mht <= 67] <- 4
smoking$mht_new[smoking$mht > 67] <- 5
smoking[,'mht_new']<-factor(smoking[,'mht_new'])

table(smoking$mht,smoking$mht_new)

model2 <- glm(premature ~ parity + mrace + mage + med + mht_new + mpregwt + inc + smoke, data = smoking
summary(model2)

rawresid2 <- residuals(model2,"resp")

#model2 with normal mht
model2_mht <- glm(premature ~ parity + mrace + mage + med + mht + mpregwt + inc + smoke, data = smoking
summary(model2_mht)

rawresid2 <- residuals(model2,"resp")



##binned residual plots for continuous variables
binnedplot(x=fitted(model2),y=rawresid2,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good

#mage
binnedplot(x=smoking$mage,y=rawresid2,xlab="Mother's age",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good

#mpregwt
binnedplot(x=smoking$mpregwt,y=rawresid2,xlab="Mother's pre-pregnancy weight",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good

#parity
binnedplot(x=smoking$parity,y=rawresid2,xlab="Parity",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#not enough data
```

```
#roc
roc(smoking$premature,fitted(model2_mht),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
#AUC = 67.5% - small improvement

#let's do the same for parity - DID NOT WORK
smoking$p_new <- rep(0,nrow(smoking))
smoking$p_new[smoking$parity <= 1] <- 1
smoking$p_new[smoking$parity > 1 & smoking$parity <= 3] <- 2
smoking$p_new[smoking$parity > 3 & smoking$parity <= 4] <- 3
smoking$p_new[smoking$parity > 4 & smoking$parity <= 6] <- 4
smoking$p_new[smoking$parity>6] <- 5
smoking[,'p_new']<-factor(smoking[,'p_new'])

model3 <- glm(premature ~ p_new + mrace + mage + med + mht_new + mpregwt + inc + smoke, data = smoking,
summary(model3)

rawresid3 <- residuals(model3,"resp")

##binned residual plots for continuous variables
binnedplot(x=fitted(model3),y=rawresid3,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")
#looks good

#roc
roc(smoking$premature,fitted(model2),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
#AUC = 67.5% - almost the same but makes the binned residuals overall plot look worse so will stick to

####Interactions
#Now let's see interactions with the smoke variable
# parity vs premature by smoke
ggplot(smoking,aes(x=premature, y=parity, fill=premature)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Parity vs premature baby, by smoke",
       x="Had premature baby?",y="Parity") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ smoke)
#median is different

# mage vs premature by smoke
ggplot(smoking,aes(x=premature, y=mage, fill=premature)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Mother's age vs premature baby, by smoke",
       x="Had premature baby?",y="Mother's age") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ smoke)
#some interaction might be there
```

```
# mht vs premature by smoke
ggplot(smoking,aes(x=premature, y=mht, fill=premature)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Mother's height vs premature baby, by smoke",
       x="Had premature baby?",y="Mother's height") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ smoke)
#some interaction might be there

# mpregwt vs premature by smoke
ggplot(smoking,aes(x=premature, y=mpregwt, fill=premature)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Mother's pre-pregnancy weight vs premature baby, by smoke",
       x="Had premature baby?",y="Mother's pre-pregnancy weight") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ smoke)
#some interaction might be there

# mrace vs premature by smoke - we want to check cos asked in question
# med vs premature by inc - we want to check cos intuitively this seems like interaction and vif for me

#mrace and smoke
model_inter1 <- glm(premature ~ inc + med  + smoke * mrace + mht + parity + mpregwt + mage, data = smok
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between mrace and smoke is not significant

#med and inc - WARNING
model_inter1 <- glm(premature ~ inc * med  + smoke + mrace + mht + parity + mpregwt + mage, data = smok
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")

#parity and smoke
model_inter1 <- glm(premature ~ inc + med  + smoke * parity + mht + mrace + mpregwt + mage, data = smok
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between parity and smoke is not significant

#mht_new and smoke
model_inter1 <- glm(premature ~ inc + med  + smoke * mht + parity + mrace + mpregwt + mage, data = smok
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between mht_new and smoke is significant at 0.1

#mage and smoke
model_inter1 <- glm(premature ~ inc + med  + smoke * mage + parity + mrace + mpregwt + mht, data = smok
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between mage and smoke is not significant
```

```
#mpregwt and smoke
model_inter1 <- glm(premature ~ inc + med  + smoke * mpregwt + parity + mrace + mage + mht, data = smoki
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between mpregwt and smoke is not significant

#inc and smoke
model_inter1 <- glm(premature ~ mpregwt + med  + smoke * inc + parity + mrace + mage + mht, data = smoki
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction between inc and smoke is not significant

#med and smoke
model_inter_final <- glm(premature ~ mpregwt + inc  + smoke * med + parity + mrace + mage + mht, data =
summary(model_inter_final)
anova(model_inter_final, model2_mht, test= "Chisq")
vif(model_inter_final)
#interaction between med and smoke is significant
#after including this interaction the vif is also in control for all variables so this is our final mod

#med * smoke and med * inc - WARNING
model_inter1 <- glm(premature ~ mpregwt + med * (inc  + smoke) + parity + mrace + mage + mht_new, data =
summary(model_inter1)
anova(model_inter1, model2, test= "Chisq")
#interaction between med and smoke is significant

#med * smoke and mht * smoke
model_inter1 <- glm(premature ~ mpregwt + smoke * (med  + mht) + parity + mrace + mage + inc, data = smo
summary(model_inter1)
anova(model_inter1, model2_mht, test= "Chisq")
#interaction is significant

#but now let's check that given the interaction between med and smoke is the interaction between mht an
model2_check <- glm(premature ~ mpregwt + smoke * med  + mht + parity + mrace + mage + inc, data = smoki
model_inter2 <- glm(premature ~ mpregwt + smoke * (med  + mht) + parity + mrace + mage + inc, data = smo
summary(model_inter1)
anova(model_inter2, model2_check, test= "Chisq")
#interaction is significant at 0.1 so we will not include it

#let's use the stepwise function to do model selection (using BIC)
n <- nrow(smoking)
null_model <- glm(premature ~ smoke + mrace,data=smoking,family=binomial)
model_step <- step(null_model,scope=formula(model_inter_final),direction="both",
     trace=0,k = log(n))
model_step$call

#let's use the stepwise function to do model selection (using AIC)
n <- nrow(smoking)
null_model <- glm(premature ~ smoke + mrace,data=smoking,family=binomial)
model_step_AIC <- step(null_model,scope=formula(model_inter_final),direction="both",
                   trace=0)
                   #,k = log(n))
model_step_AIC$call
```

```
#glm(formula = premature ~ smoke + mrace + med + mpregwt + smoke:med,
#family = binomial, data = smoking)

model_final <- glm(formula = premature ~ smoke + mrace + med + mpregwt + smoke:med,
                   family = binomial, data = smoking)
summary(model_final)

#model assessment for final model
#save the raw residuals
rawresid_final <- residuals(model_final,"resp")

##binned residual plots for continuous variables
binnedplot(x=fitted(model_final),y=rawresid_final,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Binned residual plot",col.pts="navy")


###### Model validation
#let's do the confusion matrix with .5 threshold
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model_final) >= 0.5, "1","0")),
                            as.factor(smoking$premature),positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[c("Sensitivity","Specificity")] #True positive rate and True negative rate
#we're not able to identify premature birth in mothers but we are identifying non premature births well
#for premature births?

#first, let's repeat with the marginal percentage in the data
mean(smoking$prem_inter)
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(model_final) >= mean(smoking$prem_inter), "1","0"))
                            as.factor(smoking$premature),positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[c("Sensitivity","Specificity")]
#this looks more balanced

#look at ROC curve
roc(smoking$premature,fitted(model_final),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
#AUC = 67.4%

#why interaction b/w these vars is giving a warning
table(smoking[,c("med","inc")])
```