Data Analysis Assignment 2

Himangshu Raj Bhantana

9/22/2021

Summary

This study looks at maternal smoking and pre-term birth. I examined the data to see if women who smoke had a higher risk of having a pre-term birth than mothers who do not smoke. I used exploratory data analysis, a Backward selection AIC model, to determine the logistic regression to ensure if there was a significant relationship between smoking and pre-term birth. After study, we came to conclusion that we did not have enough evidence to conclude that mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke. This affect did not vary with race.

Introduction

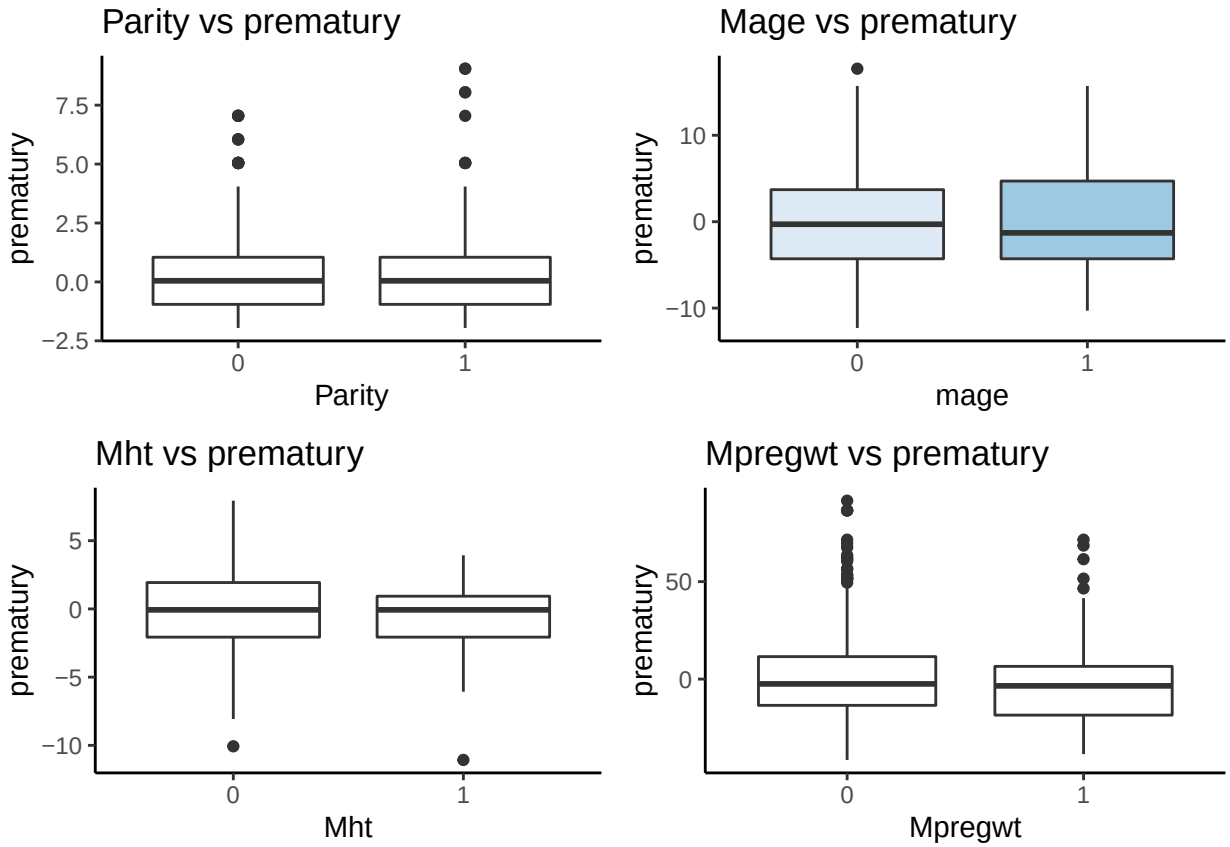
It is well known in today's society that pregnant moms who smoke risk exposing their babies to a variety of health issues. Perhaps not in the early years, because pregnant women were not aware that smoking had a harmful influence on their pre-term birth. As a result, in this study, we will look at whether smoking has a detrimental influence on pre-term birth, as evidenced by the link between smoking and pre-term birth. The following are some of the particular questions that will be addressed in this study:

- 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

Before delving into the data analysis for this study, the modified version of Smoking Data set, which comprised 869 observations on 12 variables with a mix of categorical and continuous variables, was examined. Because we are primarily interested in smoking and non-smoking, not all factors were utilized in this study, such as gestation, date, id, birth weight and time. Variables with the most sub levels of data were condensed into one, such as race=white, for our convenience and to keep the data short. For this study our response variable is Pre-mature.

Now starting with Exploratory data Analysis which is performed on the modified smoking dataset to determine the model and predictors. Our model is presented into box plot for continuous predictor in comparison with our categorical response variable and table with CHI-test for comparing both of our categorical variables. Doing this Exploratory data Analysis gives us comparison of if the predictors are independent or not. The following box plots displays the interaction between continuous variable and categorical response variable:



The box plots for the interaction effects between all numeric and categorical variables seem good because they are all independent and there isn't much of a change in median.

CHI Test

As we work with tables and the CHI test for association for factor variables, we can see whether or not the factor variables are significant. After completing the CHI test for factor variables, it was determined that the response variable premature association with race, and education is statistically significant and independent because the P value is less than 5%. We also need to determine whether smoking affects premature birth or not, and we must conduct a CHI test to determine whether the association between smoking and race is significant or not. For premature connection with smoking and income is not significant for the response variable and is dependent since the p value exceeds 5%.

Model

Before fitting the model, we must determine if the predictors are random or not, and if not, a transformation must be applied. We use a binned plot to verify the randomness of continuous predictors. The transformation was unnecessary because continuous predictors were random and no such strange patterns were created. Beginning by fitting the first logistic regression model, which contains all categorical and factor variables, including the significant correlations smoking and race, to check if it influences the model or not.

Since we require significant predictors to answer our query with the best estimation, I chose the backward AIC model, which produced a new logistic regression model with significant predictors. After attempting alternative selection models such as forward and stepwise, which produced the same significant predictors, I chose to continue with the AIC Model. As a result of our model selection, our final logistic regression

model includes prematury1 as a response variable and morace1, meduc1, mpregwt(weight), and smoke1 as predictors. The mathematical regression equation for the final model is:

$$Pr(Y_i = 1|X_i) = \frac{exp(\beta_0 + \beta_1 meduc1 + \beta_2 morace1 + \beta_3 mpregwt + \beta_4 smoke1)}{1 + exp(\beta_0 + \beta_1 meduc1 + \beta_2 morace1 + \beta_3 mpregwt + \beta_4 smoke1)} \quad (1)$$

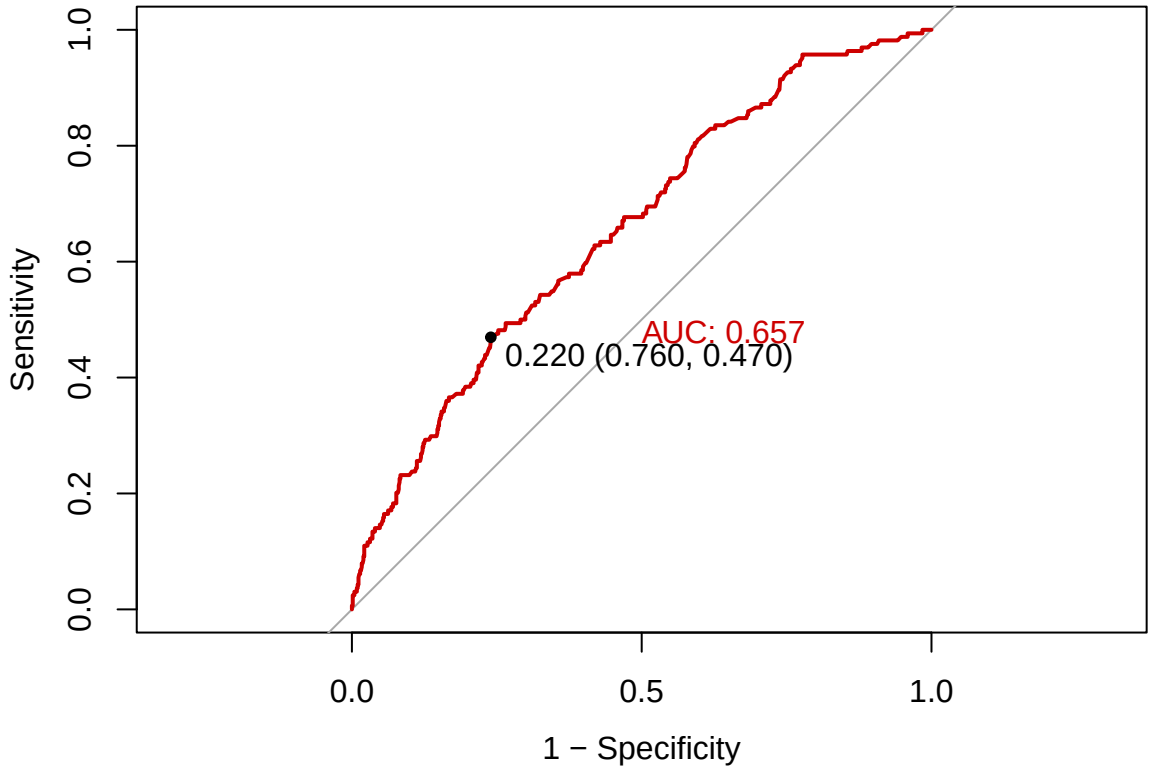
The following is a summary of our model selection for our new final logistic regression model: glm(formula = prematury1 ~ morace1 + meduc1 + mpregwt + smoke1, family = binomial, data = smoking)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.51	0.46	-1.12	0.26
morace1black	-0.14	0.45	-0.30	0.76
morace1mexican	-0.75	0.64	-1.18	0.24
morace1mix	-1.66	1.12	-1.48	0.14
morace1white	-0.91	0.41	-2.22	0.03
meduc1college	-0.52	0.31	-1.69	0.09
meduc1high school + college	-1.01	0.30	-3.33	0.00
meduc1high school + trade school	-0.17	0.41	-0.41	0.68
meduc1high school but no other schooling	-0.35	0.25	-1.38	0.17
meduc1less than 8th grade	0.54	0.95	0.57	0.57
meduc1trade school	2.37	1.18	2.00	0.05
mpregwt	-0.01	0.00	-2.51	0.01
smoke11	0.29	0.18	1.57	0.12

To ensure that our newly assigned significant factor predictors are correct, we will do model evaluation using resudial plots for our categorical predictors to determine whether randomness exists and if the function of predictors is adequately described using binned residuals. If there is no randomness, we will begin with transformation; however, after inspecting the residual plots, it is safe to conclude that randomness exists and that the majority of the 95 percent of observations are within the red lines; only a few points were outside the red lines, which is acceptable because it does not exceed the 5 percent level. To validate the randomness plot are present in our appendix.

We may now proceed with model validation. We will use the confusion matrix to validate the model. It indicates that the model's accuracy, sensitivity, and specificity are all within the range of 61 percent, 58 percent, and 62 percent. The ROC curve shows that we can maximize predictions from our model by using a 0.220 probability threshold. These numbers appear to be alright, even if it appears that they haven't balanced out entirely, which isn't conceivable.

The following shows the ROC curve and discover that the area of the curve is 0.657, which indicates how well



our model predicts.

We still need to investigate whether or not there is multicollinearity. Checking VIP of our final model for multicollinearity demonstrates that the value for each variable does not pass the value of 10. This ensures that we do not need to be concerned about multicollinearity.

Coefficient Interpretation

Taking the coefficients for the significant predictors, the odds ratio of having a pre-term birth for white mothers is 0.4 as compared to Asian mothers. The odds of having a pre-term birth mother who went to high school and college is 0.37 as compared to a mother who went to 8th to 12th grade.

The odds ratio of having a pre-term birth for a mother who only went to trade school is 10.66 as compared to mothers who went from 8th to 12th grade. One unit increase in weight will decrease the odds of having pre-term birth by 0.99.

Conclusion

In order to sum up our questions for this study, we do not have enough data to infer that women who smoke have a higher risk of pre-term delivery than mothers who do not smoke. When compared to nonsmoking moms, the chances ratio of a smoking mother having a pre-term delivery ranges from 0.93 to 1.9. Because the interaction between smoke and race, as well as the smoke variable alone, is not significant, we cannot draw any conclusions regarding their influence from our model. I don't find any intriguing correlations worth highlighting after reviewing all of the data in this study because all notable associations are significant and reported in the study.

Limitations are approaching. I noticed that the data collection is too tiny, because logistic regression models require a large data set to get significant findings. Another drawback is that the residual deviation is very big.

Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(arm)
library(pROC)
library(e1071)
library(caret)
library(ggplot2)
require(gridExtra)
library(readr)
smoking <- read.csv("Downloads/smoking.csv")

smoking <-smoking %>%
  mutate(
    prematury = case_when(
    gestation < 270 ~ '1',
    gestation >= 270 ~ '0',
    TRUE ~ "unknown"
   )
 )


smoking <-smoking %>%
  mutate(
    morace = case_when(
    mrace %in% c(0,1,2,3,4,5) ~ 'white',
    mrace == 6 ~ 'mexican',
    mrace == 7 ~ 'black',
    mrace == 8 ~ 'asian',
    mrace == 9 ~ 'mix',
    TRUE ~ 'other'
    )
  )
smoking <-smoking %>%
  mutate(
    meduc = case_when(
      med == 0 ~ 'less than 8th grade',
      med == 1 ~  '8th to 12th',
      med == 2 ~ 'high school but no other schooling',
      med == 3 ~ 'high school + trade school',
      med == 4 ~ 'high school + college',
      med == 5 ~ 'college',
      med %in% c(6,7) ~ 'trade school',
      TRUE ~ 'unknown'
    )
  )

#factor=boxplot, continous= scatter
#factoring categorical variables
smoking$morace1<-factor(smoking$morace)
smoking$meduc1<-factor(smoking$meduc)
smoking$smoke1<-factor(smoking$smoke)
```

```
smoking$prematury1<-factor(smoking$prematury)
smoking$inc1<-factor(smoking$inc)


#convering into integers

smoking$prematury2<-as.numeric(smoking$prematury1)
smoking$prematury2 <- smoking$prematury2-1



#Let's mean center the numerical predictors to avoid Mutli
smoking$parity <- smoking$parity - mean(smoking$parity)
smoking$mage <- smoking$mage - mean(smoking$mage)
smoking$mht<- smoking$mht - mean(smoking$mht)
smoking$mpregwt <- smoking$mpregwt - mean(smoking$mpregwt)

smoking$morace1<-factor(smoking$morace)
smoking$meduc1<-factor(smoking$meduc)
smoking$smoke1<-factor(smoking$smoke)
smoking$prematury1<-factor(smoking$prematury)
smoking$inc1<-factor(smoking$inc)

a<-ggplot(smoking,aes(x=prematury1, y=parity, fill=parity)) +
  geom_boxplot()+
  scale_fill_brewer(palette="Blues") +
  labs(title="Parity vs prematury",x="Parity",y="prematury")  +
  theme_classic() + theme(legend.position="none")


b<-ggplot(smoking,aes(x=prematury1, y=mage, fill=prematury)) +
  geom_boxplot()+
  scale_fill_brewer(palette="Blues") +
  labs(title="Mage vs prematury",x="mage",y="prematury")  +
  theme_classic() + theme(legend.position="none")



c<-ggplot(smoking,aes(x=prematury1, y=mht, fill=mht)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mht vs prematury",x="Mht",y="prematury")  +
  theme_classic() + theme(legend.position="none")


d<-ggplot(smoking,aes(x=prematury1, y=mpregwt, fill=mpregwt)) + #not much difference
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mpregwt vs prematury",x="Mpregwt",y="prematury")  +
  theme_classic() + theme(legend.position="none")
grid.arrange(grobs = list(a,b,c,d), ncol = 2, main = "Main title")
library(xtable)
options(xtable.comment = FALSE)
```

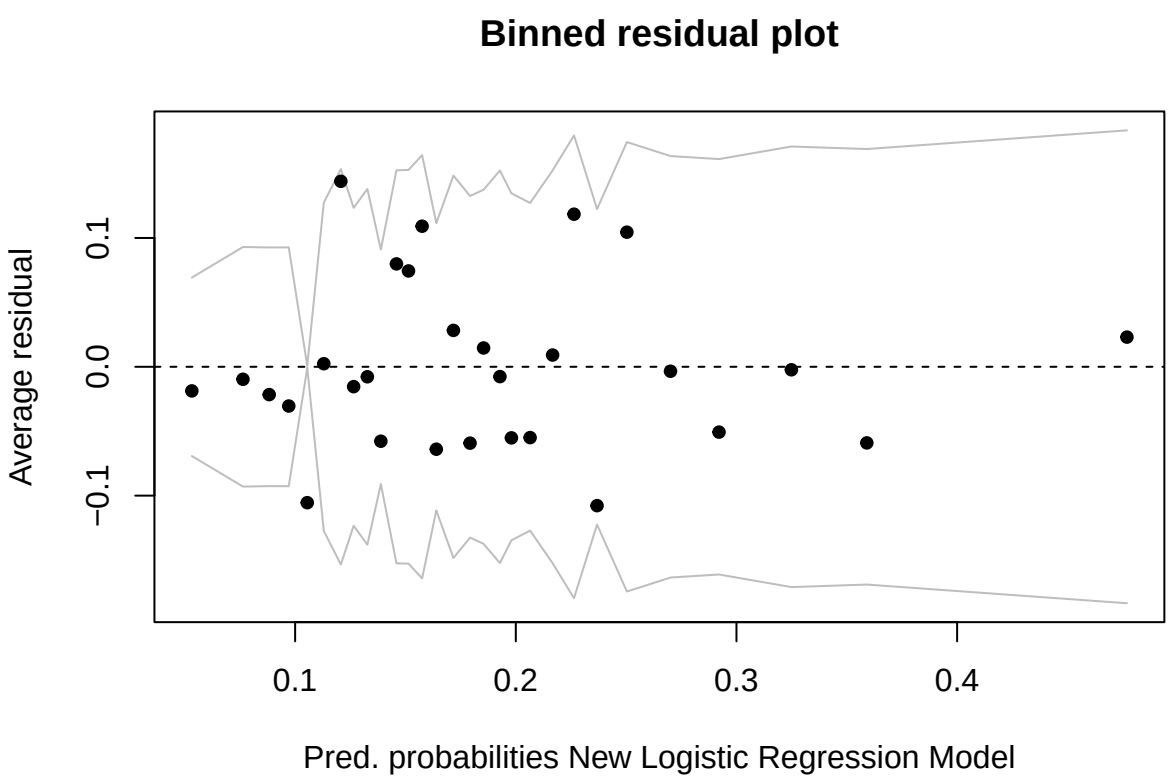
```
smokyreg <- glm(prematury1 ~ parity + morace1 + mage +  meduc1 + mpregwt + smoke1+inc1+smoke1:morace1, 
#aic
back<-step(smokyreg, direction = "both", trace=FALSE)
back$call
newregy<-glm(formula = prematury1 ~ morace1 + meduc1 + mpregwt + smoke1,
             family = binomial, data = smoking)

xtable(newregy, type='latex', title='Baseline Model Summary', header= FALSE, digits=2, no.space = TRUE)

a <-roc(smoking$prematury1,fitted(newregy),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")

library(arm)
library(pROC)
library(e1071)
library(caret)
library(ggplot2)
require(gridExtra)


smoking <-smoking %>%
  mutate(
    prematury = case_when(
    gestation < 270 ~ '1',
    gestation >= 270 ~ '0',
    TRUE ~ "unknown"
   )
 )

smoking <-smoking %>%
  mutate(
    morace = case_when(
    mrace %in% c(0,1,2,3,4,5) ~ 'white',
    mrace == 6 ~ 'mexican',
    mrace == 7 ~ 'black',
    mrace == 8 ~ 'asian',
    mrace == 9 ~ 'mix',
    TRUE ~ 'other'
    )
  )
smoking <-smoking %>%
  mutate(
    meduc = case_when(
      med == 0 ~ 'less than 8th grade',
      med == 1 ~  '8th to 12th',
      med == 2 ~ 'high school but no other schooling',
      med == 3 ~ 'high school + trade school',
      med == 4 ~ 'high school + college',
      med == 5 ~ 'college',
      med %in% c(6,7) ~ 'trade school',
      TRUE ~ 'unknown'
    )
  )
```

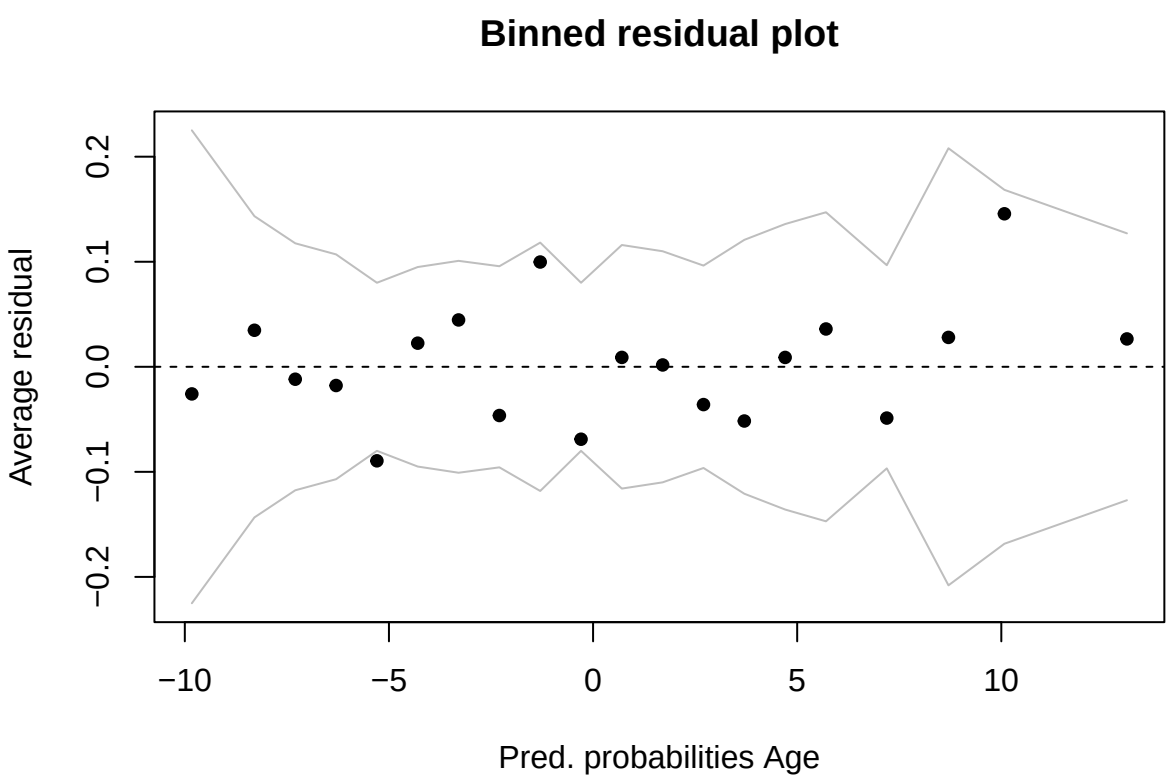
```
#factor=boxplot, continous= scatter
#factoring categorical variables
smoking$morace1<-factor(smoking$morace)
smoking$meduc1<-factor(smoking$meduc)
smoking$smoke1<-factor(smoking$smoke)
smoking$prematury1<-factor(smoking$prematury)
smoking$inc1<-factor(smoking$inc)
smoking
str(smoking)

#convering into integers

smoking$prematury2<-as.numeric(smoking$prematury1)
smoking$prematury2 <- smoking$prematury2-1
str(smoking)


#Let's mean center the numerical predictors to avoid Mutli
smoking$parity <- smoking$parity - mean(smoking$parity)
smoking$mage <- smoking$mage - mean(smoking$mage)
smoking$mht<- smoking$mht - mean(smoking$mht)
smoking$mpregwt <- smoking$mpregwt - mean(smoking$mpregwt)

#EDA
ggplot(smoking,aes(x=prematury1, y=parity, fill=parity)) +
  geom_boxplot()+
  scale_fill_brewer(palette="Blues") +
  labs(title="Parity vs prematury",x="Parity",y="prematury")  +
  theme_classic() + theme(legend.position="none")


ggplot(smoking,aes(x=prematury1, y=mage, fill=prematury)) +
  geom_boxplot()+
  scale_fill_brewer(palette="Blues") +
  labs(title="Mage vs prematury",x="mage",y="prematury")  +
  theme_classic() + theme(legend.position="none")



ggplot(smoking,aes(x=prematury1, y=mht, fill=mht)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mht vs prematury",x="Mht",y="prematury")  +
  theme_classic() + theme(legend.position="none")


ggplot(smoking,aes(x=prematury1, y=mpregwt, fill=mpregwt)) + #not much difference
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="Mpregwt vs prematury",x="Mpregwt",y="prematury")  +
  theme_classic() + theme(legend.position="none")
```

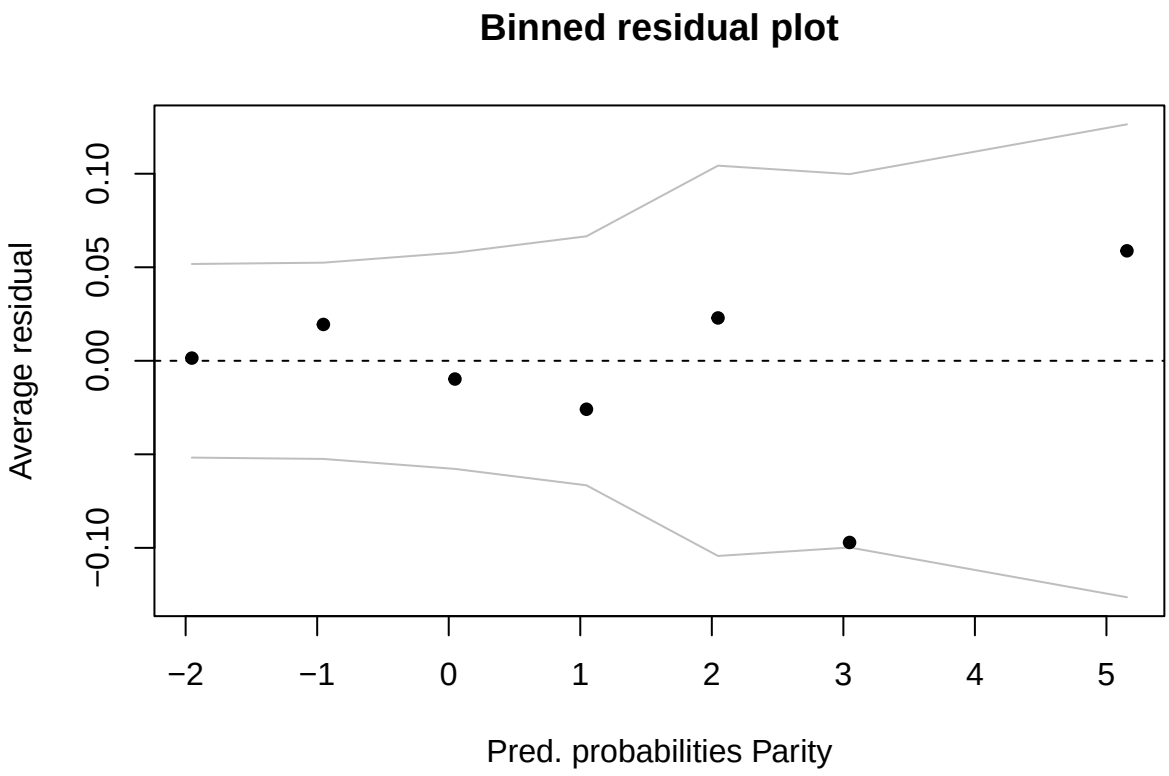
```
## We can do tables for the factor variables
#less than 0.05 two variable dependent hay
table(smoking[,c("morace1","prematury1")])
table(smoking[,c("morace1","prematury1")])/sum(table(smoking[,c("morace1","prematury1")]))

apply(table(smoking[,c("morace1","prematury1")])/sum(table(smoking[,c("morace1","prematury1")])),
      2,function(x) x/sum(x))
# You can also use the tapply command for the same thing
tapply(smoking$morace1, smoking$prematury1, function(x) table(x)/sum(table(x)))
# Finally, we can even try a chi-squared test for independence.
chisq.test(table(smoking[,c("morace1","prematury1")]))

table(smoking[,c("meduc1","prematury1")])
table(smoking[,c("meduc1","prematury1")])/sum(table(smoking[,c("meduc1","prematury1")]))

apply(table(smoking[,c("meduc1","prematury1")])/sum(table(smoking[,c("meduc1","prematury1")])),
      2,function(x) x/sum(x))
# You can also use the tapply command for the same thing
tapply(smoking$meduc1, smoking$prematury1, function(x) table(x)/sum(table(x)))
# Finally, we can even try a chi-squared test for independence.
chisq.test(table(smoking[,c("meduc1","prematury1")]))

table(smoking[,c("smoke1","prematury1")])
table(smoking[,c("smoke1","prematury1")])/sum(table(smoking[,c("smoke1","prematury1")]))

apply(table(smoking[,c("smoke1","prematury1")])/sum(table(smoking[,c("smoke1","prematury1")])),
      2,function(x) x/sum(x))
# You can also use the tapply command for the same thing
tapply(smoking$smoke1, smoking$prematury1, function(x) table(x)/sum(table(x)))
# Finally, we can even try a chi-squared test for independence.
chisq.test(table(smoking[,c("smoke1","prematury1")]))

table(smoking[,c("inc1","prematury1")])
table(smoking[,c("inc1","prematury1")])/sum(table(smoking[,c("inc1","prematury1")]))

apply(table(smoking[,c("inc1","prematury")])/sum(table(smoking[,c("inc1","prematury1")])),
      2,function(x) x/sum(x))
# You can also use the tapply command for the same thing
tapply(smoking$inc1, smoking$prematury1, function(x) table(x)/sum(table(x)))
# Finally, we can even try a chi-squared test for independence.
chisq.test(table(smoking[,c("inc1","prematury1")]))

table(smoking[,c("smoke1","morace1")])
table(smoking[,c("smoke1","morace1")])/sum(table(smoking[,c("smoke1","morace1")]))

apply(table(smoking[,c("smoke1","morace1")])/sum(table(smoking[,c("smoke1","morace1")])),
      2,function(x) x/sum(x))
# You can also use the tapply command for the same thing
tapply(smoking$smoke1, smoking$morace1, function(x) table(x)/sum(table(x)))
# Finally, we can even try a chi-squared test for independence.
chisq.test(table(smoking[,c("smoke1","morace1")]))

#binned plots for continious variable # looking if predictors are random or not
```

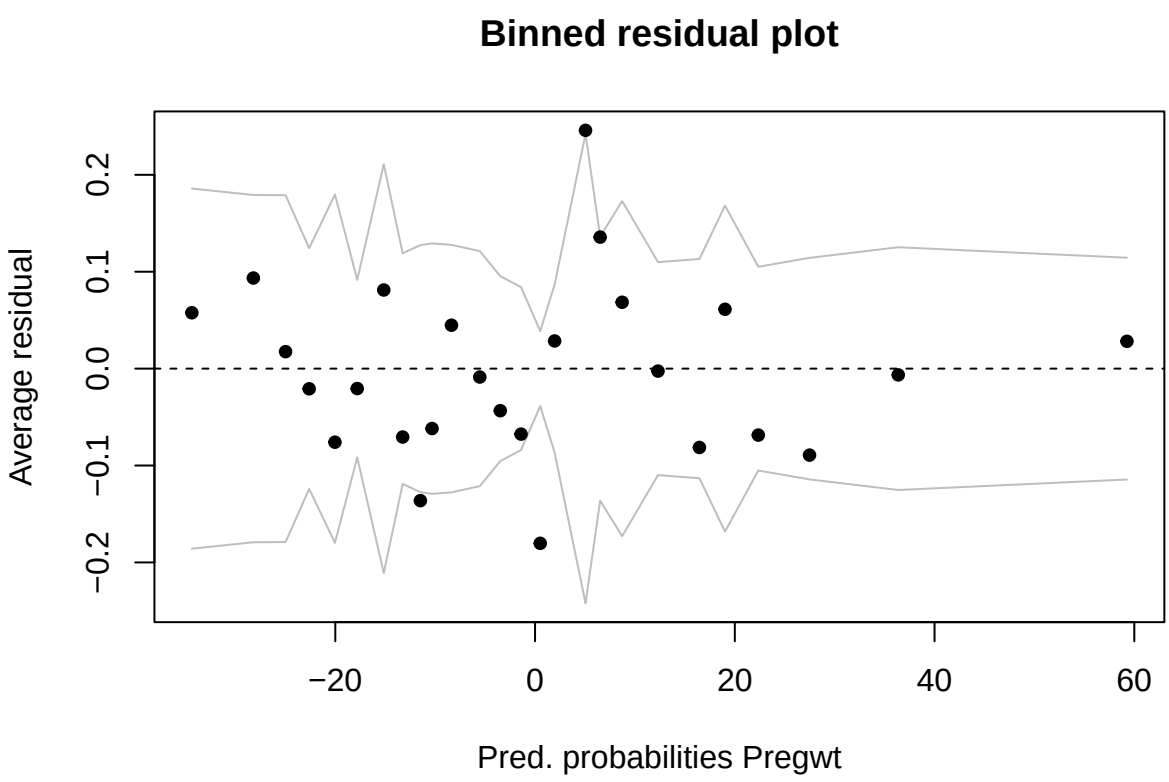
```
par(mfrow=c(1,1))
binnedplot(y=smoking$prematury2,smoking$mage,xlab="mage",ylim=c(0,1),col.pts="navy",
           ylab ="Premature",main="Binned Premature and Mage",
           col.int="white")

par(mfrow=c(1,1))
binnedplot(y=smoking$prematury2,smoking$parity,xlab="parity",ylim=c(0,1),col.pts="navy",
           ylab ="Premature",main="Binned Premature and Mage",
           col.int="white")
par(mfrow=c(1,1))
binnedplot(y=smoking$prematury2,smoking$mpregwt,xlab="pregwt",ylim=c(0,1),col.pts="navy",
           ylab ="Premature",main="Binned Premature and Mage",
           col.int="white")

str(smoking)
#model fit
smokyreg <- glm(prematury1 ~ parity + morace1 + mage +  meduc1 + mpregwt + smoke1+income1+smoke1:morace
summary(smokyreg)
#aic
back<-step(smokyreg, direction = "both", trace=FALSE)
back$call
summary(smokyreg)
anova(smokyreg)
str(smoking)
#save the raw residuals

newregy<-glm(formula = prematury1 ~ morace1 + meduc1 + mpregwt + smoke1,
             family = binomial, data = smoking)
#model assesment resiudals,
residy <- residuals(newregy,"resp")

#binned residual plots # transformation k jarurat hay ya nahi hay
binnedplot(x=fitted(newregy),y=residy,xlab="Pred. probabilities")
binnedplot(x=smoking$mage,y=residy,xlab="Pred. probabilities")
binnedplot(x=smoking$parity,y=residy,xlab="Pred. probabilities")
binnedplot(x=smoking$mpregwt,y=residy,xlab="Pred. probabilities")


###### Model validation
#let's do the confusion matrix with .5 threshold
# tell us accuraacy of the model
#sensi positive weight kitne 1 sahi predict karre hay, specifi: true negative weight Os
Confy <- confusionMatrix(as.factor(ifelse(fitted(newregy) >= 0.5, "1","0")),
                            as.factor(smoking$prematury1),positive = "1")
Confy$table
Confy$overall["Accuracy"];
Confy$byClass[c("Sensitivity","Specificity")] #True positive rate and True negative rate
#Maybe we can try to increase that accuracy.
#Also, the TNR looks low here.

#first, let's repeat with the marginal percentage in the data# we do
mean(smoking$prematury2)
Confy <- confusionMatrix(as.factor(ifelse(fitted(newregy) >= mean(smoking$prematury2), "1","0")),
```

```
                              as.factor(smoking$prematury1),positive = "1")
Confy$table
Confy$overall["Accuracy"];
Confy$byClass[c("Sensitivity","Specificity")]
#huge difference!  seems a lot of predicted probabilities are in the .5 yo .58  range, so cutoff matter
#either way, we have large off-diagonal numbers. specificity is sensitive to the cutoff

#look at ROC curve # model ko ptimize karney
roc(smoking$prematury1,fitted(newregy),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")
exp(newregy$coefficients)

###model interpretations
are the odds of having premature birth for non smoking asian mother with education 8 to 12 grade.

confint.default(newregy)   #on log odds scale
exp(confint.default(newregy))   #on odds scale
summary(newregy)
vif(newregy)


residy <- residuals(newregy,"resp")
a1<-binnedplot(x=fitted(newregy),y=residy,xlab="Pred. probabilities New Logistic Regression Model")
a2<-binnedplot(x=smoking$mage,y=residy,xlab="Pred. probabilities Age")
a3<-binnedplot(x=smoking$parity,y=residy,xlab="Pred. probabilities Parity")
a4<-binnedplot(x=smoking$mpregwt,y=residy,xlab="Pred. probabilities Pregwt")
```


Binned residual plot
Average residual
0.1
0.0
−0.1
0.1
0.2
0.3
0.4
Pred. probabilities New Logistic Regression Model


Binned residual plot
Average residual
0.2
0.1
0.0
−0.1
−0.2
−10
−5
0
5
10
Pred. probabilities Age


Binned residual plot
Average residual
0.10
0.05
0.00
−0.10
−2
−1
0
1
2
3
4
5
Pred. probabilities Parity


Binned residual plot
Average residual
0.2
0.1
0.0
−0.1
−0.2
−20
0
20
40
60
Pred. probabilities Pregwt